NOVEMBER 30, 2018

THE MECHANISMS OF EVOLUTIONARY FLEXIBILITY IN EARTHWORM GENOMES

A THESIS SUBMITTED TO CARDIFF UNIVERSITY FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

OLIVER JOHN RIMINGTON
CARDIFF UNIVERSITY
2018

Supervisory Group:
Prof. P. Kille[1], Prof. A. J. Weightman[1], Prof. D. Spurgeon[2].

[1] Cardiff School of Biosciences, BIOSI 1, University of Cardiff, P.O. Box 915, Cardiff, CF10 3TL, UK.
[2] Centre for Ecology and Hydrology, Maclean Building, Benson Lane, Wallingford, Oxfordshire, OX10 8BB, UK.

Thesis Summary

Background

Many individual organisms have latent phenotypic potentials which are never realised within their lifespans. This potential can include a huge diversity of dormant adaptations across the tree of life, such as the ability to tolerate radical changes in temperature, survive restricted nutrient availability, and resist toxins and parasites. Prior to unrealised phenotypic potentials are necessarily information potentials residing in a dormant state also. This thesis investigates the systematic interactions of facultative morphologies and atavistic adaptivity with the evolutionary systems which propagate them. Earthworms as models are for these purposes an almost archetypal form of a high-latent-potential organism. Examples abound of their thriving as peregrine species with near-global ranges.

Methods

Investigation of mechanistic context of evolutionary flexibility was pursued via three channels of inquiry. The first was to utilise modern genomics to query a pair of genomes with a demonstrably high complexity, given their remarkable allelic divergence. The second was to use large scale sequencing experiments to analyse a model system which, via our knowledge of its life history, has an apparent need for facultative morphology changes. The third was to develop novel tools for the more precise description of the information structure behind environmental adaptivity and phenotypic plasticity.

Results

Unprecedented base sequence divergence was discovered in the Earthworm *Lumbricus rubellus,* which led to the furthering of evolutionary perspectives, particularly in relation to recombination, on the role of allelic divergence in information latency. A novel sequence signature tool and the foundational mathematics supporting it were developed, this was used to discover the intriguing qualities of information structure in various test sets. The Earthworm *Amynthas gracilis* living in volcanic soils was used as model to study intrinsic information sources and the mechanisms of activation in a multi-stressor environment.

DECLARATION

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed …… (candidate) Date ……

STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD

Signed …… (candidate) Date ……

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated.

Other sources are acknowledged by explicit references. The views expressed are my own.

Signed …… (candidate) Date ……

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed …… (candidate) Date ……

STATEMENT 4: PREVIOUSLY APPROVED BAR ON ACCESS

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loans **after expiry of a bar on access previously approved by the Academic Standards & Quality Committee.**

Signed …… (candidate) Date ……

Acknowledgements

I would like to thank Prof. Pete Kille for his unendingly buoyant enthusiasm for all things that can both crawl in the dirt and run on dedicated servers. And of course, for keeping me on track!

I must thank my partner Jayde who will now spend the rest of her life with a lot of latin names for things she would rather not touch lodged in the back of her mind. This would not have been possible without her support and encouragement.

I would like to thank my mother who sadly passed away during my PhD. My memory of her conscientiousness and intellect continues to inspire me to push forward in life. And I must also thank my father who has been a great source of stability and lunch throughout this process.

Dr Dave Stanton has helped me immensely with structuring my thoughts and methods sections. Dr Luis Cunha and Dr Marta Novo deserve huge credit for going to places and digging things up, their data generation has been of great value. In addition, their prior studies have been key to the foundational knowledge basis for this work. Thanks also go to Prof. Dave Spurgeon for his input and paper reviewing! I am also grateful to Dr Craig Anderson for making his data available, and Dr Daniel Pass for his miscellaneous help, particularly when I first started.

This PhD was funded by the National Environmental Research Council (NERC) via the GW4 Doctoral Training Program.

Table of Contents

0. Glossary of terms

- **Plasticity**
  - (environmental/ w.r.t. the environment) ability to adapt physiologically to meet variables demands made by the ecosystem or natural environment
- **Scaffolds**
  - (genomic) large pieces of physically linked genetic material assembled into putative fragments, often with gaps, by a computation method.
- **Proteome**
  - A collection of the entire known set of proteins produced by an organism's genome, described in terms of its peptide sequences
- **Methylated/methylation**
  - (Genomic) 5-Cytosine Methylation: the property of a single methyl group being attached to the fifth carbon in its aromatic ring
- **Enrichment**
  - (statistical)
- **miRNAs**
  - Micro-RNA – refers to small non-coding RNAs which, in their mature form (to which this term usually refers), are approximately 22bp in length
- ***l*-mer**
  - a member of the set of possible substrings of a fixed size k-mer
- ***k*-mer**
  - a fixed length substring, or 'word' of a large sequence collection
- **UTRs**
  - Untranslated regions preceding (5') or trailing (3') a gene's protein coding regions
- **CDSs**
  - Coding Sequences – long stretches of DNA in which one of the possible six protein coding reading frames has no stop codon.
- **WSD**
  - Weighted standard deviation. A measurement of the deviation where the contributions of each data point are weighted by their scale.
- **Read depth**
  - (short reads) The property of depth refers to number of short reads from a sequencing library simultaneously aligned to the same region of DNA
- **N-mask**
  - A binary layer of sequence location flexibilities which can be applied to set of k-mers to collapse them into single objects whereby their only points of variation lie within the N-masked characters
- **Bp/Kbp/Mbp/Gbp**
  - (sequence size metric) base pairs, kilo (1,000) base pairs, mega (1,000,000) base pairs, or giga (1,000,000,000) base pairs respectively
- **Morphotype**
  - One of a group of possible different physiological modes of existence in a population. A phenotype may comprise multiple potential morphotypes, which may either be fixed for the lifetime of the individual, or variable in response to stimuli such as reproductive options or stress
- **Parent/child**

    - (computing terminology) Hierarchical relationship between nodes in a tree or hierarchical graph of some form. Of any two connected nodes, the parent is defined as the one closest (by degree) to the root, whilst the child is closes to the leaves)
- **Mosaic**
    - (Genomic) Refers to the case whereby, as the result of a hybridisation, the genome of an individual organism is a stochastic arrangement of alleles from different populations – particularly when those alleles are substantially divergent
- **Pathways**
    - (Functional Analysis) Refers to the metabolic 'pathway' by which a chain of reactions take place to facilitate a single biological effect, or mode of effects. Grouping genes by their pathway associations has the potential to strengthen statistical associations with a tested factor or variable.
- **Pileup**
    - (short read sequence mapping) refers to the 'piles' of reads present in alignment maps, where many overlapping reads of DNA align to the same region of a reference object
- **Vector**
    - (C++/programming) a one-dimensional array of data points stored in an iterable container
- **GPCR**
    - G-protein coupled receptor – a cell surface/transmembrane signalling receptor
- **Introgression**
    - Transfer of a smaller region of genetic material between species by hybridization following by multiple backcrosses with original populations
- **Kernel**
    - (matrix) the frequency matrix of an engineered set of features by their occurrence in each sample
- **DFS**
    - (algorithm) A depth-first search is a manner of navigating an information space, commencing at a root, and exploring all branches to their maximum depth before retracing to explore the next available branch
- **Capacitance**
    - (information) Used to refer to the property of information retention within a biological system – particularly when it is underutilised for long periods of time without being lost
- **Cryptic**
    - (lineages) refers to the property of genetic diversity with an apparently physically uniform population being hard or impossible to detect without targeted sequencing efforts
- **MeDIP-Seq**
    - Methylated DNA immunoprecipitation combined with next generation sequencing of the precipitated DNA. A 5-methylcytosine sensitive antibody is used to precipitate DNA fragments in suspended in solution which are methylated. The mapping of the sequences generated can be used to profile the distribution of epigenetic modifications throughout the genome of an organism.
- **miRNA-Seq**

    - A form of RNA sequencing whereby the input material is enriched for small RNA fragments, with the intention of capturing the mature forms of regulatory micro-RNAs.
- **Knowledge-free**
    - (modes of analysis) A distinction with respect to the treatment of sequence data whereby a prior association with databases of known genetic material is possible but not necessarily desirable.
- **SNPs**
    - Single nucleotide polymorphism, indicating a point of sequence variation between instances of otherwise fixed sequences of DNA.
- allopolyploid 5
- **k-/r- selection**
    - reproductive strategy distinction. R-selected organisms produce many offspring at low cost, often with greater environmental instability or range. K-selection occurs where the offspring are few and energetically very expensive for the organism to produce, typically found in smaller ranges or more stable environments.
- **Broadcast spawner**
    - A reproductive strategy employed by r-selected organisms which distribute thousands of gametes or embryos into a natural system such as the wind or ocean currents which can achieve a very wide range distribution
- **RAD-Seq**
    - Restriction site-associated DNA sequencing. DNA is fragmented by a restriction enzyme which binds consistently at short sequence motif locations, this is followed by library preparation, amplification and sequencing.
- **TSS**
    - Transcription start site – position within a gene where RNA polymerase initially begins to transcribe the gene's DNA sequence into RNA
- **WGDs**
    - Whole genome duplications – ancestral events during which the entire complement of genetic material in an organism is doubled
- **PSMC**
    - Pairwise Sequentially Markovian Coalescent – population size estimation algorithm which uses the distribution of distances between genomic variants to infer variable mutation rates throughout time, originating from estimated effective population sizes.
- **Facultative parthenogenesis**
    - The capacity within an organism for both sexual reproduction, and some form of either self-fertilisation or the production of offspring from unfertilized embryos.

1. Chapter 1: Introduction

1.0. Genomic Evolution

The genome of an organism has been widely recognised for a long time as the primary source of information which facilitates its existence. Although dependent upon a vertically available molecular context (no isolated genome in a test tube will spontaneously sprout into a human being), the genome yet contains the information required to propagate and recreate that context (Eisenberg & Levanon 2013).

Increasingly, in the modern era, to study evolution means to study genomic evolution. As early as the 70s, although the language may refer to 'sequence changes to homologous macro-molecules' (Sarich & Wilson 1973) rather than single nucleotide polymorphisms (SNPs), or a litany of more specific terms, the notion of information evolving from a 'readable' source formed a stable conceptual basis for the model of genomic evolution we use today.

Small changes that accrue over time to the 'macro-molecule' are widely observed to propagate meaningfully to the context which it describes and propagates. This can mean single SNPs causing genetic diseases – such as a nonsense mutation to a single gene resulting in cystic fibrosis (Cordovado et al. 2012), or several hundred sets of variants with complex (and as yet not fully understood) interactions culminating in probabilistic disease risk, such as the case with Crohn's Disease (Lee et al. 2017). However, these changes it can also confer positive phenotypic attributes too. The 'short-sleep' phenotype: a change in a transcriptional repressor can confer increased vigilance with less sleep in mammals (He et al. 2009). The 'supertasting' phenotype, caused at least in part by mutations to the TAS2R38 gene possessed by approximately a quarter of the population confers heightened sensitivity to sense bitterness in food (Hayes et al. 2008).

Genomic DNA can be damaged and mutated through the process of repair. However, evolution also has a base-rate of generative variation which arises naturally from the error present in the molecular machinery which copies DNA during the creation of gametes, and during their early cell divisions in the blastocyst (Johnson et al. 2000). The sum of these effects can be described as a background mutation rate, which for humans is approximately 3×10^{-8} mutations per nucleotide per generation (Xue et al. 2009).

Although the focus in humans tends toward near term large phenotypic alterations from small variants (small relative to the 3Gb genome size), there are a huge variety of alternative genomic evolutionary stories which demonstrate how massive retained changes to the whole molecule manifest in diverse ways. For example, the bdelloid rotifer, *Adineta vaga* has been shown to have an

ancestrally tetraploid genome, with local rearrangements incompatible with ordinary meiotic evolution. Copies of allelic fragments of DNA may be in high sequence identity, but with low collinearity of gene order – or the inverse, lower sequence identity, but with collinearity preserved (Flot et al. 2013). The genome continues to evolve by, it seems, becoming desiccated during the organism's life, and repairing itself – often with rearrangements and duplications (Gladyshev & Arkhipova 2010).

An opposite case comes from Cephalopods, particularly Octopi and Squid – whose genomic evolution has been slowed substantially by the positive selection for RNA-editing genes. The ability to edit post-transcriptional copies of the genes, in order to introduce flexibility and plasticity to the proteome, comes at the cost of limiting the ability of the edited genes to be successfully mutated in viable organisms (else the editing machinery should fail to function) (Liscovitch-Brauer et al. 2017).

In the lab, rather than in the wild, many-generation genomic evolution experiments have been conducted on the rapidly evolving *Escherichia coli* bacteria. In a 50,000-generation experiment – it was found that evolutionary effects which occurred on a whole genome level were more prominent than any single point mutation. The genome size drifted downward by 0.5-1% of the total size – hypermutability was initially present in some populations but decreased as the consequences of genetic load outweighed any adaptive benefits, and the populations tested split evenly by the original hypermutability property. Six populations accrued 60-100 mutations over 50K generations, whilst the other six accrued 1,000-2,000 (Tenaillon et al. 2016).

From the above examples we can see that whilst the story of individual phenotypes is often described in terms of small sets variants, the long term evolutionary trajectory of an organism can be thought of as a process which originates from the effects applied to the information contained in the genome as a whole.

1.1. Phenotype determinants

The conceptual models of genotype to phenotype, and the causal relationships therein, have been long studied (Cavalli-Sforza & Feldman 1976). Through the medium of eQTL studies (expression quantitative trait loci) in humans and other model organisms, the genetic basis for many phenotypic traits have been clearly established (Gilad et al. 2008) (West et al. 2007), of particular interest are those studies of human disease (Westra et al. 2013). However, genetic phenotypic association studies have also revealed a substantial limitation in the form of missing heritability (Manolio et al. 2009). Missing heritability is simply the unaccountability of apparently genetically describable diseases with inheritance. Put differently, many human traits, despite conventional SNP-based genetic correlations are also found to have non-mendelian inheritance patterns. If the only

information available is the genomic sequence of the individuals, and the phenotypes produced, it may seem that there are occasions where traits inexplicably manifest at dramatically higher or lower rates than simple meiotic recombination ought to prescribe.

As work with DNA methylation and large scale epigenomics is now revealing (Barros & Offenbacher 2009), the exact DNA sequence is only one determinant of phenotype. The dramatic differences in phenotypic outcomes can be seen in the differences in lion x tiger hybrids, where the weight and coloration of offspring varies radically depending on the species of the mother (Mckinnell & Wessel 2012). *Brassica oleracea* (commonly known as both broccoli and/or cauliflower), has such a remarkable plasticity in phenotype that members of the public would intuitively regard the morphotypes to be different species. Its genome has been found to be highly polymorphic, and highly cytosine methylated (Salmon et al. 2008). DNA methylation is now studied extensively in related to cancer (Jones & Laird 1999) (Dawson & Kouzarides 2012), and other conditions such as diabetes and obesity (Slomko et al. 2012). This expansion of the domain of study is an example of how additional extra-genomic intrinsic information sources are being discovered and integrated into our understanding of the determinants of phenotype.

DNA sequence, epigenetic marks such as DNA methylation, or histone modifications, pre-existing metabolic and signalling pathway activities in the maternal cell line are all information inputs to the allometric synthesis. However, as invertebrates, bacteria and plants all show, there is a distinction to be made between phenotype and morphotype (Padilla & Savedo 2013), whereby morphotypes are typically presented as the distinct modes of physical organisation availed by the phenotypic outcome. For example, colonial invertebrates such as the coral *Pseudopterogorgia bipinnata* live in highly structurally modular arrangements, great morphotype variety is possible between colonies in different environments despite an absence of genomic differences (Sánchez et al. 2007). Freshwater snails *Physa acuta* have different shell morphotypes (spire length and aperture size) which appear to be plastic responses to extrinsic environmental factors, however a single generation's offspring reared in laboratory conditions will result in uniform shell types regardless of the parental shell type (Gustafson et al. 2014).

The considerations posited in assessment of the evolutionary role participation of phenotype and morphotype are ones of environmental response chronicity. Put directly, the selective advantages of morphotype changes are associated with individual acute single generation plasticity, whereas phenotypic advantages concern chronic multi-generational population level adaptivity. Adaptivity itself is also a nebulous term as this can refer to the 60 million years nature took to arrive at fungal wood decomposition (Floudas et al. 2012), or it can refer to a selective sweep over several dozen

generations which confers toxin immunity to an entire population, as has been the case with insecticide resistance (Oakeshott et al. 2003). We can provisionally separate these two adaptations into two categories: 1) info-genic; and 2) info-latent. Info-genic adaptation (the hard way) is the eventual creation of new biological mechanisms which interface with the environment in entirely novel ways. Info-latent adaptivity is the availability evolutionary of short-cuts available to restore previously lost function, or to experience the stochastic gain of new function from previously neutrally selected copies of functional genes.

Given that much of adaptivity can be thought of as the gradual deployment of prior genetic reserves which hold sudden new adaptive advantages, we might consider that this is not particularly different conceptually to the genetic or epigenetic information which encodes morphotype variation. This has been the basis for a theory of adaptive phenotype evolution as a consequence of mutations which limit organism plasticity to a single morphotype (Hughes 2012). Plasticity-relaxation-mutations may remove competitive purifying selection on alternative morphotype genes by eliminating initiation of the developmental pathways which lead to them. This could allow us to think of plasticity in morphotype as a potential precursor to adaptivity in phenotype. On the flip-side this may suggest that single-morphotype organisms may yet contain the evolutionary pathway to genetic reserves required to re-activate ancestral plasticity.

The information inputs to biogenesis are therefore not just the source of the eventual form, but the information which may be redundant in the typical individual lifestyle case can be thought of as the latent potential for that organism's flexibility towards an environmental range that consequently may permit its survival in times of change, both near and long term. The selective advantages gained from this are discussed in the next section.

1.2. Diversity, Range and Plasticity

Studies of invasive species often reveal that the capacity for an organism to produce targeted responses to different environmental conditions is a predictor of the population's range (Lee & Gelembiuk 2008). The manner in which the eventual successful phenotype produces these responses is varied, it may occur via rapid adaptive evolution (Prentis et al. 2008), particularly in polyploid species (Te Beest et al. 2012), through the increased fitness conferred by heterosis (hybrid vigour) (Facon et al. 2005), through pre-existing plasticity in the phenotype (Weber & D'Antonio 2000), and through admixture with lineages carrying the genetic basis of adaptive traits (Keller & Taylor 2010).

Consequently, more general theories abound regarding the relationship between invasiveness and phenotype plasticity, evolutionary history or incidental hybridisation events. In plants, review studies

posit that hybridisation events are a potential trigger for invasiveness (Ellstrand & Schierenbeck 2000)(Ellstrand 2009). This does not exclude other theories which suggest the plasticity of an invasive phenotype may be equipped as some combination of both a 'jack-of-all-trades' hardiness, and the master of a select number of condition response types (Richards et al. 2006).

The origins of the invasiveness potential are multi-faceted, and there appears to be no singular one-size-fits all description for the emergence of this trait, even plasticity-relaxation-mutations as discussion above are pointed to as cues for explosive radiations (Hughes 2012). The genesis, acquisition of, or ancestral re-emergence of adaptive traits in range expansion are all descriptions of the pathways by which useful systems of information come to be active in a context whereby that activity has a self-propagating effect. The corollaries of large reserves of genetic material under neutral selection in invasive species, are its selective advantageousness and thus favourable cost-benefit ratio. We might think of such information capacitance itself therefore as an adaptive trait.

But what is *capacity*? In basic circuitry, a capacitor is a passive element which stores potential energy. In the modern age, thoughts of a 'capacity' for information might go towards the idea of a pre-defined neutral substrate onto which content may be written. In biology this could only be said to be true to a limited extent in epigenetics, whilst in most other sequence-based capacities information *is* the substrate. Entirely new information needs arise via the physical addition of new substrate, or by the replacement of previous substrate. However, it is not so much the mechanical availability of this occurrence (which seems common in molecular biology) as it is the evolutionary tolerance of it which governs the permissibility of information entering and/or leaving the system. We could think of capacitance as the ability to receive, retain, and even alter sequence substrate of potentially unknown immediate selective value. Some part of a genome which is under no selective pressure to adhere to specific sequence, yet which is perpetually kept in existence by an evolutionary history containing regular incidents of 'unpredictable' advantageous environmental interaction with the latent potential contained within.

This is description is quite abstract, however there are many studies which describe the same thing in different terms. Plants, for example, have been often cited for their frequent incidence of polyploidy, and the benefits gained from it. A review of angiosperm genomics concludes that full duplications and allopolyploid hybrids (combining copies of two full genomes from difference species) were a major driver of the plasticity and adaptivity which these plants used to generate their substantial global biodiversity we see today (Leitch & Leitch 2008). Although a duplication event is not a perpetually 'free' form of information capacitance, for the organisms which survive it an incredibly large sequence space initially unconstrained by purifying selection is created. In the

water flea *Daphnia pulex,* a mutation accumulation (MA) study found that the large-scale duplication rate in the genome was substantially elevated relative to other MA studies, despite a comparable single base mutation rate (Keith et al. 2016). Although many duplicated regions had also been selected against, the copy number variation between populations was high. The researchers posit that this mechanism acts as a generator for the rapid evolution, which it has been shown to be capable of in other contexts (Colbourne et al. 2011)(Geerts et al. 2015). One such gene duplication occurring as a result of these copy number variations (CNVs) has been demonstrated to be of ecological importance in *Daphnia's* tolerance of protease inhibitors (Schwarzenberger et al. 2017).

These two examples involve the duplication of DNA in some way. The example of *Daphnia's* CNV generation shows a potential case where the constant creation of new substrate over its evolutionary history is an example of its capacitance for information (Keith et al. 2016). Whilst the polyploidy-biodiversity of angiosperms shows how the benefits of the this free-information domain manifest (Leitch & Leitch 2008), even if the events which create them are much larger and further between. However, these information types are both quite similar, and their advantages arrive via the creation of alterable copies of pre-existing structures, which research has indicated to be of adaptive merit. The way in which we might identify organisms as creating/selecting for information capacitance as an evolutionary feature, therefore, hinges on how we might dissect and classify the advantageous traits of the information itself.

1.3. Axes of Genomic Evolutionary Flexibility

To return to the angiosperms, as described earlier, much of their biodiversity in modern times is in part attributable to polyploidy, however the ratio of successful whole genome duplications (WGD) originating from the mid-Cretaceous into the Cenozoic, compared to the abundance of detectable events from earlier eras suggests that if WGDs events were occurring at this time, most of those polyploid lineages have become extinct (Van de Peer et al. 2009). In contrast land-based vertebrates have an incredibly scant comparable history of WGDs at all. One example, *Xenopus laevis,* is an exceptionally rare example of a recent WGD in a terrestrial vertebrate. It may not be a co-incidence that it is also extremely invasive, currently present and expanding its range on four major continents, and is regarded by some as biosecurity risk given its further global invasive potential (Measey et al. 2012). It might simply be the case that the molecular biological landscape in which a genome resides has become more forgiving of WGDs than in earlier evolutionary eras. That in all genomes there is a collection of genes which are not duplicated, and that their non-duplication persists across much of the metazoan tree of life (Simão et al. 2015), suggests that the potential for phenotypic instability given an entire genome duplication is incredibly high. It is also the case that whilst the absolute genome size range in metazoans is incredibly large, most organisms of high tertiary phenotypic

complexity possess similarly sized and relatively small genomes 0.5-4Gb (Gregory et al. 2007). Study of bioenergetics which proposes the massive jump in genome size between prokaryotes and eukaryotes to be a function of mitochondrial energy availability (Lane & Martin 2010) also has the corollary of increased demand for endosymbiotic energy generation with genome size. Simply put, it is unlikely that a WGD event *de-facto* expands the mitochondrial energy generation of otherwise equivalent cells to meet the newly elevated demands of genome copying in mitosis and requisite protein synthesis rates. For these reasons we can consider genome duplication one of the least viable ongoing strategies for dynamic information capacitance.

Gene family expansion, as a result of local duplications within a single chromosome (like the *Daphnia* example) (Keith et al. 2016), is another instance by which the diversity and complexity of genomes expand. Given than the genome size increases with this kind of mutation are far smaller than WGDs, this strategy is already liberated from almost all energetic consequences. It also does not require single-copy gene duplication, and as a result is far less likely to result in a non-viable phenotype. This is consequently a far more common mechanism, can be seen in many life forms, and is the subject of large-scale study in humans where the genomic maps of 'Copy Number Variation' (Sudmant et al. 2010) can be tied to genetic disease and other phenotypic traits. One of many examples might be the variable copy number in the human AMY1 gene, whereby the higher copy number appears to have been subject to positive selection for the diversity benefits of multiple amylases in digesting starch (Perry et al. 2007). More general clues as to the adaptive value of expanded gene families can be found in their functional associations, for example the genome of the pacific oyster *Crassostrea gigas* shows that under stress conditions, differentially expressed genes are far more likely to have paralogs (χ^2 test, $p < 1 \times 10^{-10}$) than unaffected genes (Gerdol et al. 2015). Inbred domesticated maize, the results of 10,000 years of selective breeding for desirable traits, have genomes which, when contrasted between strains, vary by upwards of 2,000 copy number variation events. At least an order of magnitude higher than can be found between human genomes (Springer et al. 2009). This suggests that the intense selective manipulation these strains have been subject to, has selected for individuals with ideal gene copy number variations, suggesting these have a far stronger generation-on-generation impact on the organism's phenotype than smaller mutations. CNVs, insofar as they involve functional material, seem to be a highly pervasive and sufficiently effective mechanism of information capacitance.

So far, the timescales involved in the previous two mechanisms have been very different, but still long term. Genome duplications appear to be events whose effects can play out over millions of years (Dehal & Boore 2005), whilst CNVs can guide evolutionary history on similar time scales (Perry et al. 2008)(Springer et al. 2009), it appears these duplications can also facilitate rapid evolution over

hundreds down to even merely dozens of generations (Geerts et al. 2015)(Schwarzenberger et al. 2017). However, organisms often have adaptive needs more pressing than can be resolved by fifty grandfathers. A far more readily accessible store of functional genetics may be also available in the form of allelic variation within the population. The degree to which information is exchanged between metazoan organism populations is often linguistically binned into specific research categories. For example, evidence of information exchange between populations of closely related species, detectable over longer periods of evolutionary history is introgression. The same process after one or two generations is a hybridisation, unless the sometimes very nebulous 'species boundary' between the populations has not been described, in which case it is admixture. Over a few dozen generations of a hybrid population, the genome is described as a mosaic. If the lineage separation history of the organisms performing the sexual information exchange is far shorter, it is merely intra-population gene flow. Here we will attempt to describe these events more simply in terms of the substrate structure which supports their existence: the allele.

To return to basic concepts, the notion of the allele is simply that two or more copies of the same chromosome exist in a genome and are subject to meiotic recombination. The capacity therein for plasticity lies with the notion of divergence between an individual's copies, and evolutionary adaptivity with the divergences between copies held within the total set of organisms capable of exchanging those copies between themselves. This divergence is principally in DNA sequence, but may also exist in the form of epigenetic signature (McDaniell et al. 2010). The point at which the allele concept breaks down occurs as the alleles cease to recombine. However, whilst the 'genotype of recombining alleles driving phenotype' paradigm holds under recombinant meiosis, the phenotypic difference also arguably expands with less recombination, with the extreme end of this scale being sexual dimorphism resultant from chromosomes with only a very small pseudo-autosomal region (Kauppi et al. 2011). This exists as a scale too. In conventional genetics a lesser example may be the localised historical minimisation of recombination in a chromosomal region, which would be described as a linkage disequilibrium block (Kawakami et al. 2014). For example, a study of cultivated maize finds the explanatory power for phenotypic variation derived from SNPs can be increased from ~5% up to 23-34% when they are considered as haplotype blocks described via LD (Lu et al. 2010). Between these two ends of the recombination-phenotype-impact scale we find another terminologically distinct set of studies: those which involve 'divergence hitchhiking' (Via 2012) and 'genomic islands of speciation/divergence' (Feder & Nosil 2010). These theories propose regionally reduced recombination between two alleles in a population due to environmental isolation and/or positive selection causes the phenotypes they generate to become so distinct that speciation may occur. Such effects have notably been observed in *Heliconius* butterflies (Pardo-Diaz

et al. 2012), in Atlantic Cod (Karlsen et al. 2013), and many more in an expanding field of speciation study (Wolf & Ellegren 2017).

The adaptive merit of multiple haplotype availability arises from the diversity of molecular responses available to varied environmental challenges. It cannot really be said that any evolutionary system 'knows' that the diversity of its alleles will be relevant to future challenges, yet that they may be tolerant of high levels of diversity may function as a selective or performative trait which in the long term ensures survival. Whether the diversity of these alleles originates from hybridisation, sheer mutagenesis rates in massive population sizes, or long-term linkage disequilibrium effects, there appears to be some consistent set of benefits to be found in the divergence. However, allelic diversity does not come without its own costs. Heterosis is not inevitable, and outbreeding depression as a result of allelic incompatibility (Frankham et al. 2011) is also a common result of hybridisation. In some cases, such as the Ambrosia Beetles, outbreeding depression is far more likely than inbreeding depression (Peer & Taborsky 2005), suggesting some species are extremely intolerant of allelic diversity. Principally we may regard the organism's mitigation of these costs as its tolerance, and therefore as a measure of its information capacitance with respect to allelic divergence.

1.4. The Aims of this Thesis

This thesis is divided into three data chapters each of which address two major aims. This structure was chosen, as opposed to six separated chapters, for reasons of continuity are delineated in the following outline.

1.4.1. Chapter 2

Allelic diversity is identified as one of the most dynamic mechanisms of information latency in genomics. The boundary between autosomal alleles and partially non-recombining haplodiploid chromosomes might not always be absolute. Investigating the mechanical and systematic limits of allelic information latency was best performed by looking at cases where the divergence between alleles has been pushed to the extreme.

1.4.2. Aim 1

To assess the allelic diversity metrics of two highly divergent invasive global species

This involves discovering the tolerances of absolute base sequence divergence between alleles within the same genomes. Due to the fraught nature of genome assemblies with divergent alleles, making these assessments requires some development of bioinformatic methods. Two genomes are

investigated: A terrestrial earthworm *Lumbricus rubellus*, and a marine brachiopod *Lingula anatina.* This aim is addressed by assembling two genomes, aligning their divergent alleles, and measuring the distribution of the allelic sequence diversity between them.

1.4.3. Aim 2

To describe the potentially acclimative or adaptive information present in hyper-divergent alleles

This aim requires the functional characterisation of the sequences subject to extreme allelic divergence. Protein families with clear potential for dynamic environmental interactions will be described, and the bi-allelic regulatory mechanisms of these proteins further investigated.

1.4.4. Chapter 4

Much of the generation of new genes is duplicative, and the historical rate of duplication is likely to be variable. Different gene families expand at different rates, and to different upper limits. Overall, the total relationship between duplicative information redundancy and specialised complexity in biological sequence is something which might be described in a generalised formal manner.

1.4.5. Aim 3

To develop a general theory of redundant information structures

This aim is to create a mathematical model of information structure as contained within sequence information. This model aims to expand the dimensionality of current *k*-mer based methods, whilst avoiding the obfuscation which occurs in solution space reductions used by other entropic signature methods.

1.4.6. Aim 4

To implement and apply the developed theory to different sequence types

The information structure model measurement tool is implemented in C++. The software is applied to different singular sequence domains of known characteristics, and the results are assessed for their reflection of those characteristics.

1.4.7. Chapter 3

An organism which is relatively new to NGS research and, given its life history, in a position to be maximally in need of varied genetic flexibility sources was analysed. The objective being to observe a situation in which information redundancy, or latent adaptive potential, ought to be found, and to describe what occurs.

1.4.8. Aim 5

To assemble the genome of a species with high theoretical need of both adaptive and acclimative mechanisms to cope with its environment

Specific methods required to obtain a high contiguity assembly with low allelic inflation is needed to accurately obtain a picture of the information structures present in the genome. This aim involves solving this methodological challenge and characterising the allelic diversity content of the genome. It is also necessary to perform gene prediction and produce a high-quality reference gene annotation file for usage with Aim 2.

1.4.9. Aim 6

To discover the simultaneous roles of acclimative plasticity and atavistic adaptivity mechanisms in an organism under high environmental stress

This aim involves obtaining a systems-level description of molecular mechanisms by functional flexibility is achieved by an invasive earthworm. The endpoints investigated are microRNA and gene expression patterns, and differential methylation between conditions. Many organisms have been shown to exhibit DNA methylation in different manners, therefore it is also crucial to the understanding of methylation as a system process to create gene-models of methylation distribution, to contextually inform the differential analysis.

2. Chapter 2: Genomes of Two Invertebrates Suggest Unprecedented Degrees of Divergence May Exist as an Ongoing Intraspecific Adaptive Strategy

2.0. Introduction

Despite being a focus of evolutionary research for centuries, the genetic basis for speciation is still poorly understood (Arnegard et al. 2014). Understanding how reproductive barriers evolve between populations remains one of the most fundamental challenges in evolutionary biology (Coyne & Allen Orr 1998)(Gavrilets 2004).

The ultimate test of hybridization potential is whether the alleles of the two organisms are compatible, because if this is not the case, they will not produce viable offspring. The consensus is that as populations become genetically differentiated, homologous alleles are no longer able to recombine, and hybrids between these populations start to become inviable (Pogson 2016). This assumption has led to the prediction of "islands of divergence", which are regions of the genome that have gradually spread out from alleles under selection, via genomic hitchhiking (Cruickshank & Hahn 2014). Taxa in which these 'islands' have now been identified include Atlantic Cod (Sodeland et al. 2016), Mosquitos of genus *Anopheles* (Turner et al. 2005), Sunflowers (Renaut et al. 2013), and Butterflies (Martin et al. 2013). One hypothesis is that these islands may develop sympatrically, due to either non-recombinant regions of DNA, such as in the case of large-scale inversions, or via more sophisticated mating relationships, for example due to wing pattern differences in *Heliconius* butterfly populations. This hypothesis is usually presented in contrast to one of allopatry, whereby differences between populations accumulate under divergent selection in isolation (Rieseberg et al. 1999).

For some species, it appears that divergent selection acts additively on unlinked loci, inferring adaptation of different populations to different environments. These populations become specialised to different niches, and hybrids between them have a lower fitness than would be expected from an intermediate phenotype (Arnegard et al. 2014). However, recent studies of hybridisation have demonstrated a different effect: In the case of two populations of *Ciona intestinalis*, 3 million years was still not enough to prevent the viability of their offspring(Roux et al. 2013). Typically, it seems, of the broadcast spawning marine invertebrate, the pacific oysters *Crassostrea gigas* and *Crassostrea angulate*(Huvet et al. 2002) are known to frequently hybridize, and further hybridisations are possible within the genus (Y. Zhang et al. 2012)(Allen & Gaffney 1993), as are the blue mussels *Mytilus edulis* and *Mytilus trossulus* (Shields et al. 2010). As such, it has been observed that the species barrier in

broadcast spawning invertebrates is particularly porous. This has been assumed to be a peculiarity of the clade (Shields et al. 2010), however, as sequencing technologies and bioinformatics genomic analyses improve, and more invertebrate genomes are sequenced, it remains to be seen if these cases are indeed the exceptions or the rule. Broadly speaking, hybridization does appear to be able to act as a mechanism for gaining evolutionary adaptive potential through a selective advantage of introgressed genes; whether that be from Neanderthal to Human(Ding et al. 2014), within Conifer taxa(Ru et al. 2016), or between species of mice (Song et al. 2011).

A recent study has attempted to combine the islands of divergence and hybridization/introgression fields of research, theorizing that in some situations gene-flow may be possible between speciating populations within islands of divergence, implying that these islands have the potential to contribute to adaptive introgression (Martin et al. 2013). Evidence is presented here for far more extreme introgression behaviour: Full genomes which are comprised of a dense mosaic of divergent alleles. Rather than the typical mosaic of a recent hybrid, these genomes suggest evidence of long-term allele frequency homogenization between genomes of previously isolated cryptic 'species' on a scale of 10-30% absolute divergence. It can be argued that this genome characteristic is a feature of the common aspects of their respective life histories (short-lived, highly fecund) (Romiguier et al. 2014).

This phenomenon has been observed in two organisms with highly contrasting evolutionary and ecological backgrounds: *Lingula anatina* is a marine broadcast spawning brachiopod, referred to by Darwin as a 'living fossil' due to it superficially appearing to have remained morphologically un-changed since the Silurian (Way et al. 1994). More recent whole-genome analysis concluded that *Lingula anatina* have in fact been rapidly evolving, including the rapid duplication of many genes(Luo et al. 2015). The published genome is re-analysed, and the assembly of a new genome is presented: the European earthworm *Lumbricus rubellus*. The results identify that due to this previously undescribed mechanism, these two organisms appear to be able to sustain extraordinary levels of diversity within their genomes, likely due to hybridization between cryptic lineages or species. This chapter then begins to investigate the mechanical genetic capacity for such tolerance and seek to define a broader set of lifestyle conditions for such species boundary porosity.

2.1. Materials and Methods

2.1.1. Earthworm (*Lumbricus rubellus*) sample collection and sequencing

A draft *L. rubellus* genome was assembled from a single individual (referred to hereafter as S18). Samples were collected by Prof A.J. Morgan, DNA was isolated by Dr L. Cunha and provided to Leiden Genome Technology for long read PacBio sequencing and Wellcome Trust Centre for Human

Genetics for Illumina short read sequencing. Briefly, *L. rubellus* genome was assembled from a single individual (referred to hereafter as S18) sampled from Cwmystwyth (52°21'23.1"N 3°45'37.9"W: 52.356407, -3.760527), a former lead mining site in South-West Wales. The earthworm was collected and maintained at room temperature from soil harvested from the collection site for transportation to the Cardiff laboratories. The sample was depurated by placing the live earthworm on moist filter paper 14^{o}C for 48 hours to remove all soil or organic material from the gut. The earthworm was then snap frozen in liquid nitrogen and stored at -70^{o}C for long term archival. DNA was extracted using the material ~5 segments posterior of the clitellum to the tail of the organism. The segment was ground under liquid nitrogen and DNA was then prepared using phenol method for isolation very-high-molecular-weight DNA (Wood 1983). DNA was quality controlled using absorption spectroscopy with integrity and size being determined using agarose gel electrophoresis (0.4%) using Lambda HindIII and undigested lambda (New England Biolabs) as size markers. DNA (6 µg) was prepared for long read sequencing using the PacBio C2 chemistry with the XL enzyme (Pacific Biosciences, Inc.) and 8.6 Gbp of sequence generated using over 12 SMRT cells. A secondary aliquot of DNA (2 µg) was used to generate TruSeq PCR-Free pair end library with insert size of ~450 bp (Illumina Inc.). This library was sequenced on a HiSeq 2500 platform (Illumina Inc.) using a rapid run mode and HiSeq V2 chemistry yielding 71 Gbp 150 bp pair end data.

2.1.2. Lumbricus rubellus Genome assembly

Short read adapter removal and read trimming was performed with the software 'Trimmomatic' (Bolger et al. 2014). K-mer spectrum error correction was then performed on the short read library with 'Musket' program (Liu et al. 2013). The initial draft genome assembly was then performed with the 'Platanus' assembler(Kajitani et al. 2014). This yielded an assembly with an N50 of 5.8 Kb, and 531 k scaffolds. The assembly was then re-scaffolded by employing the PacBio 1 Kb+ read library with the program 'SSPACE-Longread' (Boetzer & Pirovano 2014), finally gap closing was performed using the 'GapCloser' program from SOAPdenovo2 assembler package (Luo et al. 2012). This improved the N50 to 6.3k, and reduced the scaffold count to 231 k, with a total length of 831 Mb. Although preferable assemblies could be obtained with alternative pipelines, the 'bubble collapse' graph method employed in the Platanus pipeline results in an assembly in which inflation by allelic variation only occurs in the case of more extreme divergence. Broadly speaking, only the highly variable alleles will split into separate graphs. Repeatmodeller (Smit & Hubley n.d.) was used to model genomic repeats in both *Lumbricus rubellus* and *Lingula anatina*. Repeatmasker (Smit et al. 1996) was then further employed to predict the abundance of those repeats in both, and to produce masked versions of the genomes. Finally the software Freebayes (Garrison & Marth 2012) was used to call variants in the assembled genome sequence.

2.1.3. The limitations of string graphs

Conventional assessment methods employed to measure allelic divergence rely on a 'haplotype' genome assembly. This is a genome sequence which represents a consensus of all the alleles present, onto which the smaller variations can be mapped. However, as the alleles become more and more divergent, achieving consensus between them becomes a process of increasingly insufficient compromise. With 1 in 10, 1 in 5, or even 1 in 3 bases altered between alleles, including major structural changes, the haplotype genome is no longer a realistic proposition.

There are various tools available to reconcile divergent alleles. One is HaploMerger2 (Huang et al. 2017), which is described as suitable for diploid assemblies with "high heterozygosity (3%)". It requires a pre-improved assembly with an N50 above 100 Kb, which no attempt at the *L. rubellus* assembly came close to meeting (N50 ~6 Kb). Another is dipSPAdes (Safonova et al. 2014), which has been demonstrated to improve assemblies of highly polymorphic genomes varying from "0.4 – 10%". An *L. rubellus* assembly with dipSPAdes was running on a 96 CPU-core 2 TB RAM compute node, the process was terminated after it did not complete within two months.

The 'Platanus' assembler was also designed to assemble highly heterozygous genomes(Kajitani et al. 2014), and was used as the baseline assembler for *L. rubellus* in this case. Figure 5 suggests that regions with 1-2% allelic divergence were successfully collapsed into a consensus haplotype. The methodology suggested in Vinson et al (2005)(Vinson et al. 2005) was also attempted, but again, the *L. anatina* and *L. rubellus* genomes were too polymorphic.

The problems with genome assembly evidenced here are brought about due to the extreme difficulty of solving the *de Bruijn*-graph for highly polymorphic regions, whereby both alleles exist undifferentiated within the same sequencing library. Although the compromises of haplotype consensus are usually enough for most higher vertebrates, for invertebrates tolerant of such extremes of divergence it may be necessary for future projects to aim towards the separate assembly of allelic sequences.

Although there are presently haplotype 'phasing' software available (such as haptree (Berger et al. 2014), and hapcompass (Aguiar & Istrail 2012)) – these rely on extrapolation from a collapsed consensus sequence, and are not likely to be effective in this case. The alternative – to assemble separate haplotypes – could be achieved with molecular methods such as "linked read" library preparation. Zheng *et al*, describe successfully using a linked-read library preparation method to capture haplotype specific information from human cancer cell genomes (Zheng et al. 2016).

2.1.4. Preliminary genome characterisation of *Lumbricus rubellus*

k-mer frequency histograms and read coverage frequency histograms for *Lumbricus rubellus* were generated (Figure 5). These histograms clearly show a collapsed and a non-collapsed portion of the genome. A distribution of the polymorphism rate for the collapsed component of the genome was generated using the variant information from Freebayes (Garrison & Marth 2012), across 5 kb and 40 kb window sizes using a custom Perl script (Figure 6). The proposed hypothesis is that these non-collapsed and collapsed regions are due to extreme variations in polymorphism rate between large portions of each chromosomal pair, that are leading to an inflated genome assembly, and as such will be referred to as 'divergent' and 'non-divergent' regions respectively.

Preparatory scripts mentioned here are available on GitHub: (https://github.com/OliverCardiff/Useful_R_and_Perl_Scripts/tree/master/Variants)

2.1.5. Analysis of *Lingula anatina* genome assembly

To investigate if this extreme allelic variation in *Lumbricus rubellus* is indicative of a wider phenomenon, a preliminary assessment of allelic diversity on published Lophotrochozoan genomes was carried out. Seventeen genomes, and their respective read libraries were downloaded from the NCBI short-read archive (SRA). Reads were re-mapped with bowtie2 (Langmead et al. 2013), subsequently FreeBayes (Garrison & Marth 2012) was used to call variants. Rolling means of 50 bp and 10 kb of allelic diversity were plotted alongside read depth to give a summary of allelic diversity. *Lingula anatina* was chosen to be analysed further in the present study, due to the extremely bimodal read depth, and the correspondence of this read depth to the polymorphism rate.

Data used in this paper for *Lingula anatina* was made public by Luo, et al, along with their publication titled: 'The *Lingula* genome provides insights into brachiopod evolution and the origin of phosphate biomineralization' (Luo et al. 2015). The PacBio sequence library was downloaded from the SRA (short-read archive) at NCBI (https://www.ncbi.nlm.nih.gov/sra) the accession code is SRX1119733. It consisted of 1.1 M sequences, 8.5G bases. The short-paired sequence library was also downloaded from the NCBI SRA. The accession code was SRX1118889. It consisted of 54.1 M read pairs, 23.1 G bases. A set of transcripts representing the draft transcriptome of *Lingula anatina* was retrieved from the NCBI 'Nucleotide' database (https://www.ncbi.nlm.nih.gov/nuccore). It consisted of 43,670 annotated mRNA sequences.

Multi-mapping divergent genome regions for both *Lumbricus rubellus and Lingula anatina* was attempted, using Bowtie2 (Langmead et al. 2013) parameterised for increased flexibility (Figure 7). *K*-mer frequency histograms and read coverage frequency histograms were also generated for *Lingula anatina*, as described above, to compare to *Lumbricus rubellus*. The draft *anatina* genome

was fragmented into regions of consistent read coverage as described in Figure 1. Qualitative assessment of the integrity of the draft assembly was performed as described in 2.3.2.

2.1.6. Identification and Characterisation of the 'divergent alleles'

The proposed hypothesis was that the bi-modal distributions found in Figure 5 are due to highly allelically diverged, regions across the genome in both organisms. If this were the case, it ought to be possible to find a single allelic match for the fragments of each genome that are at half-coverage. The matched sequence being a separate half-depth coverage scaffold elsewhere in the assembly. A set of 5739 candidate pairs were identified for *Lumbricus rubellus,* and 5836 pairs for *Lingula anatina*. The pair-matching method was based on: 1. Unique shared PacBio read alignments with blasr (Chaisson & Tesler 2012), 2. End-to-end alignment identity scores, with Clustal Omega (Sievers & Higgins 2014), and 3. Set filtering based on divergence distributions (can be seen visually in Figures 2, 3 and 4). To validate this approach and demonstrate the mis-assembly of the original *Lingula anatina* genome, open reading frames and protein sequence alignments between allelic pairs were visualised alongside read-depth (see Figure 8). PacBio reads were used solely for sequenced matching, and not for divergence rate estimation, due to their inherent error rate. The full pipeline for the selection and description of these allelic pairs is as follows:

Read depth fragmentation.

The *L. anatina* genome was fragmented based on read-depth coverage, using a custom algorithm implemented in C# (https://github.com/OliverCardiff/HollowScaffolds). This program determined whether a region was half- or full-coverage based on a combination of 2 rolling-means of large and small window size (5 kb, and 500 bp). The 10-bp bin read-coverage data (from Figure 5 (right half)) was flagged as half-depth if either of the rolling means fell below a cut-off point that was halfway between the half and full coverage levels. Stretches of continuously flagged bins were defined as half-depth regions, with fragmentation points (one scaffold split into two) created at their edges. *L. rubellus* draft genome did not need to be fragmented because the scaffolds were already of sufficiently small size to have a mono-modal read-depth coverage (I.e. there were a negligible number of changes in read-depth coverage within scaffolds).

PacBio library alignment with 'blasr'

PacBio reads were mapped to the draft *L. rubellus* and the fragmented *L. anatina* assemblies using the long read alignment algorithm 'blasr' (Chaisson & Tesler 2012). Blasr was parameterised to retain the three best 'hits' per read.

Read depth filtering

Scaffolds from *L. anatina* and from *L. rubellus* that had a read-depth between 75 – 125% the median value of the low-coverage read-depth peak (from Figure 5 (right half)) were selected.

Candidate allele pair selection

All PacBio reads that had anything other than exactly two 'hits' were removed. Additionally, partially mapped reads (alignments that were 80% of read length or lower) were also filtered from the dataset. The pairs of scaffolds that were repeatedly aligned to by the same PacBio read(s) were selected as candidate pairs. For a candidate pair to be selected they had to both be mapped to by at least three PacBio reads. This produced a set of 9708 candidate pairs for *L. rubellus* and 8018 pairs for *L. anatina*.

Candidate pairs aligned with Clustal Omega

Each candidate pair was end-to-end aligned with Clustal Omega (Sievers & Higgins 2014). The reverse complement of one sequence was also aligned to its counterpart, the highest scoring alignment was kept. Distribution of pair sequence divergence was calculated using a custom Perl script ('vienna_toperc.pl', see GitHub link at end of section) and summarised in R (see Figure 3).

Set Filtering

Candidate pairs above 25% divergence in L. anatina, and above 50% divergence in *L. rubellus* were filtered from the dataset. These cut-offs were reflective of a conservative separation between false positive and signal peaks in the divergence distribution. Cut-off is shown as a vertical black line against the distributions in Figure 3. This left a final set of 5739 candidate pairs for *L. rubellus* and 5836 pairs for *L. anatina*.

Successful *L. anatina* pairs used to demonstrate mis-assembly

Allele pair selection (above) was repeated on the non-fragmented *L. anatina* genome. The same read filtering as in 4. was also applied. This was done to visualise the mis-assembly of *L. anatina* draft genome (Luo et al. 2015). This visualisation of the mapping can be seen in Figure 9.

Predicted CDSs within 'successful pairs' extracted with Transdecoder

Predicted CDSs above 300 bp (as recommended by the Transdecoder manual) in both sequences of the candidate allele pairs were extracted with Transdecoder. If here were overlaps between reading

frames, then the longest was selected, and the other removed. Transdecoder (Brian J Haas et al. 2013) also generated the amino acid conversions of these CDSs.

Protein sequences within allele pairs aligned with Clustal Omega

The amino acid conversions from each allele (from 9, above), were aligned with the program 'Clustal Omega' (Sievers & Higgins 2014) to every other amino acid conversion from its corresponding pair using a custom Perl script (aln_proteins.pl, see GitHub link at end of section). The best match was saved for each amino acid sequence and a summary of this data was visualised in R. Some examples of these amino acid sequence pairings relative to the fragment pairs are shown in Figures 15 and 16.

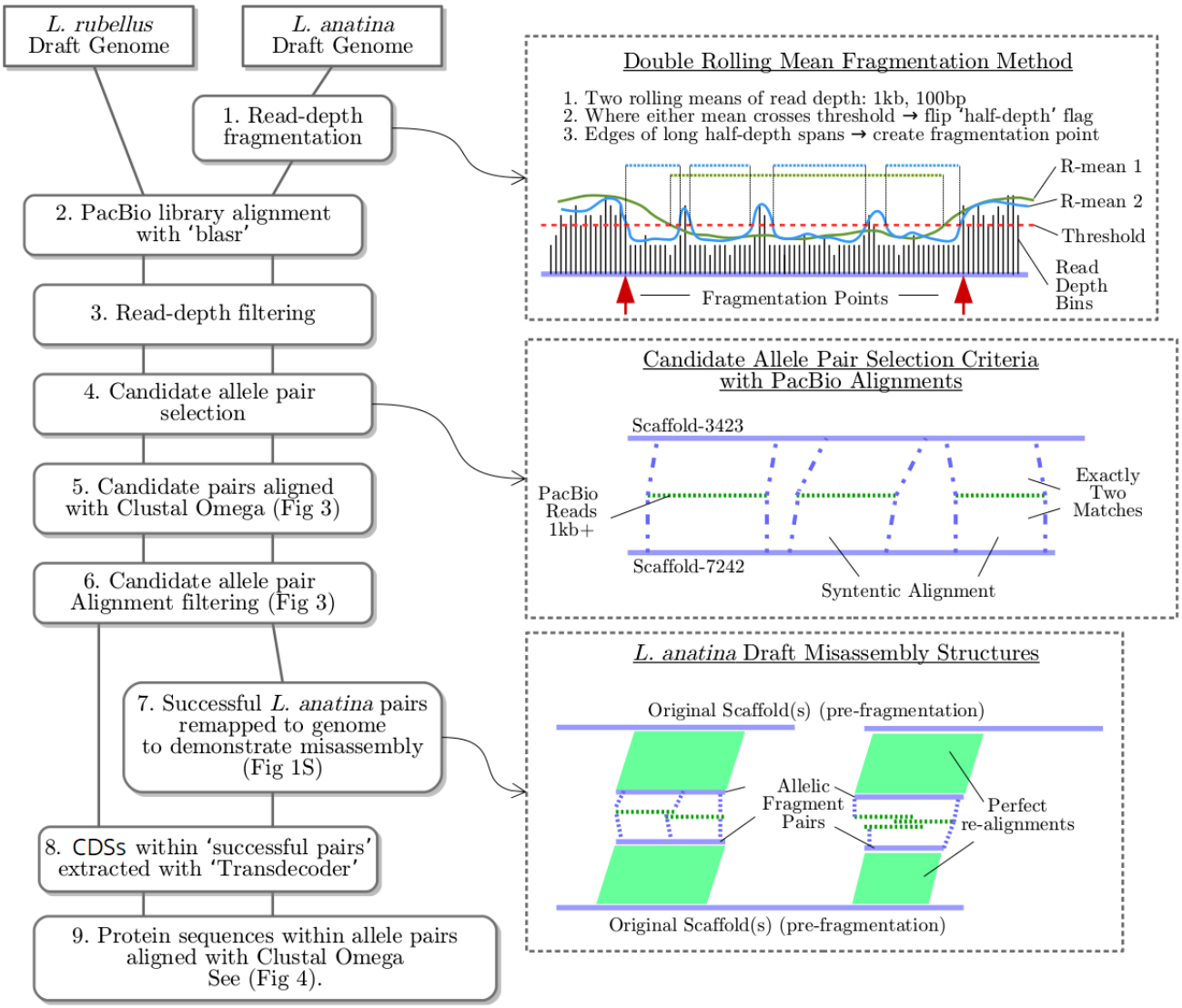


Figure 1. Candidate allele fragment pair identification and analysis flowchart.

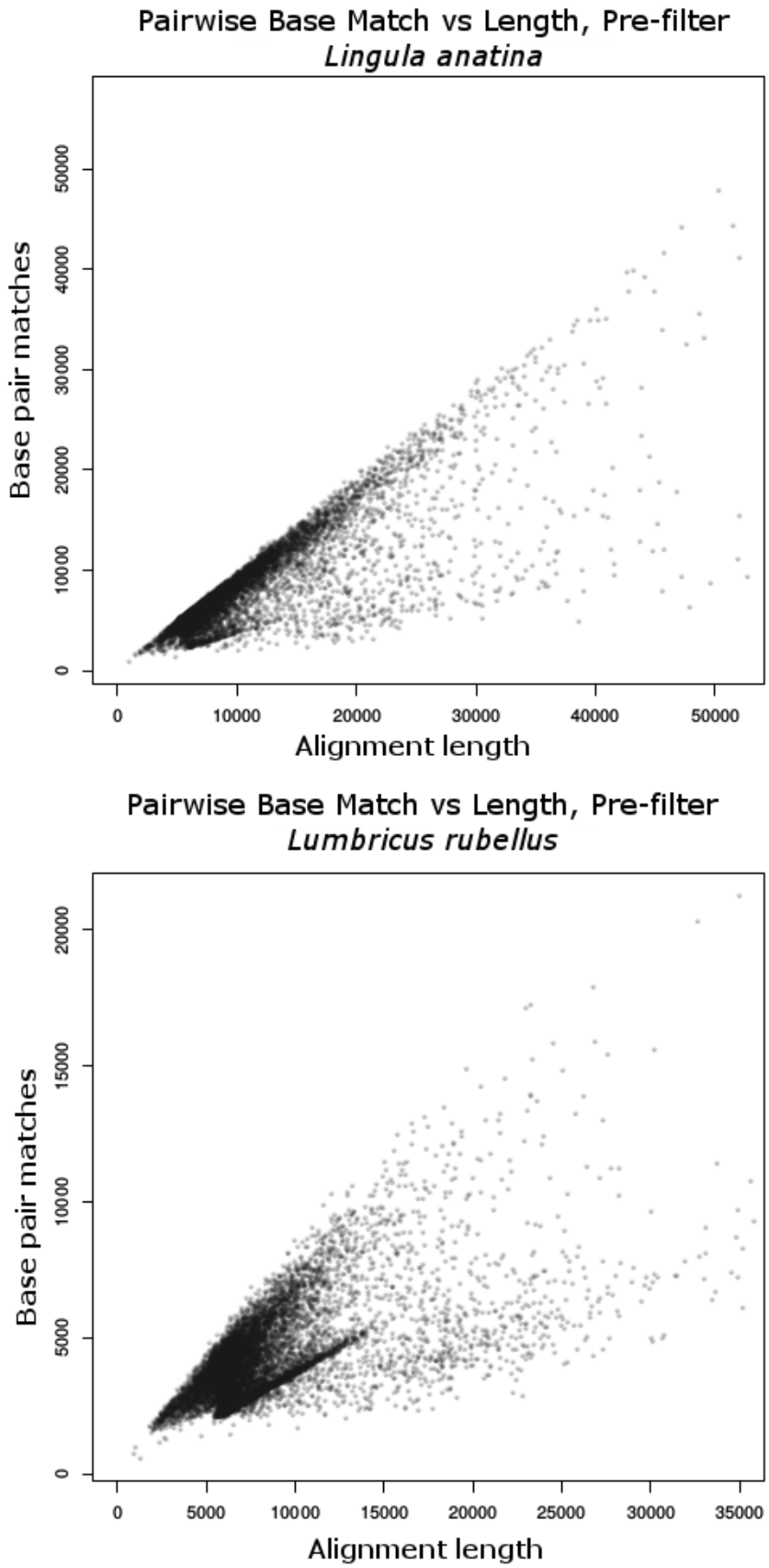


Figure 2. Candidate Pair selection pre-filter results, tracking number of matched base pairs per fragment against fragment length, no filtering.

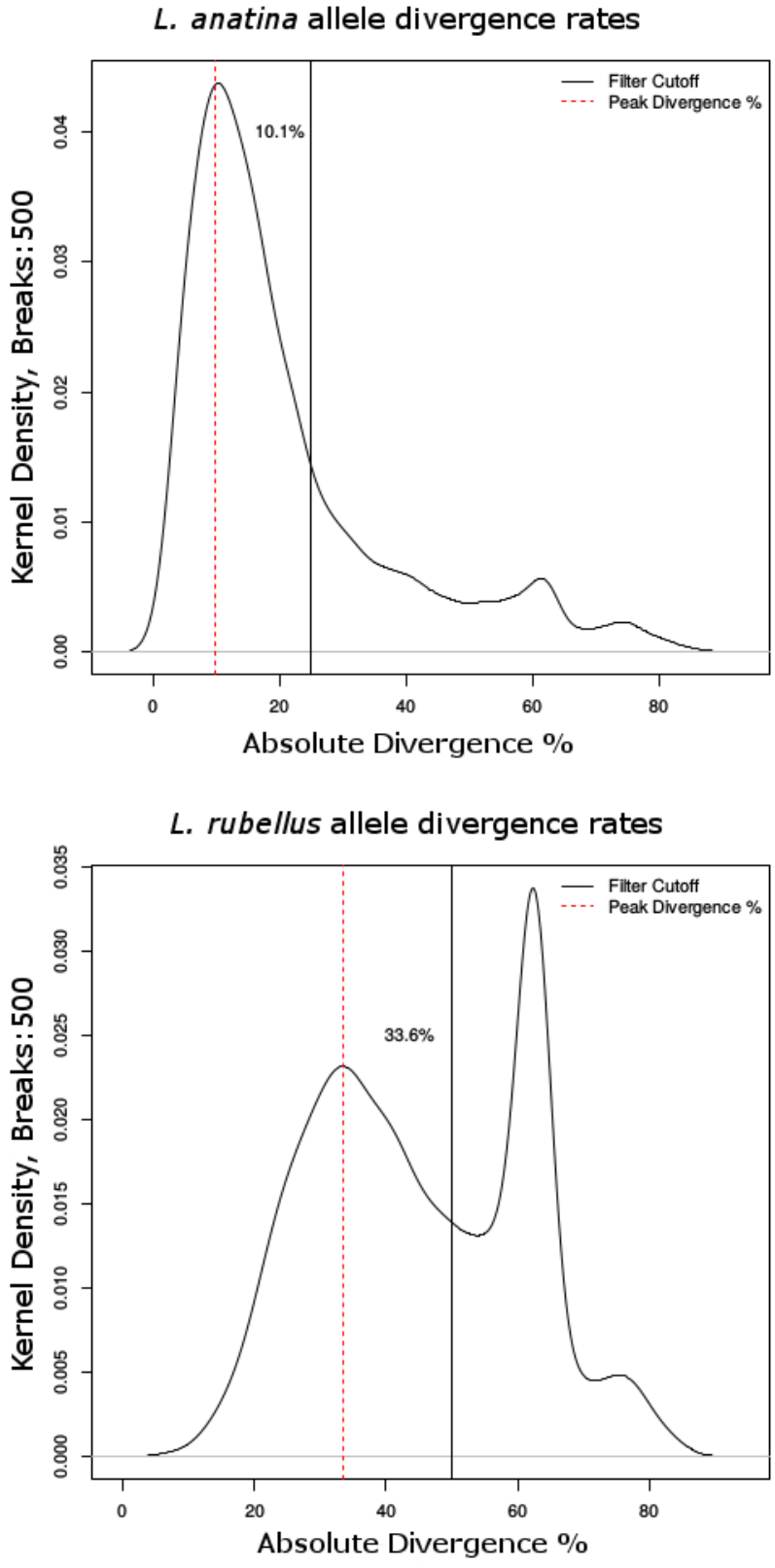


Figure 3. Fragment pair divergence kernel density functions, pre-filtration.

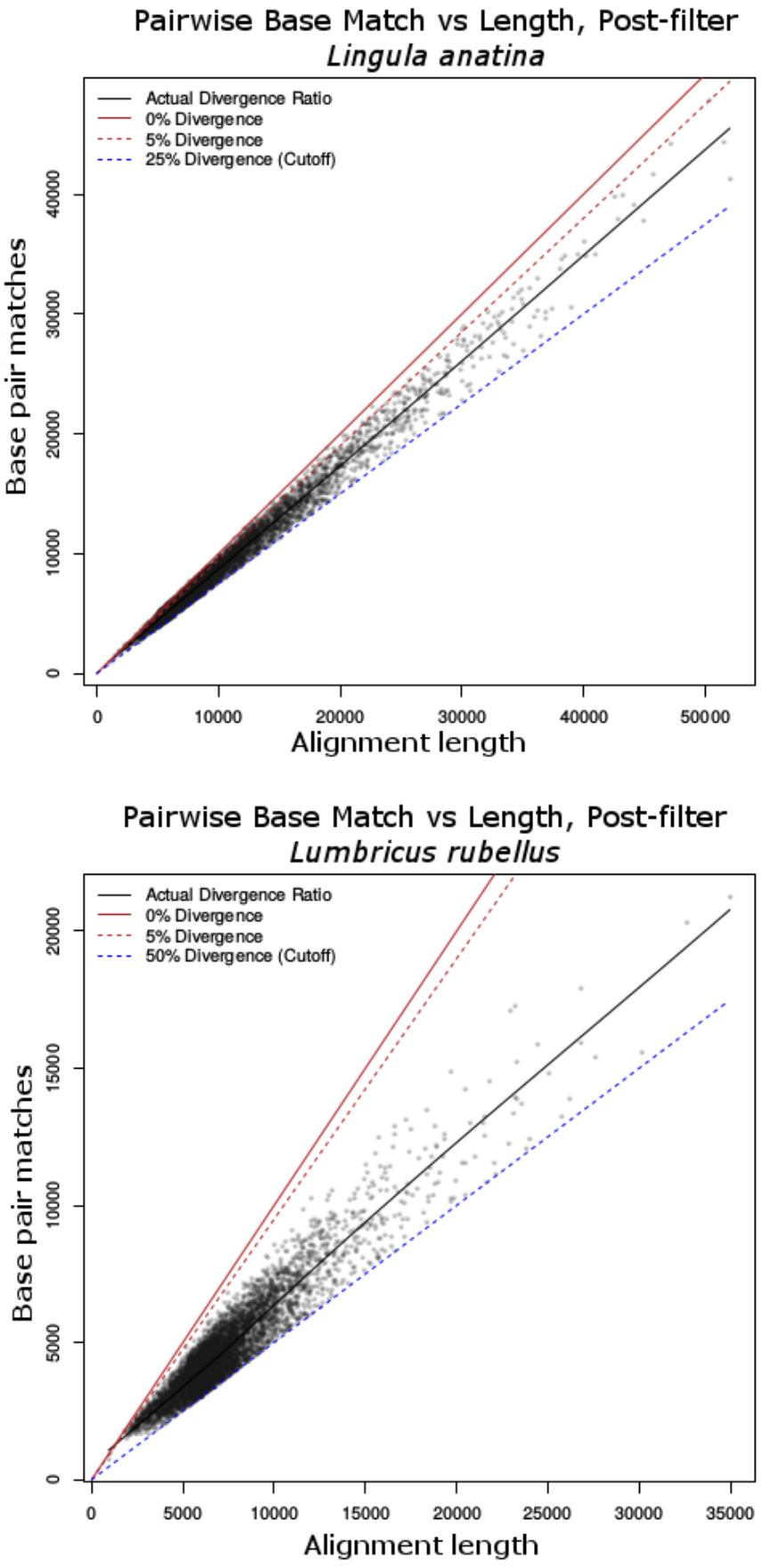


Figure 4. Candidate allele fragment pairs post-filtering, base pair matches per fragment against fragment alignment length.

2.1.7. Nucleotide divergence

To describe the composition of the sequence divergence seen in Figure 11, 12,13 and 14, mutations were categorised as either substitutions, insertions/deletions or transposon-like features. The alignments used in Figures 2-4, were re-analysed by a custom Perl script which sequentially processed the alignment files, categorising sequence changes as substitution or indel, and recorded indel length. Indels longer than 10 bp were categorised as 'transposon-like' features. The gaussian kernel density function of substitutions and indel occurrence rates were calculated over a set of window sizes, from 100 bp to 5 kb, in 200 bp increments.

2.1.8. Protein divergence

As described in Figure 1, converted protein sequences from candidate allele pairs for both organisms were aligned with their counterparts with Clustal Omega (Sievers & Higgins 2014). The set of protein alignments was then filtered for alignments longer than 200AA, with identities of at least 60%, with no indels sequences longer than 100AA within that alignment. Although these are lenient parameters, there were also many cases for both genomes whereby an ORF's opposing sequence had decayed and was no longer capable of encoding protein, or where an ORF appeared to be a divergent duplicate of another, with no equivalent duplication having occurred on the opposing allele. In these cases, the next best possible protein alignment was achieved for the unmatched sequence – which often was of a very low identity (~20-40% the same). This filtering method removed most of these cases. To investigate the distributions of larger scale changes, incidents such as duplications, deletions, inversions, and large indels were also tracked in the protein alignments using a separate Perl script (aln_proteins.pl, see GitHub link at end of section), see Figure 19. The alignments were sorted for relative substitution rate before visualising. Alignments pairs were also annotated by protein family (Figure 20) and six protein families were selected as interesting examples (Figure 21). All visualisation scripts were written in R (R Development Core Team 2016).

2.1.9. Validation of Allelic Nature of *Lumbricus rubellus* divergent regions

The expectation that paired fragments behave like alleles is derived from both shared sequence homology, and the half-depth read coverage observed when a short-read library is re-mapped. Evidence for the hypothetical mendelian behaviour of these fragments as alleles would therefore be seen in the read-depths of a short-read library derived from a second individual from the same population, when a high stringency alignment is performed. If fragment A from a fragment pair receives full read depth coverage from a second individual, the allele hypothesis would expect fragment B to receive a depth within an error tolerance zone around zero. Similarly, if one allele is discovered to be half-depth again, so should its opposite.

To test this theory, a low-coverage genome sequence library was generated (8.6 Gb, 100 bp paired end, approx. 14-fold coverage) from a second *Lumbricus rubellus* individual (S20) from the same population as the full genome. Reads were mapped from individual S20 individual to the original repeat masked genome using bowtie2 (Langmead et al. 2013). Read depths for the set of 5739 fragment pairs previously identified were extracted and examined for congruity with the allele hypothesis using the following binning rules: **Bin A**) The read depths of one fragment are more than twice the level of the other, and the lower coverage fragment's depth is also lower than a quarter of the full coverage level. **Bin B**) Both fragment's read depths are greater than a quarter of the full coverage level. The same analysis was also performed on the fraction of the *L. rubellus* genome where high levels of polymorphism were not observed. This fraction is referred to as the "non-divergent genome". This results of these validation are shown in Figure 25.

2.1.10. Population Genetics of *Lumbricus rubellus* 'divergent regions'

Forty *L. rubellus* worms were sampled from Cwmystwyth mines in south Wales (the same location as individuals S18 and S20) and underwent RAD-tag sequencing. This RAD-tag sequencing included the two highly divergent mitochondrial lineages of *L. rubellus*. RAD-Seq restriction enzymes rely on a specific 8bp sequence to bind with the sample DNA. If any base in that sequence is altered, it is highly unlikely that the restriction site will remain. Since the estimated divergence between rubellus alleles is around 33% (Figure 3), it follows that in many of the cases a restriction site on one allele will not be matched by restriction site at the corresponding location on another. As a result, the RAD-Seq process for this organism will not produce an effective means for 'calling' variants between alleles, due to missing data. However, the missing data itself may yet be informative.

L. rubellus is known to have two major mitochondrial lineages, A and B (Spurgeon et al. 2016)(Anderson et al. 2017)(Giska et al. 2015a). The RAD-Seq individuals were categorised by the unity of two methods:

1. A presence/absence matrix of RAD-tag sequence alignment against all assembled contigs in the genome was generated (Andre et al. 2010). This was then used to produce PCA estimation of population separation (Figure 22 (top)), and a merged matrix of pattern similarities (Figure 24).

2. RAD-Seq data was aligned to two other *L. rubellus* genome assemblies, with known mitochondrial lineages. The per-library genomic alignment rates were treated as two separating dimensions (Figure 22 (bottom)).

The presence/absence dataset used the PCA analysis, was then filtered for sparse entries ($n < 9$), and short sequences (< 5 Kbp). It was then sorted by a lineage-difference metric defined as the sum of

positive lineage A columns minus the sum of positive lineage B columns, the result was plotted in R. (Figure 23)

2.1.11. Motifs Conserved in Divergent Alleles

With absolute sequence divergence evidently high and given the proposed hypothesis that these divergent regions are still able to act as Mendelian alleles, there is the question as to how these pairs of alleles can maintain compatibility. This question was addressed by identifying and describing conserved sequence motifs within divergent regions. This was done by:

1. Using a k-mer analysis (via the Jellyfish (Marcais & Kingsford 2012) software package) that identified and aligned the most frequent 32-mers (excluding microsatellites; Figure 26),

2. Describing the allelic mutation rate within each of these identified motifs (at each divergent allele locus, in a pairwise manner; Figure 27),

3. Aligning these motifs to transcriptomes of both organisms to determine where they occur (Figure 28).

Motifs are DNA patterns which re-occur regularly, they may be internally variable, yet follow some consistent structure. *K*-mer analysis is involved with assessing sub-string frequencies within a larger string, or set of strings, as in DNA sequence. This makes k-mer analysis a suitable starting point for searching, de novo for conserved motifs.

The sets of filtered candidate allele pairs described in Figure 3 were collapsed into consensus sequence. Gaps and SNPs in the alignment resulted in an ‘N’ in the consensus sequence. The consensus sequences were then fragmented into their ‘N’-free sub-sequences (strings split by ‘N’). All sequences longer than 32-bp were retained, the rest eliminated from the set. This produced a set of fragments representing the points of perfect conservation between alleles. This process was conducted separately for both organisms.

The reduced sub-sequence sets then had their absolute 32-mer frequencies counted by Jellyfish (Marcais & Kingsford 2012). The 100 most frequent 32-mers were then extracted from these sets for both genomes.

The k-mers and their reverse complements were then aligned using Clustal Omega (Sievers & Higgins 2014), overlapping k-mers were then collapsed into singular representative motifs. These motifs were then re-aligned to all the candidate pairs using BLAST. All matching sequences and their reverse complements were then aligned in Clustal Omega to create a large multiple sequence alignment file of several hundred sequences for each motif. This file was trimmed at the edges to

include only loci where 80% of the sequences had base-pairs present. These motif alignments were then used to create motif diagrams with WebLogo (Crooks et al. 2004).

After looking for the consistency of the motif overall, the allelic mutation rate within each motif (at each divergent allele locus, in a pairwise manner) was analysed. The two sets of motifs found using the procedure in were re-aligned with candidate allele fragment pairs, using BLAST+(Camacho et al. 2009). Using the local BLAST+ coordinates of where motifs aligned to a single allele pair, a custom Perl script (pairs_filt_and_blast.pl) was used to extract the corresponding region from the alignment of the two allelic fragments. The result in each case being a short two-sequence alignment file containing two allelic copies of that incidence of the given motif. Each of these short alignment files was then assessed for per-loci mutation rates. The total mutation rate for each locus, for each motif was then summarised in R.

The results show that each motif has a central region of substantially lower allelic mutation rates that corresponds to the higher consistency regions seen in the motif letter visualisation. Overall the mutation rates along each motif are always consistently below the mean allelic divergence for the corresponding organism.

To investigate is the conserved motifs had any relationship with transcribed regions, a mapping was performed between the motifs and assembled transcriptomes of *L. rubellus* and *L. anatina*. The Lingula anatina transcriptome was retrieved as described in 'Data Sources'. The *Lumbricus rubellus* transcriptome was provided by Prof. P. Kille from other work in preparation. Bowtie2 (Langmead et al. 2013) indices were built for each transcriptome. The motifs were then aligned to the transcriptomes with the '-all' flag, to keep all possible alignments. A custom C# program was then used to intersect motif mappings with transcriptome annotation files (GFF3 format). There was no notable presence of motif alignments in coding sequences, or 5' UTRs. As Figure 28 shows, the motifs piled up very consistently on the terminal 60bp of mRNA transcripts, in the 3' UTR. Except for just one of *L. anatina's* motifs, all motif aligned in this manner, suggesting strongly that conserved motifs of this length perform similar roles across diverse taxa, and may be crucial components of maintaining regulatory compatibility between divergent alleles

Perl scripts mentioned in this Methods section can be found at the following Github link:

(https://github.com/OliverCardiff/Useful_R_and_Perl_Scripts/blob/master/allele_matches/)

2.2. Results

2.2.1. Preliminary Genome Characterisations

In both organisms there is a bimodal distribution, both for the 17-mer frequency histograms (Figure 5 (top)) and for the read-coverage frequency histograms (Figure 5 (bottom)). For both organisms, the mono-allelic (or lower frequency) peak is higher than the bi-allelic peak for the 17-mer frequency histograms, demonstrating a high proportion of single-allele sequences. In the case of *Lumbricus rubellus* there are approximately 50% more mono-allelic sequences. Clear peaks indicate that the libraries were of adequate quality. Both organisms' assemblies also show highly bi-modal read-coverage, a pattern usually seen when allelic copies have not been collapsed into a single haplotype during the assembly process. These analyses indicate that most of both genomes are subject to extremely high levels of variance. By analysing only regions of full-coverage, it was possible to determine an average polymorphism rate of ~1% for the *Lumbricus rubellus* individual S18 and ~2% the *Lingula anatina* individual (see Figure 6), in the fractions of the genome that are referred to as the "non-divergent" regions.

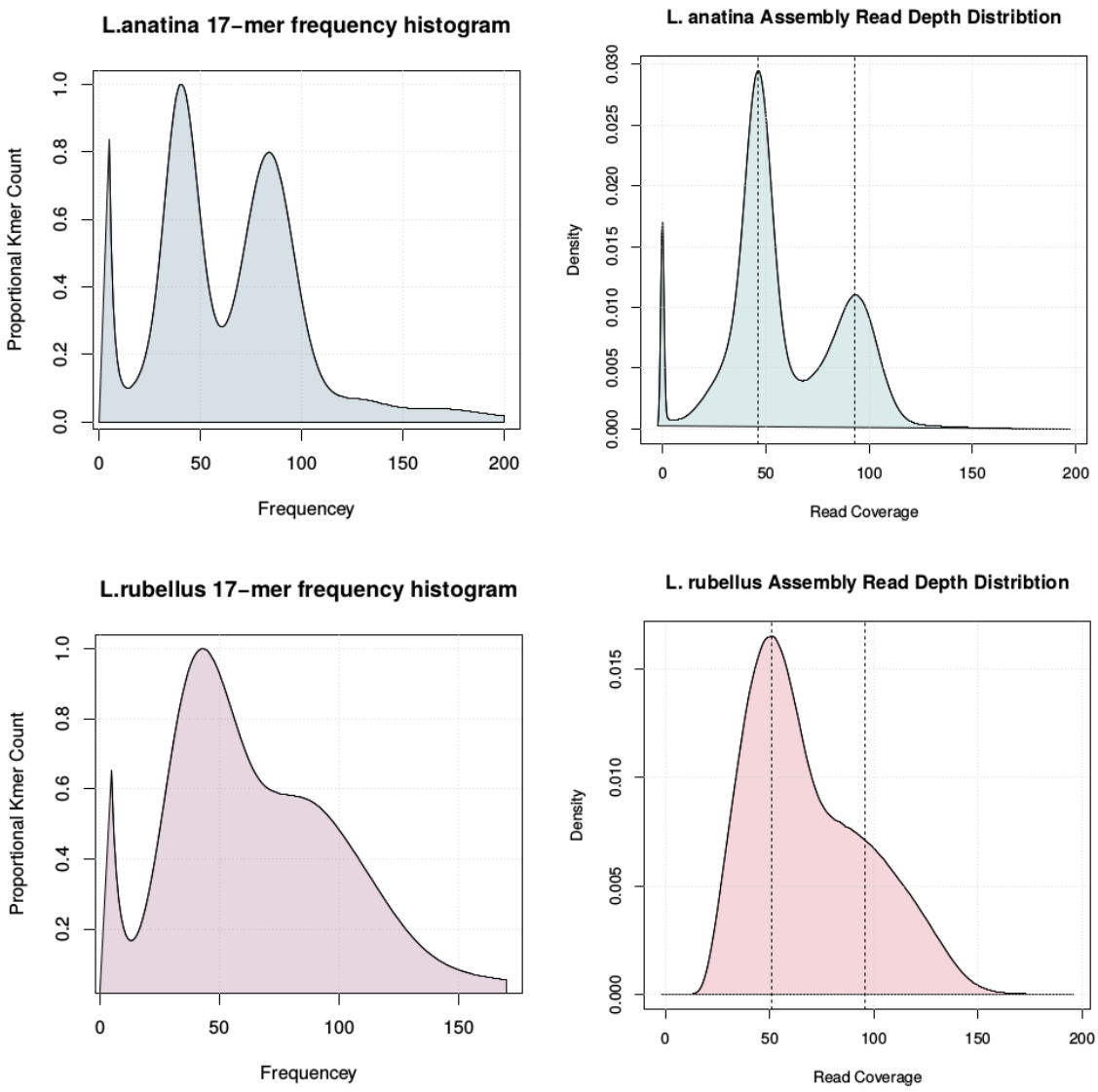


Figure 5. (Left) 17-mer frequency histograms derived from short read sequence libraries, (right) read depth distributions of libraries mapped to assembly.

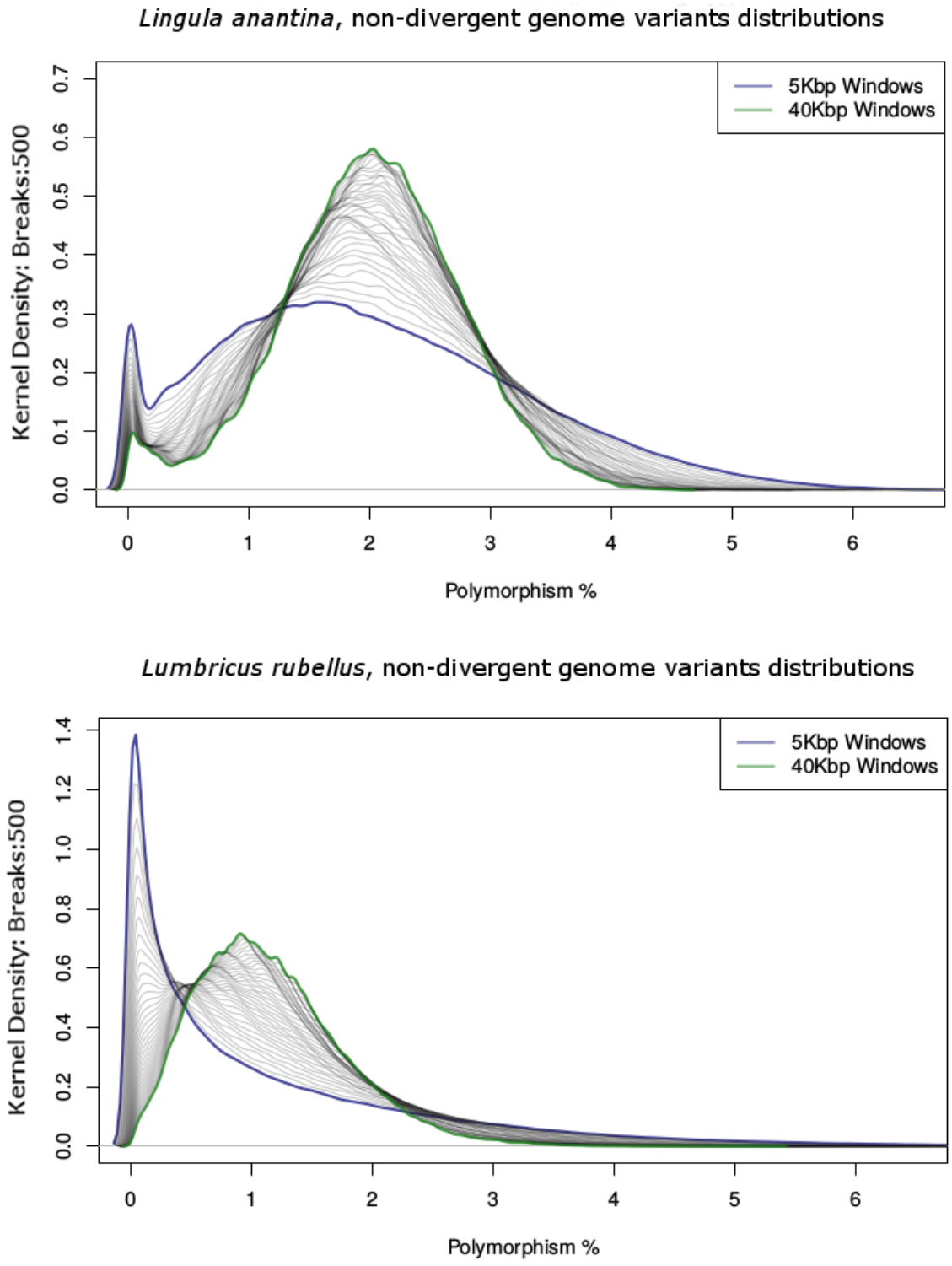


Figure 6. Multiple rolling window allelic variant distributions in full-read-depth fractions of both genomes. (top) Lingula anatina, (bottom) Lumbricus rubellus

2.2.2. Analysis of Draft *Lingula anatina* Assembly

There was no change in read depth distributions throughout the assembly when short-read mapping was attempted with k=2 allowed mapping per read ('single map' Figure 7 (top), versus 'double map' Figure 7 (bottom)). In a 'medium allelic divergence' diploid assembly (~1-3%) that is subject to size inflation due to non-collapse of string graphs, a flexible re-alignment at k=2 would be capable of matching polymorphic short reads to both their primary and opposing alleles, correcting scaffold read-depth disparity. There was no meaningful change in the read pileup pattern when the 'double map' procedure was used (Figure 8 (top)) compared to the 'single map' procedure (Figure 8 (bottom)), as would be expected given the distributions in Figure 7. As this does not happen, it seems to be the most likely explanation that the extent of the allelic divergence within this draft *Lingula anatina* genome is beyond the tolerance of conventional short read aligners (even when using the very sensitive and alignment parameters). When coverage and polymorphism rate is visualised along assembly scaffolds (Figure 8; examples shown here for *Lingula anatina*), the polymorphism rate can clearly be seen to track the coverage, with regions of low, or zero polymorphism corresponding to the half-coverage regions. This characteristic bimodal pattern was seen in the read pileups of almost all scaffolds analysed in this way (approx. 12 of the largest scaffolds; see Appendix 1.1 'LA Scaffold Visualisation').

Polymorphism and Read Depth

Figure 8 was generated using a simple variant rate rolling mean of two sizes (50 bp and 10 kb) of allelic diversity and plotted alongside read depth (using R) to give a summary of the relationship of allelic diversity to read depth. B1 was generated using 'single-map' read coverage levels, B2 was generated using 'double-map' coverage levels.

There was a characteristic bimodal pattern seen in the read pileups of almost all scaffolds analysed in this way (approximately 50 of the largest scaffolds). Read pileup depth (green) corresponded to polymorphism rate (across both a 50 bp [black bars] and 10 kb [orange line] mean), with an approximately half-level of read pileup corresponding to zero, or very low polymorphism rate, and full read pileup corresponding to a greatly increased polymorphism rate. There was no meaningful change in the read pileup pattern when the double map procedure was used (top) compared to the single map procedure (bottom).

The correspondence of read pileup with polymorphism suggests that the full depth level is the result of two alleles' reads mapped to the same scaffold whilst the single depth sections likely represent just a single allele. The short read alignment incompatibility between alleles would also suggest that the same sequences would have been assembled separately by the Newbler assembler in the

original paper (Luo et al. 2015), as these differences would also have made the formation of a collapsed string graph impossible.

Structural Mapping of Mis-assembly

The running hypothesis was that the half depth pileup scores represent regions of the genome that have been mis-assembled due to extreme allelic divergence. If this hypothesis is true, it would be expected that most half depth regions in the genome to have a single collinear match with another half depth region from a separate scaffold elsewhere in the assembly.

Half depth fragment pairs were identified using the method described in ‘Fragment Pair Selection’. Pairs were further filtered using a custom C# program (‘Reconstruct’, see Github at end of Section) that employed the non-redundant anchor method for pairing divergent alleles as employed in the divergent *Ciona intestinalis* assembly(Vinson et al. 2005). This anchor method was modified slightly so that if multiple anchor points lay in a colinear series, they would be merged into a single 'anchor span'. Pacbio read alignments made on the read depth fragmented assembly were lifted over onto the original assembly (filtering out Pacbio reads with coverage of <2 to reduce noise) and visualised in R.

Structures visualised in Figure 9 are characteristic of 14 scaffolds analysed in the same way (see Appendix 1.2.'Misassembly Structures). Purple bars and black lines show scaffolds and annotated genes (respectively) from Luo *et al*. Green bars show read depth pileup. Brown connecting lines show local alignments between two regions with no local overlap, which map uniquely to each other (non-redundant anchor method described above). Blue connecting lines show Pacbio long reads aligned with blasr (Chaisson & Tesler 2012), which align exactly twice within the whole assembly.

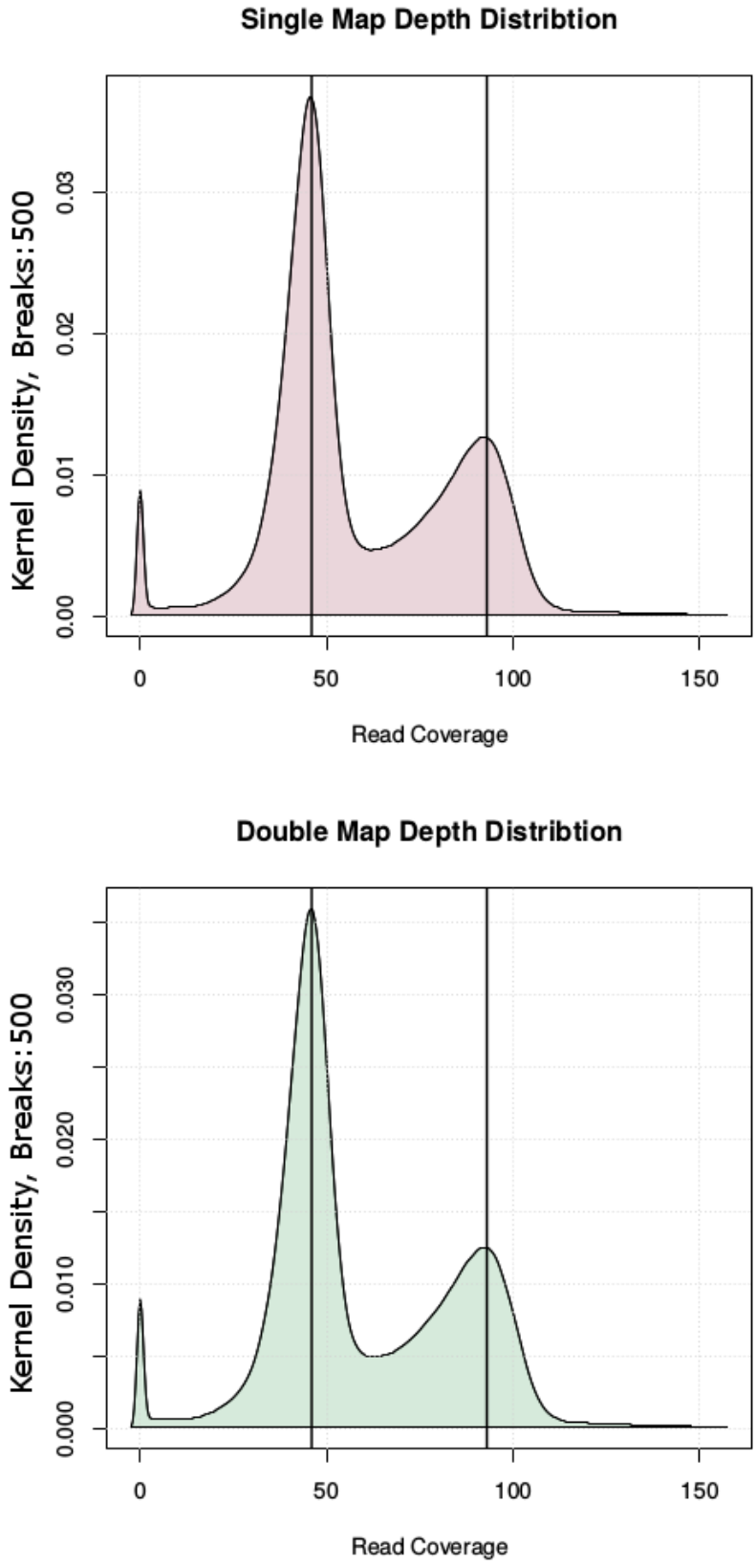


Figure 7. Lingula anatina genome, read depth distributions after single-mapping and double-mapping sequencing library (one alignment per read (top) vs two alignments per read (bottom)).

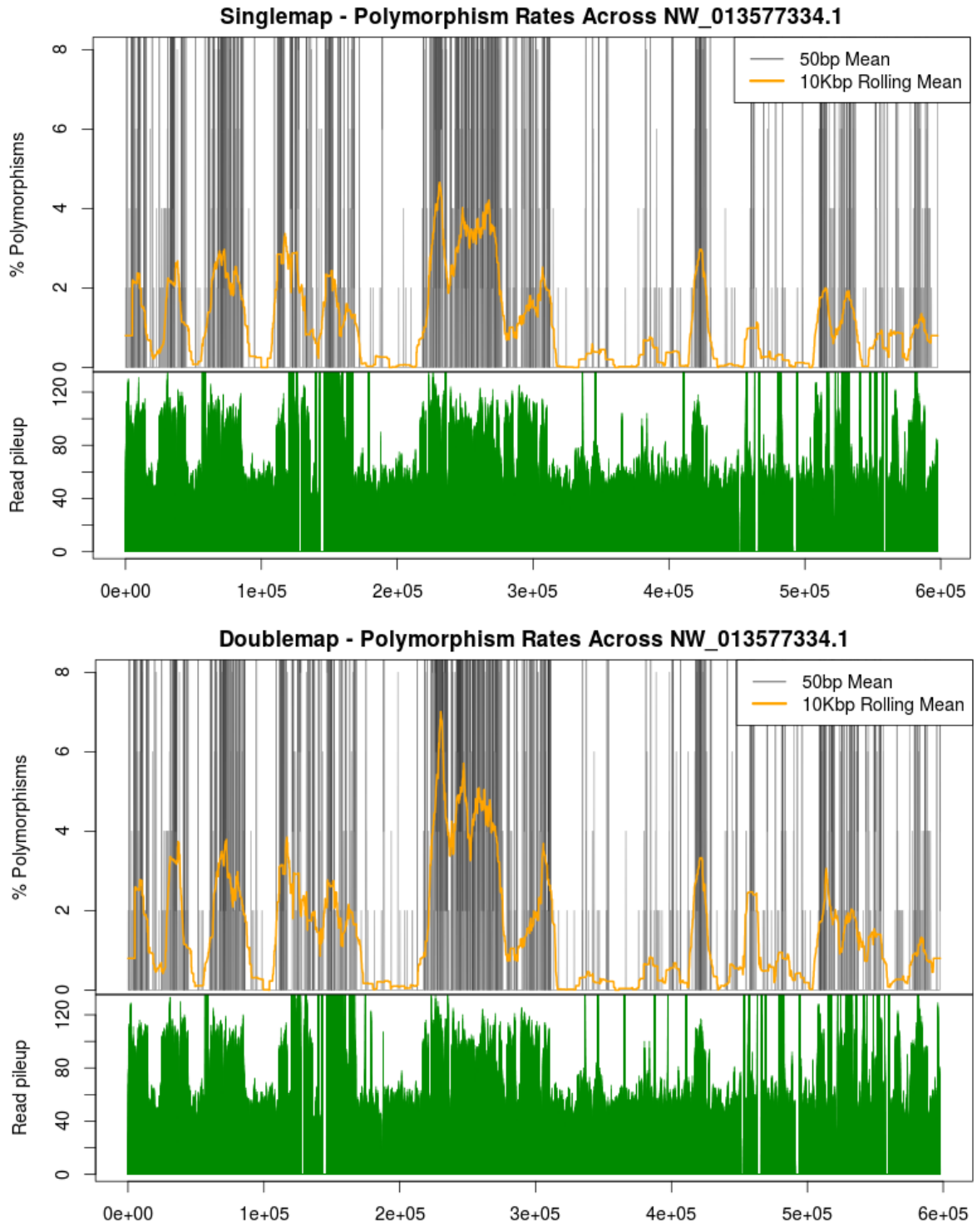


Figure 8. Read pileup against spatial polymorphism rates in Lingula anatina assembly scaffolds. (top) mapping performed with a single match retained per read, (bottom) mapping performed with two matches retained per read.

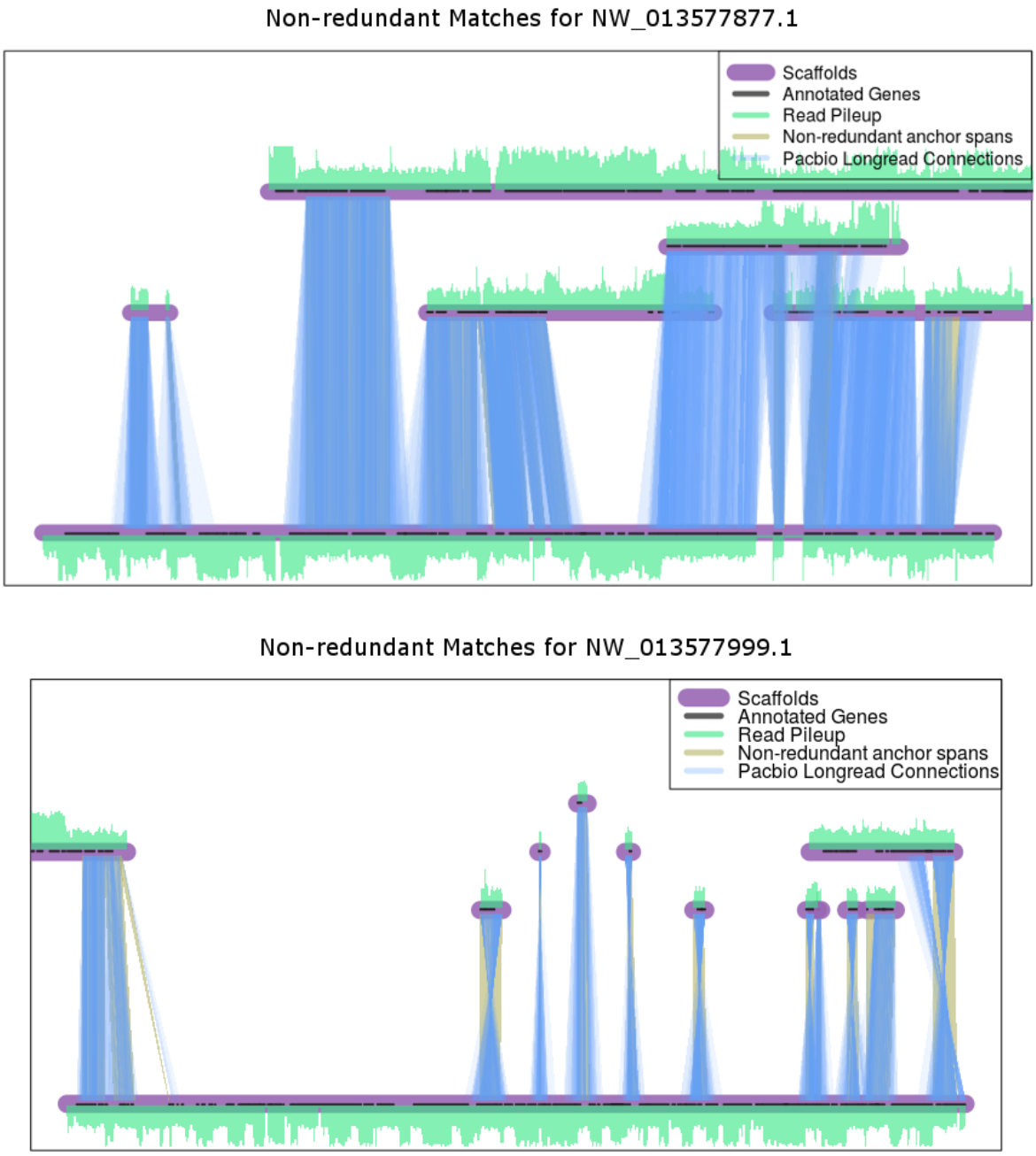


Figure 9. Allelic inflation in Lingula anatina genome visualised. Two long scaffolds (200kb+) with mapped read depths and non-redundant connection to allelic fragment pairs.

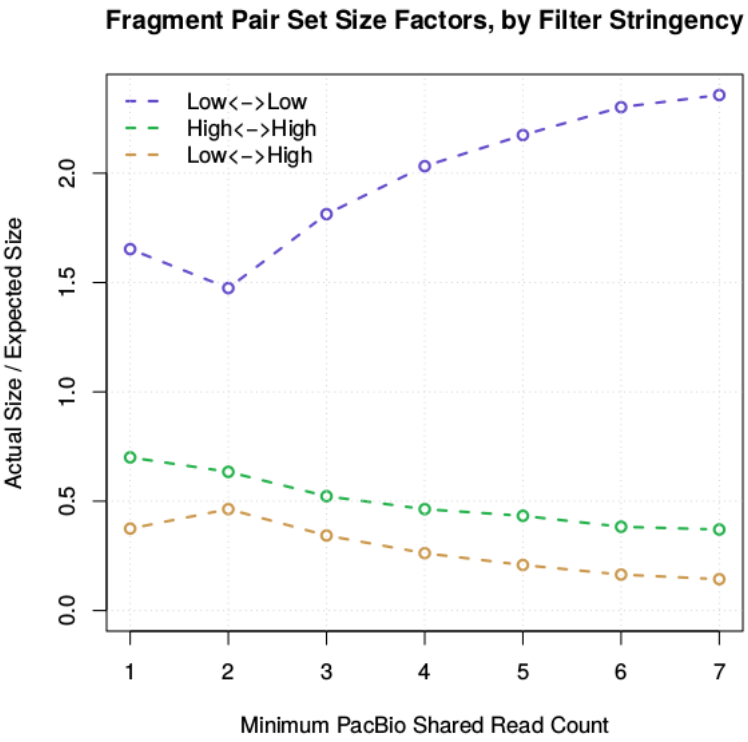
Statistical Evidence for Mis-assembly

To add a level of empirical support to the assertion that the draft *anatina* genome is mis-assembled, a further non-redundant fragment pair selection was undertaken. The central hypothesis was that low read-depth regions had real identifiable allelic pairs discoverable within the draft assembly, and high read-depth regions did not. If this were the case, fragments with low read-depth ought to pair together at a substantially higher rate than fragments of high or mixed read depth.

The *anatina* assembly was fragmented, however in this case each fragment was tagged with a binary classification that indicated the read depth peak (see Figure 5 (top)) it fell under. The classifier simply selected the read depth peak which the fragment depth had the smallest difference to. Fragments were thus classified as 'high' or 'low'.

The Fragment Pairing procedure described above was then iterated over the integer range 1:7, for a minimum shared read parameter. In each case, pairs were segregated into three sets corresponding to the three possible pair combinations of fragment classification. For each set of pairs, high-high (HH), low-high (LH), and low-low (LL), the expected set size given a null distribution of pairings were calculated. An enrichment factor for each pair type was then derived by dividing the actual set size by the expected set size. The variation in these factors was then mapped over the range of the minimum shared read parameter. The result can be seen in Figure 10.

In the unfiltered sets, the LL group factor is 1.64, whilst the other groups factors are LH=0.37 and HH=0.704. As the filtering becomes more stringent, groups LH and HH experience a continuous reduction in factor level, whilst the LL group's factor increases over this range. This indicates that as noise is reduced, the signal of the low-low fragment pair set strengthens. The over-representation of the LL group is made even more significant by the fact that it is also the largest group by far in the highest stringency set, making up 82% of all fragment pairings. From this it can be concluded that fragment read-depths are a strong predictor of non-redundant homology matches between fragments; the result which would be expected if the genome mis-assembly hypothesis was true.



No Filtering

Pair Classes	Metric	
HH	*Pair Count*	1902
LH	*Pair Count*	4953
LL	*Pair Count*	9184
HH	*Exp. Count*	2698
LH	*Exp. Count*	13158
LL	*Exp. Count*	5579
HH	*Factor*	0.704
LH	*Factor*	0.376
LL	*Factor*	1.646

>6 Shared reads per pair

HH	*Pair Count*	257
LH	*Pair Count*	485
LL	*Pair Count*	3382
HH	*Exp. Count*	693
LH	*Exp. Count*	3383
LL	*Exp. Count*	1434
HH	*Factor*	0.37
LH	*Factor*	0.1433
LL	*Factor*	2.35

Figure 10. Allele pair discovery frequency tests between low and high coverage fragments.

2.2.3. Identification and Characterisation of 'divergent alleles'

For both draft assemblies, half read-depth fragments in the genome frequently have a single collinear match with another half read-depth fragment from a separate scaffold elsewhere in the assembly. Figure 9 shows 2 characteristic *Lingula anatina* scaffolds analysed with the half depth allelic fragment matches re-attached (see Appendix 1.2. 'Mis-assembly Structures' for more examples). The previously published draft *L. anatina* genome therefore appears to have been assembled into fragments of two read-coverage sizes, and then scaffolded into blocks of both half and full coverage DNA. This would mean that this genome was subject to substantial mis-assembly and allelic over-inflation in size. The mis-assembly hypothesis was further evidenced by an empirical approach which found that half-depth fragments formed unique pairings at far higher rates than full depth or mixed pairs (see Figure 9). The knock-on effect being that over-estimation of gene family sizes and misrepresentation of gene family diversity metrics will have occurred in a widespread manner in the original publication in Nature Communications (Luo et al. 2015).

This is also very strong supporting evidence for the hypothesis that large (30-40%+) portions of the genome of these two organisms show extreme levels of allelic diversity. To further back-up this conclusion and gain a qualitative view of the synteny between these divergent alleles, ORF alignments, modelled repeats, and alignment identities were visualised between some example

pairs of allelically diverged regions for both *Lumbricus rubellus* and *Lingula anatina* (Figure 16 and 17). The images show the extent of re-arrangement and sequence divergence between allele pairs. Divergent regions contain long non-overlapping open reading frames. Open reading frames may or may not be in perfect synteny across the two alleles. Protein alignment identities between the two alleles is also highly variable, with some of the values at the lowest extreme probably reflecting the loss of one of the protein allele copies.

2.2.4. Nucleotide divergence

Base substitutions and small indels appear to occur at an equal rate in the *Lingula anatina* genome whereas in *Lumbricus rubellus* base substitutions appear to occur more frequently (see Figure 11). In both cases transposon-like features (loosely classified as indels longer than 10bp) seem to account for around 5% of the total allelic sequence divergence.

The breakdown of both organisms' allelic divergence rates for both indels and substitutions across a range of test window sizes are displayed in Figure 12, 13 and 14. *Lumbricus rubellus* exhibits the greater divergence rates, but also the more well-defined divergence rate peaks, likely a reflection of two-lineage origin of diversity. *Lingula anatina* appears to have a far more skewed distribution of allelic divergence, possibly a reflection of the marine broadcast-spawner's capacity to interconnect various diverse and ancient sub-species across a broad geographic range in the near-term reproductive life history of an individual's genetic lineage.

Small far right peaks on the *Lumbricus rubellus* substitution rate graph appear to indicate larger scale movements of DNA, which during the evolutionary process have resulted in incomparable sequence at the same loci in the alleles of the two lineages. These are absent in the *Lingula anatina* graphs.

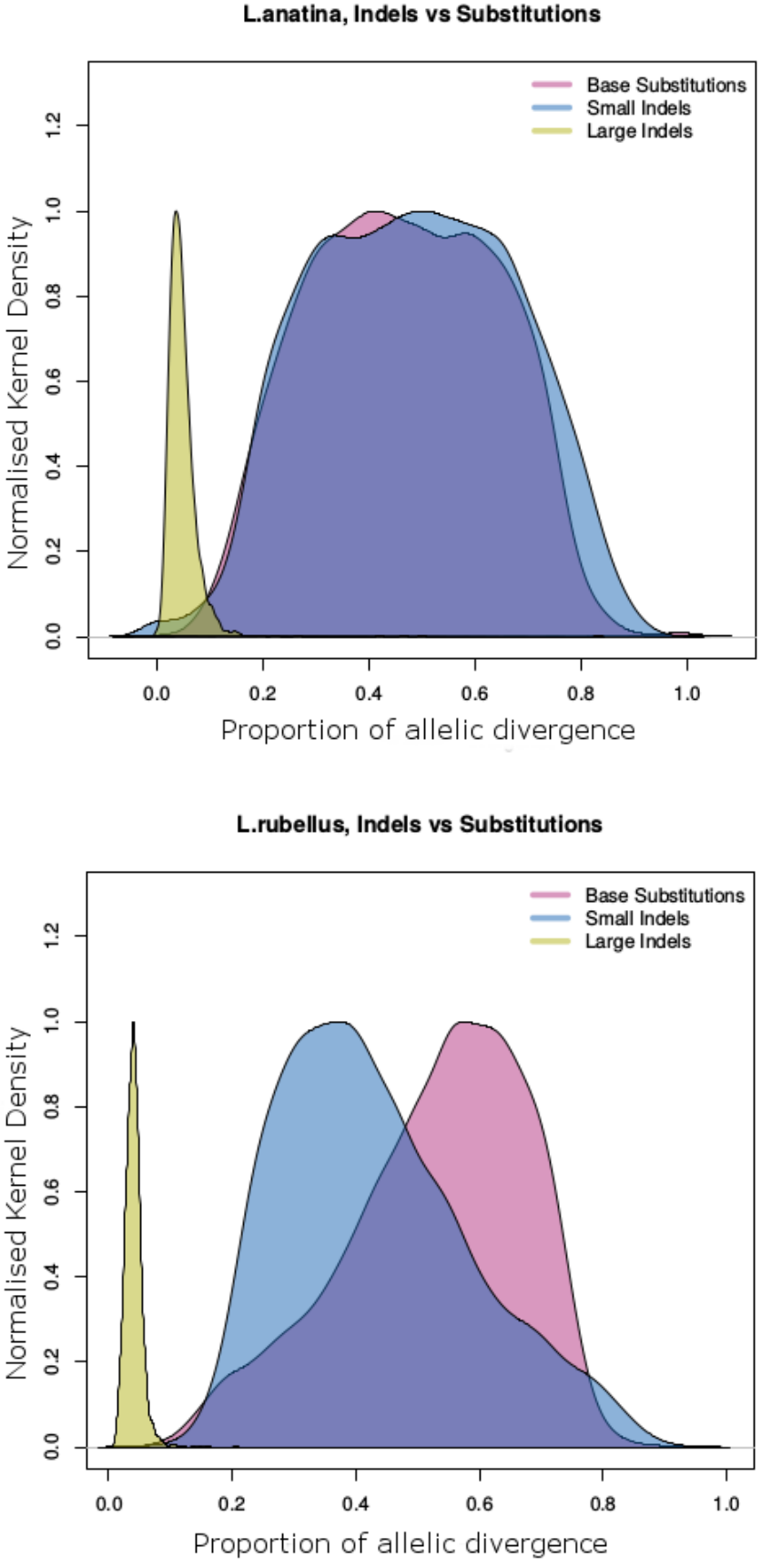


Figure 11. Components of allelic divergence between fragment pairs. (top) Lingula anatina, (bottom) Lumbricus rubellus.

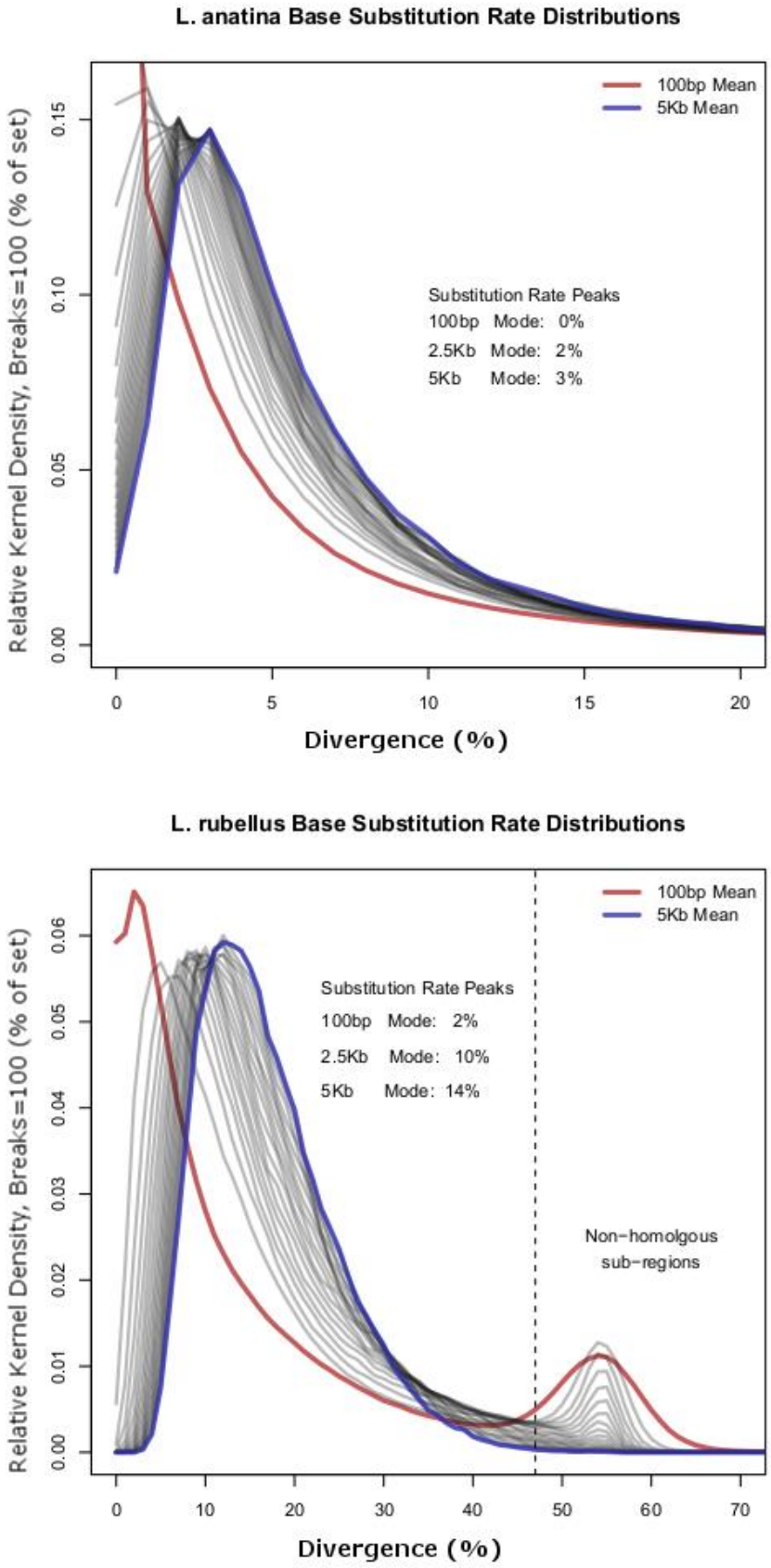


Figure 12. Base substitution rate distributions for both sets of fragment pairs. (top) Lingula anatina, (bottom) Lumbricus rubellus.

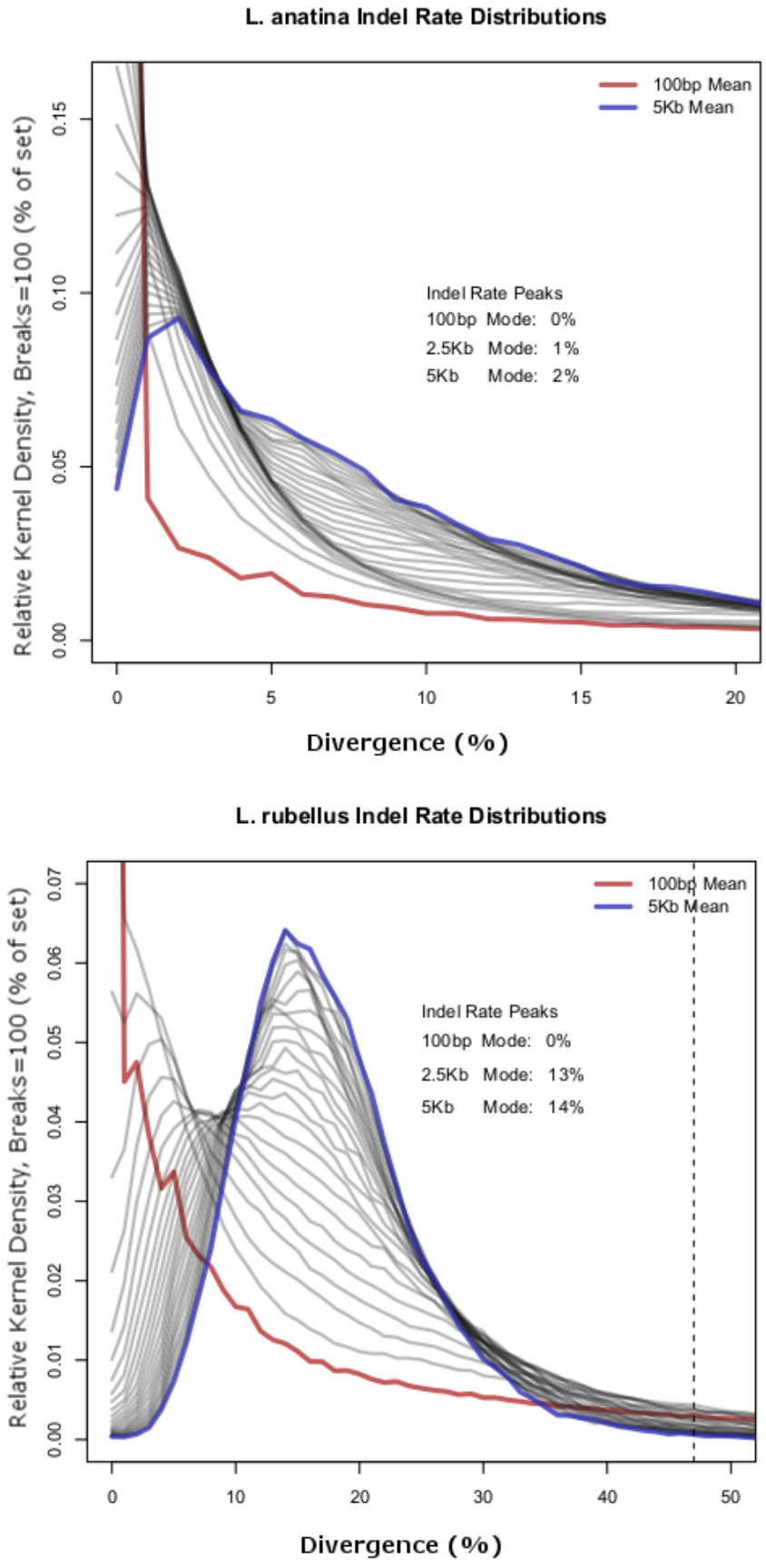


Figure 13. Indel rate distributions for both sets of fragment pairs. (top) Lingula anatina, (bottom) Lumbricus rubellus.

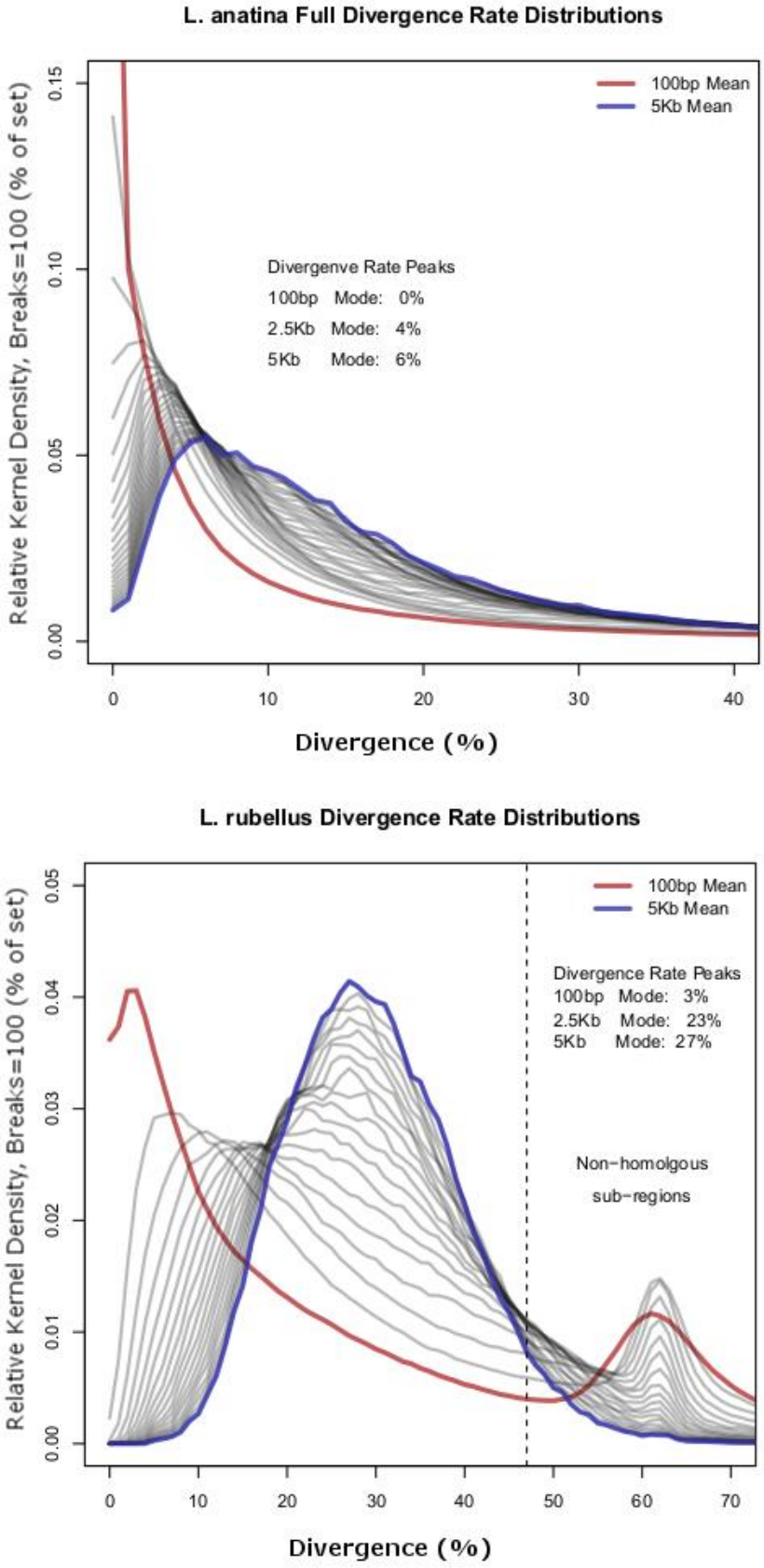


Figure 14. Full divergence profiles for both sets of fragment pairs. (top) Lingula anatina, (bottom) Lumbricus rubellus.

RepeatModeller identified that over 20.11% of the *L. anatina* and over 37.66% of the *L. rubellus* genomes were made up of repeats (see Tables 1 and 2). Most of these repeats were unclassified, particularly for *L. anatina*. In both cases, transposon-like features (indels longer than 10 bp) seem to account for around 5% of the allelic sequence divergence (data not shown). Graphs showing pairwise divergence for base substitutions (Figure 12), indels (Figure 13) and combined indel and substitution rate (Figure 14) for both *L. anatina* and *L. rubellus* are shown above. The 5 kb mean base substitution rate distribution mode was 3% and 14%, and indel rate distribution mode was 2% and 14% for *L. anatina* and *L. rubellus* respectively. Combined (indels and substitutions) 5 kb mean divergence rates were 6% and 27% for *L. anatina* and *L. rubellus* respectively.

2.2.5. Qualitative Overview of Allele Pairs

The panels in Figure 15 and 16 show the extent of re-arrangement and sequence divergence between allele pairs. Divergent regions contain long non-overlapping open reading frames. These open reading frames may or may not be in perfect synteny across the two alleles. The results of converted protein sequence alignments between alleles are also displayed within each image.

There are several obvious examples within these visualisations of ORFs or repeat regions being matched with a gap in the alignment with the opposing allele. This is suggestive of how re-arrangements may be occurring. For a larger collection of these images see Appendix 1.3. 'Fragment Pairs'.

There is also a clear variation in the extent of sequence divergence across the examples. For example, in Figure 15 (bottom) the base substitution rate appears to spike the most around two short ORFS. In Figure 15 (top) the substitution rate is exceptionally high around the first annotated repeat. This suggests that there may be functional drivers of both positive purifying selection which have modulated the rate of divergence within these sequences.

Table 1. Lumbricus rubellus genome, repeat modeller output table.

Lumbricus rubellus

Element Type	Number of Elements	Length Occupied (bp)	Percentage of Sequence (%)
SINEs:	**90,188**	**14,285,537**	**1.69**
ALUs	0	0	0.00
MIRs	56,320	8,977,740	1.06
LINEs:	**343,733**	**52,107,319**	**6.15**
LINE1	301	29,577	0.00
LINE2	196,114	28,722,524	3.39
L3/CR1	3,686	724,623	0.09
LTR elements:	**38,650**	**7,458,476**	**0.88**
ERVL	0	0	0.00
ERVL-MaLRs	0	0	0.00
ERV_class I	249	67,396	0.01
ERV_class II	0	0	0.00
DNA elements:	**228,840**	**49,418,328**	**5.83**
HAT-Charlie	36,781	8,748,082	1.03
TcMar-Tigger	695	68,016	0.01
Unclassified	**881,099**	**143,366,744**	**16.91**
Total Interspersed Repeats	-	266,636,404	31.46
Small RNA	**1,337**	**273,587**	**0.03**
Satellites	**1**	**336**	**0.00**
Simple Repeats	**625,910**	**47,783,941**	**5.64**
Low complexity	**40,044**	**4,464,876**	**0.53**
Totals	**2,249,801**	**319,158,808**	**37.66**

Table 2. Lingula anatina genome, repeat modeller results.

Lingula anatina

Element Type	Number of Elements	Length Occupied (bp)	Percentage of Sequence (%)
SINEs:	**2,295**	**14,285,537**	**0.14**
ALUs	0	0	0.00
MIRs	204	8,977,740	0.01
LINEs:	**15,802**	**4,467,334**	**1.05**
LINE1	312	142,598	0.03
LINE2	2,682	885,140	0.21
L3/CR1	162	69,517	0.02
LTR elements:	**5,492**	**1,852,461**	**0.44**
ERVL	0	0	0.00
ERVL-MaLRs	0	0	0.00
ERV_class I	557	54,697	0.01
ERV_class II	0	0	0.00
DNA elements:	**50,658**	**13,473,980**	**3.17**
HAT-Charlie	0	0	0.00
TcMar-Tigger	0	0	0.00
Unclassified	**250,729**	**56,195,200**	**13.21**
Total Interspersed Repeats	-	76,570,012	18.00
Small RNA	**1,358**	**220,051**	**0.05**
Satellites	**1,444**	**382,735**	**0.09**
Simple Repeats	**192,174**	**7,748,564**	**1.82**
Low complexity	**20,397**	**998,335**	**0.23**
Totals	**538,905**	**99,241,462**	**20.11**

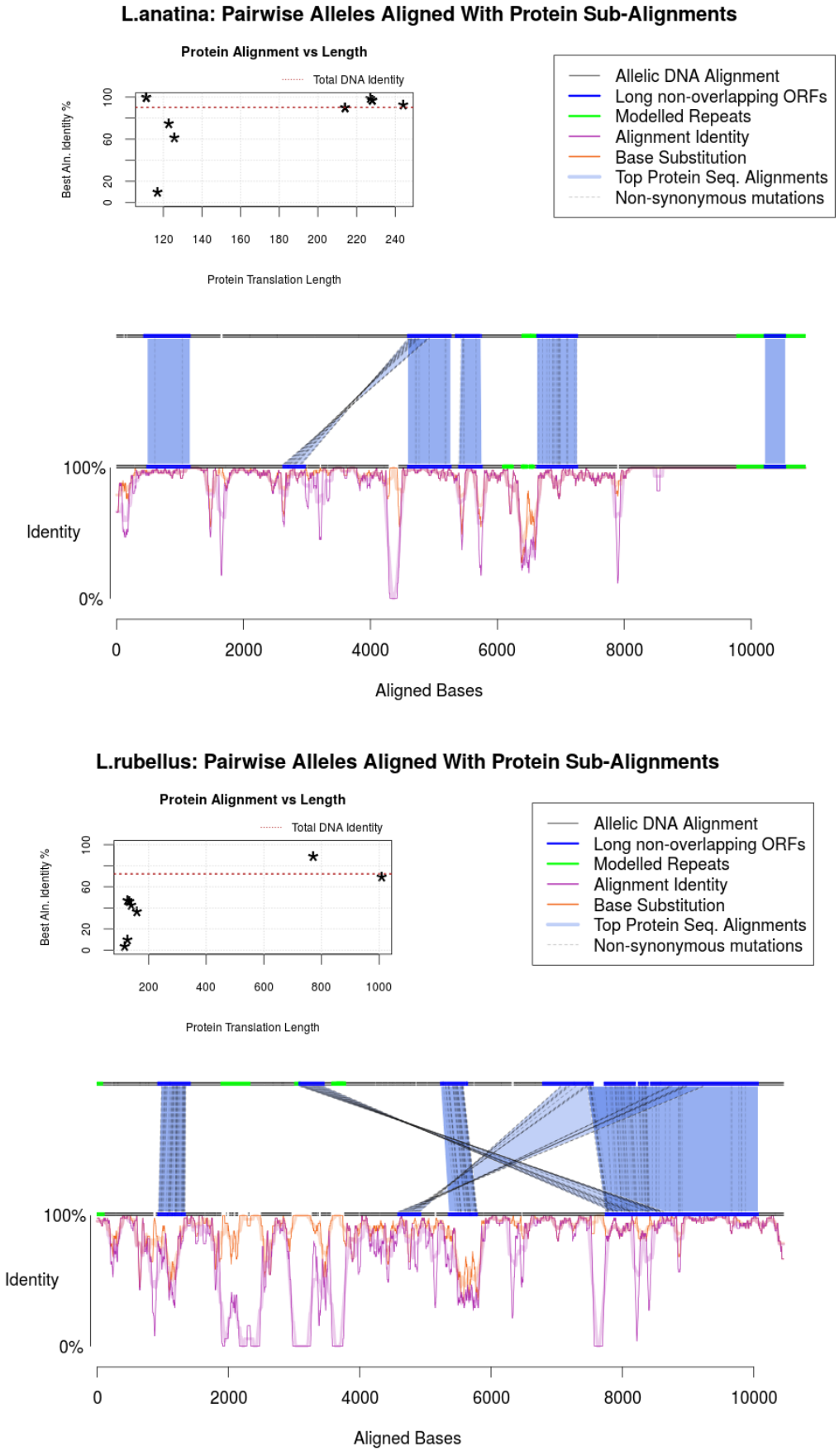


Figure 15. Examples of L. rubellus and L. anatina candidate allele fragment visualisations (A)

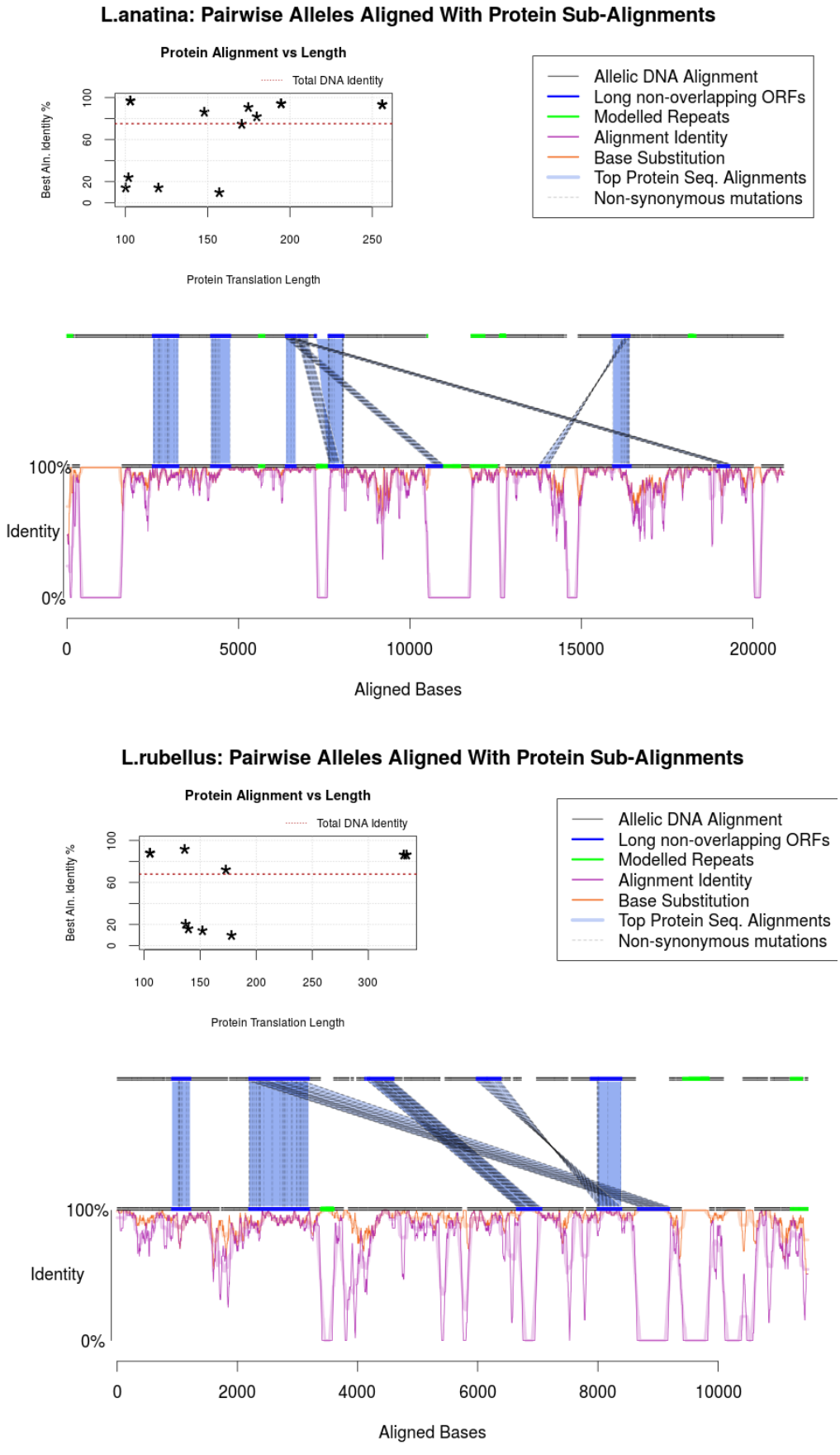


Figure 16. Examples of L. rubellus and L.anatina candidate allele fragment visualisations (B)

2.2.6. Protein Sequence Divergence

The purpose of the protein alignments is to measure the scale of function alterations which have arisen from the allelic divergence. Although absolute base sequence variation between alleles is extreme, this does not immediately lead to the conclusion that the functional content is substantially altered. Differences in encoded protein sequences however, are a clearer indication that this may be the case.

The difference in the protein sequence variation distributions between the two organisms, displayed in Figure 17, is reflective of the differences seen in Figure 14. That the peak score for both organisms is substantially lower than both the mean and median statistics for their respective sets, suggests that a range of other genetic mechanisms may have acted upon these protein sequences to cause their divergence, other than the predominantly purified alterations made by the molecular clock. This may include partial duplication of sequence, rearrangements, or inversions.

Protein Alignment Validation

Some Amino Acid residues have shared properties with various others. Hydro- and lipophilic side chains, side chain sizes, and capacity for ionic bonding all account for a residue's functional role in the peptide. Histidine, Methionine, and Isoleucine are recorded as being more changeable residues, and cysteine, tryptophan and tyrosine and notably more stable (Tourasse & Li 2000). Most importantly, it would be expected that substitution rate differences between residues should be relatively consistent for both organisms. Since the scale of protein polymorphism is exceptionally high, particularly in the case of *L. rubellus*, to see the chemical/biological function of residues reflected in their substitution rates would increase confidence in the legitimacy of this analysis.

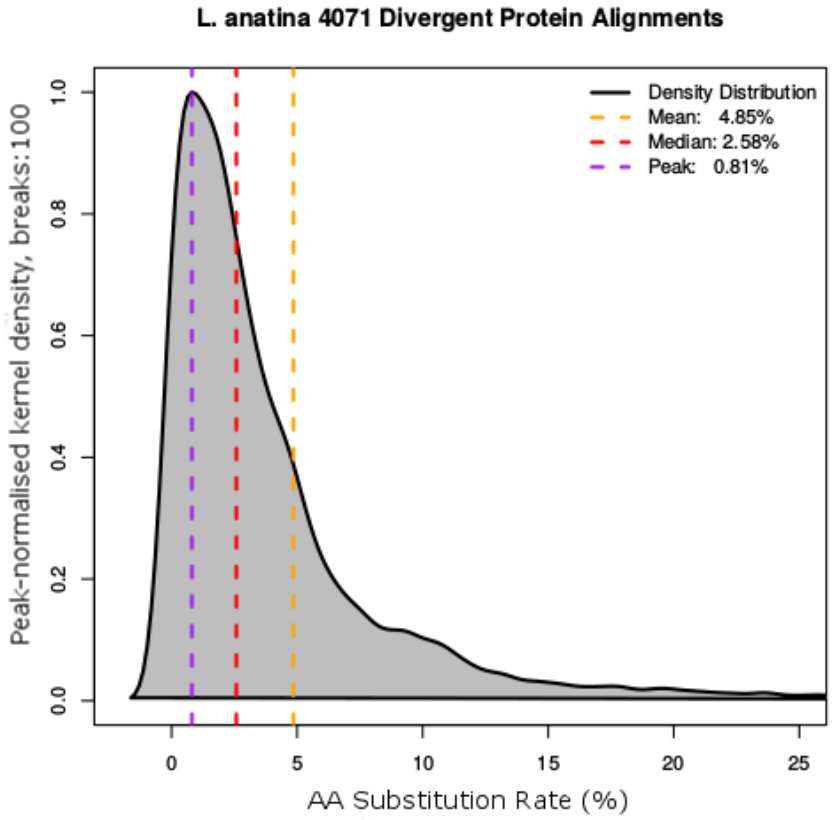
With the set of alignments used in the two panels of Figure 17, each alignment was further processed to produce the rate of per-residue sequence mismatch for each residue type. This procedure was performed for both sets of alignments. Figure 18 shows a consistent trend in the residue substitution rate differences between both organisms. In line with Tourasse and Li's results, Histidine, Methionine and Isoleucine were the top three most changeable residues in both cases. Tryptophan was the most stable, with cysteine being the 5th most stable, and tyrosine being the 10th.

The AA substitution rates observed appear to fall in line with prior investigations of core biology, thus building confidence to the biological legitimacy of the set of alignments made here.

Indels and Substitutions

Although the pure protein substitution rates can account for many gradual changes in peptide sequence, there also appears to be a broader distribution of larger scale changes, as seen in Figure 17, for which other genetic alterations may be responsible. Incidents such as duplications, deletions, inversions, or similar movement of genetic material could likely be represented as large ‘indels’ in the protein alignments. Each candidate allelic protein alignment was assessed for both substitution rates and indel features. The set of all alignments was then sorted for relative substitution rate, and visualised in R.

Figure 19 appears to show a light correlation between the indel features and the mismatch rate. Low-substitution alignments also tend to have little-to-no indel features. Across the rest of the set, almost independent of mismatch-rate, large scale indels appear to occur, reaching up to 40% of the size of the total alignment. These graphs show that although residue substitution remains the primary source of variation, there are also substantial indel features in the allelic matches of both protein sequences. This indicates that during a period of isolated divergence in these lineages, duplications, deletions and inversions have played a large role the in their base sequence evolution. Interestingly the large indel feature is more visible here than in the analysis of absolute DNA sequence change.



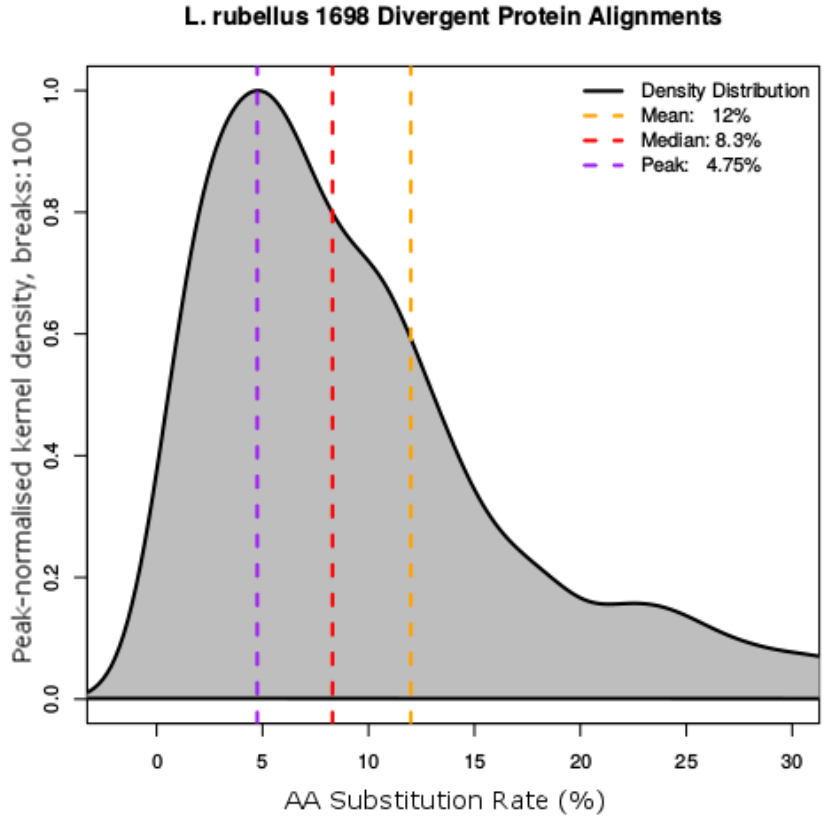


Figure 17. Candidate allele fragment pair protein sequence translation alignment distributions for both genomes.

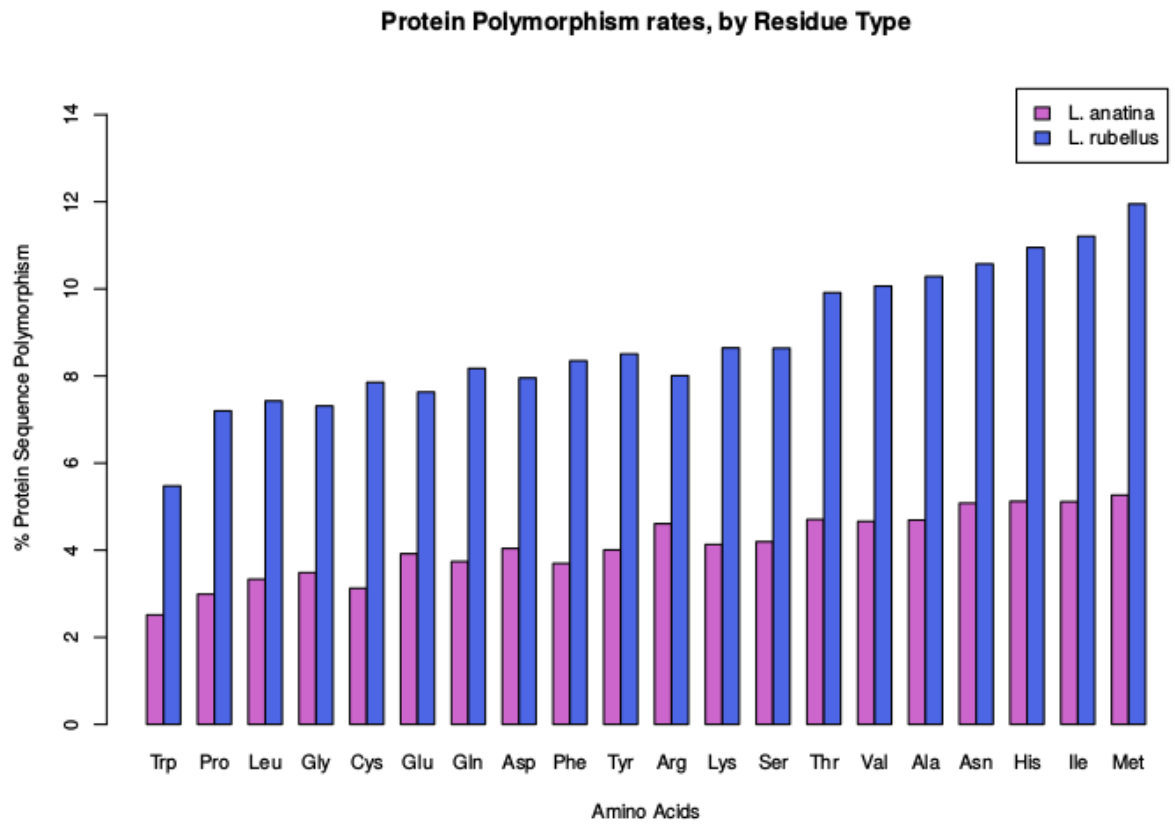


Figure 18. Amino acid sequence polymorphism rates per residue type, both protein alignment sets.

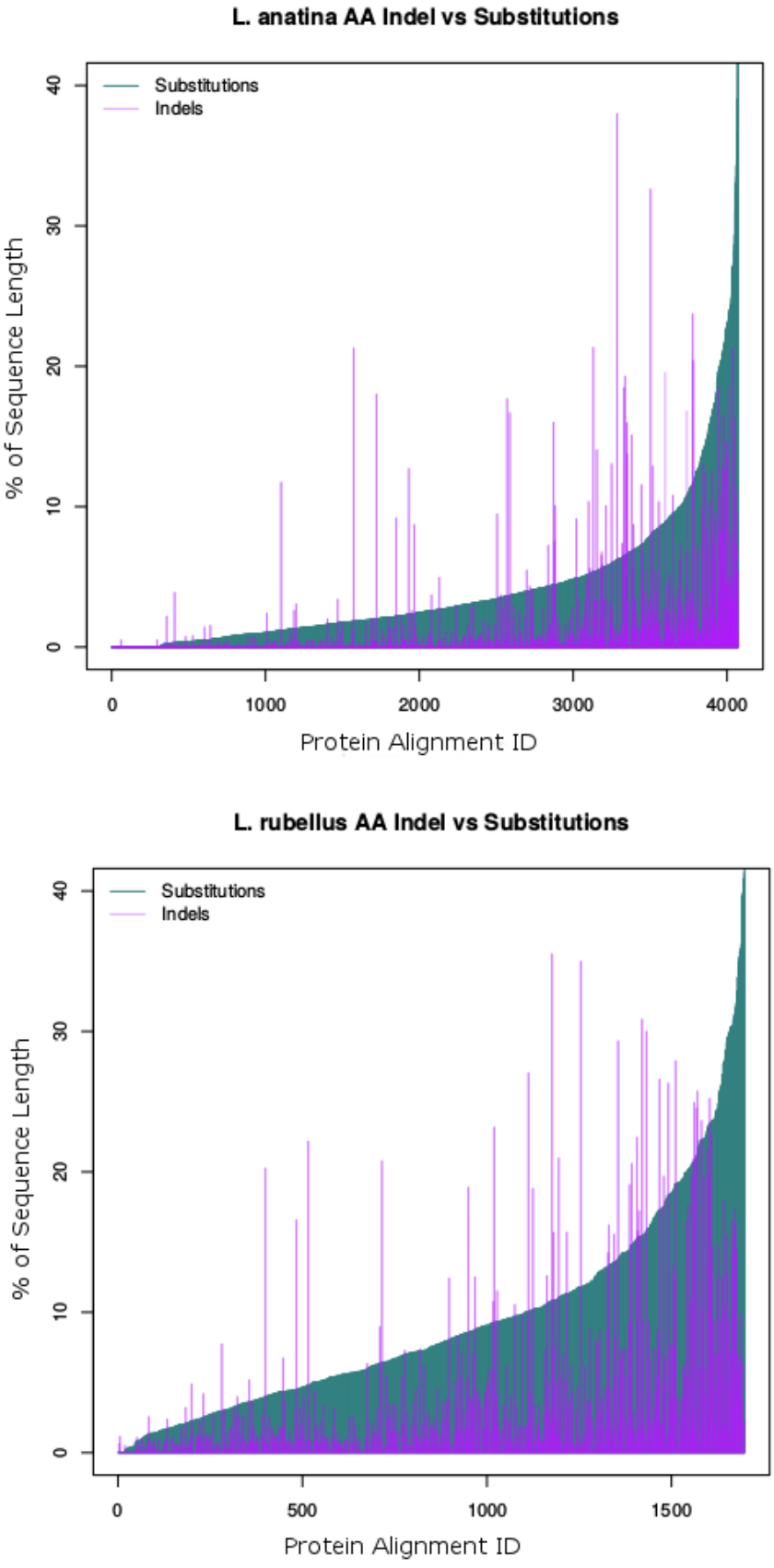


Figure 19. Base substitutions and indels as proportions of protein alignment lengths, full sets of alignments. (Top) Lingula anatina, (bottom) Lumbricus rubellus.

2.2.7. Environmental Adaptation and Protein Divergence

The protein divergence distributions for allelically matched ORFs plotted in Figures 17-19 show a polymorphism rate that is substantially lower than the absolute base sequence divergence (distribution peak at 0.81% for *L. anatina* and 4.75% for *L. rubellus*). However, these average values are only indicative of polymorphisms between protein sequences that are directly comparable and so are probably an underestimate (Figures 15 and 16 both show several ORFs without direct allelic counterparts). In both cases there is a long right-hand tail to the distribution, and the difference between peak and median scores (see Figure 17) indicates that the peak average accounts for less than half of even the filtered set of alignments.

Alignments pairs were also annotated with protein family definitions (Figure 20). There are two clear distributional peaks in protein alignment identity for 'best matches' (including unpaired ORFs) in both organisms. The lower peak (around 10% identity in both cases) likely represents cases whereby an ORF on one allele had no direct counterpart on the opposite allele – resulting in a best match being found with a highly divergent nearby ORF (i.e. a non-reciprocal best hit) and appears as the cluster of points towards the origin of the scatter plot. The higher peak corresponds with the ORFs for which a syntenic match remains and appears as the linear trend of points on the scatter plot. There is an apparent linear relationship between protein family divergence in *L. rubellus* and *L. anatina* in Figure 20 (right), as would be expected given the heterogenous sequence flexibility that occurs between different families. Classified family sizes are not consistent between organisms, however this analysis has only been carried out on the subset of the genome that makes up the divergent regions, so is not necessarily representative of the actual size of the protein families in these organisms. There are very few protein families exhibiting higher divergence in *L. anatina* than in *L. rubellus*, and there are a wide range of protein families exhibiting the inverse.

Six protein families were chosen as examples of large, well-annotated families of potential adaptive relevance. These protein families (see Figure 21) show a range of divergences and sizes and are broadly representative of the linear relationship shown in Figure 20, although ZIP metal transporters had a higher average allelic diversity in *L. rubellus* (~65%) than *L. anatina* (~30%). Of the six protein families chosen, epithelial sodium channels, GPCR chemoreceptors and glucuronosyltransferase had a higher family size in *L. anatina*. Mucin-like glycoproteins and ZIP metal transporters had a higher family size in *L. rubellus* than in *L. anatina*, and laminins had a similar family size between both. The overall rates of protein sequence divergence between alleles were exceptionally high in most cases for both organisms.

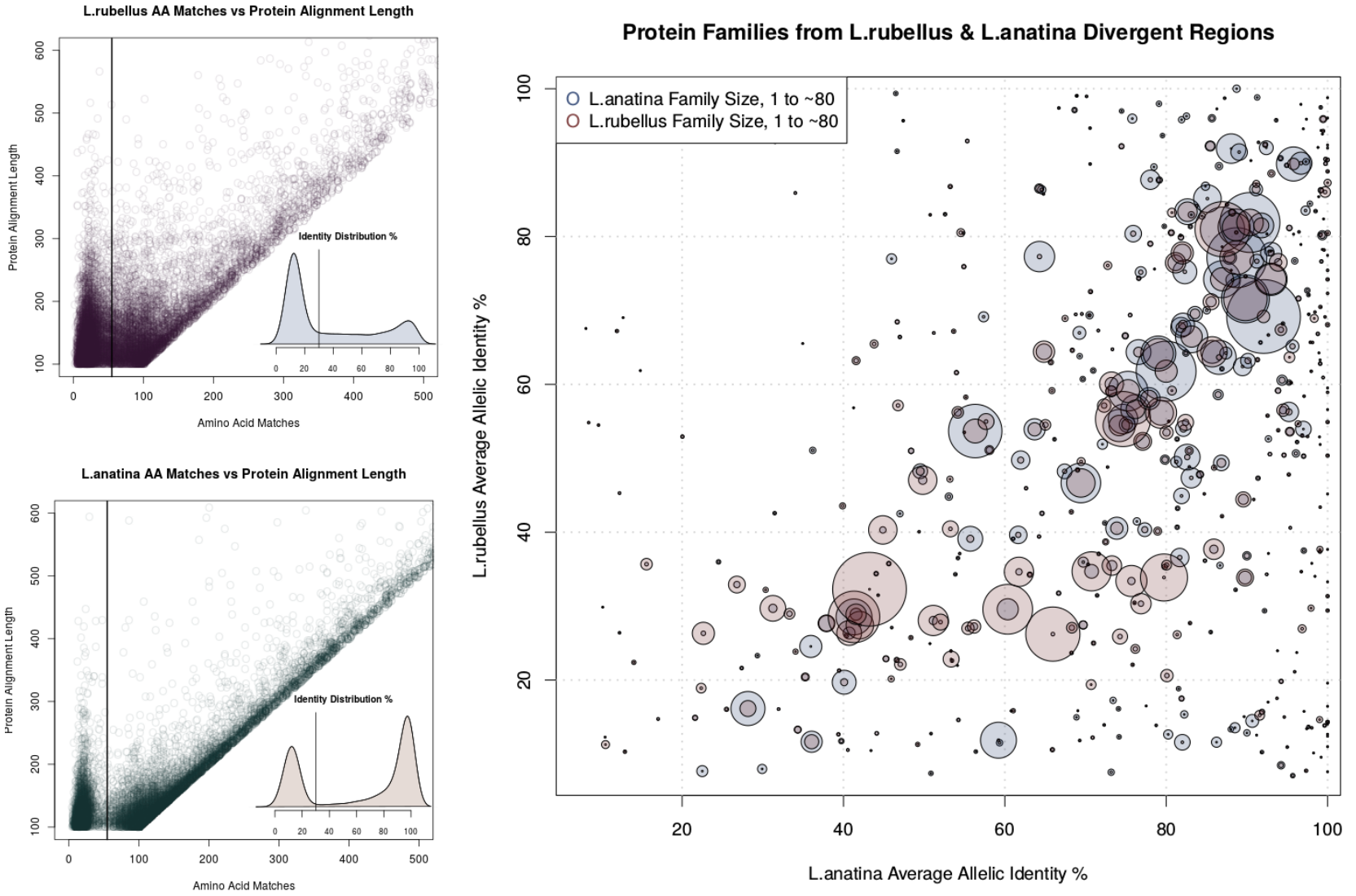


Figure 20. All identified protein families shared by both genomes (right), distributions of protein alignment residue matches against length for both full alignment sets (left).

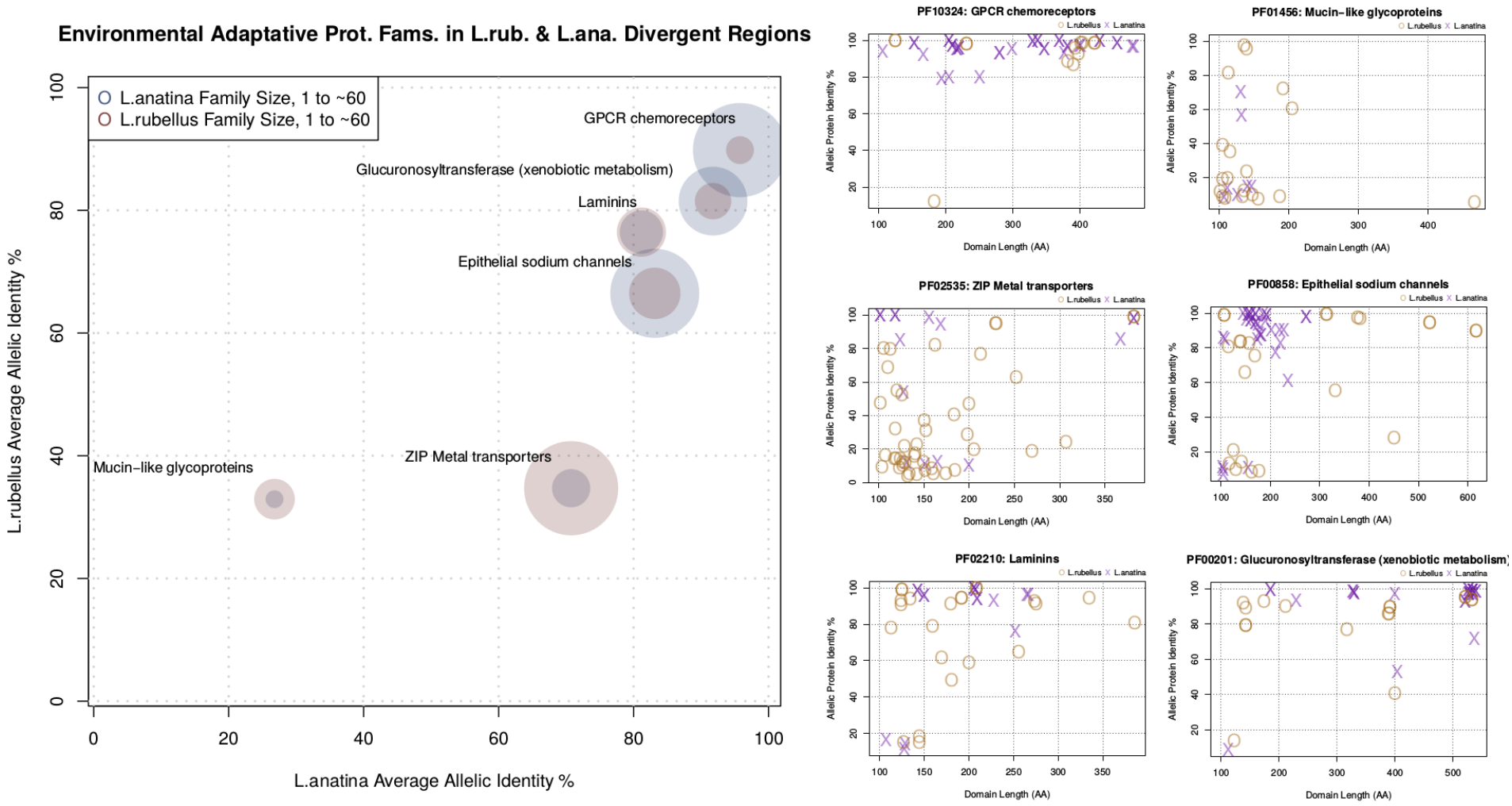


Figure 21. Protein Families of Interest, Largest shared protein families of highest divergence rates shared by both genomes.

2.2.8. Population Genetics in *Lumbricus rubellus* sample Population

The clear spatial separation of individuals into lineages produced identical groupings in both cases. The intersection of the images in Figure 22, can be seen in Table 3. The visualised presence/absence matrix for RAD-Seq scaffold stack presence shows that whilst there are marked differences in the alignment patterns between the two lineages. However even where the scaffolds in the visualisation were selected specifically for their capacity to segregate the population, the A/B lineage separation does not define the whole landscape so clearly. Despite maintaining a consistently recognisable lineage specific genomic fraction (Figures 23 and 24), the relative proportion of this DNA was very low compared to the whole genome. This analysis suggests that there are relatively few incompatible alleles between the A and B lineages.

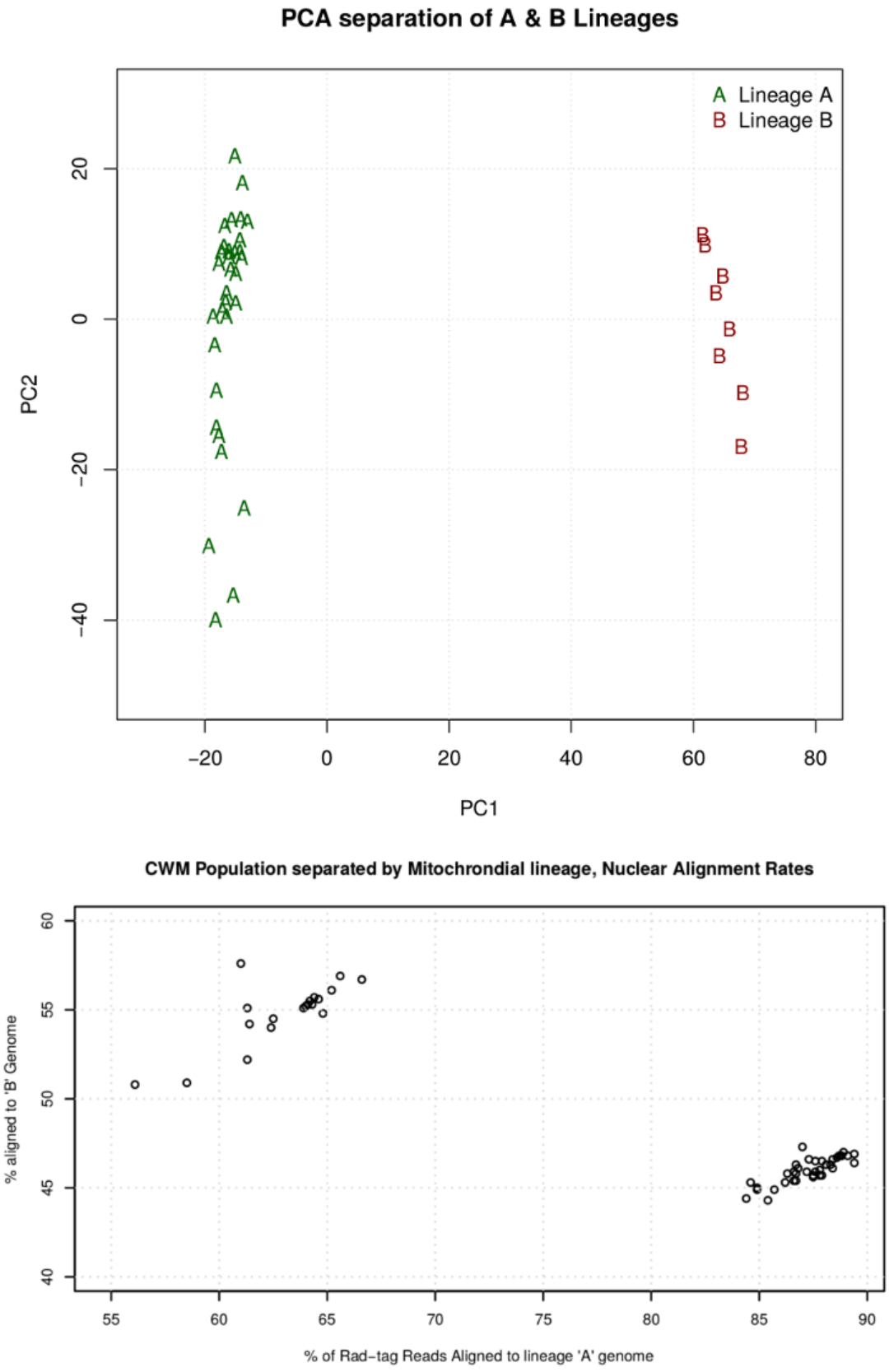


Figure 22. Lineage separation metrics. Principal component analyses of RAD-Seq alignments (top), and differential draft genome alignment rates (bottom).

Table 3. Intersection of individual scores from both lineage separation metrics in figure 22, this table shows the perfect concordance of these metrics.

Lineage Identifier Intersection

Sample ID	Principal Component		Lineage Alignment		
	PC1	PC2	'A' Rate	'B' Rate	
CWMC_1.2	**64.21193**	-4.851574	**0.64**	0.552	
CWMC_1.4	**63.69493**	3.542804	**0.652**	0.561	Worms Identified as **lineage 'B'**, identified in the same category in both analyses
CWMC_2.4	**68.07215**	-9.773377	**0.585**	0.509	
CWMC_2.5	**61.45939**	11.193412	**0.624**	0.54	
CWMC_3.3	**65.86026**	-1.283664	**0.643**	0.553	
CWMC_3.6	**64.74442**	5.712063	**0.641**	0.553	
CWMM_2.3	**67.8455**	-16.921207	**0.644**	0.557	
CWMM_3.9	**61.909**	9.868352	**0.648**	0.548	
CWMC_1.1	**-19.29933**	-30.1480973	**0.878**	0.46	
CWMC_1.3	**-18.23981**	-39.8997683	**0.854**	0.443	
CWMC_1.5	**-17.71591**	7.4527081	**0.875**	0.456	
CWMC_1.6	**-15.10786**	8.8624651	**0.875**	0.457	
CWMC_1.7	**-14.14551**	13.3799944	**0.849**	0.45	
CWMC_1.8	**-18.07431**	-9.4510231	**0.887**	0.468	
CWMC_2.1	**-13.59617**	-25.0566349	**0.846**	0.453	
CWMC_2.2	**-16.52544**	0.4291632	**0.872**	0.459	
CWMC_2.8	**-16.10405**	8.3065774	**0.873**	0.466	
CWMC_2.9	**-15.01159**	2.2175491	**0.886**	0.467	
CWMC_3.2	**-14.30385**	10.5853055	**0.868**	0.461	
CWMC_3.5	**-17.19455**	1.2274657	**0.866**	0.454	Worms Identified as **lineage 'A'**, identified in the same category in both analyses
CWMC_3.7	**-18.34074**	-3.4539132	**0.878**	0.457	
CWMC_3.8	**-14.99761**	6.068049	**0.866**	0.459	
CWMM_1.10	**-16.52652**	3.5011453	**0.891**	0.468	
CWMM_1.1	**-18.11346**	-14.3286941	**0.894**	0.469	
CWMM_1.2	**-15.34837**	-36.6257094	**0.87**	0.473	
CWMM_1.3	**-17.36745**	8.9451205	**0.876**	0.459	
CWMM_1.4	**-13.03161**	12.9758288	**0.876**	0.465	
CWMM_1.5	**-13.83473**	18.1442847	**0.867**	0.463	
CWMM_1.7	**-15.59906**	13.236664	**0.894**	0.464	
CWMM_1.8	**-17.29132**	-17.6467317	**0.883**	0.463	
CWMM_1.9	**-18.67533**	0.3825125	**0.879**	0.457	
CWMM_3.1	**-16.8167**	12.3894567	**0.884**	0.461	
CWMM_3.2	**-16.92392**	9.6137812	**0.884**	0.466	
CWMM_3.3	**-13.96111**	8.1540588	**0.862**	0.453	
CWMM_3.4	**-16.05824**	9.0678384	**0.857**	0.449	
CWMM_3.7	**-16.63925**	2.2693346	**0.863**	0.458	

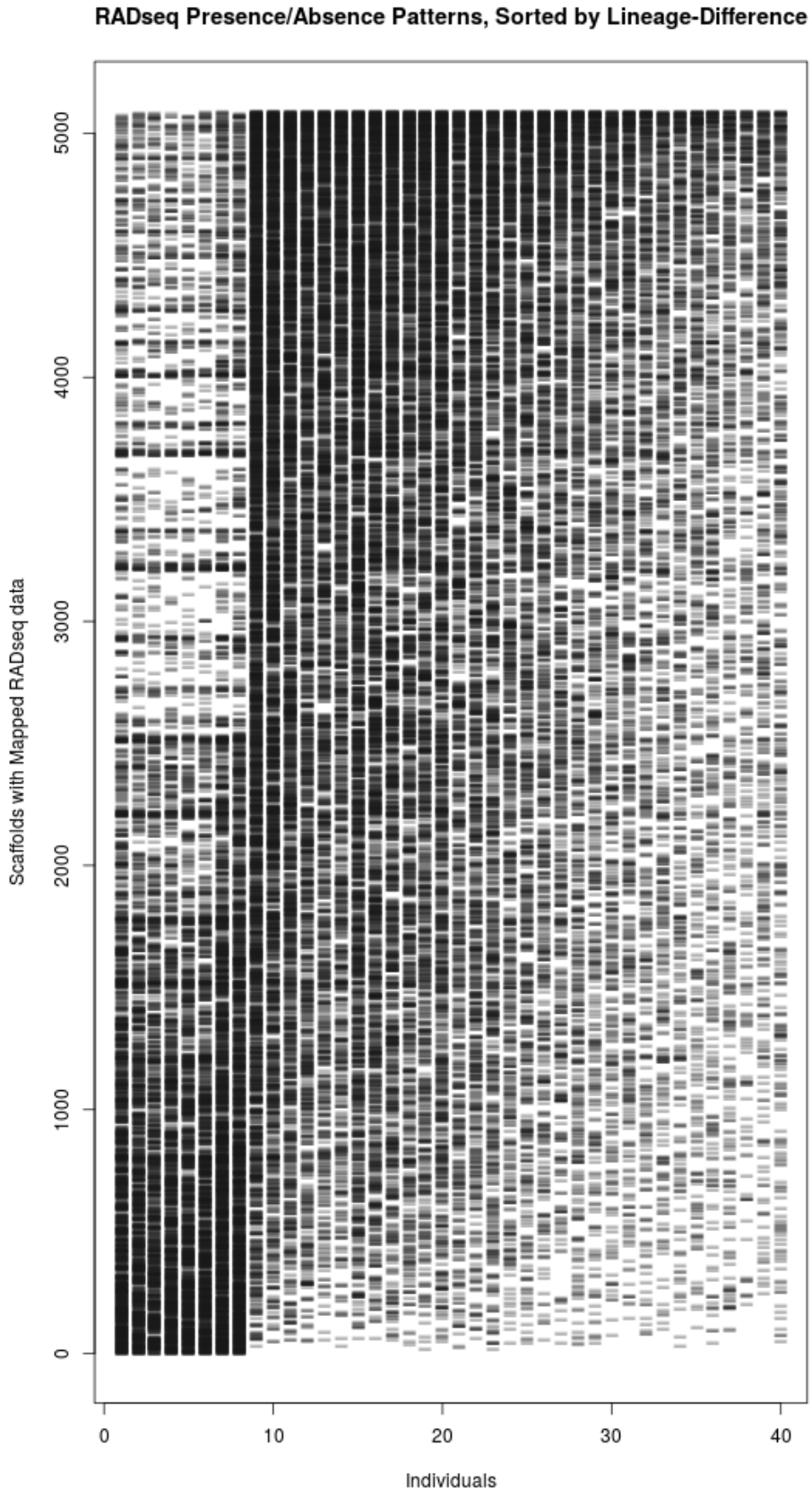


Figure 23. RAD-Seq stack presence/absence patterns as a lineage separator visualised. Input data also used in Figure 22 (top)

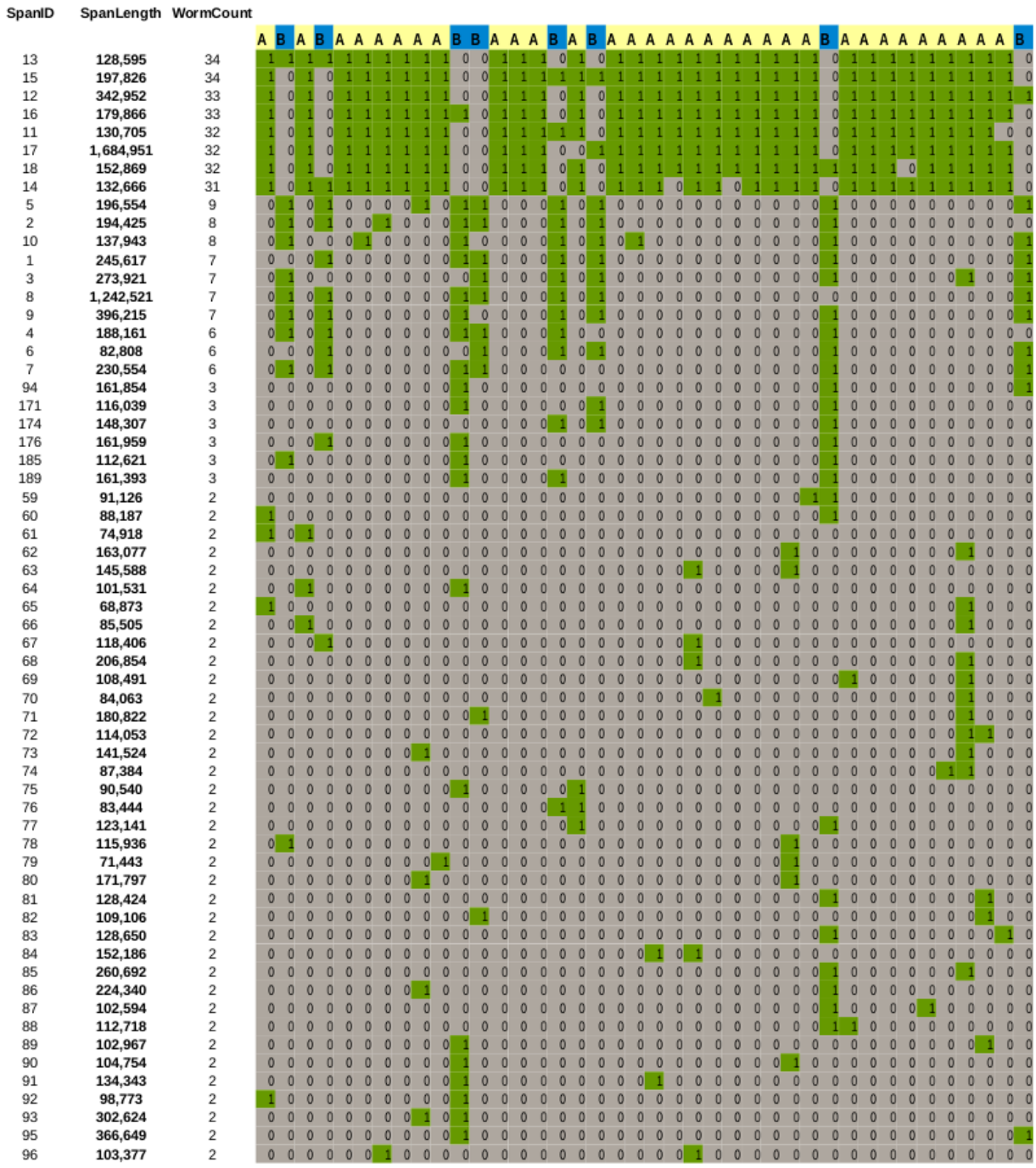


Figure 24. Merged scaffold collections (Spans) with identical RAD-Seq presence/absence patterns. Lineage classification from figure 22/table 3 is show as column headers.

2.2.9. Validation of Alleles in *Lumbricus rubellus* Assembly

When reads from *L. rubellus* individual S20 were mapped to the original repeat masked genome, the fragment pair read-depths for divided into bins that are consistent with presence/presence (full coverage), presence/absence (half coverage) and absence/absence of the mapped alleles (Figure 25). As would be expected, the reads from the reference genome individual (sample S18) mapped back onto itself show a consistently high coverage level. Those from the test individual (S20) show a proportion at half, and zero coverage, suggesting that those locations also have high polymorphism in the test individual.

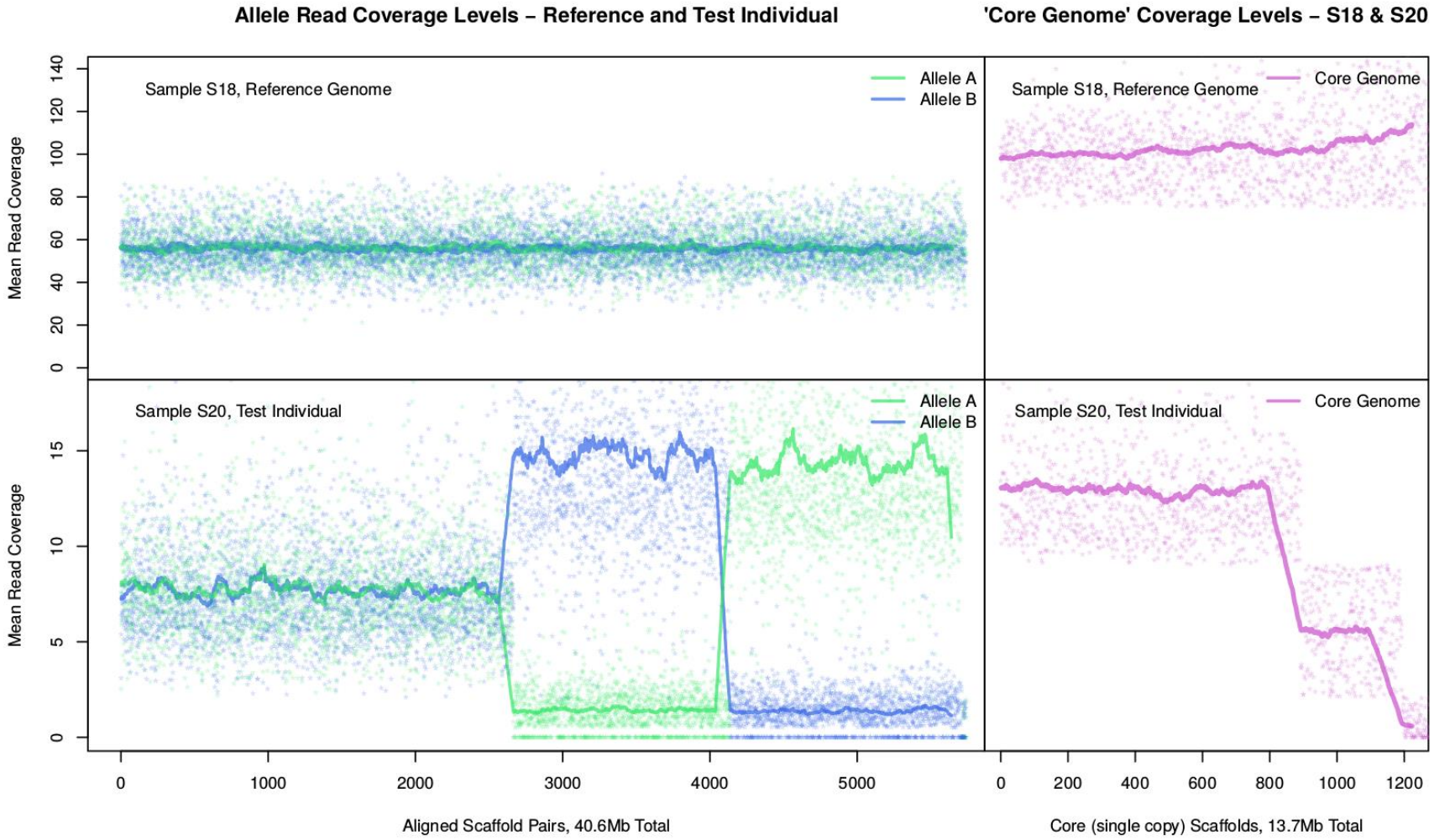


Figure 25. Validation of Lumbricus rubellus alleles. Divergent allele pairs (left) vs Non-divergent collapsed scaffolds (right), comparison between reference individual S18 (top) and test individual S20 (bottom).

2.2.10. Motifs Conserved in Divergent Alleles

To investigate the question of how such divergent alleles could maintain compatibility, analysis was conducted to identify conserved sequence motifs within divergent regions. The most frequent 32-mers were identified and aligned: For each organism, 4 archetypal groups of ~60bp long motifs were extracted from the top 250 unique sequences. Each of these has a central ~25bp region which is highly consistent. These motifs are shown in Figure 26, with the size of the nucleotide letter proportional to how conserved that base is across all motifs in that group. The flanking regions on each motif have variable ranges of consistency. The allelic mutation rate within each of these identified motifs was described as simple per-base frequencies in Figure 27: Each motif has a central region of substantially

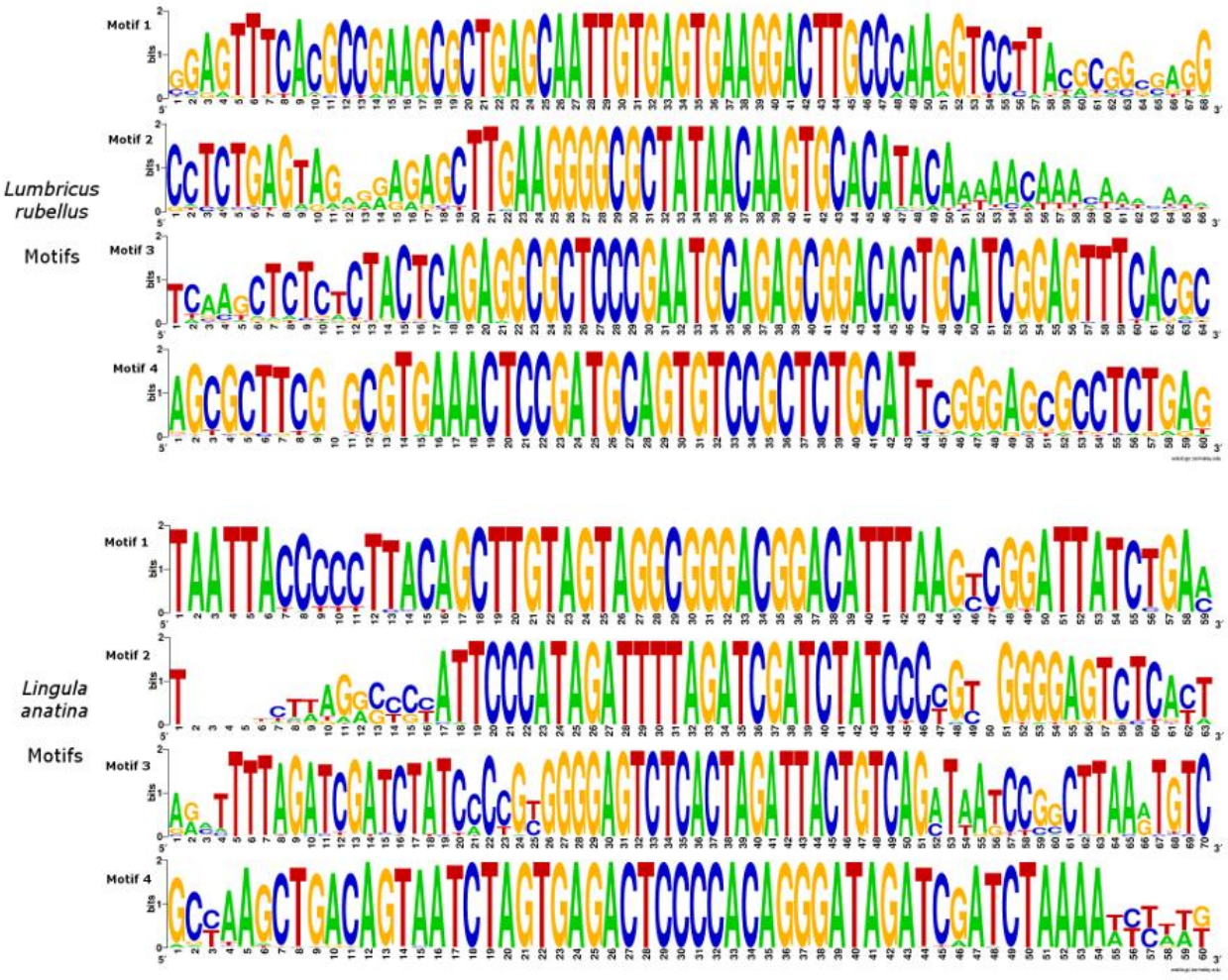


Figure 26. Two sets of top four maximally conserved long motifs in hyper-divergent allelic fragment pairs. Lumbricus rubellus (top), and Lingula anatina (bottom). Weblogo format: Letter size equates to proportion of motif set which exhibits that base.

lower allelic mutation rates that corresponds to the high consistency regions seen in Figure 26. Overall the mutation rates along each motif are always consistently below the mean allelic divergence for the corresponding organism.

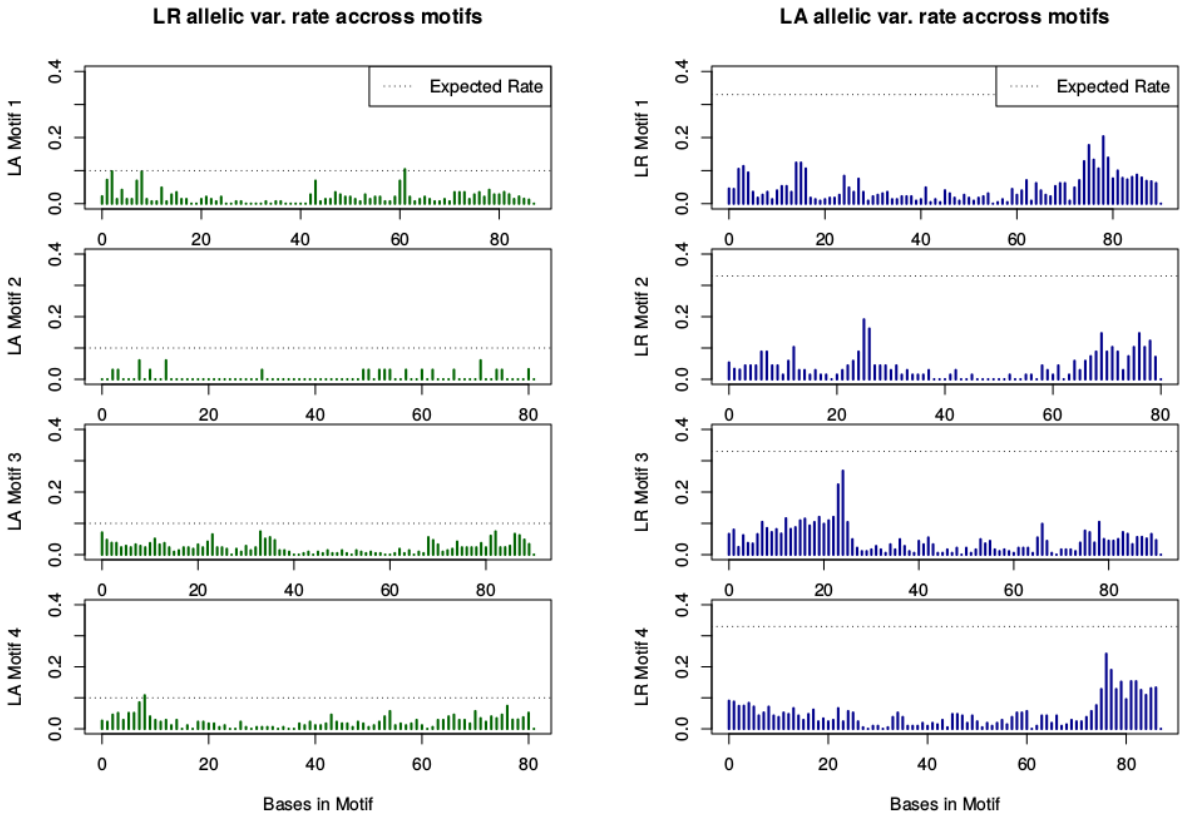


Figure 27. Presence of motifs plotted across the genome was indicator of divergence. Motifs which make up the diagrams in Figure 26. Lumbricus rubellus (left) and Lingula anatina (right). Y-axis values are base substitution frequencies.

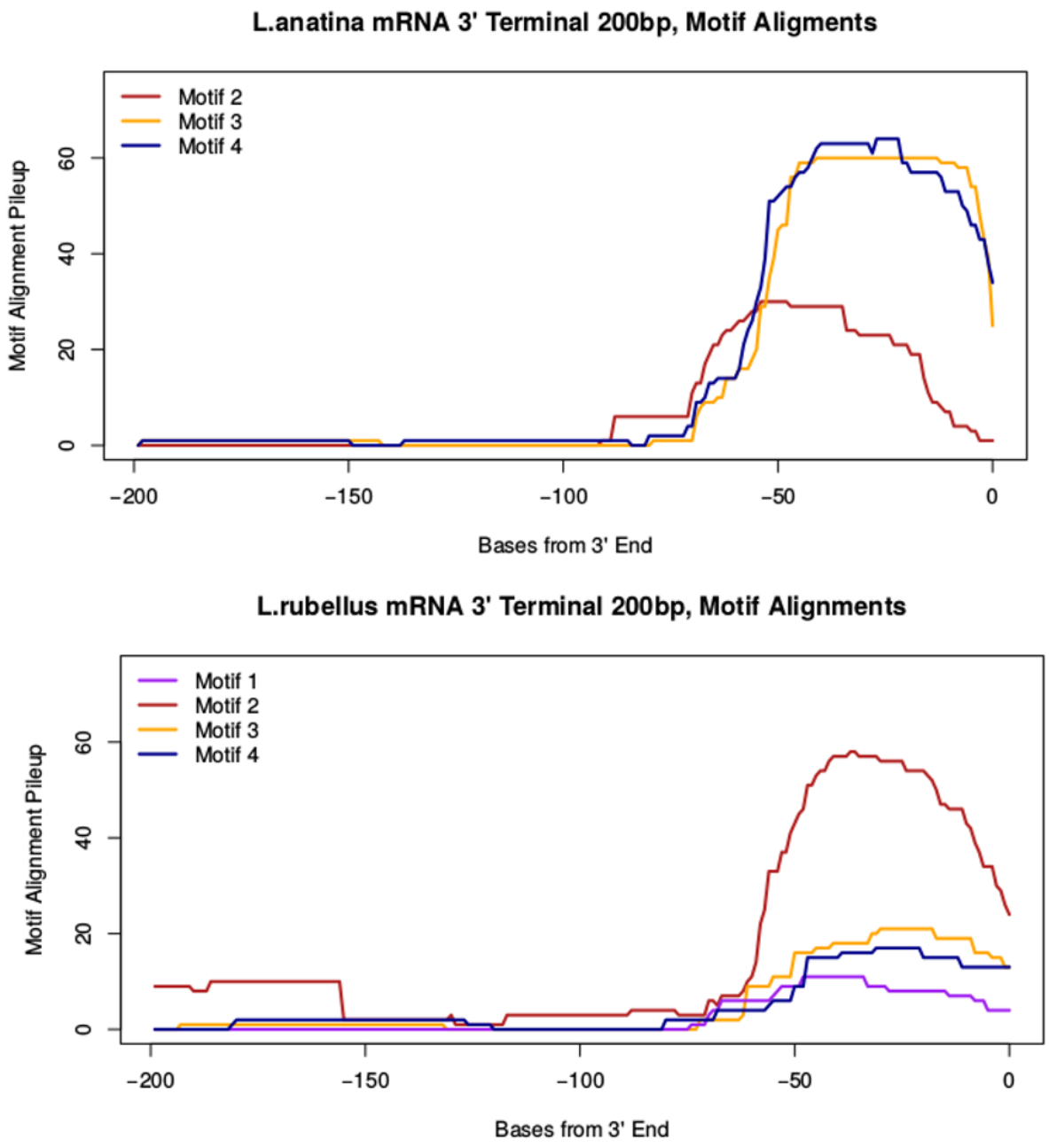


Figure 28. Alignment pileup of motifs in figure 26 against 3' UTRs of expressed mRNAs, Lingula anatina (top), Lumbricus rubellus (bottom).

2.3. Discussion

2.3.1. Summary of Key Findings

The results presented demonstrate that both the *L. anatina* and *L. rubellus* contain a large proportion of their genome that is highly polymorphic (see Figure 5 and 13), with combined (indel and SNP) absolute sequence divergence rates of 10.1% and 33.6% respectively. The levels of polymorphism detected within single individuals from these organisms is extreme, and in the case of the earthworm *L. rubellus*, unprecedented modern genomics. In the case of both organisms, the identification of allelic pairs was also contingent upon some degree of sequence identity between them. Since, in both cases, a proportion of the half-depth fragments could not be well mapped to an allelic counterpart, there is also the possibility that some alleles are either physically unpaired (lone chromosome arms) or have no meaningful homologous counterpart because of large scale inversions, duplications, or transposon events.

Several other studies have identified highly polymorphic genomes, for example:

- Homologous chromosomes of *Pinot Noir* differ by 11.2% (Velasco et al. 2007)
- 3.5% diversity has been detected between two haplotypes of a heterozygous diploid breeding line of tuber crop potato, *Solanum tuberosum* (Xu et al. 2011)
- The genomes of inbred and outbred Pacific oyster *Crassostrea gigas* were 0.73% and 1.3% respectively(G. Zhang et al. 2012)
- The sea urchin *Strongylocentrotus purpuratus* has a SNP polymorphism rate of 4-5% (Sodergren et al. 2006)
- Genome-wide average SNP heterozygosity in the urochordate, *Ciona savignyi* is 4.5% (Small et al. 2007b)
- The freshwater cnidarian *Hydra magnipapillata* showed ~0.7% single nucleotide polymorphism between alleles (Chapman et al. 2010)
- Amphioxus showed 450 Mb with 4% heterozygosity (Huang et al. 2014), and hookworms showed 330 Mb with <1% heterozygosity (Schwarz et al. 2015)

Until now, the most extreme case of between-individual genome diversity is probably *Schizophyllum commune* (Safonova et al. 2014), the model organism, wood-degrading mushroom where genomes of two different individuals of the same species have been shown to differ by 7–12% (and up to 25% if collected on different continents). However, none of these come close to the *within*-individual diversity that has been identified in *L. rubellus.* This level of polymorphism would be typical of inter-Genera/Family/Order divergences in vertebrates/insects/plants. This extreme nucleotide sequence divergence exists as a mosaic across the whole genome (see Figures 8 and 9). This divergence is also

functional, corresponding to allelic variation at the amino acid level (see Figures 17, 18 and 19). It appears that, rather than occurring as isolated islands that are under differential selection and a starting point for speciation (Martin et al. 2013), these highly divergent regions can move as discrete genetic units, or divergent alleles, between both individuals (see Figure 25) and "lineages" (see Figures 23 and 24).

These divergent alleles can have huge functional differences that can be tolerated *within* individuals (see Figure 20 and 21). This genome characteristic would allow for populations to be able to adapt to extreme changes in environment by maintaining heterosis and introgression potential between populations, lineages, or species complexes. There are conserved sequence motifs that are associated with the 3' ends of genes in these divergent regions (see Figures 26, 27 and 28), which are likely to be involved in maintaining the compatibility of these alleles which in many cases would otherwise be likely be non-complimentary. These results suggest that a paradigmatic shift may be needed in the way that we analyse evolution in some non-model organisms. Rather than focussing on nucleotide diversity, a wider perspective may be needed by considering the wide-scale changing of active/inactive status of genes and pseudo-genes and finding ways to characterise large-scale variations in genomic structure within and between populations. One such approach could be the modified RAD-tag methodology that was used to summarise the genome-wide diversity in a population containing two *L. rubellus* lineages at Cwmythswyth mine in North Wales (see Figures 23 and 24). This analysis shows that while there are lineage-based differences within this population, there is also a continuous gradient of sequence divergence between the two.

It is also probable that many other sequencing projects that encounter organisms with this degree of divergence are unlikely to have been successful with shotgun-style NGS techniques, especially with most assemblers being geared towards lower divergence genomes. Those that achieve statistically 'good' assembly (high N50, low fragment counts) are also at risk of having been incorrectly assembled (good example is *L. anatina*), leading to highly inflated genomes and inaccurate gene family analysis.

2.3.2. Divergent Gene Families

The considerable differences in sizes of homologous gene families between the two genomes analysed here point to key differences in the lifestyles of the two organisms. This is true of large, well-annotated families that would be of high importance for adaptation to these organisms' environment, such as epithelial sodium channels, GPCR chemoreceptors, glucuronosyltransferase, mucin-like glycoproteins and ZIP metal transporters (see Figure 21). Here follows a detailed consideration of the adaptive merits of the large protein families identified to by highly divergent.

- **GPCR Chemoreceptors:** Transmembrane proteins that are essential components of the olfaction/gustation mechanism (Skoufos et al. 2000). In both *L. rubellus* and *L. anatina*, chemoreception would be an essential biological tool for sensing the surrounding environment. The substantially larger number of domains associated with chemoreception in *L. anatina* may be due to their feeding behaviour. As a scavenger, *L. anatina* needs to have highly adapted gustative senses to detect the biological matter available in their marine environment. As a saprophage, *L. rubellus* still has substantial chemoreceptive needs, however the smaller range of developed proteins in this family may represent a reduced nutrient seeking need, relative to *L. anatina*. This protein family is the least divergent of those in the curated set. Functionally, the reception of small molecules may be highly sensitive to small changes in active site conformation (Simpson et al. 2011), something that may be significantly affected by even one altered residue. As a result, even though the allelic protein sequences remain in the 95%+ identity range, this could constitute a huge variety of functional variation.

- **ZIP Metal Transporters:** First discovered in plants (Guerinot 2000), these are a core part of the mechanism by which nutrients may be distributed around the body of an organism. This protein family is much bigger in *L. rubellus* than in *L. anatina*. This may reflect the fact that the *L. rubellus* draft genome individual was taken from Cymstwyth Mines, an old Welsh Zinc, Silver and Lead mine, with high metal content in the surrounding soils. Additionally, it is also the case that the metal geochemistry of terrestrial soils in general is substantially elevated compared to the surface level sea bed sediments in which *L. anatina* resides (Savazzi 1991). In the case of the saprophagic *L. rubellus*, the regulatory need in response to variable metal concentrations is likely to be very high. This diverse range of transporters may allow earthworms such as *L. rubellus* to regulate consistent levels of metal concentrations in their tissues when met with highly variable metal concentrations in the highly degraded biological material that they consume (Frouz et al. 2006). Although metal uptake may be highly variable in *L. anatina*, it is unlikely to see the extremes encountered by *L. rubellus*.

- **Laminins:** Structural extracellular proteins that make up the basal lamina, the non-collagenous structural foundation for most tissues (Durbeej 2010). These genes are present in similar quantities, and at a similar level of divergence in both organisms. In the case of an earthworm living in volcanic soils, it has been observed that significant restructuring of the epithelium may occur as a plastic response to high temperature and pH differences in the volcanic soils (Cunha et al. 2011a). Given the order of magnitude differences in soil geochemistry that worms are required to tolerate to be able to survive in their environment, compared to the relatively homogenous medium of sea water in which marine brachiopods live, it is perhaps surprising that

the laminin families are not more different (in terms of both allelic diversity and family size) between *L. anatina* and *L. rubellus*. This could represent a fundamental limit on the extent to which a laminin-like protein is able to evolve, whilst retaining its functional role.

- **Glucuronosyltransferases:** These enzymes are recognised as highly relevant to drug metabolism in vertebrates (Oda et al. 2015). Their functional pathways are involved with the metabolism of xenobiotics. Dealing with unexpected xenobiotic interference encountered during nutrient intake is of course an evolutionary challenge for most organisms at some point or another. When considered in the context of an organism with global scale range, the accompanying range of xenobiotics is likely also vast.

- **Epithelial Sodium Channels:** Ion channel family involved in a wide variety of sodium pathways, in various tissues across metazoan lifeforms (Kellenberger & Schild 2002). Although the functional relevance of the sodium gateways which are employed is broad, in a similar vein to ZIP-transporters, the homeostatic maintenance of ion-channel balances is highly relevant to an organism which may experience variable ion-content in the substrates it inhabits.

- **Mucin-like glycoproteins:** Mucins are a primary constituent of the mucus layer found on luminal tissue surfaces (Gendler & Spicer 1995). Mucus forms a selective barrier with the external world, which is often protective of the underlaying tissues. *Lumbricus rubellus* has dedicated mucus cells which secrete mucus across the epithelial surface , the cuticle (Lavelle 1997a). This mucus layer has been shown to be environmentally interactive with soil pH levels(Schrader 1994), and metal toxicity (Sizmur et al. 2011), although more broadly mucosal layers are known to have many environmentally interactive functions, from digestion to immune defence. Mucus is also incredibly important as a point of environmental interaction for *Lingula anatina* and most reborrowing sediment dwellers. In addition to the above stated generalities, these organisms used mucus to give structural support to the burrows they create in soft sediments. The mucosal interaction with the sediment will vary based on sediment physical qualities chemical composition. The entire lifestyle of *Lingula* hinges on its successful creation of reusable burrows, central to which is the mucosal lining (Savazzi 1991). The mucin-like domains found to be highly divergent are also relatively small, which is perhaps not surprising as mucins are, save for the larger tandem repeat domains, typically encoded by many smaller exons (Ferez-Vilar & Hill 1999).

2.3.3. Failure to Speciate; Failure to Homogenize

Research in butterfly genomics has demonstrated that during the divergence and speciation of several strains, back-crossing and introgression events continuously occurred (Martin et al. 2013). Even between separate 'species' hybridisations may still occur, presumably as a means of adaptive introgression (Pardo-Diaz et al. 2012). The Lepidopterans also give us a clear example of what has become known as 'genomic islands of speciation'. These genomic regions are thought to be responsible for the emergence of reproductively isolated sub-species (Cruickshank & Hahn 2014), although their origins are not necessarily defined by their permeability. These regions may in fact be the cradle of functional genes which cause speciation with some commonality across species (Nosil & Schluter 2011), however other studies have shown that the functional genes for speciation may not occur in islands at all, and that the drivers of speciation may also be singular and widespread(Michel et al. 2010). The permeability of a species boundary is therefore tied to the definition of the species: If the boundary was completely permeable, there could not be said to be any speciation.

The analysed genomes appear to exhibit a relatively extreme manifestation of this idea of boundary permeability, to the point where the case could be made that there is no real species boundary, despite the high absolute base divergence. However, this does not discount phenotypic and behavioural differences between sub-species. There is evidence that the 'A/B' mitochondrial lineages of *L. rubellus* tend towards insular mate choices (Jones et al. 2016). When these observations are considered in the context of greater than 40% divergent allelic content in the assembled genome, the *rubellus* sub-species complex genome could be considered functionally 'lineage agnostic' (despite having 'lineage specific' origins).

Additionally, as *L. rubellus* often lives in stationary communities, which may inbreed and even self-fertilise, the question of how its heterozygosity is maintained is not trivial. In genome of the flatworm, *Schmidtea mediterranea* (Grohme et al. 2018), consistent levels of heterozygosity were maintained in the face of multiple generations of inbreeding, and inheritance was demonstrated to be non-mendelian in nature (Guo et al. 2016). It is currently unknown how this can occur, and whether this is even the case in other organisms, but *that it can* occur may help to guide research which explains how the extreme levels of heterozygosity observed have been maintained, rather than being gradually homogenised in in *L. rubellus* communities.

This novel presence/absence RAD-Seq analysis demonstrates that the *rubellus* genome possesses only 3-5 Mb of DNA exclusively present in other members of its mitochondrial lineage (B). Similarly, the genome also contains 3-5 Mb of DNA possessed solely by the other lineage (A) (see Figure 24). These

sizes closely mirror studies of sub-species reproductive isolation in *Anopheles gambiae* (Turner et al. 2005). In these cases, pericentromeric islands appear to be genetically linked to species despite physical non-linkage, however it is not necessarily true that these islands are causal in speciation (White et al. 2010). It is therefore uncertain whether the lineage specific DNA seen in *Lumbricus rubellus* is lineage-defining, of cross-lineage incompatibility, or simply a product of low pericentromeric recombination rates and chance.

Given the sequence distance between the lineages, it may well be that gene mutations and/or islands of DNA that might cause reproductive isolation in these genomes are strongly selected against. In a broadcast spawner such as *anatina*, the ability to flexibly hybridize across the breadth of the oceanic population is likely key to the reproductive success of any given individual. For the earthworm, dispersal ability is so short that genotypes leading to phenotypic reproductive isolation could lead to debilitating range restriction for that nascent sub-species. As far as the consequences for species classification, it could be suggested that a separate species label perhaps ought to be a last resort when no pre-existing sub-species complex will accommodate the lineage in question. A case in point where Annelids again show a hybridisation capacity that is challenging their classical taxonomic species labels is *Eisenia fetida* vs *Eisenia andrei* (Plytycz et al. 2018), which readily produce fertile offspring.

This failure to speciate also seems to extend to morphology. Both organisms are already recognised as examples of 'morphological stasis' in the animal kingdom. *L. anatina* belongs to the *Brachiopoda* sub-phyla *Linguliformea*, which has evidence of phenotypic consistency in fossil records dating back to the Cambrian era (Williams et al. 1996). *L. rubellus* and many other Annelida earthworms have also been reported to possess huge cryptic diversity within many previously identified single 'morphospecies' (King et al. 2008)(Novo et al. 2012)(Novo et al. 2010). Both these organisms have been subject to an environmental canalisation of their phenotypes: There has been little scope for nature to improve on the sand-burying shell structure of the Lingulate for half a billion years. Although comparatively poor, the fossil record for Lumbricid worms suggest that there has been little scope for further optimisation of their soil-burrowing (Savazzi 1991) musculature since the Cretaceous (Domínguez et al. 2015). In both cases, species' morphotypes respond flexibly only around the range of demands presented by their chosen substrate (Williams et al. 1994). It seems a straightforward suggestion therefore that the 'undisruptability' of the base morphotype by small, or even significant genotypic alterations would be an extremely valuable evolutionary trait in such an organism, particularly when this grants it the capacity to support, allelically, wide divergence in genetic response to the varied chemical challenges encountered throughout its global reach.

2.3.4. Extension of the ‘Meselsen Effect’

The 'Meselsen Effect' theory of post asexual evolution, originating with bdelloid rotifers, posits that in the absence of conventional meiotic recombination, selective pressure is released from one copy of a pair of alleles/genes/heritable units (Welch 2000). This is in part because both are now guaranteed to be co-inherited, and in part because they no longer must re-combine. One might describe this as the emergence of two haploid selective spaces in the place of a diploid one.

It could be suggested here that the ancient sub-species complex achieves a similar effect. It appears that two copies of an allele within certain single species populations may end up functionally divergent to the extent that recombination between them is no longer functionally viable, or even physically possible. It also appears to be the case that both alleles may yet remain compatible in a single individual. In this case it could be proposed that lineage-heterologous individuals benefit from functional diversity, whilst lineage-homologous individuals may yet, through meiotic recombination, allow the otherwise non-recombinant allelic copies to avoid ‘Muller’s Ratchet’ (Haigh 1978). The evolutionary process thus allows for two diploid selective spaces in the place of just one.

The intraspecific existence of separate selective spaces at the same genomic location however, is likely reliant on at least two factors: 1) Individual tolerance of extremely divergence allelic base sequence, and 2) Populations that are large, wide ranging and adequately structured such that both selective spaces are maintained and may co-evolve. These factors occur in the case of the two organisms present in this study.

Previously evidence for genetic exchange between subspecies suggests there to be residual ‘introgressions’ between lineages on the path to speciation (Martin et al. 2013). The researcher is asked to consider that in some species lineages may not be on the path to speciation at all and may instead be utilising the vast genetic reserve of a sub-species complex as an intraspecific mechanism to remain evolutionarily robust to global environmental changes, and with a myriad of selective options.

2.3.5. Allelic Compatibility

Organisms that are genotypically outbred often show some form of heterosis (Comings & MacMurray 2000). Outbreeding also has its limits, and more extreme forms can result in outbreeding depression (Frankham et al. 2011). As the organisms in the present study might be considered examples of extreme outbreeding from a genotypic perspective, the consequent inquiry perhaps ought to be of

how outbreeding depression is mitigated and/or overcome at the molecular level. If tolerance of extreme diversity exists, it would not be a huge leap to assume the corollary to be some retention of allelic compatibility in the allometric process of phenotype. At the sequence level it is possible to investigate this in terms of gene expression regulation. The regulation of gene expression occurs in a variety of manners, and whilst some of these are epigenetic the most well-known are in some way tied to base sequence. Pre-translatory regulatory protein binding sites are predominantly sequence based, as are miRNA and other post-translatory binding sites. With absolute allelic sequence divergence as high as it is in *L. anatina* and *L. rubellus*, one could expect that even short regulatory motifs should stand out in a manner that would not be visible in an organism with a lower absolute divergence. Taxa with this tolerance for sequence divergence are therefore important biological models for furthering the understanding of genetic regulatory systems.

The results presented here using the motif discovery method suggest that 3' UTR regulatory structures in both these organisms are of particular importance to their gene regulatory networks, as they have been conserved between lineages in both genomes. This suggests that there is a 'static' component of genetic regulation, which must remain under purifying selection in all sub-species during allopatric lineage divergence. This is of course the case with all organisms with respect to house-keeping genes and the constituents of our shared molecular biology. These conserved motifs may maintain reproductive compatibility between lineages, providing a regulatory feature which maintains allelic compatibility.

Roux et al (2013) posit that researchers should attempt to "identify the level of sequence divergence above which introgression is definitely impossible" (Roux et al. 2013). The results suggest that absolute sequence divergence is less likely to be the causal preventative factor of introgression/hybridization. Instead, it could be argued that taxa will often impose a self-selected pre-zygotic mechanism to avoid hybridization. The adaptability and evolutionary potential that cross-lineage hybridization may give to an earthworm or a brachiopod will come at the expense of specialisation and could be highly detrimental in other organisms. For example, reproductive isolation has been observed in only three generations in Darwin's finches, following a homoploid hybridisation event (Lamichhaney et al. 2018).

2.4. Conclusion

Divergent regions of *L. anatina* genome were found to have an average of 10.1% absolute divergence. In *L. rubellus* the number was higher, at 33.6%, demonstrating substantial regular outbreeding. This divergence in sequence is shown also to affect functional components of these genomes and probably provides substantial environmental plasticity, with important environmentally adaptive protein families being highly functionally diverged in the alleles of both organisms. However, the activity of transposable repeats only appeared to play a small role in the sequence divergence, accounting for only ~5% of allelic sequence mismatch in both cases. These genomes contain core regulatory sequences that must consistently survive population divergence events and may be key to the organisms' management of extreme functional changes within highly divergent alleles.

One of the biggest advantages of hybridization in the face of extreme sequence divergence would be range expansion. In plants particularly, range has been shown to expand following hybridisation (Ellstrand & Schierenbeck 2000)(Culley & Hardiman 2009). The plasticity and adaptability granted by outbreeding has often been associated with invasiveness. Indeed, several of the morphotypically stable organisms (e.g. *Crassostrea gigas* and *L. rubellus*) are known to be invasive (Moehler et al. 2011). In the longer term, maintaining a supply of highly functionally divergent genes for any one genomic loci would act as a great insurance policy against large scale environmental changes. However, to maintain the allelic diversity, it would also be imperative that extensive interbreeding does not limit diversity by over-purifying the functional content of the alleles. The results from *Lumbricus rubellus* population data demonstrated that a high degree of ongoing genetic exchange between two interbreeding mitochondrial lineages has not resulted in the homogenisation of the corresponding nuclear lineages. Each lineage appears to maintain some degree of distinctiveness, which allows genetic exchange whilst also allowing their “reserves” of genetic potential to co-evolve. How this extreme polymorphism is not rapidly eroded by genetic drift is still uncertain. The hypothesis proposed is that these genome-wide mosaics of extreme sequence divergence represent a strategy that allows species with limited post-zygotic dispersal (such as brachiopods and earthworms) to avoid evolutionary dead-ends. It would therefore be possible to predict that this structure will increasingly be observed in other non-model taxa with these life-history traits, as sequencing technologies and analysis pipelines are further improved.

3. Chapter 3: Genomic Diversity, Epigenetics and Gene Expression: Signatures of Plasticity and Stress in an Invasive Earthworm

3.0. Introduction

3.0.1. Earthworm Diversity and Range

Globally, earthworms are a crucial component of soil's functional harbour for countless forms of life. Their impact on soil health has deep relevance to agriculture and most terrestrial plants. Their population levels also impact many small predators for which they can be an essential or secondary, nutrient reserve. *Amynthas gracilis* is a coloniser. Originating in East Asia, it has spread from there to the Mediterranean, and North and South America (Blakemore 2012). It is a highly successful invasive species and with a current pan-tropical distribution. In modern history this earthworm has been introduced into the Azorean Island of Sao Miguel through anthropogenic activities most likely associated with multiple agricultural transplantations.

In plants, hybridisation events have been found to be associated with the development of invasiveness on multiple occasions (Pfennig et al. 2016)(Seehausen 2004)(Aïnouche et al. 2009). It is also a commonly observed issue whereby invading species hybridise with native populations (Hurka et al. 2003)(Shields et al. 2010). In these cases, it is posited that the genetic diversity gained in hybridisation is a significant factor in their ability to adapt to new environments, and previously untested selective pressures. Meta-analysis have also suggested a combination of environmental plasticity and the speed of responsiveness to selection as primary factors in a species effectiveness as an invader (Ellstrand & Schierenbeck 2000)(Ellstrand 2009)(Schierenbeck & Ellstrand 2008).

It has also been observed that earthworm, across a wide variety of taxa, appear to exhibit a wide variety of cryptic species. This has been observed as a wide range in mitochondrial lineages (King et al. 2008). For example, specimens of invasive *Amynthas* worms in north American forests seem to have very flexible digestive capabilities, allowing it to thrive in new environments, possibly as a result (Zhang et al. 2010). The high, and highly variable rate of polymorphism in the sequenced *A. gracilis* genome seems to suggest that the exceptionally high variation maintained in its natural population has been sustained by regular hybridisation between these cryptic lineages.

3.0.2. Stress Responses in the Soil

Earthworm adaptive and stress responses to soil contaminants are of interest for various reasons. The health of the soil as an agricultural medium, as the bedrock of an ecosystem, or as a marker of the impact of human activity are directly related to the biophysical tolerances of the earthworm. Stressors experienced by soil-dwelling organisms can be highly multivariate, comprising changes by

many orders of magnitude in soil pH, salt concentrations, heavy mental abundance, concentrations of microplastics and pesticides. Further complexity is added by the intersection of soil types, i.e. clay, loam, peat, sandy or silty, and other variables such as temperature and moisture availability. As a result, many of these stressors have been studied either singularly, in pairs, or all-together under the umbrella of 'multi-stressor' environments. No unified model of their stress response interactivity between species is available, however there are many clues from both past and recent studies as to how earthworms might be expected to respond to a given alteration in environment.

For example, it has been shown that the signatures of coelomocyte metabolic processes and DNA damage metrics in A*mynthas gracilis* provide sub-lethal indicators of a stress response to agricultural pollutants (Parelho et al. 2017). A review of pesticide toxicity to earthworms found fungicides and insecticides to have the most pernicious effects, although most were lifestyle stressors (Pelosi et al. 2014). A study has shown that the worm possesses functional responses which limit photooxidation by exposure to UV radiation (Chuang & Chen 2013). A multi-species population abundance study of terrestrial earthworms demonstrated that motor oil as a soil pollutant was highly toxic to most species, resulting in lower populations in roadside areas (Ramadass et al. 2015).

As far as abiotic factors in earthworm survival are concerned, by far the largest area of study appears to have been responses to elevated metal content in the soils. Heavy metals are often studied for their stress response effects. In a study of the terrestrial earthworm *Lumbricus rubellus* it was shown that gene expression changes may be the most sensitive measurable response types in metal-toxicity stress related cascades (Spurgeon et al. 2005), when compared to higher order changes such as fecundity and community diversity. *Lumbricus terrestris* has also been studied for sub-lethal metal toxicity responses, these studies find that glutathione reductase and metallothionein concentration increases were detected as stress responses to the oxidative stress created by mercury contamination (Colacevich et al. 2011). More generally, one study was able to show that *Eisenia fetida* increased the bioavailability and environmental motility of heavy metals in certain soils (Wen et al. 2004). However, the interactions that *Eisenia fetdia* has with metals in the soil have been shown in multiple studies to be substantially modulated by other factors, in particular soil pH (Spurgeon & Hopkin 1996) (Heggelund et al. 2014). These relationships suggest that the intersection of multiple stressors should probably not be considered as limited to summative action.

It should also be noted that broader surveys of earthworm vs soil type distributions indicate quite strongly that certain species are far more likely to naturally inhabit soils which another might experience as a stressor, such as those which range from 3-9 pH (Jänsch et al. 2013). As a result, we

ought to consider whether an environment might be stressful to an earthworm in context of the range within its typical habitats.

3.0.3. Epigenetics and Plasticity

Epigenetics is a broad term which describes the range of modifications possible to the genome of an organism. The ongoing research into epigenetics can be divided between considerations of their germ-line heritability and their acute somatic functionality. The scope of this work pertains solely to somatic cells. Of the various epigenetic features identified in the genome DNA methylation is probably the most well-studied at this point. DNA methylation in this context concerns the covalent bonding of methyl group to the base cytosine's 5^{th} carbon atom in its aromatic ring, to create 5-methyl-cytosine (Schübeler 2015). These groups are added to the cytosine base by methyltransferases, with the most significant players in eukaryotes recognised as being DNMT1 and DNMT3 (Goll & Bestor 2005).

Although initially associated with gene expression silencing in humans (Cedar & Bergman 2012), and of great functional relevance to various human pathologies including cancer research (Jones & Laird 1999), work is only recently beginning to be carried out on its roles in non-model organism DNA methylation. The interest which epigenetics might hold for studies of environmental response is largely in the phenomenon of 'plasticity' – this is the ability of an organism's genetics to dynamically alter its phenotype in response to certain environmental triggers (Pigliucci et al. 2006). This moves beyond the description merely of stress responses, to evolutionarily assembled developmental bifurcations which can be selected for their adaptive qualities in suitable environments. These alterations can occur simply within an organism's early development, altering the form of the adult in a predictive adaptive response (Gluckman et al. 2005), or can occur as labile plasticity throughout the organisms life. This study is concerned with the labile plasticity of *Amynthas gracilis* as it responds to different soils, and the difference between the deployment of existing adaptive developmental equipment, and the acute features of a stress response.

The transitive nature of epigenetic change makes it a suitable candidate for the short-term modification of environmentally plastic genetics. Meta-analysis of literature regarding invasiveness suggests strongly that successful species commonly exhibit substantial environmental plasticity (Davidson et al. 2011), although it does not always lead to the same niche competitive advantage achieved by less-plastic native species. *Amynthas gracilis* exhibits substantial phenotypic plasticity and could reasonably be described as widely invasive as a result, it follows that investigation of DNA methylation as supporting mechanism may yield functional clues about the success of this organism.

3.0.4. *Amynthas* in São Miguel's Volcanic Soils

São Miguel is an island in the Portuguese archipelago of the Azores. Its rich soils have been a great boon for local agriculture, with Portuguese colonists having arrived approximately 1500-1600AD. The Island has two quiescent central volcanoes, names Fogo and Furnas. Environmental stressors for include soil pH changes, CO_2 soil degassing (Viveiros et al. 2008), elevated temperatures and raised heavy metal content (Novo et al. 2015).

Earlier work on this population of *Amynthas gracilis* on São Miguel has suggested multiple introductions of the species were likely, due to the evolutionary distance between the mitochondrial lineages discovered. Another interesting highlight of this work was to discover that *Amynthas cortisis* population abundances negative correlated with the abundances of *gracilis*, suggesting their direct competition, with *cortisis* better adapted to soils less affected by the volcano (Novo et al. 2015). The corollary to this relationship might be that the selective advantage is had by *gracilis* in the higher-stress environments via a physiology more equipped for environmental plasticity. These results strengthened the position of an earlier study of the epidermis of *gracilis* worms sampled from active versus inactive volcanic soils. This had demonstrated, with detailed tissue analysis, that the worm is capable of substantial morphogenic restructuring in response to these stressors (Cunha et al. 2011b). A clear example of an organism-level plasticity response.

3.0.5. Primary Aims

The biochemical mechanisms by which plastic responses to environmental stressors are deployed have been a subject of extensive research. Whilst cell-signalling in response to variables such as oxidative stress (Martindale & Holbrook 2002) or osmotic stress (Schachtman & Goodger 2008) demonstrates the acute communicative ability of an organism to deploy physiological responses, the study of epigenetics and non-coding RNAs suggest a chronic coding schema which allows an organism to display non-mendelian adaptive trait inheritance (Geeleher et al. 2012) (Liebers et al. 2014). By encoding the cohort of adaptive responses to environmental stressors in flexible yet persistent regulatory systems, it is postulated that the stability of the phenotypic response might be maintained and persisted in sub-Darwinian timeframes. To this end, two extra-genomic regulatory mechanisms were investigated in addition to RNA-Seq: micro-RNA abundance, and DNA (5-cytosine) methylation.

The main aims of this study were to use high throughput sequencing experiments to infer the intersection of regulatory mechanisms in the per-trait adaptation of *A. gracilis* to multi-stressor volcanic soils. We aimed for the identification of a physiological and biochemical trait change cohort, and to establish the contributions to each from different regulatory mechanisms. To

achieve this a reciprocal-design, mesocosm-based, *in situ* transplant experiment was deployed with the intention of measuring three conceptually defined categories of environmental response. 1) Acute general environmental change/stress response, 2) Chronic adaptive trait changes, 3) Specific acclimative changes

By identifying the trait-contribution matrix across the above three test categories against the above three sequencing experiments, this study sought to produce a systematic view of the molecular biology involved in different modes of adaptive plasticity.

3.1. Materials and Methods

3.1.1. Transplant Experimental conditions

The following section was provided by Dr Luis Cunha and Dr Marta Novo. The Azores archipelago comprises nine islands and is in the North Atlantic Ocean, between 36º45'–39º43'N and 24º45'–31º17'W, at the triple junction of Eurasian, African and North American plates, characterized by a complex tectonic settlement, where seismic and volcanic phenomena are common (Cole et al. 1999). São Miguel island belongs to the most eastern group together with Santa Maria, and the latter is the oldest of all nine. São Miguel is the largest island (757 km^2), which presents several active volcanic spots including fumarolic fields, cold and thermal springs and soil diffuse degassing (Viveiros et al. 2008). Two field sites on São Miguel, differing in their contemporary volcanic activity (thermal and degassing outputs), were selected for microcosm exposures: (a) Furnas, which displays the most conspicuous degassing and geothermal activity in the entire Azores archipelago, and (b) Macela, which does not presently display any thermal and degassing phenomena (Table 1). See Figure 102 for an explanatory map of the area.

A group of adult (clitellated) *Amynthas gracilis* from Furnas, 37º 46' 24.6'' N 25º 18' 10.3'' W (S. Miguel, volcanically inactive site) and another group of from Macela, 36° 46' 04.0" N 25° 31' 46.7" W (S. Miguel, volcanically inactive site) were collected by digging and hand-sorting during Spring of 2012, and were assigned to factorial-design treatments (with earthworm sources and exposure sites as factors) within 24 h of collection (Figure 102). Ten individual worms were placed in perforated, cube-shaped, mesh bags (volume 15 L). Twelve bags were used per site, with six bags per 'treatment' (Furnas- or Macela-derived worms, respectively, n=240). Soil from the given exposure site was used as substrate. They were exposed for 31 days.

Figure 102 provides a schematic representation of the experimental design. After sampling, the earthworms were immediately transferred to the laboratory, where they were processed as described in the following section

3.1.2. Sampling and Sequencing

3.1.2.1. Environmental characterisation

The following was performed by Dr Mark Hodson, who then made the data available. The soils were air-dried and sieved to < 2 mm. Subsamples of the soils were oven dried at 105 °C and their moisture content determined to allow all concentrations to be expressed on an oven dried basis. Organic matter content was measured by loss on ignition, soil being oven dried overnight at 105 C and then ignited overnight at 500 C in a muffle furnace (Rowell 1994). pH was measured on suspensions of 10 g air-dried soil in 25 mL deionised water following shaking for 15 minutes. Texture was calculated from the percentage of soil particles in the size ranges < 2 mm, 2 – 63 mm and > 63 mm as determined using a Malvern MasterSizer 200 with a Hydro2000MU wet dispersion unit. One to two grams of air-dried, < 2 mm soil were analysed with an obscuration of between 5 and 25 %. Instrument performance was checked using a Malvern 15- 150 mm quality audit standard and "general purpose sand, 40 – 100 mesh" purchased from Fisher. Rather than measuring soil moisture content in the field during sampling, water holding capacity of the soil samples was determined following the method in ISO guideline 11274 (ISO 1992).

3.1.2.2. Histological processing and morphometry

The following was performed by Dr Luis Cunha, who then provided the dataset for analysis. Two earthworms from each bag were depurated for 36h and fixed in neutral-buffered formaldehyde for 24h, dehydrated in graded ethanol series, and embedded in paraffin wax. Histological sections (4 µm thickness) were cut on a Leitz 1512 microtome (Leica Microsystems, Wetzlar, Germany), mounted on albumin-coated slides (Menzel-Glaser, Braunscheig, Germany), dried at 40°C for 24 h, and stored at room temperature until staining.

Sections were stained with hematoxylin-eosin (Martoja & Martoja-Pierson 1970). Epidermis thickness was measured in 3 sections (4 fields per section), 40 µm apart, in each individual worm. Images were captured using a CoolSNAP-cf camera (Photometrics GmbH, Munich) coupled to a light microscope, and analysed with Image Pro-Plus 5.0 software (Media Cybernetics, Silver Springs). For statistical analysis the average value from 12 measurements per individual earthworm was considered as the true replicate (n= 12 per treatment). Epidermal thickness measurements were analysed (with or without $\log_e$ transformation, as appropriate) by two-way ANOVA and Tukey *post hoc* pairwise comparisons, with $p \leq 0.05$ considered the level of significance.

3.1.2.3. RNA-seq preparation and sequencing

The following was performed by Dr Luis Cunha. Three earthworms from each bag (n=72) were flash frozen in the field with liquid nitrogen and posteriorly powdered in the laboratory using mortar and pestle. RNA was extracted from ca. 50 mg of powder by combining Trizol extraction with column

purification according to manufacture instructions (RNeasy mini kit QIAGEN). Briefly, the powder was homogenized in 1.5 ml of Trizol, and then centrifuged at 12,000g for 5 min. The supernatant (1.2 ml) was transferred to a new Eppendorf tube and incubated for 5 min (RT), then 240 µl of chloroform was added and the mixture incubated for 3 min (RT). The sample was then centrifuged at 15,000 rpm for 15 min (4^{o}C). Upper phase was mixed with 250 µl of absolute ethanol and transferred to the column. After centrifugation at 10,000 rpm for 10s, RNeasy mini kit protocol was followed, finishing after two elution steps with 30 µl of H_2O. RNA quality, integrity and quantity were checked in Bioanalyzer (RNA Nano Chip) and checked for contamination using a Nanodrop. Fourteen samples were poly-A selected and prepared for cDNA library construction, each of them being a pool of three individuals from the same bag or origin (n=12). Truseq RNA paired-end libraries were prepared at GenePool (Edinburgh), multiplexed and sequenced in two lanes of an Illumina Hiseq 2000 (100 cycles).

3.1.2.4. miRNA-Seq and MEDIP-Seq preparation

Samples for epigenetic analysis were prepared by Dr M. Novo and provided to NERC Biomolecular Analysis Facility Edinburgh for miRNA and Zymo Research Ltd for methylation analysis. Briefly samples of *A. gracilis* were collected from the transplantation mesocosms (see sections above) and immediately placed in RNAlaterTM-ICE (ThermoFisher Scientific, AM7030) to preserved DNA integrity. Samples were transported in the preservative at room temperature prior to be placed at -70^{o}C for long term archival. Archived specimens stored in RNAlaterTM-ICE were slowly thawed. DNA was then prepared from 10 segments ~5 segments posterior of the clitellum using DNeasy (Qiagen Ltd). DNA was quality controlled using adsorption spectroscopy with integrity and size being determined using agarose gel electrophoresis (0.4%) using Lambda HindIII and undigested lambda (New England Biolabs) as size markers. Equal amounts of DNA from 5 individuals, for which parallel RNA-Seq had been derived (see section above), were pooled and MeDIP analysis performed (Zymo Research Ltd). Libraries for MeDIP-Seq were prepared following immunoprecipitation using the DNA Methylation IP Kit (Cat #D5101, Zymo Research). Immunoprecipitated DNA was subjected to amplification with a primer that contained part of the adapter sequence in addition to four random nucleotides, followed by two additional steps of amplification to add on the remaining adapter sequence and to barcode the fragments, respectively. All PCR products were purified using the DNA Clean & Concentrator-5TM (Cat#: D4003, Zymo Research). The input DNA library was prepared from pooled sample DNA that was fragmented and denatured. Libraries were quantified using the Agilent 2200 TapeStation and by qPCR. Samples concentrations were normalized to 4 nM, then sequenced on the Illumina HiSeq 2500.

Parallel tissues from 3 of the samples used for RNAseq were processed with miRNeasy (Qiagen Ltd) and quality assess using TapeStation. Samples were supplied to NERC Biomolecular Analysis Facility Edinburgh where libraries were prepared using TruSeq Small RNA library kit (Illumina Inc) and 50 bp single end sequencing performed on an Illumina 2500 HiSeq platform using HiSeq V3 chemistry (Illumina Inc).

3.1.3. Genome Assembly

3.1.3.1. Library Preparation and Processing

DNA was isolated by Dr L. Cunha and provided to NERC Biomolecular Analysis Facility Edinburgh for sequencing. Briefly samples of *A. gracilis* were collected proximal to the Furnas caldera (37°46'23.0"N 25°18'15.7"W) and immediately placed in RNA*later*™-ICE (ThermoFisher Scinetific, AM7030) to preserved DNA integrity. Sample were transported in the preservative at room temperature prior to be placed at -70ºC for long term archival. Archived specimens stored in RNAlaterTM-ICE were slowly thawed and muscle tissue removed by dissection. DNA was then prepared using phenol method for isolation very-high-molecular-weight DNA (Wood 1983). DNA was quality controlled using absorption spectroscopy with integrity and size being determined using agarose gel electrophoresis (0.4%) using Lambda HindIII and undigested lambda (New England Biolabs) as size markers.

Table 4. Genome sequencing library statistics

		Raw Reads		**Post-Trimming & QC**		**Post Error Correction**	
	Insert Size	*Pair Count (Million)*	*Base Count (Gb)*	*Pair Count (Million)*	*Base Count (Gb)*	*Pair Count (Million)*	*Base Count (Gb)*
Library 1 (MP)	3kb	50.7	15.2	49.5	14.84	32.7	9.8
	5kb	67.3	20.2	66.2	19.8	45.4	13.6
Library 2 (PE)	350	128.5	38.6	124.6	37.44	113.2	33.9
	550	133.1	39.9	129.01	38.7	117.1	35.1

Short read libraries were processed with Trimmomatic (Bolger et al. 2014). This removed residual Illumina adapters from the reads. The LEADING and TRAILING parameters were set to 3, removing low quality terminal bases from the reads. The SLIDINGWINDOW parameter was set to 5:15, making the moving average read measurement five bases wide, with a clipping threshold of 15 ('Phred' scale). Table 8, "Post trimming and QC" shows read count reduction because of this process. The 1-2% size reduction indicated that the libraries were generally of high quality.

Error correction was then performed using the software 'Musket' (Liu et al. 2013). This is a k-mer base error correction and filter for genomic sequencing libraries. It searches for kmers in reads which are represented only once or twice in the library (rather that 50-100x) and attempts to find a

next-nearest suitable depth kmer to correct them (assuming the low depth kmer is a result of sequencing error). If there is no suitable correction to be performed on the reads with low-depth kmers, they are removed from the library. This reduced the short-paired library sizes by a further 10%, however it also reduced the mate-pair libraries by nearly 40%, suggesting that the differences in library preparation method may have caused the introduction of substantial error. All libraries were processed in a single execution of Musket. Despite the loss of read-count, this step was of benefit in cleaning up the assembly graphs for an already allelically diverse organism.

3.1.3.2. Assembly QC

SGA (string-graph assembler) is a genome assembly software package, it contains a module 'preqc' which allows the exploration of pre-assembly genome characteristics (Simpson 2014). The error-corrected short-paired libraries were used as input for this program. Outputs can be seen in Figure 71. Both (A) and (B) sections of Figure 71 suggest that Amynthas gracilis is highly allelically divergent, although not to the extent of *Lumbricus rubellus.* This conclusion is drawn from (A) the variant branch rate being comparable with the *Crassostrea gigas* libraries, which has a known absolute divergence rate approximately at 4% (Gerdol et al. 2015), and the bi-modal coverage separation in the GC-coverage graph. The lower coverage group representing the divergent alleles is slightly less dense than the higher 'both allele' group. This suggested the divergence will be close to, but probably slightly lower than 4%. It is also observable that *Lumbricus rubellus* registers a far denser single-allele group that either *Amynthas* or *Crassostrea.*

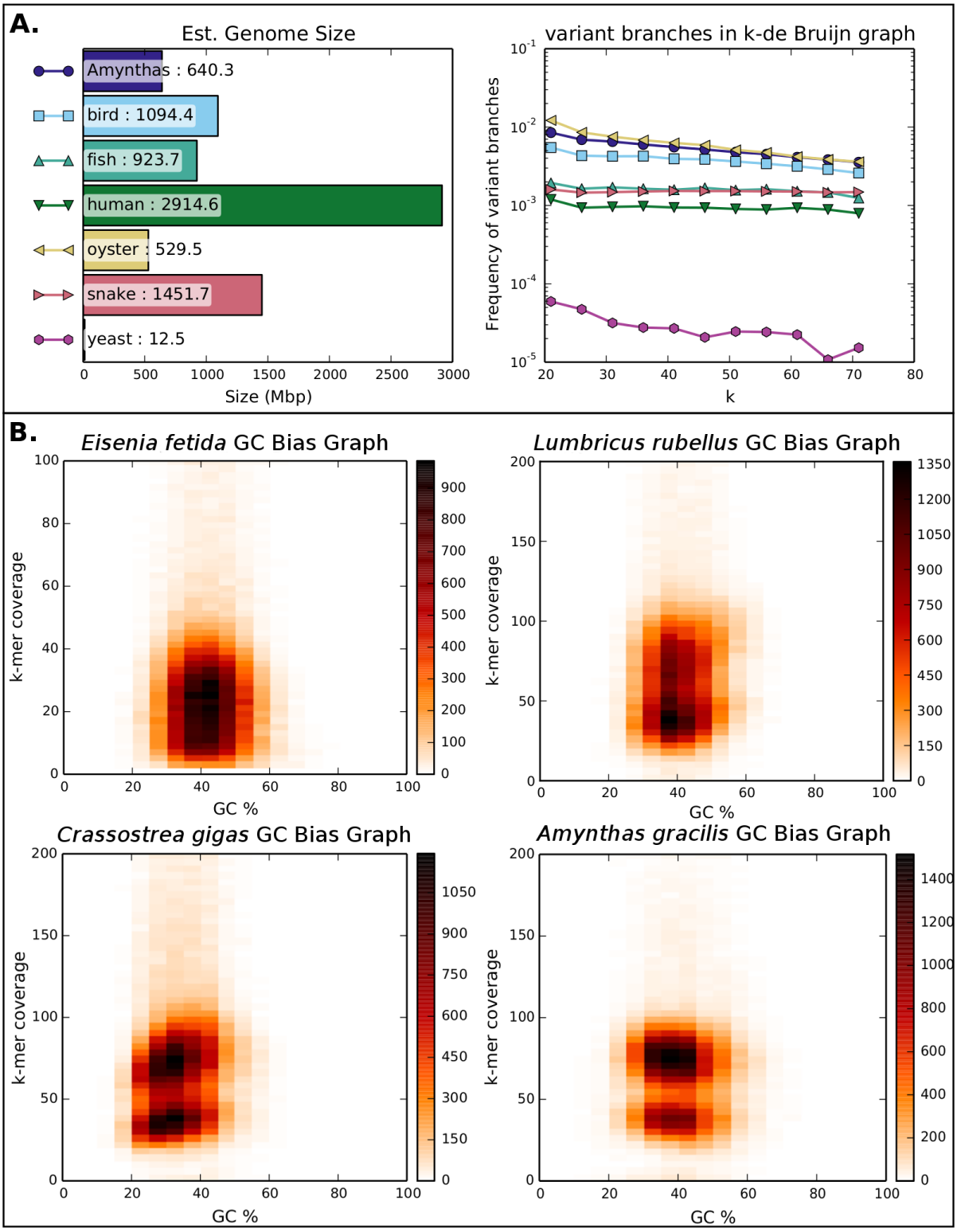


Figure 29.SGA 'preqc' output graphs. (A) Genome size estimation for Amynthas gracilis, and comparison of de Bruijn graph branch rates. (B) GC/Coverage kmer plots, A. gracilis compared to two other earthworms, and an oyster.

3.1.3.3. String graph assembly pipeline

Because of the anticipated high divergence in the genome sequence, the assembler 'Platanus' was selected to perform string graph assembly. It was designed with a fairly aggressive 'bubble-

collapsing' protocol, which seeks to merge the variant branches in string graphs created by allelic divergence (Kajitani et al. 2014). For reference however, another baseline assembly was also performed using the popular package, SOAPdenovo (Luo et al. 2012). Table 9 shows the results of several assemblies that were performed.

Table 5. Assembly Statistics During Finalisation

Stage	Size (Mb)	N50 (Kb)	N90 (Kb)	Scaffolds
SOAPdenovo	835.1	189.5	2.8	202,273
Initial Assembly	824.6	223.3	11.9	48,927
Post Collapsing	775.3	317.4	14.3	45,350
Spatial Selection	589	425.8	115.9	4,846
Finalisation	589.7	478.5	129.9	4,350

As the primary assembly shown in Figure 72 still retained substantial allelic inflation, a custom assembly correction procedure was created. The objective was principally to identify terminal regions of scaffolds which were allelic copies of other terminal regions, and to collapse them together, creating a much larger scaffold. Another issue seen in Figure 72 is that of contamination. Earthworms have vertically transmitted symbionts of the genus *Verminephrobacter* (Pinel et al. 2008). These also feature in the genome assembly; however, they are readily identifiable by their increased GC content. To eliminate the parasites, and other uncollapsed fragments from the main assembly, a spatial selection optimiser was deployed. A summary of the total assembly pipeline is shown in Figure 73. The 'custom collapser' and 'custom optimiser' entries in Figure 73 will be detailed in subsequent sections.

Other software used in the pipeline included SSPACE scaffolder (Boetzer et al. 2011), 'bowtie2' short-read mapper (Langmead et al. 2013) and BLAST+ (Camacho et al. 2009). At each assembly stage the genome was searched for the core metazoan single-copy gene set, and core arthropoda gene-set as described by BUSCO (Simão et al. 2015). The core gene duplication rate was used as a measure of how well the allelic inflation in the assembly has been managed, alongside 'bubble' style plots such as Figures 72 and 74 Tracking of the BUSCO completeness and duplication rates can been seen in Table 3.

Table 6. BUSCO genome completeness at various assembly stages, including other invertebrates

	Genome	N50 (Kbp)	BUSCO assessment
Metazoa (BUSCO examples)	Caenorhabditis elegans	17494	C:85% [D:6.9%], F:2.8%, M:11%, n:843
	Helobdella robusta	3060	C:74% [D:3.4%], F:10%, M:14%, n:843
	Schistosoma mansoni	34464	C:56% [D:4.3%], F:8.3%, M:34%, n:843
	Lottia gigantea	1870	C:89% [D:2.3%], F:4.3%, M:5.8%, n:843
	Trichoplax adhaerens	5978	C:81% [D:1.1%], F:7.8%, M:10%, n:843
Metazoa (Amynthas Assemblies)	*SOAPdeNovo Assembly*	61	C:85%**[D:35%]**,F:8.0%,M:6.8%,n:978
	Platanus Assembly	223	C:89%**[D:20%]**,F:4.7%,M:5.9%,n:978
	Post-Collapsing	317	C:87%**[D:19%]**,F:4.8%,M:7.3%,n:978
	Post-Spatial Selection	478	C:78%**[D:11%]**,F:4.9%,M:16%,n:978
Arthropoda (BUSCO examples)	Anopheles gambiae	49364	C:93% [D:4.7%], F:4.1%, M:2.5%, n:2675
	Linepithema humile	1402	C:92% [D:3.3%], F:7.0%, M:0.6%, n:2675
	Drosophila melanogaster	23011	C:98% [D:6.4%], F:0.6%, M:0.3%, n:2675
	Bombyx mori	4008	C:73% [D:2.2%], F:17%, M:8.3%, n:2675
	Tetranychus urticae	2993	C:61% [D:4.5%], F:12%, M:25%, n:2675
Arthropoda (Amynthas Assemblies)	*SOAPdeNovo Assembly*	61	C:42%**[D:15%]**,F:24%,M:32%,n:2675
	Platanus Assembly	223	C:45%**[D:12%]**,F:27%,M:27%,n:2675
	Post-Collapsing	317	C:45%**[D:11%]**,F:26%,M:27%,n:2675
	Post-Spatial Selection	478	C:42%**[D:9.4%]**,F:25%,M:32%,n:2675

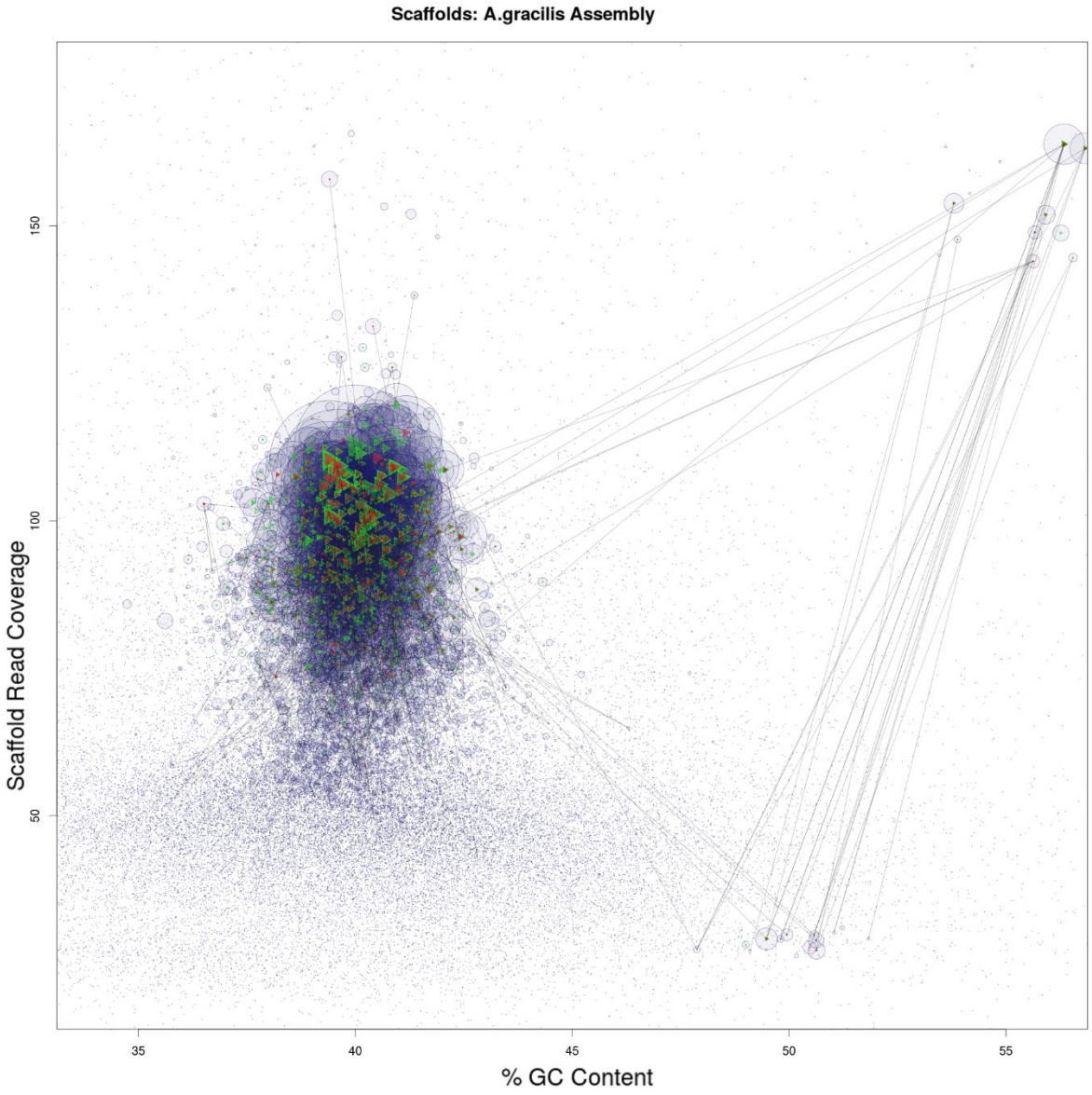


Figure 30. Amynthas gracilis assembly visualisation. Circle sizes indicate scaffold size, green triangles indicate single copy BUSCO genes, red triangles indicate duplicated versions of these genes, connections drawn between single copy gene duplications. Middle-left shows the main set of genomic contigs in the assembly. Upper-right shows the occurrence of the earthworm's vertical symbiont Verminephrobacter. To the lower right is another unidentified, likely intracellular parasite.

Figures 74 and 75 serve as an adjunct to Figure 73, visualising the changes in the genome assembly as the custom pipeline methods were applied. In the final images the non-collapsed allelic fragments have been removed, the remaining genome is no longer subject to contamination by non-host DNA, and all fragments are of a consistent read-coverage in relation to the original libraries.

In the final assembly MAKER (Holt & Yandell 2011) identified 26,951 Gene models. MAKER is a software package which merges down gene model predictions from multiple sources. To gain an optimally informed set of gene models, various other gene prediction programs were run, such that

their output could be fed into MAKER. These programs were Augustus (Keller et al. 2011), which predicts genes based on protein family knowledge, GeneMark ES (Lomsadze et al. 2005), which uses statistical modelling to predict genes *ab initio* in anonymous genome sequence, and SNAP (Korf 2004) another *ab initio* self-training algorithm.

Two other annelid proteomes (and one arthropod: A*pis mellifera)* were aligned to the genome with 'blastp' (Camacho et al. 2009) to further support exon identification. The annelid proteomes were extracted from two genomes produced by the same study (Simakov et al. 2012), whilst the *mellifera* proteome was presented by The Honeybee Genome Sequencing Consortium (Consortium 2006).

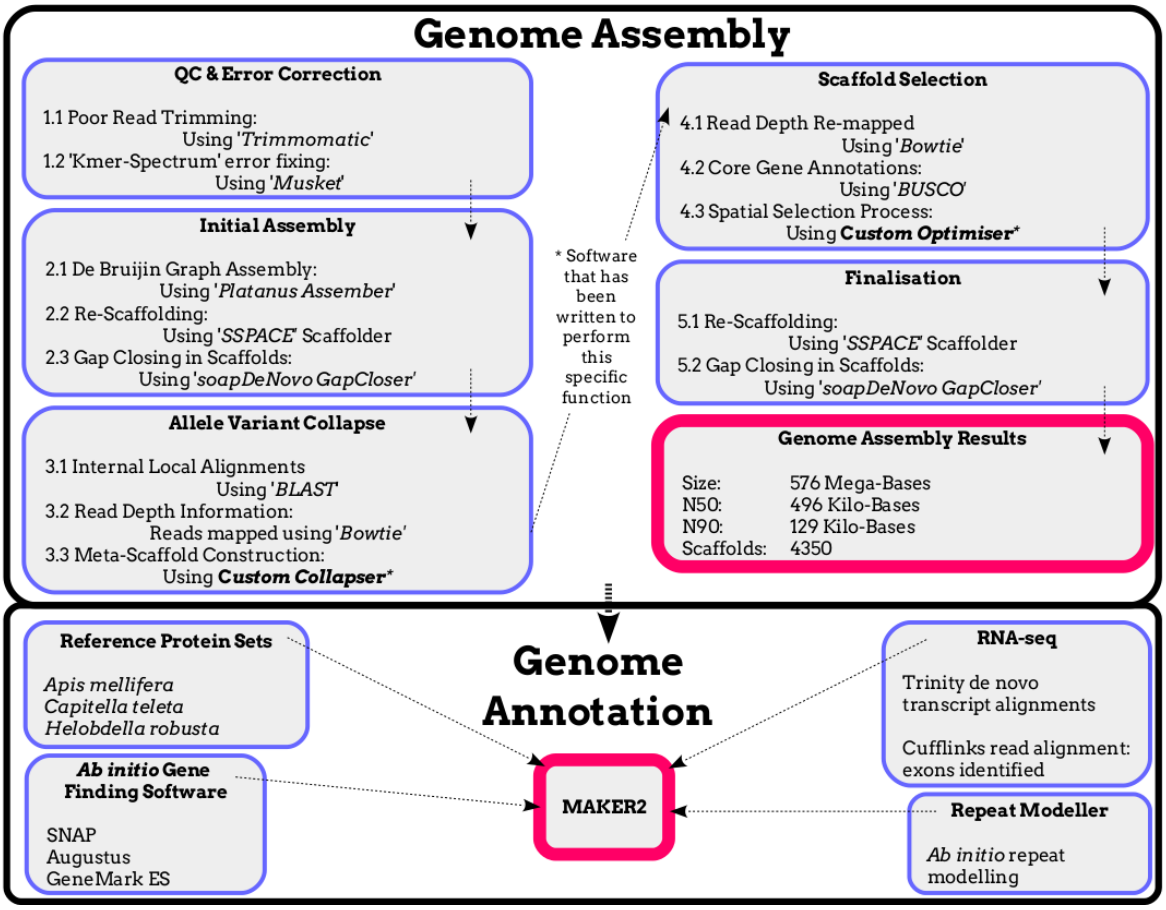


Figure 31. Amynthas gracilis, Genome Assembly Pipeline

The last input to MAKER was generated by TopHat and Cufflinks pipeline (Trapnell et al. 2012) in the form of cDNA-based genome predictions. This involved re-mapping RNA-Seq reads from the twelve samples to the genome and using paired-read information to establish co-transcription.

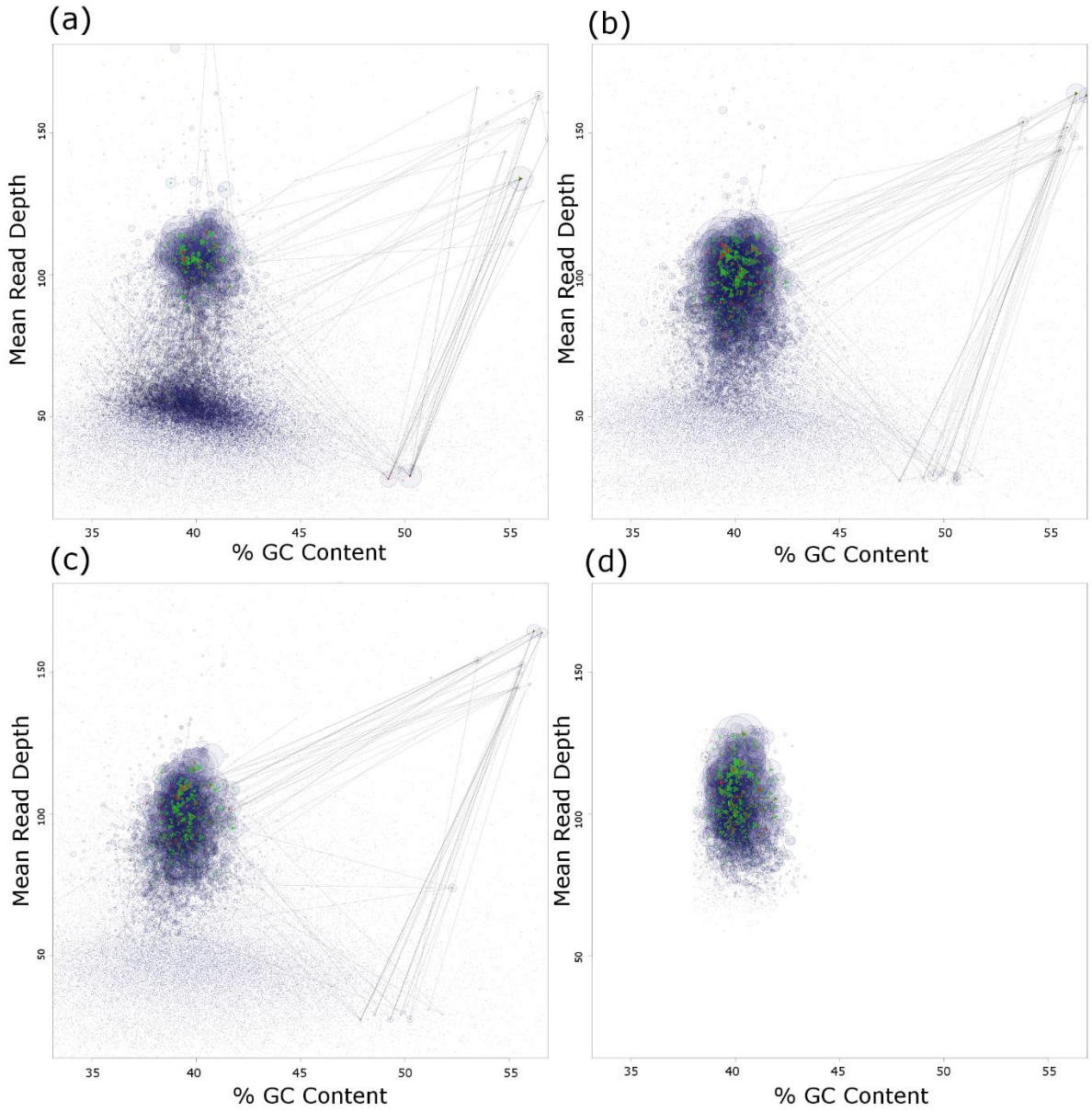


Figure 32. Visualised genome assembly progress, full assembly data with core gene duplication network. Shows BUSCO gene complete single copies (green triangle), BUSCO gene duplications (red triangle), connects scaffolds with the duplications of the same single-copy gene (black lines), shows scaffolds by size (blue circles). (a) SOAPdenovo assembly, (b) Initial Platanus assembly, (c) Post 'custom collapser' application, (d) post spatial selection application.

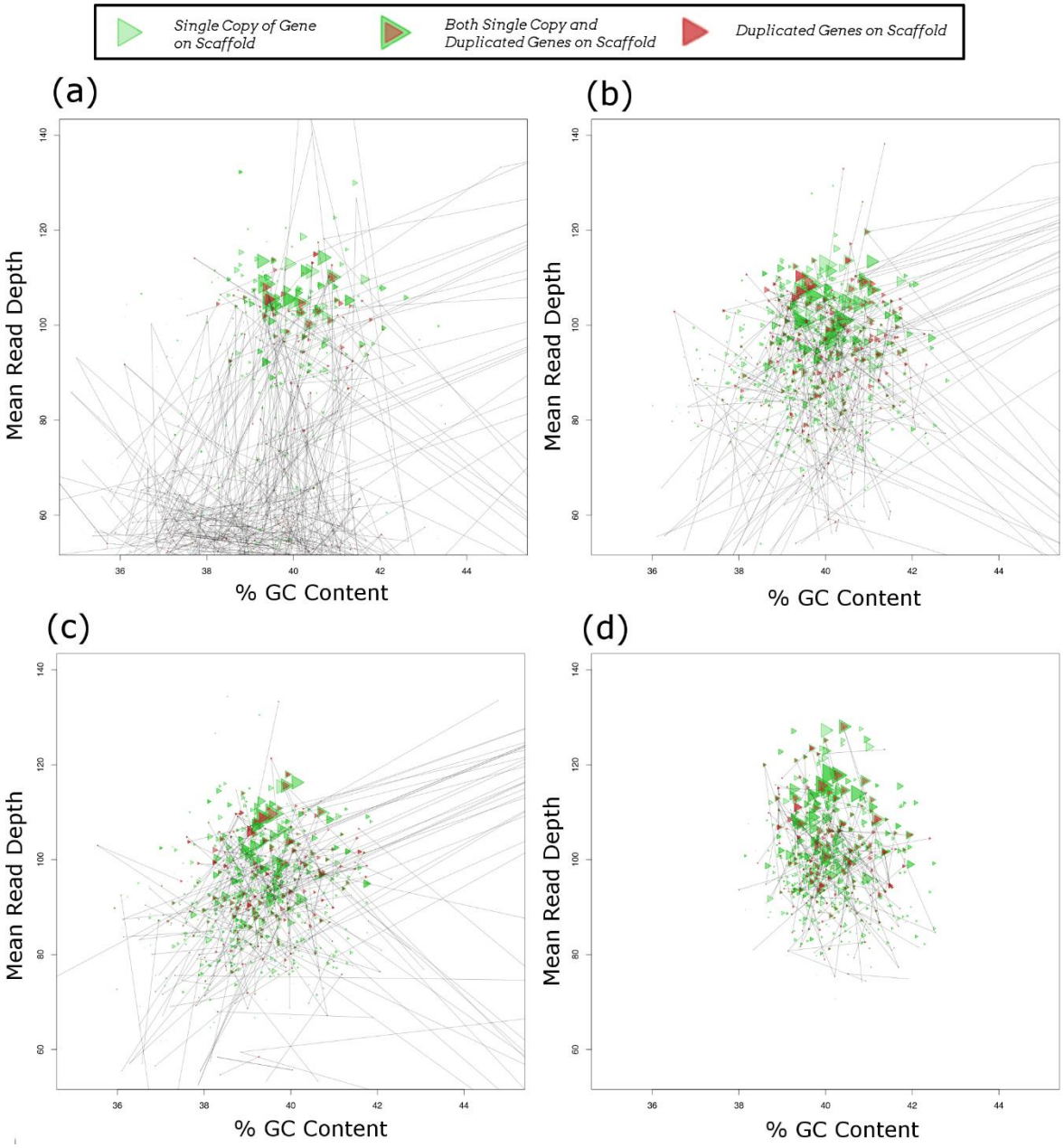


Figure 33. Visualised metazoan core-gene duplication networks with genome improvement, (a) SOAPdenovo assembly, (b) Initial Platanus assembly, (c) Post 'custom collapser' application, (d) post spatial selection application.

3.1.3.4. Custom Allele Collapser

The custom algorithm described in Figure 76 functions to piece together scaffolds which the assemblers have failed to connect due to spikes in allelic divergence. To this end it is a restricted version of an overlap-layout-consensus method (OLC) (Li et al. 2012). OLC assemblers and de Buijn graph-based assemblers constitute the two major algorithm categories in currently available software packages. As this genome was assembled using a *de Bruijn* graph algorithm, it may follow that the contiguity trade-offs extant in these assemblers might be partially annealed via an alternative method.

The custom collapser takes as input a read pile-up of the genome sequencing libraries re-mapped to the genome, and 'self vs self' BLAST-search within the genome, with all first hits removed. The application of non-redundant BLAST hits as 'anchor points' for consensus layout is a development of the methods created for the assembly of the *Cionia savignyi* genome (Small et al. 2007a). Additional features include the restriction of anchor-points to within the terminal 90/10% sequence regions of scaffolds, and the read-depth filtering of those regions, such that only uncollapsed tails may be joined together.

3.1.3.5. Custom Spatial Selection Optimiser

A naïve optimisation algorithm was used to determine an ovoid region in GC Ratio vs Read coverage space to select contigs as members of the final genome assembly. This involved stepwise alterations to the location of the ovoid's centre, and x/y dimensional scaling, and measuring incrementally the trade-offs achieved between reduction in genome size and retention of core genes (as defined by BUSCO's metazoan gene set) for a given set of parameters. Figure 77 describes the parameters of the ovoid and the ranges across which they were modified. Figure 78 shows the best possible trade-offs per duplication achieved by the execution of the ovoid alteration procedure described in Figure 7. The 'duplicate gene score scaling factor' is described as **OptA/OptB** (unique vs total duplicate minimisation) in Figure 77, or otherwise, the optimisation terms. By altering the threshold of the term, it was possible to explore the changes in best trade-off achievable. Given the sharp falloff in complete core-genes present after 2.0 in Figure 78, it was determined that this threshold level was the most useful for selecting ovoid permutations

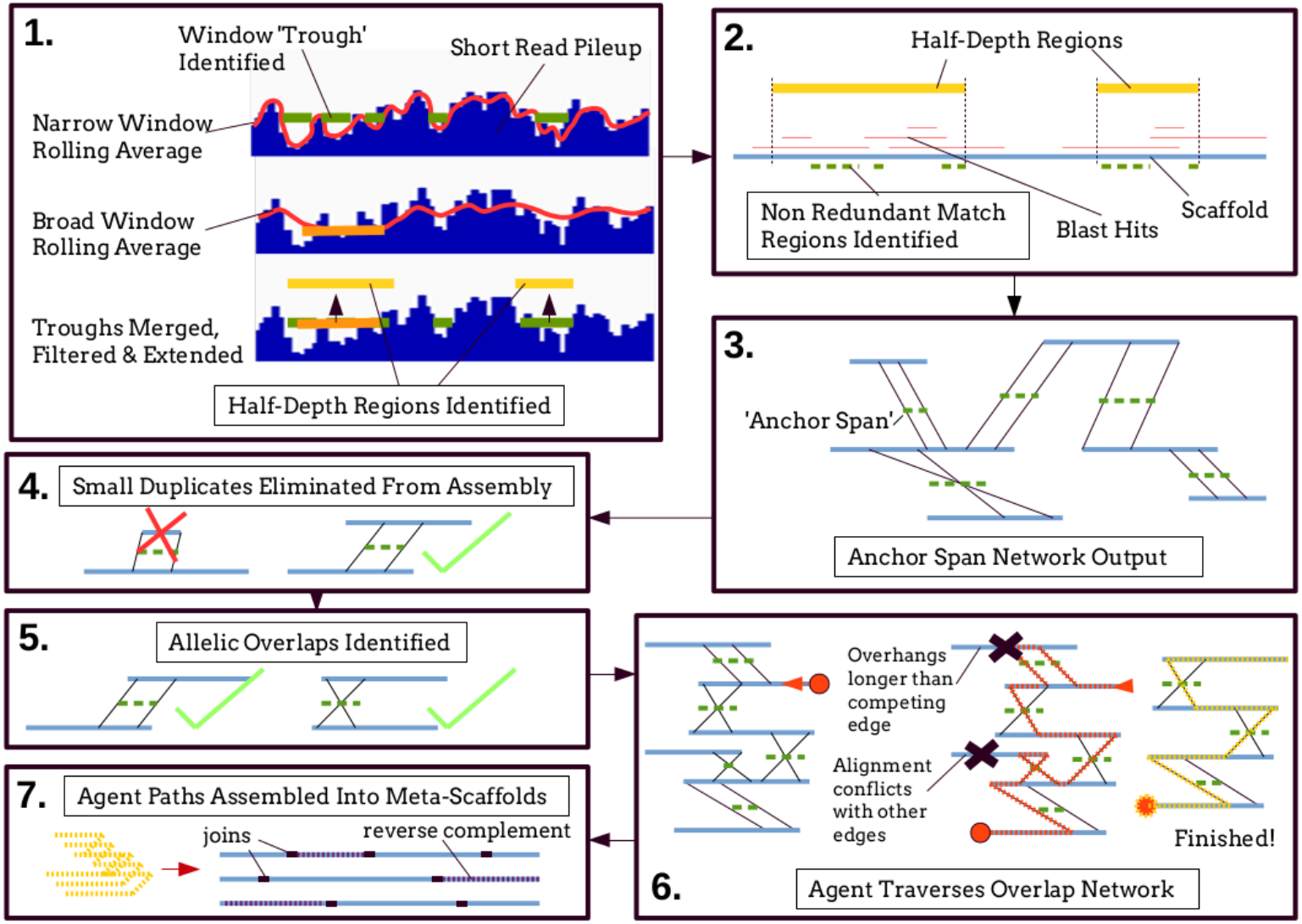


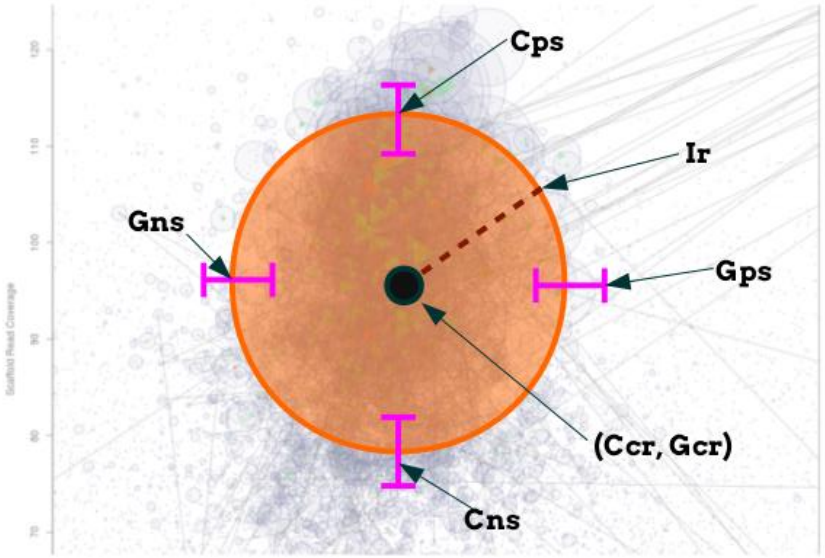
Figure 34. Custom allele collapser algorithm. (1) Double rolling average window method used to identify consistent read depth regions, (2) Identification of non-redundant BLAST hits which exist only within half-depth regions, (3) Anchor-spans created as connections between co-linear series of non-redundant BLAST hits in half depth regions, (4) Removing small allelic duplicate fragments, (5) Anchor spans connecting scaffolds via terminal 20% regions identified for potention collapsing, (6) Algorithm navigates over-lap network, (7) Agent paths are converted into meta-scaffolds with regions reverse-complemented where required.

Single Input Parameters

Core Genome Minimum:	**Gm**	5.0e8
Excess Bases Weighting:	**Bw**	3.3e-7
Expansion Resolution:	**Er**	0.05
GC Resolution:	**Gr**	0.2
Coverage Resolution:	**Cr**	1
Duplicate Weight Res.	**Wr**	0.2
Initial Radius Res.	**Rr**	1
GC Distance Weight:	**Gcw**	10

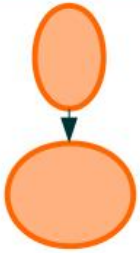
Range Input Parameters

Coverage Positive Scale:	**Cps**	0.75-1.25
Coverage Negative Scale:	**Cns**	0.75-1.25
GC Positive Scale:	**Gps**	0.75-1.25
GC Negative Scale:	**Gns**	0.75-1.25
Initial Radius:	**Ir**	16-24
Duplication Weight:	**Dw**	0.5-3
GC Centre Range:	**Gcr**	39-41
Coverage Centre Range:	**Ccr**	90-105



Optimisation Loop

1. Oval is Resized



Oval Resizing
Nested loops increment through:

Cps by **Er** **Ir** by **Rr**
Cns by **Er** **Gcr** by **Gr**
Gps by **Er** **Ccr** by **Cr**
Gns by **Er**

2. Contents Quantified

Scaffolds Assessed For:

A. Complete Core Genes
B1. Genes With Duplication
B2. Total Duplicates
C. Genome Size

3. Optimisation Terms Tested

Over range of **Dw** by **Wr:**

OptA = **A** – (**B1** * **Dw**) – (**C** – **Gm**) * **Bw**
OptB = **A** – (**B2** * **Dw**) – (**C** – **Gm**) * **Bw**

Whereby the terms maximising **OptA** and **OptB** are saved for all values of **Dw**.

Figure 35. Custom optimiser parameters, ranges and terms.

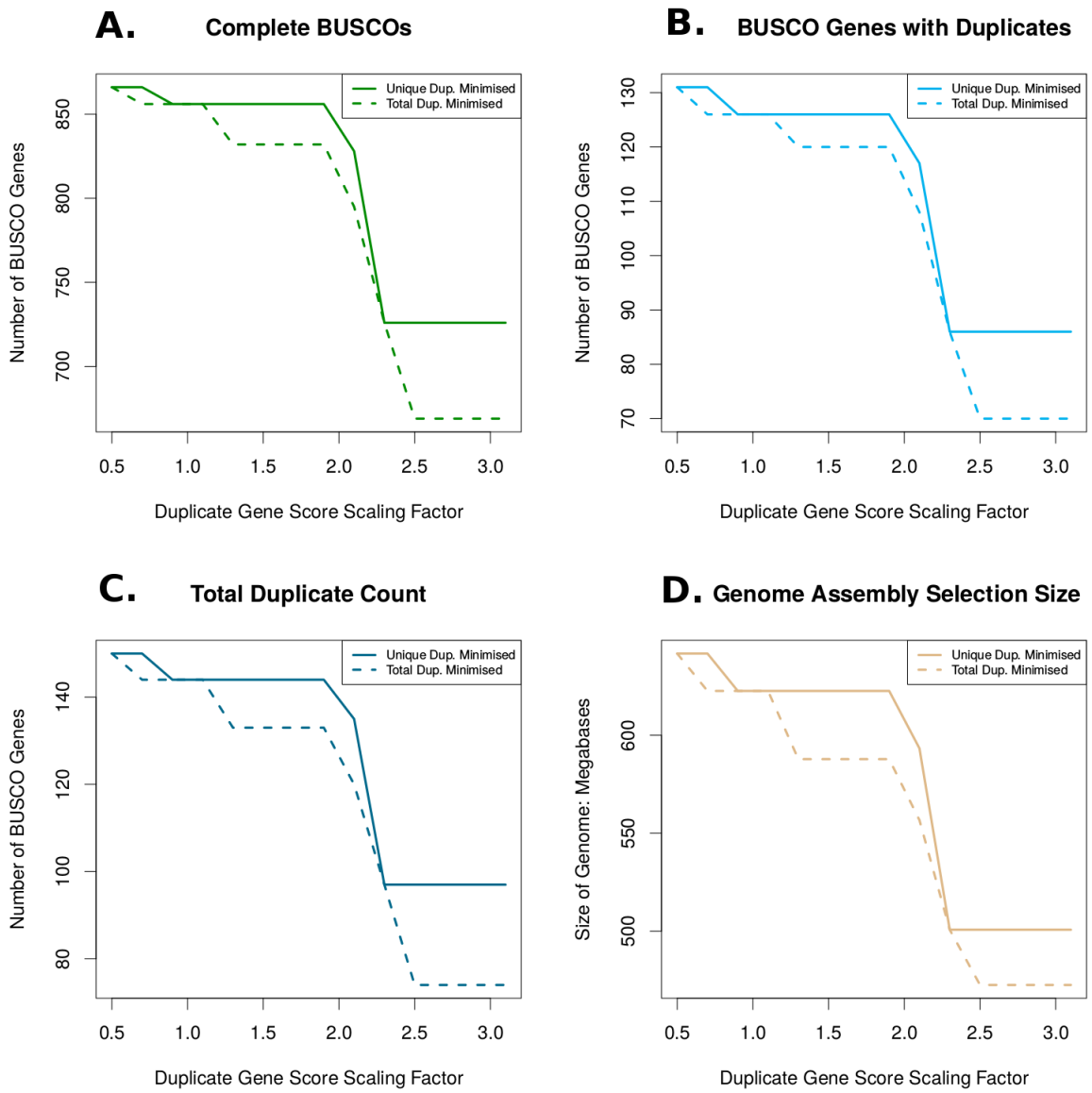


Figure 36. Optimisation of Genome size against BUSCO completeness results.

3.1.4. Mapping and Quantification

Short read libraries for all three sequencing experiments were end-clipped and filtered with Trimmomatic (Bolger et al. 2014). RNASeq and Me-DIP libraries were aligned to the genome using BBmap (Bushnell & Brian 2014) with the pre-set 'slow'. RNASeq libraries aligned consistently at a mean rate of 86.05%. MeDIP libraries aligned at 90.5%. See Table 11 for library sizes and alignment rates. Samtools (Li et al. 2009) was used to convert output SAM files to sorted and indexed BAM files 'htseq-count' function of the HTSeq python package (Anders et al. 2015) was used to generate read counts for RNA-Seq data. Samtools (Li et al. 2009) 'bedcov' was used to quantify MeDIP read-count levels within genes, and within promoter regions. Promoter/TS-factor motif binding levels were

divided into two regions: 100bp 5’ of TSS, and 1Kb 5’ of TSS. These read counts were not used for absolute methylation analysis and did not require normalisation.

Table 7 Sample sequencing libraries with trimming and alignment rates

RNA-Seq Libraries

Sample	Raw (Read Count)	Trimmed	%	Aligned	%	Gb Raw	Gb Trimmed
V→V.1	22,305,582	20,666,302	92.7	17,913,624	86.7	3.35	3.10
V→V.2	31,911,228	30,025,366	94.1	25,528,676	85.0	4.79	4.50
V→V.3	32,531,640	30,731,750	94.5	26,429,150	86.0	4.88	4.61
V→M.1	29,123,856	27,394,556	94.1	23,594,138	86.1	4.37	4.11
V→M.2	34,197,226	32,364,024	94.6	27,869,880	86.1	5.13	4.85
V→M.3	30,819,382	29,063,042	94.3	24,813,578	85.4	4.62	4.36
M→V.1	29,724,434	27,861,854	93.7	24,121,614	86.6	4.46	4.18
M→V.2	33,385,916	31,472,080	94.3	27,183,240	86.4	5.01	4.72
M→V.3	42,270,714	40,075,616	94.8	34,414,400	85.9	6.34	6.01
M→M.1	38,009,410	35,987,182	94.7	31,246,464	86.8	5.70	5.40
M→M.2	36,211,208	34,253,562	94.6	29,578,904	86.4	5.43	5.14
M→M.3	40,477,494	38,548,640	95.2	32,915,512	85.4	6.07	5.78

miRNA-Seq Libraries

Sample	Raw (Read Count)	Trimmed	%	Aligned	%	Gb Raw	Gb Trimmed
V→V.1	25,747,134	24,292,420	94.3	7,652,639	31.5	3.86	3.64
V→V.2	16,981,154	16,412,964	96.7	7,157,806	43.61	2.55	2.46
V→V.3	26,188,542	24,343,821	93.0	8,604,909	35.35	3.93	3.65
V→M.1	15,078,814	14,407,309	95.5	6,559,646	45.53	2.26	2.16
V→M.2	21,472,620	20,177,299	94.0	9,703,776	48.09	3.22	3.03
V→M.3	20,634,684	19,536,629	94.7	5,664,319	29	3.10	2.93
M→V.1	20,646,086	19,669,009	95.3	9,288,720	47.23	3.10	2.95
M→V.2	22,016,268	20,886,785	94.9	10,574,566	50.63	3.30	3.13
M→V.3	22,384,786	21,148,899	94.5	7,729,828	36.55	3.36	3.17
M→M.1	20,872,376	19,547,815	93.7	8,186,318	41.88	3.13	2.93
M→M.2	24,970,848	23,189,178	92.9	8,180,014	35.28	3.75	3.48
M→M.3	28,557,792	27,316,670	95.7	7,995,913	29.27	4.28	4.10

MeDIP-Seq Libraries

Sample	Raw (Read Count)	Trimmed	%	Aligned	%	Gb Raw	Gb Trimmed
V → V	9,150,071	8,721,337	95.3	7,888,845	90.5	1.37	1.31
V → M	12,703,975	11,969,754	94.2	10,855,396	90.7	1.91	1.80
M → V	14,175,950	13,250,605	93.5	12,225,125	92.3	2.13	1.99
M → M	15,477,553	14,754,260	95.3	13,469,337	91.3	2.32	2.21

3.1.5. Methylation Model Building

3.1.5.1. Static Modelling

Custom software was developed to facilitate the building of methylation gene-models. This software reads SAM formatted alignment files and finds coverages of genomic annotation elements across normalised intervals. This involves finding the coverage rate of relative intervals along a set of elements. For example, the 0-10%, 20-20% … 90-100% intervals along the 5’->3’ length of an exon.

The read pileup means of the set of each interval across the set of all elements of the same type allows for the profiling of 'typical' patterns of a given read-alignment map. Coupling-vector

The software also supports rank-grouping these intervals by an additional variable, in this case gene expression levels were used. This process involves ranking the set of elements by their parent gene's all-sample geometric mean expression level and calculating the Me-DIP read coverage intra-group arithmetic means per interval.

Methylation is often described in organisms as being tissue specific in humans (Lokk et al. 2014), and mice (Maegawa et al. 2010). As these samples were mixed-tissue, the results the yield are only summaries of the regions in a gene which might be methylated at *some* point, and the abundances are subject to the bias of tissue content in the sample. To facilitate a broader perspective on the methylation probabilities of gene models, a second execution of the interval-based measurement was performed, with all read coverage rates flattened into binary conditions. This was intended to demonstrate the distributional differences between the incidence of methylation across all cells, and the potential for even rare methylation to occur under the right conditions.

The final output metric types for the two test types differed, as they were measuring different things. For the model which included depth information, the 'relative read-depth' was used. This was calculated as the ratio between any given coverage interval, and the mean of all-element all-interval means. This scaling produces a neutral rate of 1. For the binary coverage model, simple frequency rates were used as the final metric, with a value of 0.2, for instance, indicating that 20% of that entire set has a non-zero Me-DIP read-depth.

Coverage intervals over a set of intra-genic annotations were then presented as miniature 'gene-models of methylation', comprising a primary promoter 34bp in length, a 5' UTR, three exonic coding sequences with two interspersed introns, a 3' UTR and a 3' flanking region of 300bp. Although not all gene models contained all these elements, a coercion process was used to allocate gene components to the model.

3.1.5.2. Differential Modelling

Once models of intra-genic methylation were developed, they were then used as a normalisation function for a secondary differential expression model. The differential model was developed in a similar manner to the static model with the source of the input values the primary exception. The inputs were absolute RPM differentials, and the same dataset again flattened into binary zero/non-zero coverages.

The metric used for the uncorrected depth-based differential output was mean RPM change per interval, scaled by the mean-of-means as above. This was then divided by the static model output to produce a set of ratios of the rate of change against the neutral expectation of coverage.

For the binary coverage output each interval was calculated as an odds ratio between the probability of coverage change and the probably of coverage.

3.1.5.3. Sequence Structural Signatures

To investigate if the regulatory sequence structures associated with methylation, the High Dimensional Signature software developed in Chapter 4, was applied to various sequence subsets. Four categories of sequence subset were created:

- Intronic splice junctions (100bp limit)
- 3'UTRs (full length)
- 5'UTRs (full length)
- 34 bp promoter regions.

Each of these subsets consisted of the full set of those elements available in the genome. Each of the subsets was then split evenly into 3 rank-groups based on their Me-DIP read depths. These were labelled: No Methylation, Low Methylation, and High Methylation. The DNA processing version of the signature software was run on all 12 groups with the parameter N=2.

Controlling the initial frequencies of the root nodes in the signature developed allowed for direct comparative measures to be made in the 2D signature outputs. This involves calculating the total structure score fold-change between the three groups, per-K, per-N. For each set of three groups, three 2D signature structure fold changes were calculated:

- High/Low
- High/None
- Low/None

The results of these differentials are intended to demonstrate how the total sequence subsets become more, or less structured as their methylation rate is considered as a condition.

3.1.6. Prediction and mapping of miRNA

Novel miRNA prediction was performed using the MiRDeep2 (Friedländer et al. 2012) package. Processed reads were aligned to the genome and collapsed into a non-redundant set with the mapper.pl script. The alignments were converted to novel mature miRNA predictions with mideep2.pl script. In total 168 novel mature miRNAs were predicted using this method. A bowtie

(Langmead et al. 2009) database was built using a combination of the predicted miRNAs and the latest version of the MiRbase database (Kozomara & Griffiths-Jones 2014). Bowtie was then used to map the collapsed reads onto the database, and a custom script was used to extract the raw read counts per miRNA from the output (which involved parsing read name suffixes). This method differs from the series of MirDeep2 processing steps and was performed to access non-normalised counts per miRNA, for the sake of keeping consistent input forms into deseq2 (maintaining the homogeneity of the statistical pipeline), for downstream comparability.

Studies in human and *drosophila* molecular biology have discovered canonical cohorts of 1917 and 2058 of active miRNAs respectively according to miRbase (Kozomara & Griffiths-Jones 2014). The outputs from this processing pipeline included upwards of 20,000 hits per sample. However, the read-counts per hit had a pareto-like distribution, with 10,313 miRNA targets achieving an average read count of less than 20. To select the hits most likely to represent active miRNAs, and with reference to typical miRNA cohort sizes, the output set was limited to the top 2000 hits by mean expression.

3.1.7. Differential Expression and Methylation

3.1.7.1. Exploratory analysis and Metric Choice

The analytic pipeline downstream of the differential analysis component of this study required *p*-values to be comparable. For this reason, all three differential analyses were performed using the same pipeline in deseq2 (Love et al. 2014), this included using the 'normal' fold change effect size shrinkage estimator in all cases. The 'normal' mode was chosen for its greater stringency in mitigating the low expression level effects in noisy data, which was particularly useful for the MeDIP-Seq analysis in which each sample was highly heterogenous (see Figure 111).

The experimental design of the differential tests performed is described in Figure 102 (bottom left). As described in the introduction, three modalities of conceptual adaptivity were mapped onto three test types in the design. These were

1. Non-specific environmental change detection/response
2. Persistent specific adaptive changes
3. Acute specific acclimative changes

These map onto the three tests in the design, via the differentials found between:

1. Transplanted worms vs Stationary worms (Static vs Change)
2. Worms of common origin despite transplant (Origin vs Origin)
3. Worms of common transplant destination (Destination vs Destination)

In each of these three tests, RNA-Seq and miRNA-Seq counts were compared in 6 vs 6 groupings, whilst MeDIP-Seq counts were compared in 2 vs 2 groupings.

The outputs produced by deseq2, along with reference heatmaps of the normalised expression matrices, are displayed for mRNA, miRNA, and methylation in Figures 103, 104 and 105 respectively.

3.1.7.2. Methylation Change

The intersection between changes in methylation and transplant gene expression for the genomic gene models were explored via the exploration of Me-DIP log FC vs RNA-Seq log FC distributions for the four tests. Early exploratory analysis revealed that no simple linear relationship existed between the two distributions. However, the limits of each distribution did appear to disproportionately dense. To investigate the underlying relationship between gene expression and methylation change, a quantile-grid bucketing method was applied.

This method entailed creating a 10x10 grid of bins for data points FC vs FC space. The x and y values of the grid lines were derived via the probabilistic quantile intervals of each distribution, with increments of 0.1. This method ensures approximately equal membership of each bin under null conditions. To investigate the extent to which the test set's bin-density fluctuations were significant, a bootstrap model was applied. Bootstrap bin densities distributions were created by independent random re-sampling of both FC columns. The distribution means, and upper and lower 5% confidence intervals of the resultant distribution, were used to measure test set fluctuation significance.

Once this procedure had been developed, it was also applied to the two promoter-fold-change category differentials against gene expression. A final procedurally identical deployment of this model was conducted with all-sample means instead of differentials.

3.1.8. Functional Annotation and Enrichment

Analysis of the functional content of differentially expressed genes was performed using the DAVID webserver (Dennis et al. 2003). The genes described in *Amynthas gracilis* are not part of a standardised gene naming ontology, as this earthworm is novel with respect to '*omic* level analysis systems. A proteome was translated from a genome-derived transcriptome with Transdecoder (Brian J. Haas et al. 2013). This proteome was used to search the Uniprot/Swissprot (EMBL et al. 2013) database (accessed 09/18) with 'blastp' (Camacho et al. 2009). All hits achieving an *e*-value below $5e^{-05}$ were retained and used as a symbol translation table for DAVID. The full list of gene annotations was also provided to the webserver as a 'background' against which to measure enrichment.

A 3x3 matrix of functional clustering analyses was performed on nine gene lists. Gene lists are defined as lists of symbol-translated genes achieving shrinkage-corrected Wald test *p*-values of < 0.05 in differential tests. In the case of miRNAs, the gene list was derived from the miRNA binding network as immediate neighbours of all miRNAs differentially expressed at a significant *p*-values. The 3x3 matrix of analyses consisted of three test types, each ran on three data types.

Test Types: Static vs Change, Origin vs Origin, Destination vs Destination

Data Types: RNA-Seq, miRNA-Seq, MeDIP-Seq

Including multiple annotation terms in clustering often resulted with inconsistent term source ratios in the output clusters. To standardise the tests and allow them to be inter-comparable, the Gene Ontology (Ashburner et al. 2000) 'Biological Process' category at Level 4 was used to generate initial clusters. The gene list memberships of each cluster were then annotated with Gene Ontology 'Cellular compartment' terms using simple enrichment. For the visualisation outputs Figures 106, 107 and 108 which describes each gene belonging to a principal cellular compartment (despite the matching of multiple terms), some simple rules were used to delimit the overlaps. This was to choose the more detailed term unless doing so caused further overlap with similarly detailed terms, in which case collapse to the next lower level term with the best *p-value.* This process was intended to retain the maximum descriptive power of the cellular compartment terms, without displaying the excessive redundancy in the ontology (for example, it is a given than an ion channel is a trans-membrane protein and does not need to be labelled twice).

The trait matrix shown in Figure 109 is a collapsed set of the most enriched clusters in each of the three tests combined. The gene lists for each cluster were gathered via an inclusive union of all gene-lists in all clusters containing the 'merged-by' term. The merged-by term in each case was the most significant p-value term in the set of similar clusters. For example, the Epithelial Development merged trait is a union of all gene lists found within clusters containing the GO Term GO:0060429, across the three test types, and across the three data sources.

3.2. Results

3.2.1. Soil Content Differences

Table 8. Gas and temperature content differences between soils

Variables in soil	Macela Transplant Site			Macela Sampling Site			Furnas Transplant Site			Furnas Sampling Site		
	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min
CO_2 (vol.%) 25 cm	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	21.5	71.7	4.3	6.7	9.6	1.8
CO_2 (vol.%) 50 cm	0.8	2.8	0.1	0.7	2.3	0.2	48.6	96.5	14.0	6.9	8.8	2.4
CO_2 flow (g m^{-2} d^{-1})	19.5	27.5	12.3	15.4	32.4	6.2	181.0	533.4	58.3	19.8	25.0	12.5
O_2 (vol.%) 25 cm	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	15.7	19.1	6.3	10.7	18.8	3.4
O_2 (vol.%) 50 cm	18.5	19.9	13.1	18.8	20.0	15.6	10.4	17.1	1.4	9.9	18.3	1.2
Temperature (ºC) 25 cm	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	35.0	45.7	28.0	18.4	19.6	17.9
Temperature (ºC) 30 cm	17.0	17.6	16.2	17.0	18.9	15.9	48.6	96.5	14.0	18.3	19.3	17.8

Table 9. Geochemistry differences between soil samples

	TEXTURE				PROPERTIES		
Sample	CLAY (< 2 micron)	SILT (2 - 63 micron)	SAND (63 micron - 2 mm)	pH	Soil moisture content (H2Og per 100g soil)	Loss on ignition (g per 100g oven dried soil)	Water Holding Capacity (%)
V->V.1	3	65	35	5.46	4.13	10.74	74.19
V->V.2	2	66	35	5.69	7.47	10.71	78.93
V->V.3	3	71	30	5.24	9.08	12.58	82.93
V->V.4	5	70	29	5.40	5.20	12.64	93.50
V->V.5	2	38	62	5.49	5.04	11.45	79.01
V->V.6	3	66	35	5.83	4.83	10.25	75.43
V->V Avg.	**3**	**62.67**	**37.67**	**5.52**	**5.96**	**11.40**	**80.67**
M->V.1	3	74	26	5.66	6.10	11.42	74.61
M->V.2	4	76	25	5.43	5.51	11.11	86.15
M->V.3	2	36	64	5.23	10.29	12.12	83.38
M->V.4	3	68	33	5.84	8.08	10.87	85.42
M->V.5	2	55	47	5.46	4.80	10.22	75.81
M->V.6	2	62	40	5.53	5.32	9.89	76.81
M->V Avg.	**2.67**	**61.83**	**39.17**	**5.53**	**6.68**	**10.94**	**80.36**
V->M.1	1	53	49	5.72	17.44	32.72	140.79
V->M.2	1	54	47	5.60	8.17	24.52	106.36
V->M.3	1	39	62	5.89	6.00	25.00	112.20
V->M.4	2	50	51	5.45	13.39	29.37	120.16
V->M.5	2	57	44	5.05	31.68	34.91	119.90
V->M.6	1	62	40	5.65	16.48	23.06	106.40
V->M Avg.	**1.33**	**52.50**	**48.83**	**5.56**	**15.53**	**28.26**	**117.63**
M->M.1	1	55	47	4.99	22.04	30.40	157.30
M->M.2	1	50	51	5.23	12.92	25.61	103.99
M->M.3	2	54	47	5.81	12.34	25.34	108.61
M->M.4	1	53	49	5.33	22.56	36.70	155.31
M->M.5	1	59	42	6.00	18.38	21.71	110.57
M->M.6	*	*	*	6.37	5.80	14.42	100.16
M->M Avg.	**1.20**	**54.20**	**47.20**	**5.62**	**15.67**	**25.70**	**122.66**

The geochemistry of the source and transplant sites were analysed separately, the following raw data was been provided by Luis Cunha. Table 12 shows a summary of the gas and temperature differences between soils. The mean temperature at a depth of 30cm at the Furnas transplant site was 48ºC, compared to a relatively low 17ºC at Macela, and 18ºC at the Furnas sampling site. The transplant CO_2 mean volume was 48.6%, compared to 6.9% at the Furnas sampling site, and 0.7-0.8

at the Macela sites. The O_2 volume % mean was consistent between sampling and transplant sites but varied between Macela and Furnas by ~10% to ~18% respectively.

Soil physical properties demonstrated in Table 13 show a consistent set of compositional differences between the two sites, but not between sampling and transplant. Specifically, Furnas soils had reduced water holding capacity, less than half the clay content, and were sandier. The organic content measured by loss on ignition also varied substantially, with Furnas soils possessing less than half the organic matter by weight when compared to Macela soils (~11g vs ~25-35g per 100g), although the Macela soils were also more variable in content (std. dev.: 0.86g vs 5.9g).

The metal content results in PPM (parts per million) of the microcosm soils is displayed in Table 14. The primary results of this analysis are that substantially different metal content profiles exist for both soils. Furnas was remarkable for elevated lead and copper content (means: 108ppm, 100ppm); at 6.1-fold and 5.02-fold respective increases compared to Macela soils. Macela was most differentiated by its Nickel content (27.4ppm), at a 3.5-fold increase compared to Furnas.

Table 10. Soil Metal Content Differences (parts per million), columns manually sorted by content differences

	Soil Elemental content, by PPM (parts per million)																
Detection limits:	1.63	12.98	1.45	18.84	0.09	0.14	15.56	4.16	0.10	0.63	0.36	109.23	1.11	9.04	2.11	0.36	2.81
Sample	**Al**	**Ca**	**Fe**	**Ga**	**Mg**	**Mn**	**Na**	**Ni**	**Sr**	**Ti**	**Cu**	**K**	**Li**	**Pb**	**Zn**	**Ba**	**Cr**
V->V.1	17302.21	2981.99	18933.25	47.31	1694.03	520.34	958.60	5.92	33.23	1124.81	74.47	2192.42	5.19	88.81	311.63	67.13	12.97
V->V.2	19364.81	3563.31	18817.08	42.09	1682.33	538.27	941.00	7.24	39.31	936.53	144.59	2384.85	5.64	103.55	385.69	83.28	22.30
V->V.3	17617.56	3065.01	16818.62	42.89	1426.00	560.68	819.36	5.41	27.06	1010.86	95.51	1926.65	5.40	96.65	296.39	64.17	12.27
V->V.4	19645.72	4202.94	21657.17	41.66	1557.16	651.18	799.40	8.59	33.86	1017.41	92.53	1919.39	4.86	133.61	398.90	76.26	16.96
V->V.5	19106.49	4372.01	21405.10	47.37	1591.29	669.06	948.93	5.33	37.59	1208.42	102.98	2168.60	5.65	117.70	351.96	112.65	20.49
V->V.6	18562.16	4372.56	19452.33	53.96	1895.53	584.46	896.54	4.88	37.04	1149.08	96.38	2053.43	5.64	110.30	365.91	77.03	17.06
M->V.1	18380.25	3374.64	18125.16	41.54	1392.23	550.11	699.65	4.43	35.72	945.24	85.62	1957.16	5.54	94.46	335.42	82.01	14.53
M->V.2	17955.92	3969.22	20101.68	23.91	3498.47	535.83	734.25	20.45	40.75	740.81	95.68	1832.56	5.55	112.64	564.06	86.74	22.91
M->V.3	17930.35	3785.84	17049.70	42.09	1367.73	561.33	873.92	5.10	33.51	1027.88	98.50	1929.52	4.94	105.42	314.50	72.21	14.59
M->V.4	19783.47	4992.15	20271.38	54.89	1552.55	647.75	1008.41	11.67	37.42	1142.20	99.43	2220.72	5.73	129.81	415.20	89.25	20.94
M->V.5	18192.97	3757.32	16920.38	48.77	1769.64	624.81	802.92	6.21	35.37	1033.80	82.41	2034.98	5.49	118.00	292.29	77.43	17.72
M->V.6	19509.13	4549.14	17081.15	46.86	1695.92	577.09	844.08	6.87	40.99	1027.17	143.12	2290.48	5.49	90.89	391.62	80.30	18.40
V->M.1	22827.11	5964.18	19476.54	124.34	3765.89	618.27	1159.25	12.86	48.77	2577.67	17.37	1395.17	4.29	18.57	202.11	61.22	15.06
V->M.2	22784.92	7837.57	24755.22	174.42	7894.90	691.42	1965.41	32.93	71.93	3611.32	17.48	1634.44	4.71	9.82	112.97	74.75	22.94
V->M.3	23380.20	11100.58	28135.04	194.86	10625.11	732.52	2546.28	42.30	106.09	4018.35	25.58	1980.01	4.57	18.68	153.04	85.45	29.64
V->M.4	22990.87	5158.89	20979.89	119.90	3891.44	648.28	1008.72	15.03	41.15	2489.47	14.36	1038.61	4.41	23.44	261.26	62.03	18.00
V->M.5	17817.31	6534.80	19480.91	120.07	5481.24	652.32	1229.59	20.61	53.55	2529.31	16.75	1297.80	3.35	15.28	94.24	57.32	14.96
V->M.6	25185.77	12904.20	29677.42	207.92	11667.23	815.23	2912.52	47.27	122.73	4385.87	28.32	2341.13	4.40	19.30	99.28	92.79	38.14
M->M.1	22314.51	4698.56	19471.42	118.84	3513.76	640.87	970.87	12.23	37.96	2458.19	11.63	1294.64	4.30	18.21	251.52	63.24	19.36
M->M.2	22560.42	8634.13	25640.86	164.64	7466.89	679.13	1946.66	27.59	77.29	3508.75	22.08	1648.38	4.12	16.33	124.58	76.64	22.37
M->M.3	23192.68	10067.99	27579.66	180.42	9365.04	718.53	2280.60	36.49	93.95	3606.90	25.64	1871.50	4.12	19.94	150.49	87.79	24.64
M->M.4	22297.44	6255.13	20015.13	116.66	4040.81	808.09	1080.55	13.34	49.72	2472.11	15.47	1277.69	4.67	15.11	308.63	65.98	13.34
M->M.5	23301.41	7020.44	23337.11	146.23	5533.41	693.67	1522.78	20.66	57.70	3144.58	16.89	1373.33	4.96	20.20	108.74	70.04	18.70
M->M.6	24939.47	11807.40	30214.02	206.73	11662.54	731.68	2762.52	47.70	109.87	4411.62	29.57	2176.25	4.35	15.86	92.26	88.31	30.63
Molecular Weight:	27	44	56	71	24	55	23	60	88	47	65	39	7	208	66	137	52

3.2.2. Morphometric Changes

Representative histological and histochemical analysis, performed by and made available by Luis Cunha, of the epidermis of earthworms from the two source populations after 31 days exposure to active (Furnas) and inactive (Macela) volcanic soils are given in Figure 79 (c). Epidermal thickness was directly representative of the destination soil after 31 days, in almost all cases regardless of origin. Mean thickness in Furnas soils was 24.02 mm (std. dev. 3.86 mm). The mean thickness in Macela soils was 42.8mm (std. dev. 8.01 mm). Furnas-origin worms transplanted to the Macela microcosm demonstrated a mean *78.3%* increase in their epidermal thickness. The reciprocally transplanted worms (Macela->Furnas) demonstrated a *44.1%* reduction in thickness.

Increases in epidermal thickness were reflected by a similar individual weight-change per bag. Individuals remaining in their destination soils also showed a decrease in weight, likely due to the microcosm condition differentials between sampling and transplant sites (Tables 12 and 13) (F: -*10.7%,* M: -*3.3%*). Worms transplanted from Furnas to Macela still showed a mean *5.4%* increase in body weight, whilst the reciprocal transplant yielded a *13.6%* mean body weight loss. Figure 79 (a) shows before and after weight distributions.

Table 11. Earthworm body matter metal content fractionated between metal rich granules, soluble material, and soft matter (issues fragments, membranes etc).

Earthworm Body Elemental content, by PPM (parts per million)

Exposure	Source	Na	Mg	Al	K	Ca	Ti	Cr	Mn	Fe	Ni	Cu	Zn	Ga	As	Sr	Cd	In	Pb	Bi	Fraction
Furnas	Furnas	19,564.48	2,571.44	2,958.38	5,987.49	12,280.83	216.91	11.44	364.75	3,219.78	10.21	93.91	376.32	2.02	7.01	200.98	9.23	0.01	35.53	0.15	
Furnas	Macela	17,702.80	3,243.86	2,383.58	5,324.93	14,235.48	141.92	9.95	430.34	3,628.37	8.42	105.75	451.42	1.72	7.69	1,167.10	11.77	0.01	37.79	0.10	TOTAL
Macela	Furnas	22,598.08	4,699.46	2,156.99	5,297.71	14,387.06	232.25	15.10	462.16	3,010.24	49.44	37.67	398.63	1.36	6.24	344.87	3.54	0.01	3.23	0.02	
Macela	Macela	22,633.12	4,461.20	1,445.61	4,454.68	14,788.90	158.33	9.51	566.40	2,575.09	38.99	32.21	432.45	1.14	7.77	914.32	4.50	0.01	2.57	0.02	
Furnas	Furnas	3,123.02	1,270.76	26.94	4,228.52	4,152.39	25.29	0.75	35.32	243.74	1.41	43.43	189.98	0.04	1.30	26.24	7.87	0.00	1.50	0.07	
Furnas	Macela	2,577.30	1,333.00	29.19	3,479.32	4,102.59	29.43	0.66	33.54	316.14	1.58	47.35	208.82	0.05	2.05	44.25	9.82	0.00	1.56	0.01	Soluble (inc. cytosol & soluble protein, i.e. metallothionein)
Macela	Furnas	2,588.72	1,736.88	25.84	3,494.73	3,673.93	33.08	0.57	44.39	263.46	0.87	17.49	211.47	0.05	1.62	37.56	3.16	0.00	0.41	0.01	
Macela	Macela	2,308.80	1,506.85	26.28	3,116.96	3,484.65	26.52	0.68	43.13	225.79	0.98	15.95	197.84	0.05	3.31	43.88	3.97	0.00	0.19	0.00	
Furnas	Furnas	1,011.36	238.67	412.76	1,209.00	604.78	7.56	0.21	13.07	73.08	0.19	5.49	46.73	0.12	4.42	7.59	0.46	0.00	1.75	0.01	
Furnas	Macela	1,195.40	192.09	343.95	1,231.90	473.42	5.71	0.20	9.76	72.94	0.23	5.26	52.40	0.11	4.08	26.16	0.55	0.00	1.56	0.00	Tissue fragments, membranes, organelles
Macela	Furnas	1,448.58	221.96	368.67	1,192.24	454.86	6.25	0.12	8.88	50.48	0.17	1.60	35.39	0.09	3.92	8.04	0.14	0.00	0.20	0.00	
Macela	Macela	1,414.90	170.44	295.70	926.75	370.07	4.04	0.10	9.54	39.63	0.13	1.59	40.23	0.08	3.62	13.22	0.20	0.00	0.15	0.00	
Furnas	Furnas	15,430.10	1,062.01	2,518.68	549.97	7,523.65	184.06	10.47	316.36	2,902.97	8.61	44.99	139.60	1.86	1.30	167.14	0.91	0.01	32.28	0.08	
Furnas	Macela	13,930.10	1,718.78	2,010.45	613.71	9,659.47	106.77	9.09	387.04	3,239.30	6.61	53.15	190.21	1.56	1.56	1,096.68	1.40	0.01	34.67	0.09	Metal Rich Granules
Macela	Furnas	18,560.78	2,740.62	1,762.48	610.74	10,258.27	192.92	14.41	408.89	2,696.30	48.40	18.59	151.77	1.21	0.70	299.27	0.24	0.01	2.62	0.01	
Macela	Macela	18,909.42	2,783.91	1,123.63	410.96	10,934.17	127.76	8.73	513.74	2,309.67	37.88	14.67	194.38	1.01	0.84	857.22	0.34	0.01	2.24	0.01	
Molecular Weight:		23	24	27	39	44	47	53	55	56	60	65	66	71	75	88	111	115	208	209	

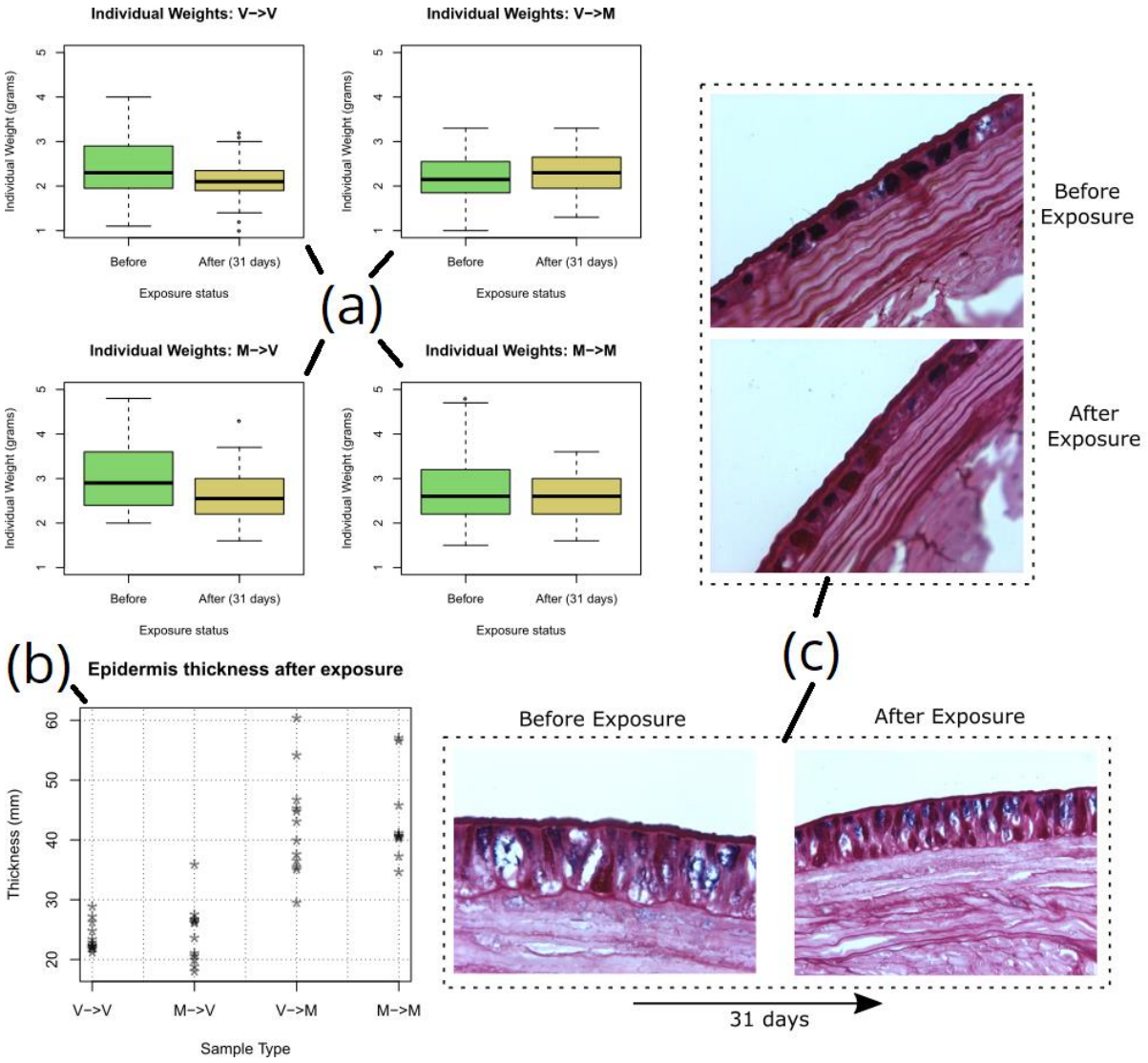


Figure 37. Morphometric changes in worms provided by Luis Cunha from paper in preparation (REF), (a) weight changes between samples before and after exposure, (b) mean epidermis thicknesses of post-exposure individuals, (c) micrograph of pre and post-exposure epi

3.2.3. Genome Assembly

Analysis of the genome began with investigation of the allelic diversity distributions, as described in 3.2.2., they were expected to peak a little below 4%. A sliding window plot of scaffold segments between demonstrated that whilst a divergence peak just over 3% did exist, the distribution of allelic divergence in the genome was also highly bimodal (Figure 80).

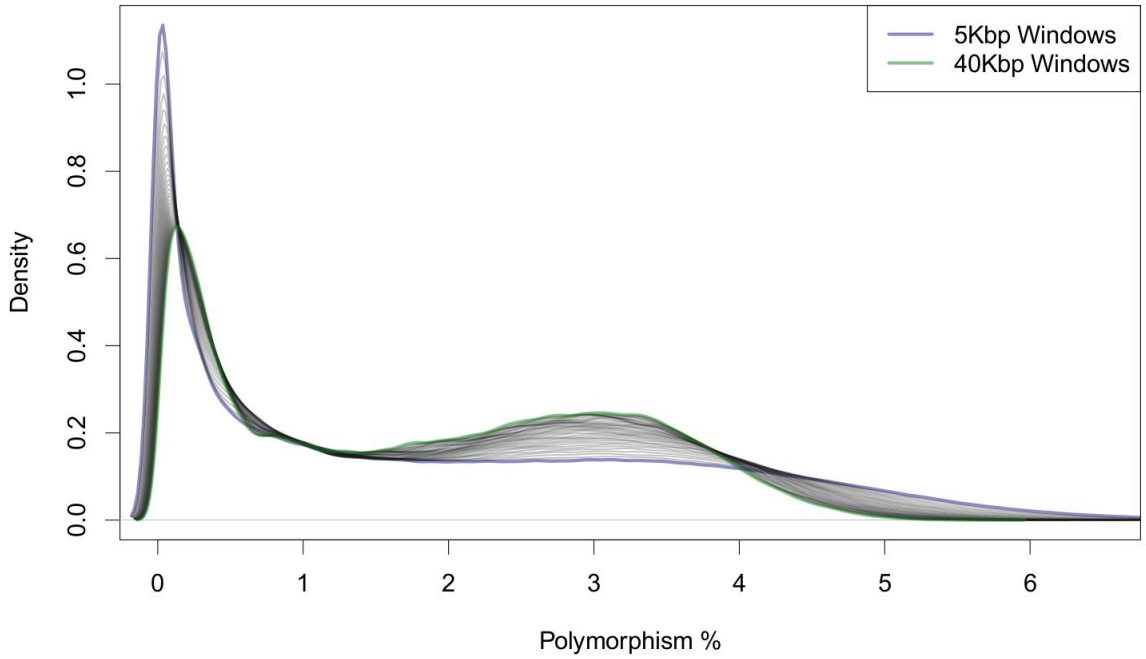


Figure 38. Allelic divergence in Amynthas gracilis genome, 5-40Kbp windows.

To investigate how this bi-modality manifested at a scaffold level, additional visualisations were developed to map moving average divergence rates along scaffolds. Figures 81 and 82 show examples of the divergences mapped along scaffolds.

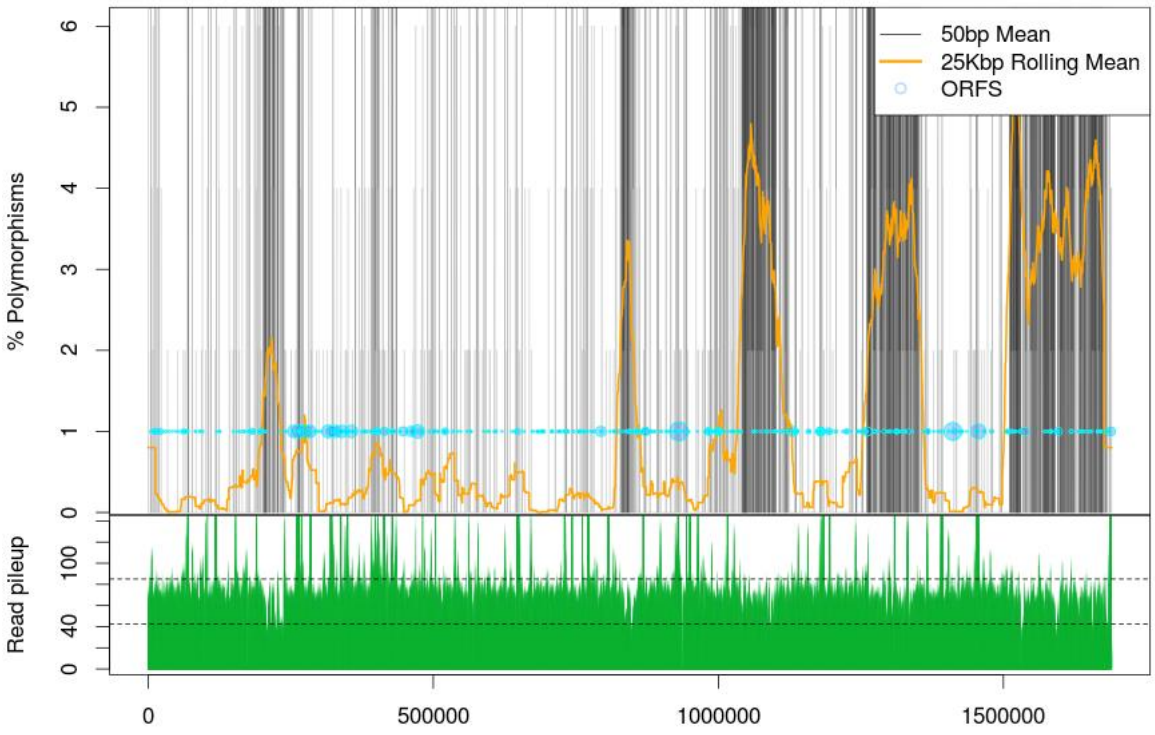


Figure 39. ~1.7 Mega-base region, (scaffold10) spatial polymorphism rate

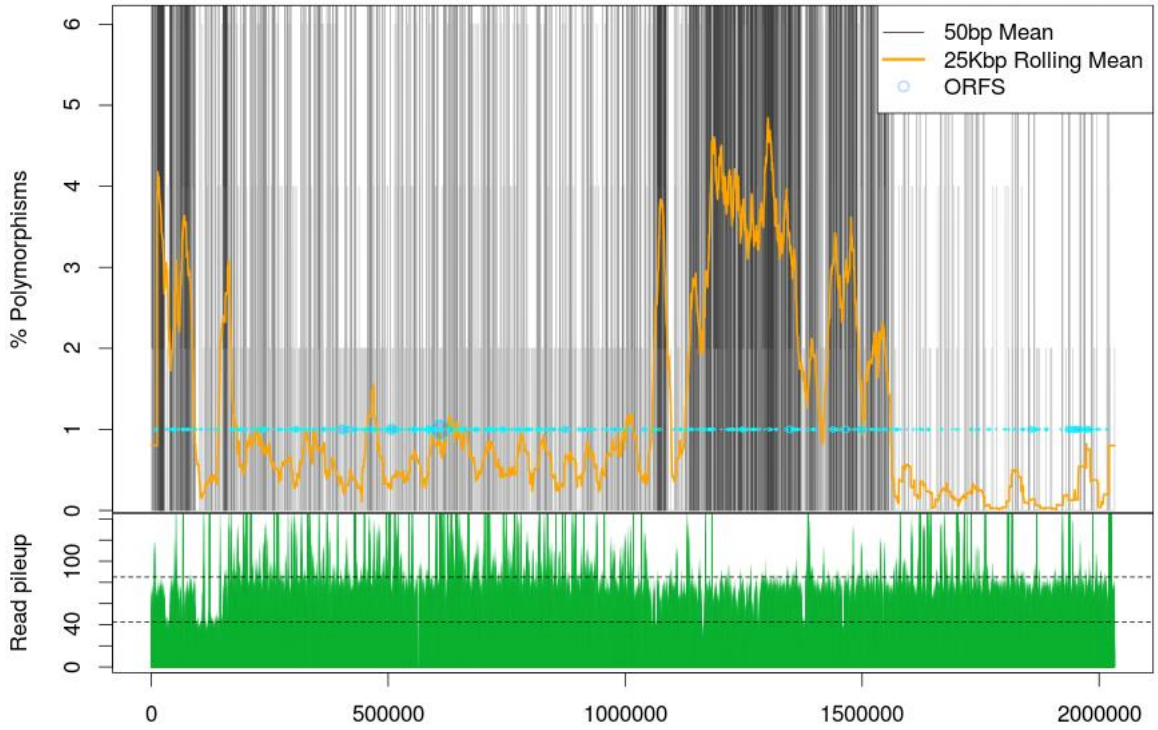


Figure 40. ~2 Mega-base region, (scaffold15) spatial polymorphism rate

These images demonstrate that a certain 'blocky' feature exists in the genome, whereby the allelic divergence changes rapidly across relatively small intervals. They also demonstrate how read-pileup remain the same modal depth along the length, although opposite allele alignment rates tend to dip slightly in the higher divergence regions.

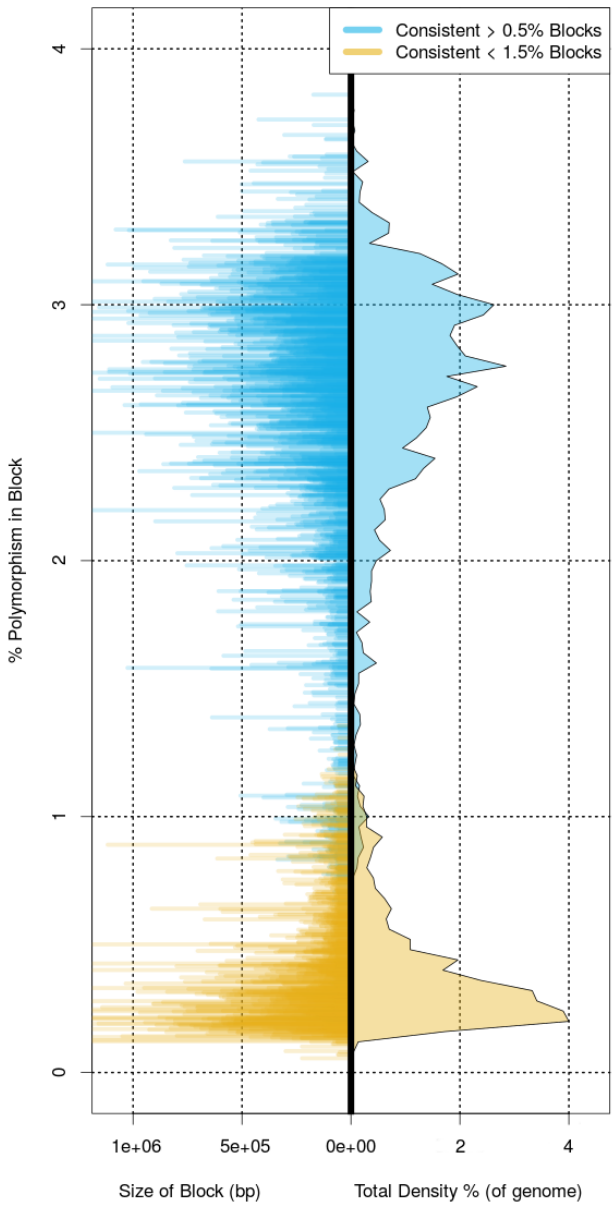


Figure 41. Consistent Polymorphism Rate blocks in Amynthas gracilis genome.

To demonstrate the pervasiveness of this bi-modality and short interval divergence rate change, a simple procedure measuring areas 'consistent' polymorphism rates was deployed. In this case 'consistent region' is defined as a stretch of DNA for which 25Kbp rolling mean rate did not cross a given threshold in either direction. Figure 83 shows the distribution of these 'consistent' regions throughout the genome.

Finally, a PSMC population history analysis was performed on the genome, to investigate how this unequal distribution of variants might reflect its recent evolutionary history. As the base mutation rate for this species has not been tested, the discovered mutagenesis rates in *C. elegans* were used as estimates (Kutscher 2014).

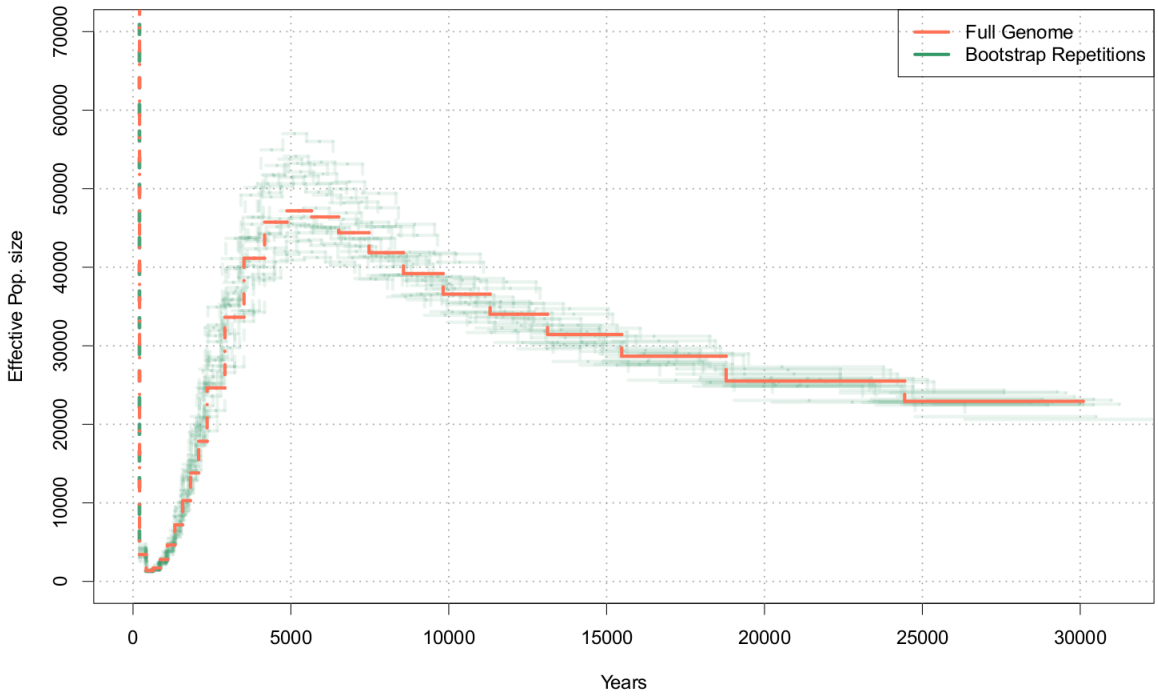


Figure 42. PSMC Population history estimation derived from the Amynthas gracilis genome, mutation rate (2.8 x 10^{-9}).

The PSMC analysis (Figure 84) suggests the genesis of the current allelic divergence to reflect a historically large effective population size, with a sharp bottleneck having occurred approximately 400 years ago, followed by a population explosion. This seems to reflect the timing of the arrival of Portuguese colonists to the Azores.

3.2.4. Methylome Models

3.2.4.1. Static Models

Despite the huge differences in the per-gene methylation rates in the four samples, Figure 85 shows a remarkable convergence in gene-body methylation structure. This suggests that the large-scale changes between the samples are genuine reflections of their biological roles (as opposed to noise). The largest differences between the two model types are the differences between exonic and intronic methylation rates. The binary model shows that there are more introns which can receive small numbers of Me-DIP read alignments, whilst the depth-based model shows that smaller number of methylation-susceptible exons are methylated with far higher frequency across tissues. Overall these models suggest that the subset of methylation-capable exons is comparatively small, relative to introns, but of that set the methylation occurs more frequently than a given methylated intron. Intronic methylation appears comparatively rarer across tissues, but more diverse in sources.

Another relationship confirmed by all samples was the increasing likelihood of methylation towards middle of the gene. The read-depth and binary coverage models both showed a directly proportional relationship between exon and intron positions and the coverage rates.

There are also methylation probability spikes towards the 3' ends of both introns, which aren't repeated in the depth-model. Suggesting that these splice junctions are the gene element most broadly interacted with by DNA methylation, even if the consistency with which any one of them is methylated is comparatively low. This might suggest a more transient set of niche epigenetic controls.

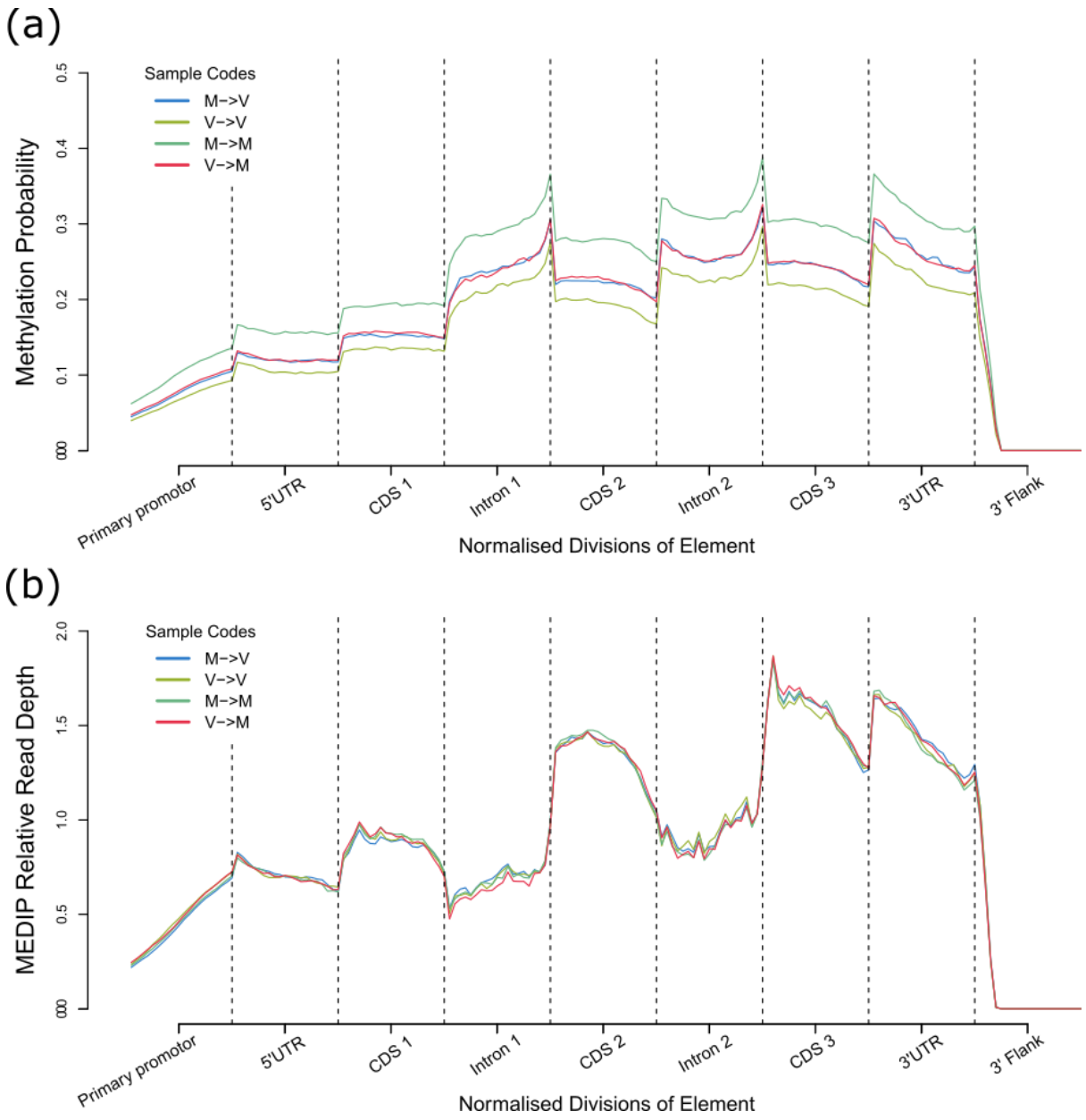


Figure 43. Initial Gene-body models (A) Binary Coverage Probabilities, (B) Relative Read depths.

All-sample mean expression rank group methylation models were effective at demonstrating how the typical expression level of a given gene affects the differential between epigenetic activity within its elements. The separation in methylation rates occurs the most strongly between the lowest five ranks groups, indicating that the lower 50% of gene by expression have a positive linear relationship between their epigenetic interactions and their expression level, however of the upper 50% of the group there seems to be little correlation. These conclusions are supported by both models shown in Figure 86.

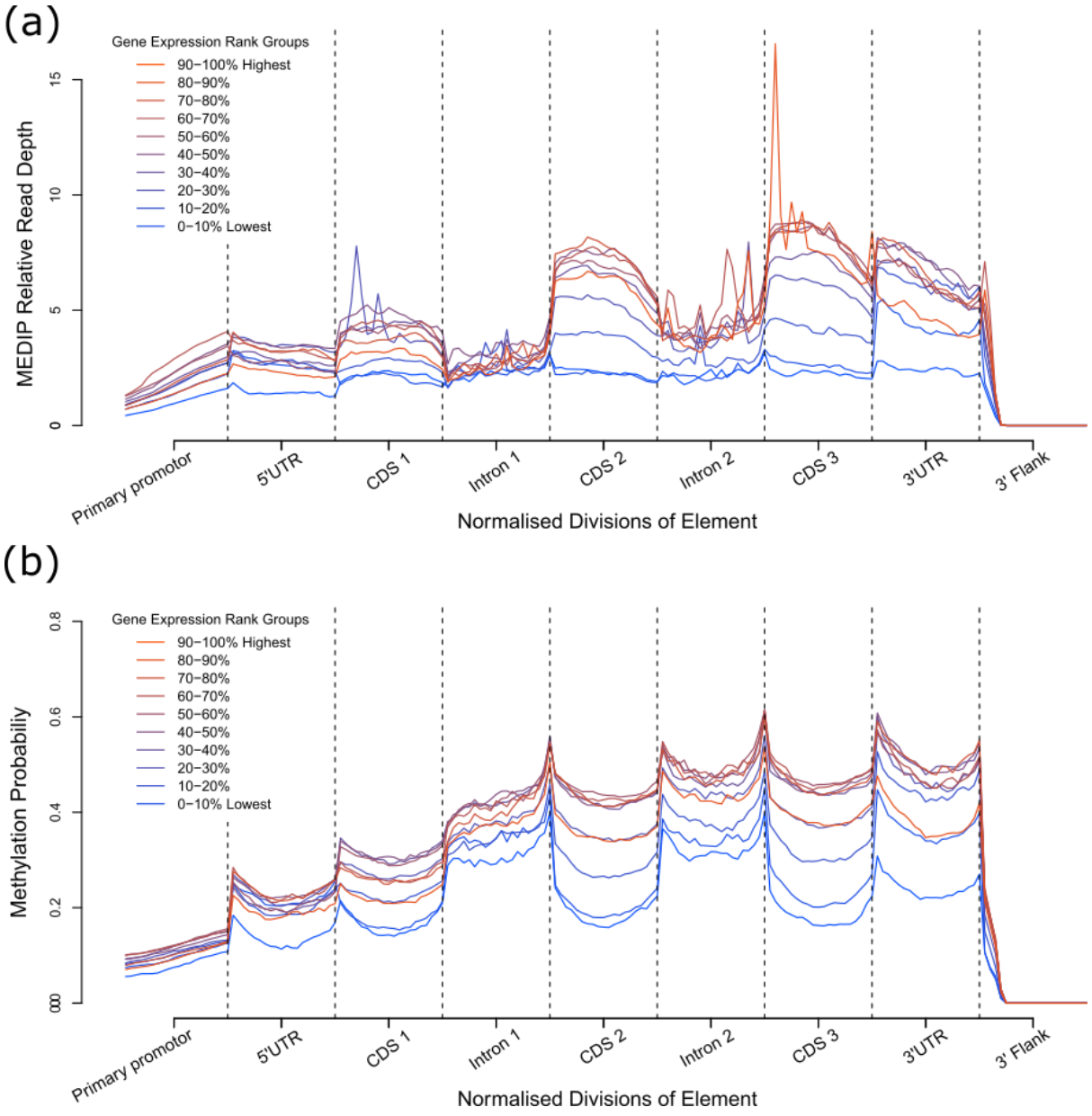


Figure 44. Expression Rank-group gene-body models, (a) Absolute read depths, (b) binary coverage probabilities.

Closer examination of the terminal regions of intronic elements showed a noticeable feature probability increase towards the 3’ terminal 50bp of the element (means across the set of all

introns). The 5’ 50bp of intron also demonstrated a slight alteration in read coverage distribution, but the probability change was relatively negligible compared to the 3’ end, see Figure 87.

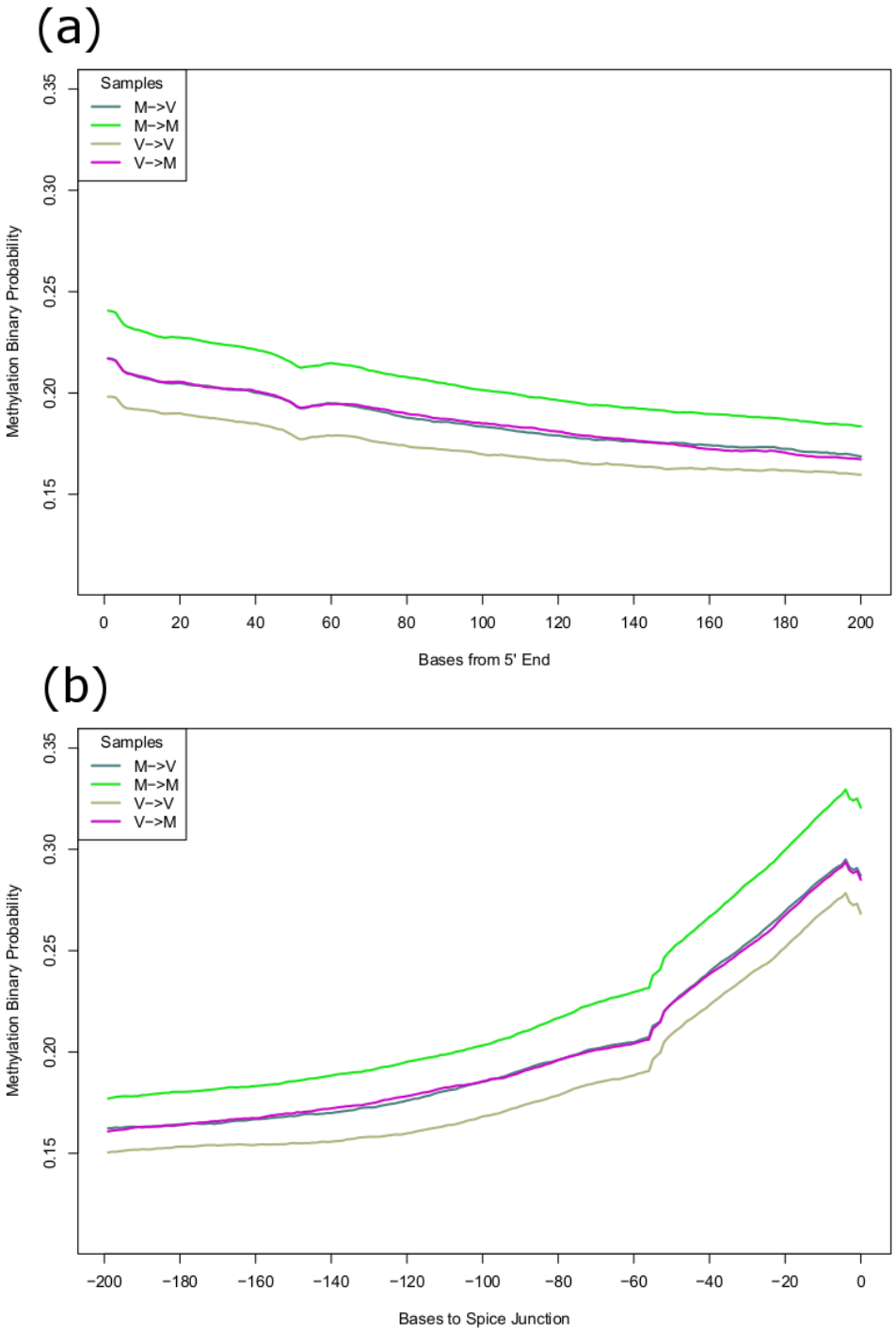


Figure 45. Intron Terminal Region MeDIP read-mapping probability, (a) terminal 200 bases from 5’ end, (b) terminal 200 bases from 3’ end.

3.2.4.2. Differential Models

Normalised differential models were generated, Figures 88 and 89 both show (a) the profile models normalised per interval, and (b) the differential model divided by abundance model. This applies to both the absolute and binary models. Whilst changes in abundance are the most marked in Figure 88 (a), they almost completely disappear when divided by the general abundance, except for promoter methylation, which continues to show the greatest rate of flux despite its lower incidence rate.

Figure 89 (b) shows that the binary odds ratio is consistently above 1 for most of the gene model. This is reflective of the high degree of difference that can be seen in the epithelial development gene methylation abundance heatmap in Figure 111. This also suggests that while intronic methylation may be more stable, most exonic methylation rates are highly changeable and may be partially stochastic. It also aligns in with the absolute model in suggesting that promoter methylation is the most unstable and prone to flux. The way in which gene-body elements separate far more clearly in the binary model than in the absolute model post-profile normalisation seems to indicate that the noise in the absolute levels actually limits accessibility to biologically informative results in the data, this might be the result of the mixed tissue sample preparation method, and could be circumvented by a single tissue sampling procedure with higher numbers of replicates.

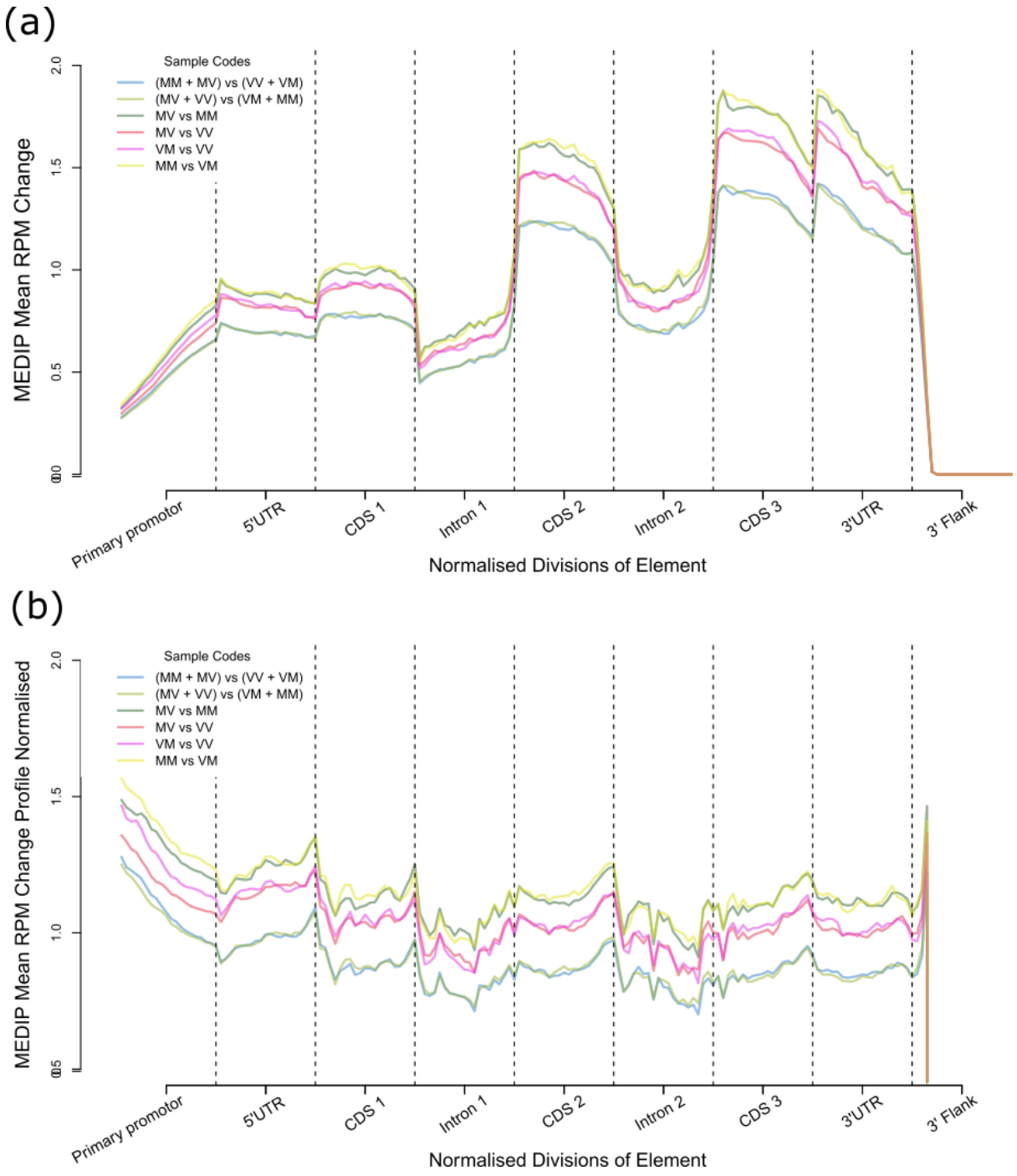


Figure 46. Read depth-based gene body differential models, (a) Interval-normalised RPM changes, (b) Model in (a) divided by abundance model in Figure 85 (b) – change normalised by abundance

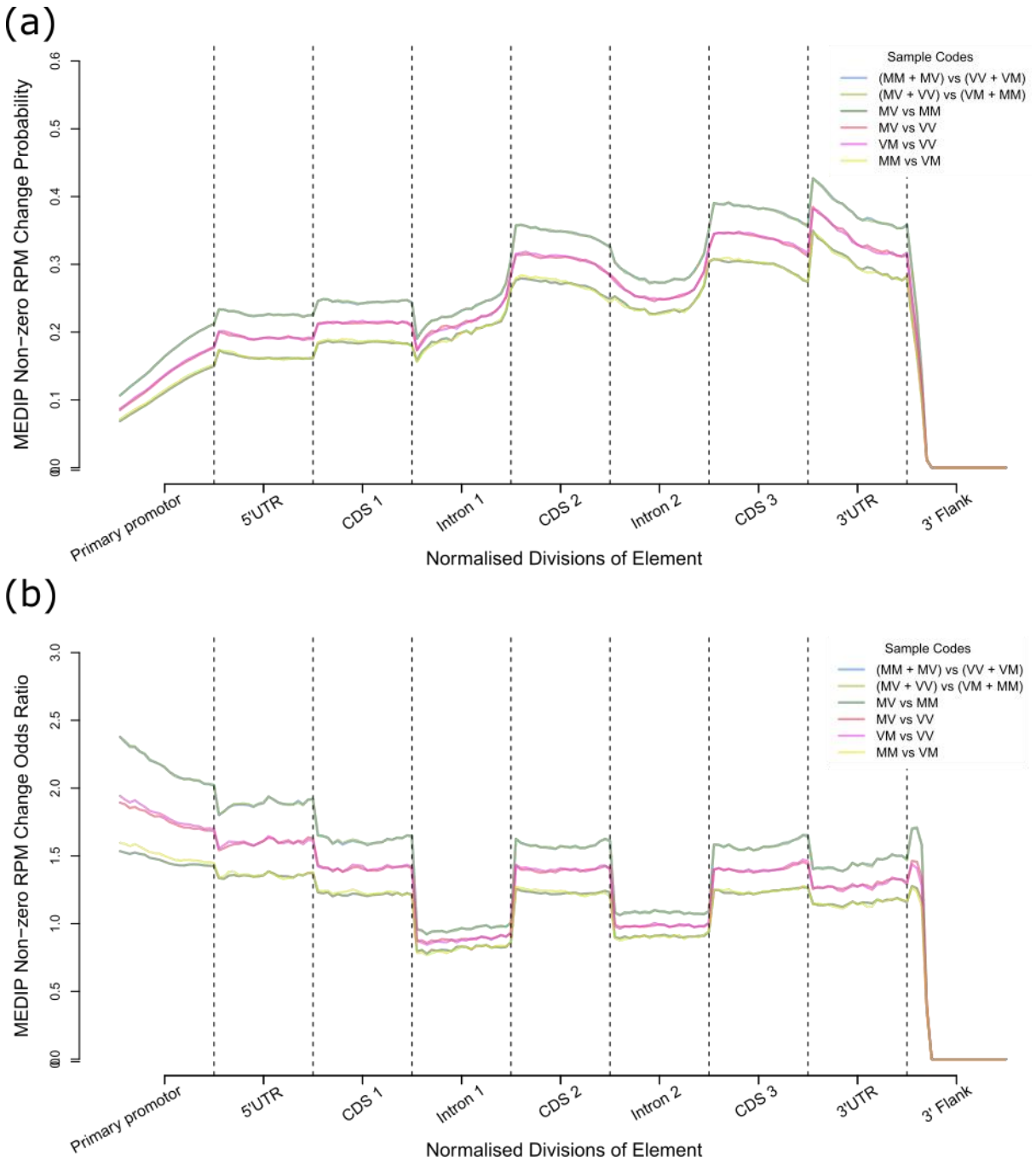


Figure 47. Binary Coverage gene body differential models, (a) Interval-based read depth change binary probability. (b) Model in (a) divided by incidence probability rate in Figure 85 (a) – Odds ratio of change probability over incidence probability.

3.2.4.3. Sequence Structure Signatures

Between Figures 90 and 91, it appears than 3' and 5' UTRs, and primary promoters has sequence structure dispersal rate differences between sets based on their methylation levels, with dispersal rates being unchanged in the intronic set. The three sub-graphs within these two Figures should be taken in context of the 2D signature fold-change differentials displayed in Figures 92-95. For example, both UTRs display substantial narrowing of the set of banding patterns in the graph, indicating that the set of dispersal rates of the motif types present becomes most homogenous, and less consistent. Referring to the Chapter 4 interpretive schema, this is suggestive of a larger numbers of similar motifs with inconsistent points of variation in the more methylated UTRs, indicating that there is a long-motif sequence set response to the variable of DNA methylation (as opposed to a simple short motif response as might be anticipated given an increased abundance of cytosines). However, whilst the banding pattern narrowing occurs in both UTRs signature plots, the fold-change graphs indicate that these changes have divergent interpretations.

Figure 93 shows that between the Low methylation and No methylation groups there is no change in structuredness along k-mer sizes 1-15, and a progressive loss of structure in the low methylated group from 15-30. This suggests that occasional/infrequent methylation in 3' UTRs behaves in a complex way which isn't tied to particular sequence structures. However, the two-fold change tests which include the 'high methylation' group show a very different story. There is an 8 kmer range, positioned differentially depending on the N value, which loses a consistent ~30% of its sequence structure in highly DNA methylation regions relative to both other groups. One option here is that these are miRNA seed alignments which are less present in the highly methylated regions, as this are typically in that range (Lewis et al. 2005). This is significant because the high methylation group then demonstrates 2-3-fold increase in long motif structure above k=15. To summarise, the correlative effects methylation therefore appears have with UTR structure appear only in the top 30% of UTRs by methylation, and comprise consistently reduced dispersal rates, loss of short k-mer structuredness and a dramatic increase in long motif abundance. We would expect therefore to find a large set of longer similar sequence structures in the methylated UTRs, each set internally diverging based on inconsistent points of sequence polymorphism.

Comparing 5' UTRs in Figure 94 with Figure 93 shows a remarkably similar short motif depletion effect, however in this case is occurs between the two methylated sequence sets and the non-methylated set. The difference being that occasional methylation in 5' UTRs now correlates the same way with sequence structure change as frequent methylation does with the unmethylated set. The same short motif seed alignment loss theory may be accurate; however, this also suggests a substantially low miRNA binding rate with the 5' UTR than the 3' end. Another difference is that the

longer k-mer motifs are not enriched in the same way as the 3’ UTR. A short region of motif length hyper abundance occurs around the sizes 13-18bp reaching up to a 3-fold differential structure increase, but substantially longer motif signatures lose structural abundance, suggesting that a different class of motifs by size is associated with methylation in 5’ UTRs than in 3’ UTRs.

The group fold change graph for splice junctions shows some small methylation effects, however the X-scale of these changes is on the order of $1/10^{th}$ that of all other signature structure variants, making them inconclusive, other than to suggest that differential splice junction methylation rate across the set is generally unrelated to their sequence content.

Promoter sequence groups by structure exhibit no variation up to k=13 in Figure 94. This suggests that the structures which emerge later causing the 2-3-fold differential are the result of many similar small sequence groups. The effect of restricting the input set to a group of sequences close in length to the depth of tree also has the effect of making the *k* sizes in the dispersal graphs in Figure 91 more reflective of the variation at specific genomic loci relative to the TSS, rather than of motif structure in the kernel space. For this reason, the promoter dispersal signatures are far more jagged, reflecting commonalities in base conservation amongst the set of all promoters. This effect is also manifesting in the higher values of the fold-change graph. That the fold-changes similar in form but greater in magnitude in the N=1, and N=2 graphs suggests that many of the highly similar sets of methylated TSS adjacent promoter sequence vary only by a couple of bases.

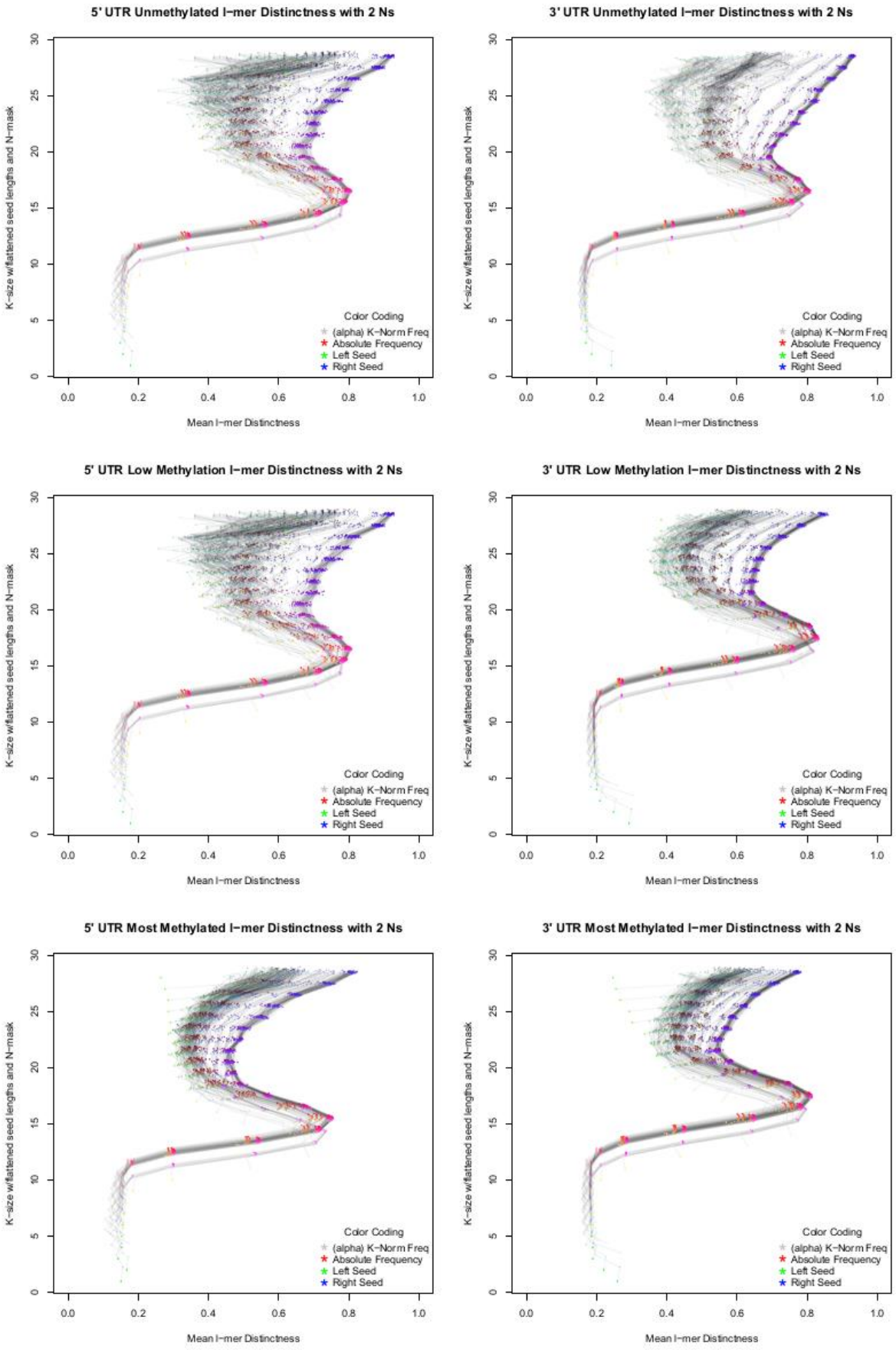


Figure 48. 5' UTR (left) and 3' UTR (right) methylation rank group sequence signatures.

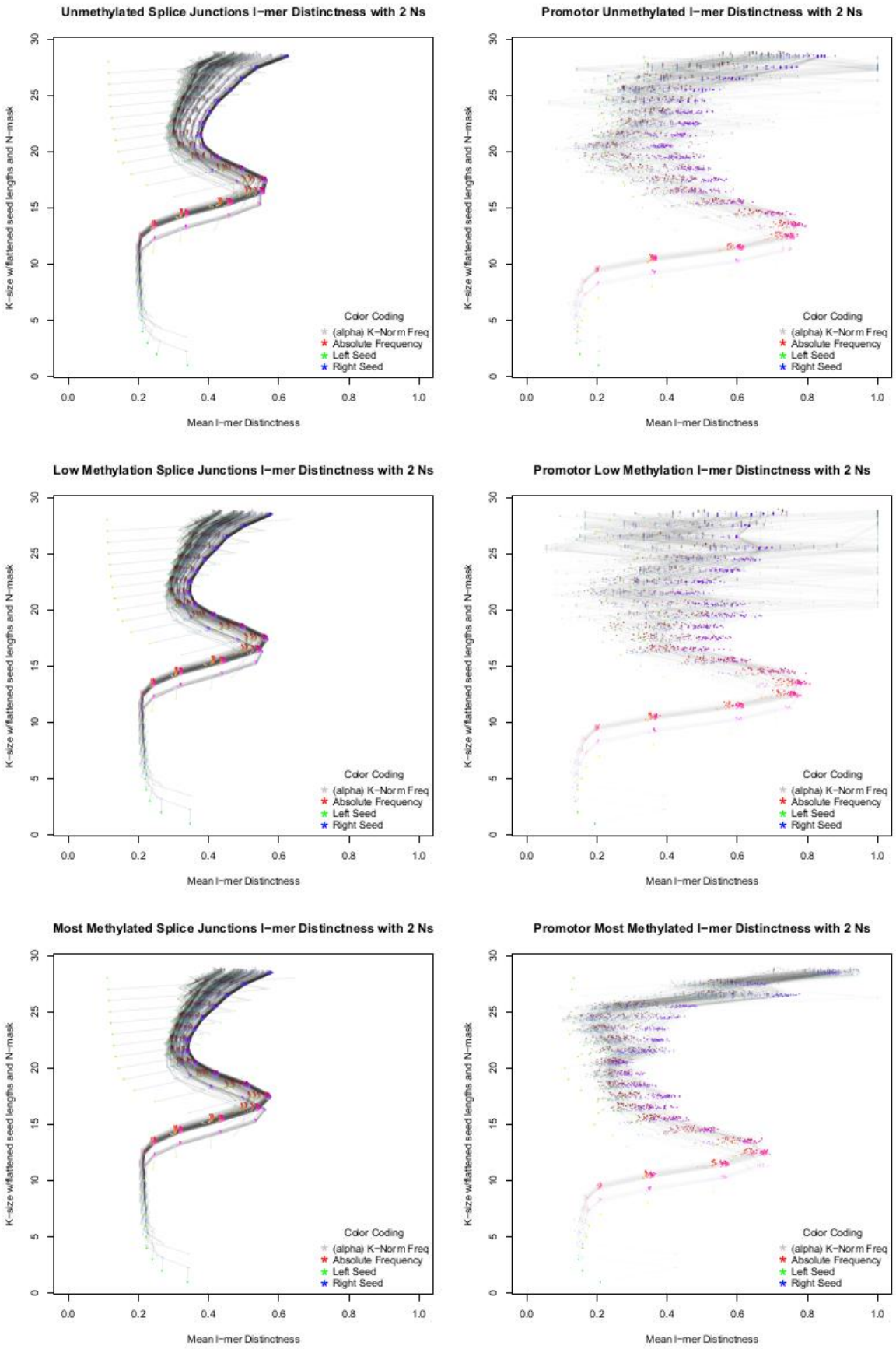


Figure 49. Splice Junction (left) and 34bp Promoter (right): methylation rank group sequence signatures.

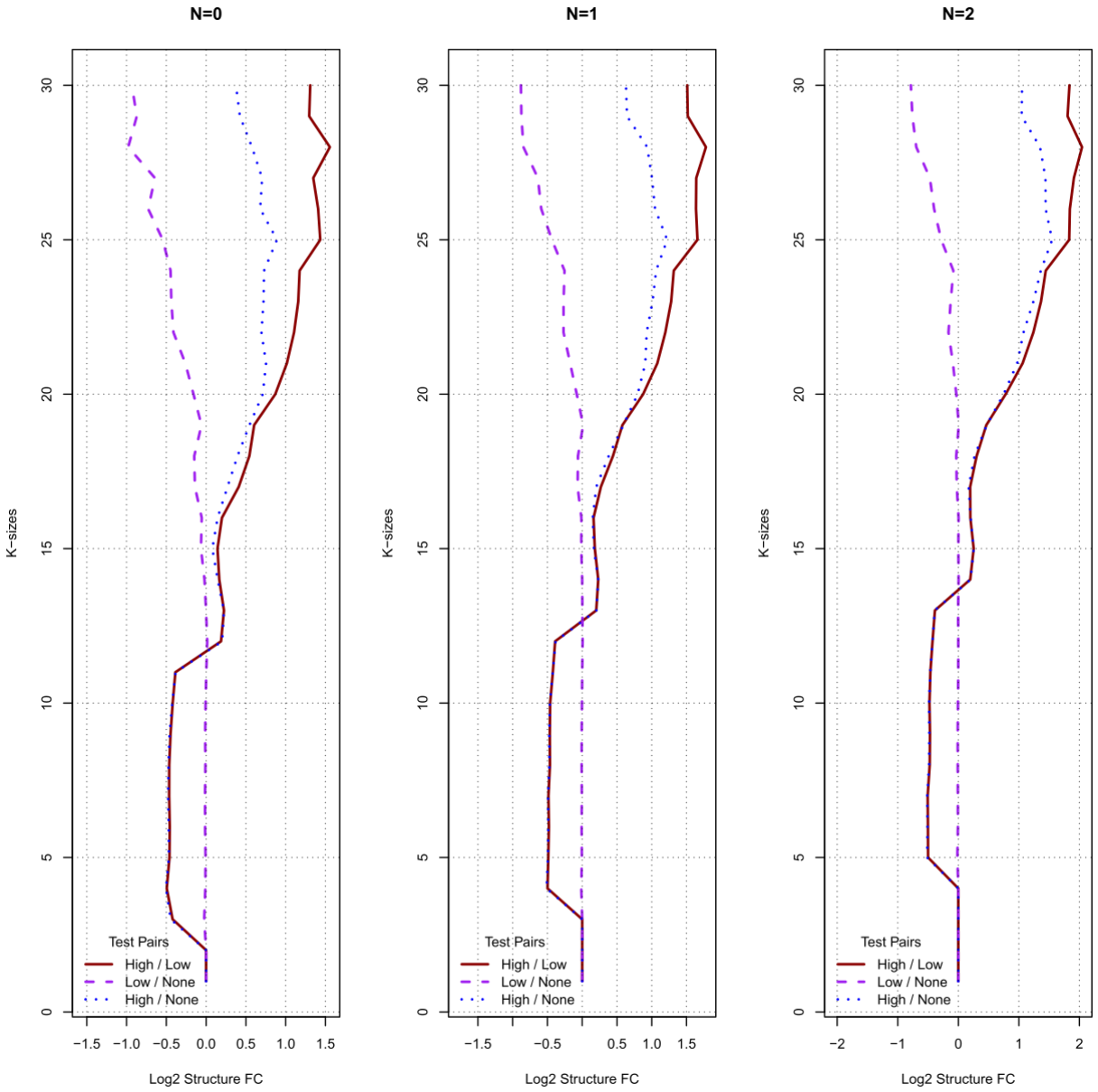


Figure 50. 3’ UTR Differential K-mer structure scores fold change between methylation groups.

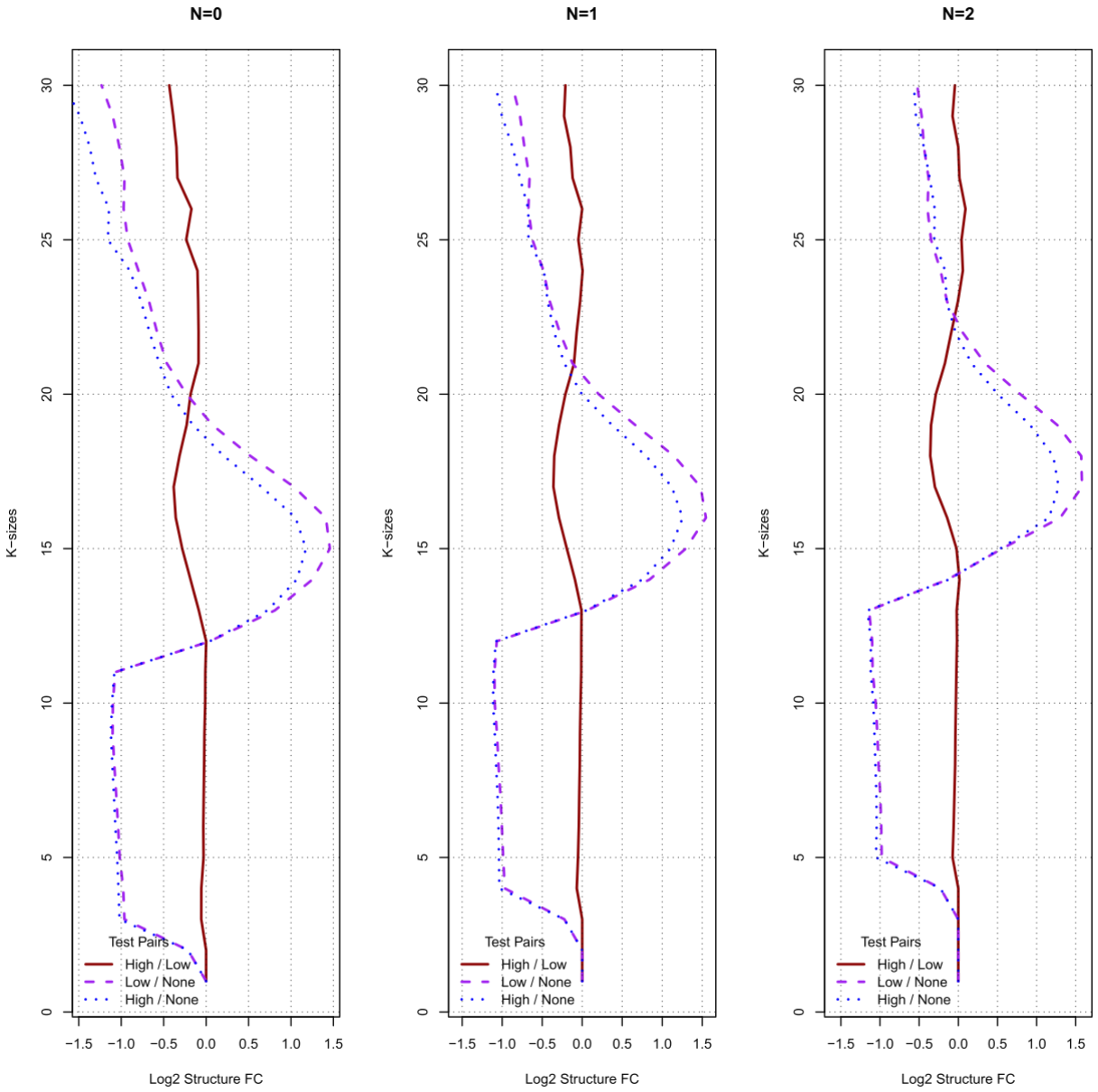


Figure 51. 5' UTR Differential K-mer structure scores fold change between methylation groups.

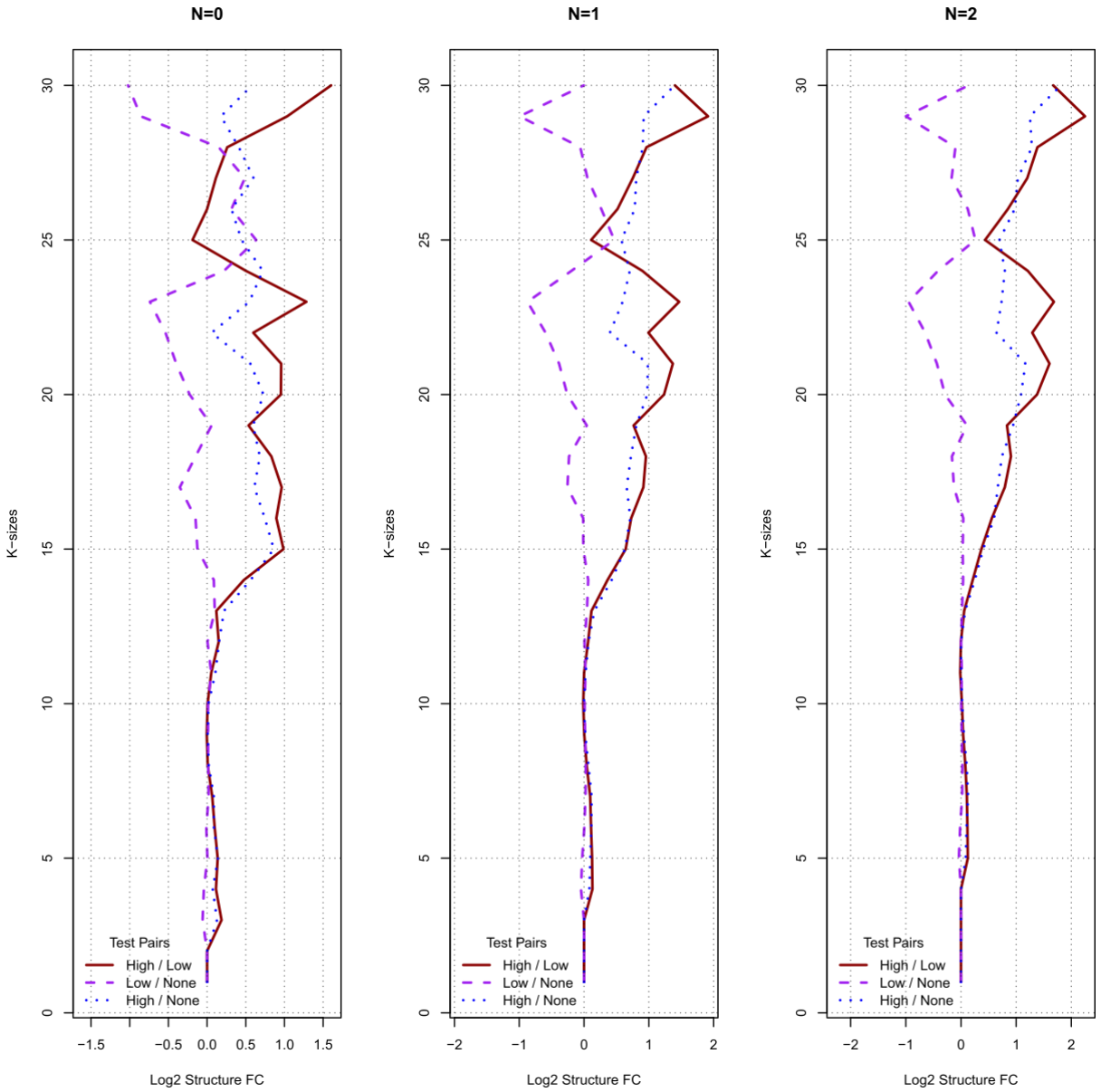


Figure 52. 34bp Promoter, differential K-mer structure scores fold change between methylation groups.

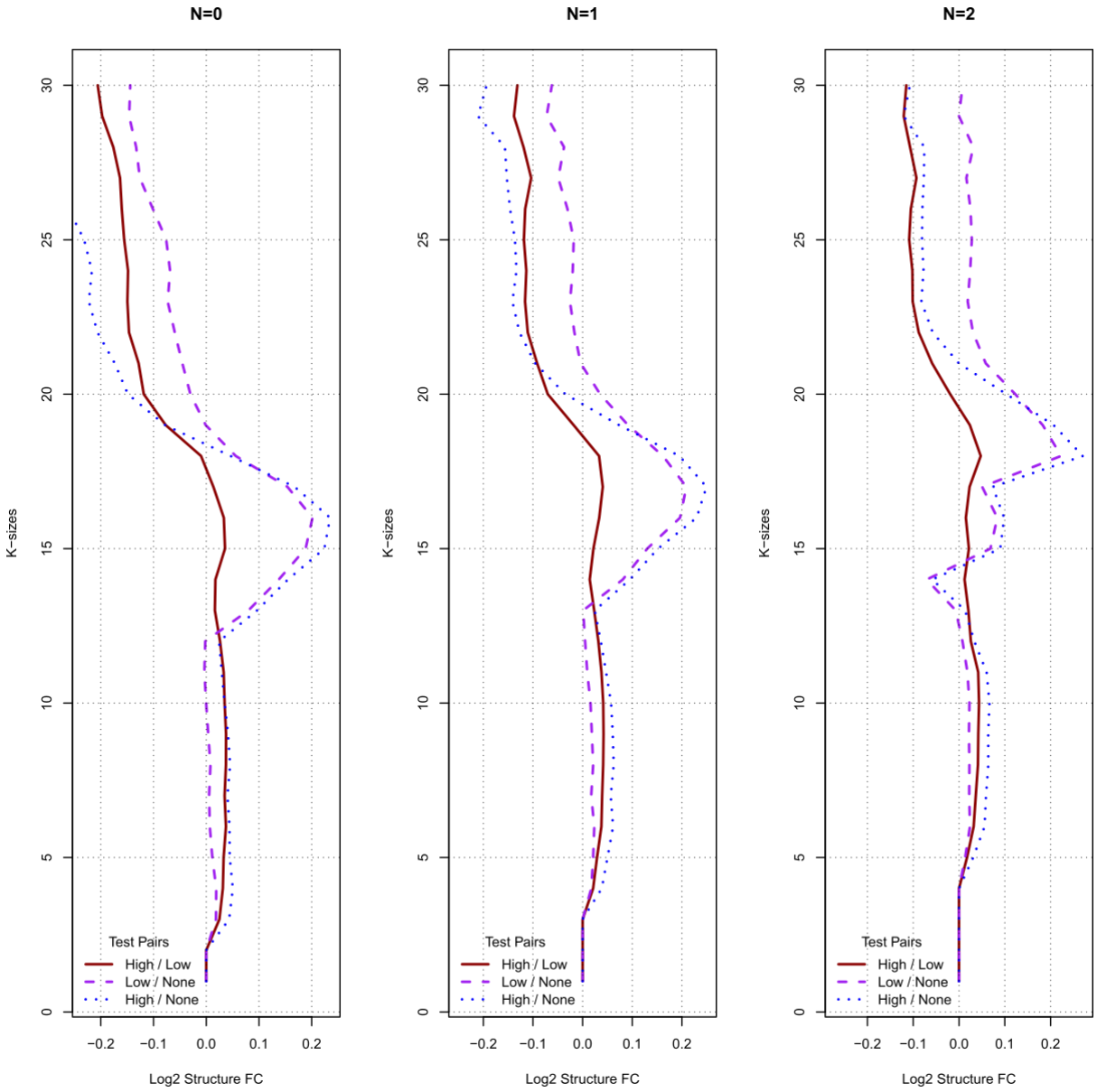


Figure 53. Splice Junction (100bp) differential K-mer structure scores fold change between methylation groups.

3.2.4.4. RNA-Seq Intersect Models

Quantile-Quantile grids show a few clear messages about the distribution of methylation amongst genes by their expression levels. Firstly, Figure 96 buttresses the interval model suggestion from Figure 96 that the lower half of the gene set by expression is the most positively correlative with mean methylation levels. In the lowest 20% of genes by methylation in 26 (a) there is a notably 40-70% over-representation of genes in the lowest category of expression by 10% quantile, and under-representation of highly expressed genes, although the top 10% of genes by expression exhibit no methylation correlative effect. The models produced to investigate the same effect in 1kb promoter regions show very little significance at all. Figure 97 (a) and (b) show gene body models developed for gene expression and methylation differentials with two different test types. The relationship does not vary by test type, and it appears that the top 20% of genes by exhibited variability in methylation have up to 2-fold enrichment in the top 20% of genes by FC differential. Quite simply it suggests that the differential x differential relationship is: ‘change begets change’ in one way or the other. The repetition of the differential model for 1kb promoter regions in Figure 98 showed a similar enrichment pattern in the top/bottom 10% categories, but with highly diminished scale and significance, suggesting this might just be an adjacency effect given the primary methylation target of the gene body.

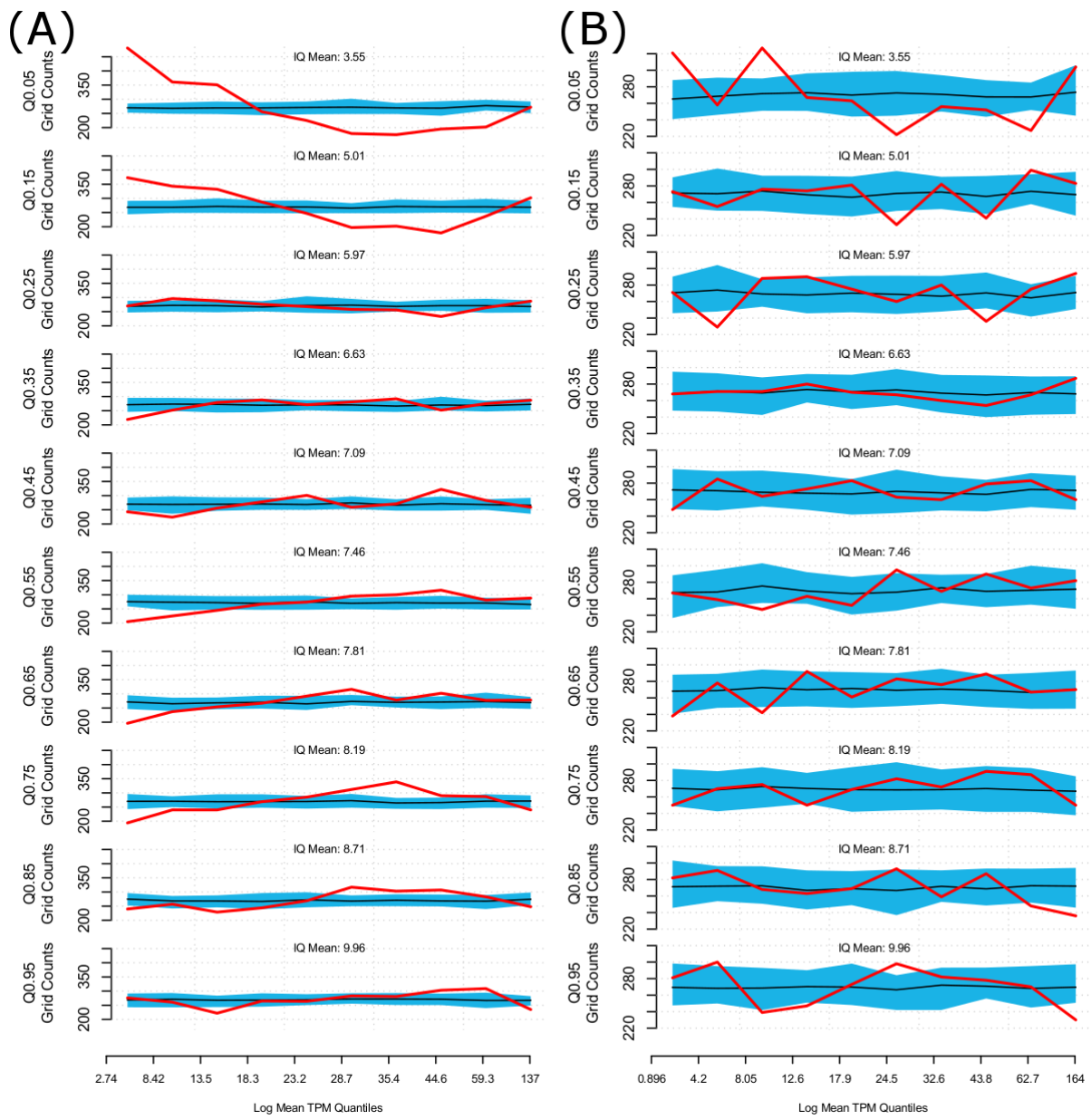


Figure 54. Quantile-Quantile map for mean gene expression. X= Log TPM Gene Expression 10% incremental Quantile densities given the current Y group. From top to bottom: Genes grouped by 'least to most' methylated (Log RPM), in 10% incremental Quantiles. This chart shows the null quantile density 5/95% confidence interval width for bootstrapped random pairings of expression and methylation read count (Blue), and the actual quantile densities given the real data pairings (Red), such that where the red line departs from the confidence interval it might be considered significant. This test repeated for associated methylation rates originating from (A) Gene Bodies, (B) 1kb Promoter.

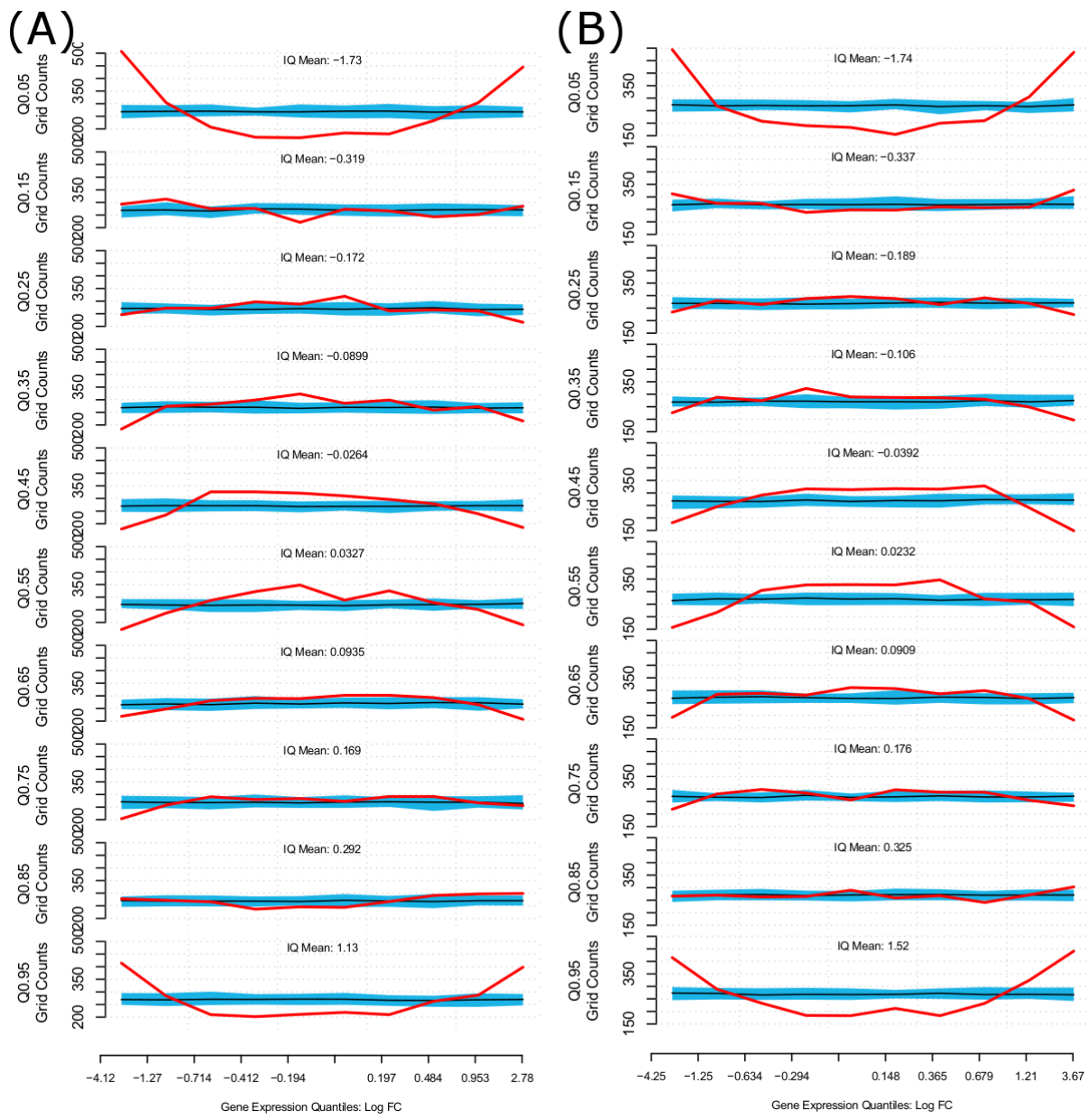


Figure 55. Quantile-Quantile map for differential gene expression vs differential methylation. X = Log FC differential gene expression 10% incremental Quantile densities given the current Y group. From top to bottom: Genes ordered by log FC gene body methylation differential, in 10% incremental Quantiles. This chart shows the null quantile density 5/95% confidence interval width for bootstrapped random pairings of expression and methylation read count (Blue), and the actual quantile densities given the real data pairings (Red), such that where the red line departs from the confidence interval it might be considered significant. This test repeated for associated methylation rates originating from (A) Test (4): (M -> V) vs (M -> M), (B) Test (5): (V -> M) vs (V -> V).

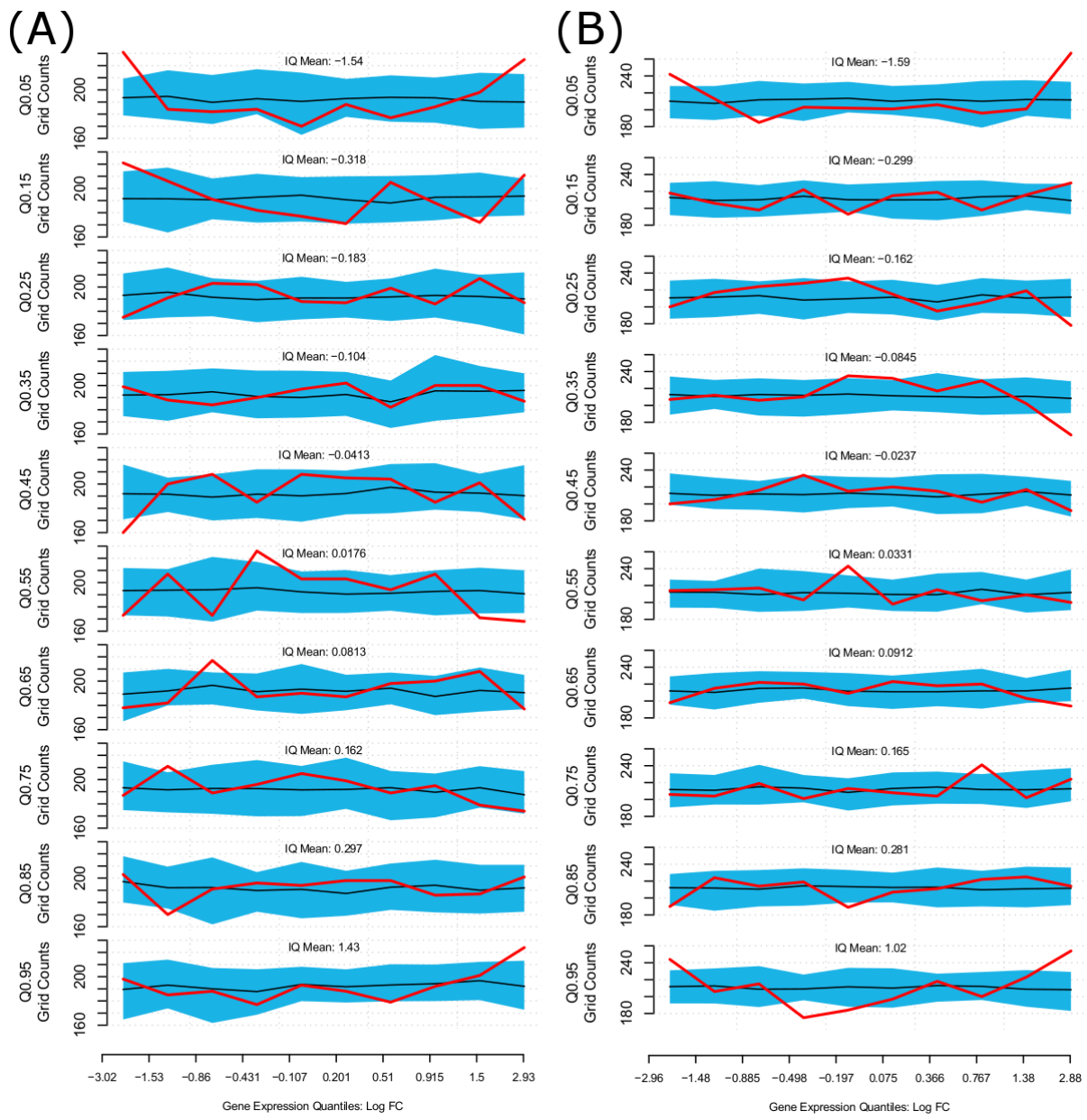


Figure 56. Quantile-Quantile map for differential gene expression vs differential 1kb promoter methylation. X = Log FC differential gene expression 10% incremental Quantile densities given the current Y group. From top to bottom: Genes ordered by log FC 1kb Promoter methylation differential, in 10% incremental Quantiles. This chart shows the null quantile density 5/95% confidence interval width for bootstrapped random pairings of expression and methylation read count (Blue), and the actual quantile densities given the real data pairings (Red), such that where the red line departs from the confidence interval it might be considered significant. This test repeated for associated methylation rates originating from (A) Test (4): (M -> V) vs (M -> M), (B) Test (5): (V -> M) vs (V -> V).

3.2.5. miRNA Summary

Mature miRNA predictions were made using MiRDeeP2 (Friedländer et al. 2012) – there were 237 novel miRNA which were assembled *de novo* from the sequencing data. 42 of the entries only occurred in two or less of the samples and were filtered out. Of those remaining 89 had no match compared to the entries in MiRbase (Kozomara & Griffiths-Jones 2014), and 106 were identical to pre-existing entries. All the retained miRNA assembled post filtering were within the top 500 miRNA mapped to by abundance – which gives confidence that they are from genuine sources. Although the Lophotrochozoan taxa has very sparse representation in the miRNA sequence database, many of the top reference hits were from the nearest taxonomic neighbours – 19 from *Capitella teleta*, 43 from *Ciona intestinalis*, 27 from *Strongyloides ratti*, 8 from *Lottia gigantea*, and 21 from *Schmidtea mediterranea.* This suggests that despite the relatively poor performance of prior knowledge in annotation in most cases – the taxonomic proximity confirms some degree of expected miRNA conservation within the clade.

3.2.6. miRNA Networks

An initial miRNA network was generated based on the set of miRNA -> Gene bindings discovered. This network visualisation and edge-distribution summary is shown in Figure 99 (a) and (b) respectively. Of the 26,951 genes, 9,363 (*34%)* were found to have at least one miRNA binding site on the conservative rule of two base-changes or less compared to the mature sequences. In the network 8026 (85.7%) of genes were bound to by one miRNA. Of the 2,000 putative miRNAs included in the alignment query, 1,554 (77%) were found to bind to one or more genes, suggesting some degree of over-inclusive error given the arbitrary selection cut-off. Since miRNA functional impact was assessed relative to the bound genes, all unbound miRNAs were not included in the functional annotation enrichment (later sections). Of the miRNA in the network 623 (37.4%) only bound to a single gene, with the rest binding to multiple.

Abundances of p-significant genes related to miRNA changes in this and later sections are gathered based on the p-significance of the miRNAs which bind to them. Functionally describing the miRNA networks involved in expression change between sample was also performed based on gene annotation, rather than miRNA annotation.

Differential miRNA expression sub-networks are shown in Figures 100 and 101, for the Origin vs Origin, and Destination vs Destination tests respectively. The Static vs Change test did not yield enough p-significant results to generate a useful network view. In both test types, functional enrichment clustering of genes annotated by Gene Ontology BP4, in DAVID, yielded similar sets of results to the main clustering experiments in later sections. Both included a highly enriched

proportion of membrane bound proteins. Neural system development was featured in multiple highly abundant clusters in the Origin test (Counts: 130, 46, 45, 11), and in two less abundant but highly significant clusters in the Destination test (Counts: 31, 19). The Destination test network also showed a cluster of 27 ion channel related proteins

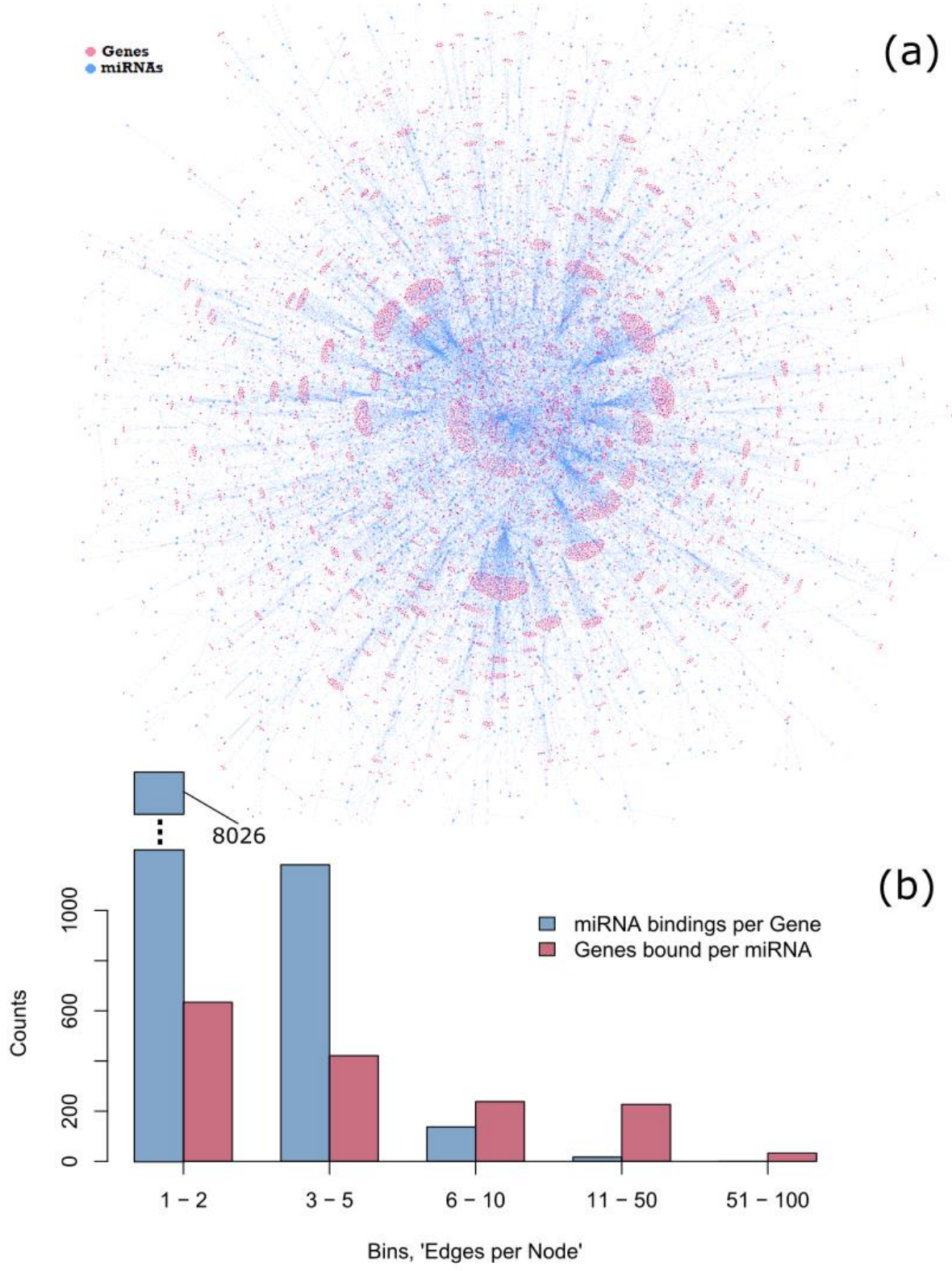
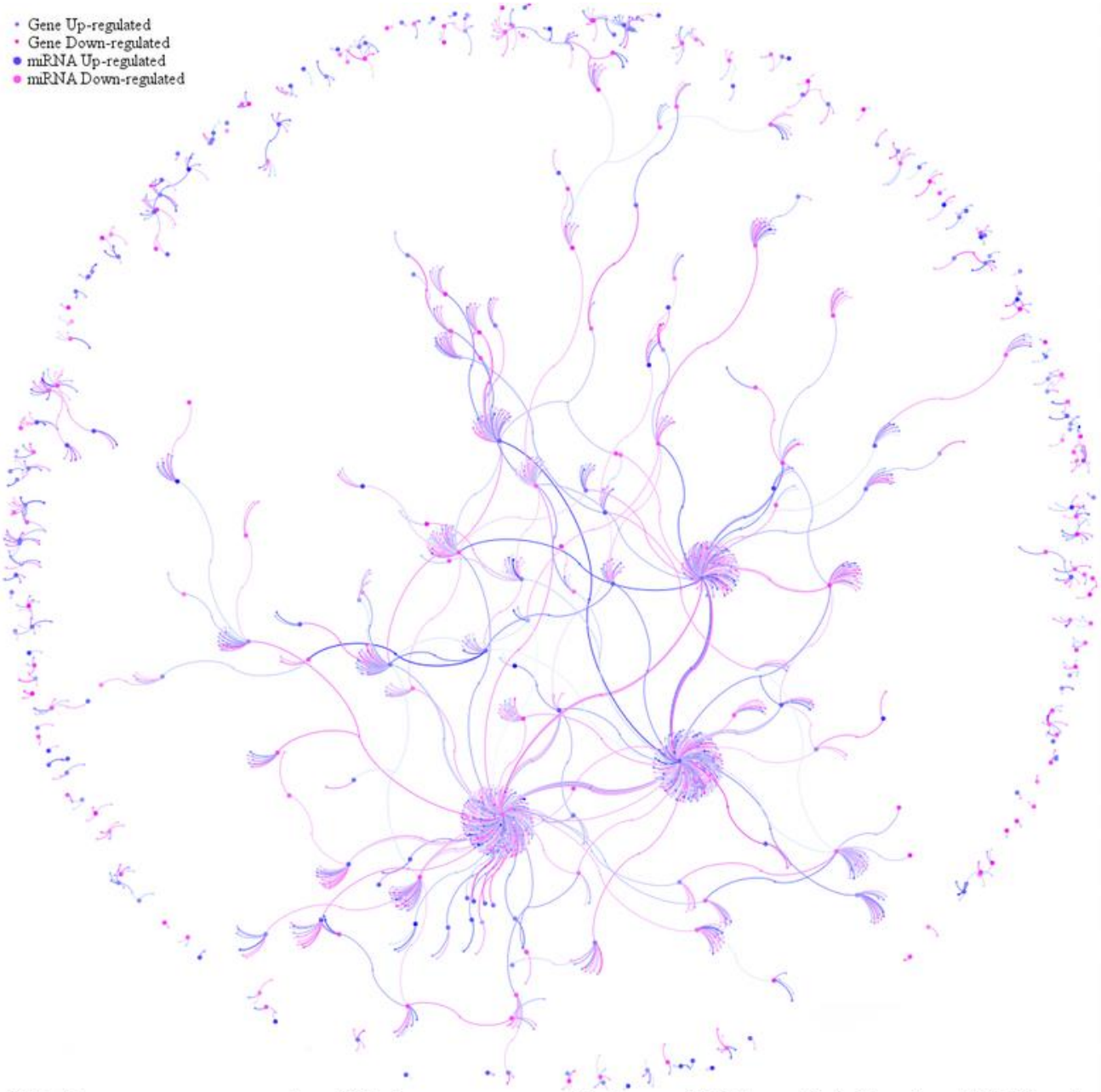
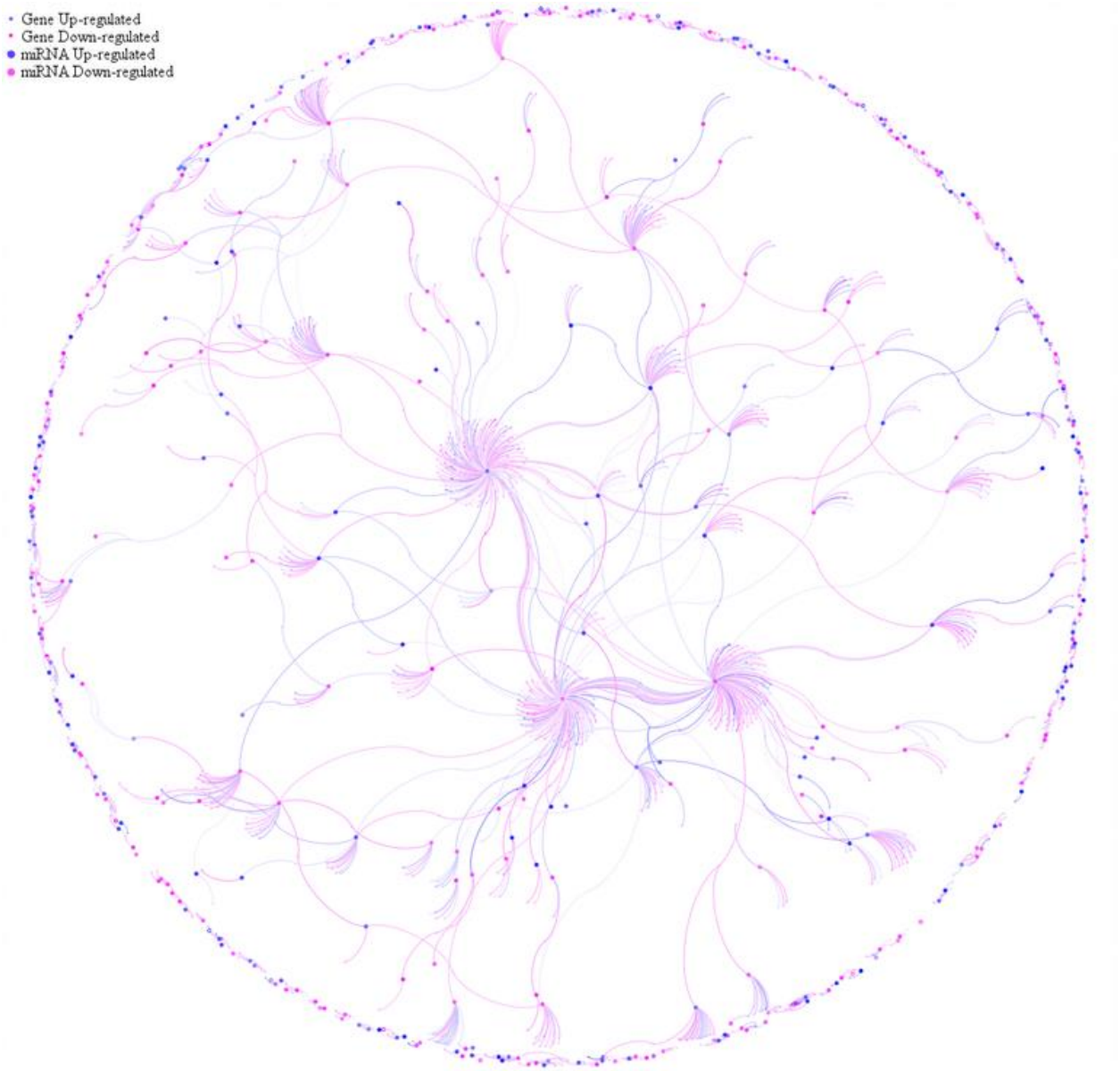


Figure 57. (a) Visualisation of the full miRNA regulatory network between genes (Pink) and miRNAS (blue). (b) Bar graph of the per-node edge count.



Cluster ID	Largest Single Term	Count	% of Set	Benjamini p-value	Fold-Enrichment
1	GO:0044459~plasma membrane part	131	16.7	7.31E-09	7.02E+00
2	GO:0007399~nervous system development	130	16.5	6.47E-06	5.53E+00
3	GO:0099537~trans-synaptic signaling	46	5.9	1.69E-04	5.10E+00
4	GO:0002009~morphogenesis of an epithelium	41	5.2	4.53E-03	3.99E+00
5	GO:0008188~neuropeptide receptor activity	11	1.4	1.42E-02	3.76E+00
6	GO:0061564~axon development	45	5.7	9.03E-05	2.94E+00
7	GO:0003002~regionalization	30	3.8	1.47E-02	2.87E+00
8	GO:0006355~regulation of transcription, DNA-templated	140	17.8	4.06E-03	2.60E+00
9	GO:0048546~digestive tract morphogenesis	11	1.4	1.15E-02	2.60E+00
10	GO:0072358~cardiovascular system development	52	6.6	3.13E-02	2.60E+00

Figure 58. Origin vs Origin differential miRNA expression network and functional enrichment clusters.Shows miRNAs (large dots) and genes (small dots) and whether or not they are upregulated (blue) or down-regulated (pink). All miRNA shown had a p<0.05 significance to its differential expression after fold-change effect size shrinking in deseq2.



Cluster ID	Largest Single Term	Count	% of Set	Benjamini p-value	Fold-Enrichment
1	GO:0031226~intrinsic component of plasma membrane	64	16.5	1.12E-14	1.29E+01
2	GO:0005216~ion channel activity	27	6.9	1.35E-05	5.04E+00
3	GO:0007417~central nervous system development	31	8.0	7.94E-03	4.05E+00
4	GO:0097060~synaptic membrane	19	4.9	4.18E-04	3.74E+00
5	GO:0048731~system development	102	26.2	1.53E-03	2.76E+00
6	GO:0008076~voltage-gated potassium channel complex	7	1.8	4.50E-02	2.28E+00
7	GO:0001944~vasculature development	20	5.1	1.07E-01	2.27E+00
8	GO:0015464~acetylcholine receptor activity	6	1.5	2.09E-02	2.08E+00
9	GO:0008015~blood circulation	16	4.1	4.02E-02	1.98E+00
10	GO:0009886~post-embryonic morphogenesis	10	2.6	3.11E-02	1.91E+00

Figure 59. Destination vs Destination differential miRNA expression network and functional enrichment clusters. Shows miRNAs (large dots) and genes (small dots) and whether they are upregulated (blue) or down-regulated (pink). All miRNA shown had a p<0.05 significance to its differential expression after fold-change effect size shrinking in deseq2.

3.2.7. Expression Patterns

The most substantial expression pattern difference between sample groups could be found in the Destination vs Destination test (see Figure 102). Both miRNA-Seq and RNA-Seq showed a similar relative set of significant p-value counts between tests. The primary difference between test significance could be found in the Origin vs Origin (2) test, where miRNAs showed a much stronger signature relative to the Destination (3) test. This is also displayed in deseq2 outputs (see Figures 103-105).

Methylation sample differences were consistently very large (Figures 105 and 111), and a roughly equal number of *p*-significant changes were found in each of the three 2v2 tests. However, the set of genes changed between in test (1) Static vs Change, were found to generate functional cluster fold enrichment scores higher than any other gene list found for that test, or any other methylation test gene list. This difference is displayed in Figure 102 (DAVID Functional Clustering). The samples suffer from low replication and high inter-sample difference which suggests a large amount of noise of either a biological or methodological origin. Despite this, the cohort of consistent changes in gene-body methylation between the distantly and locally transplanted worms occurred with a functional specificity which should not be ignored.

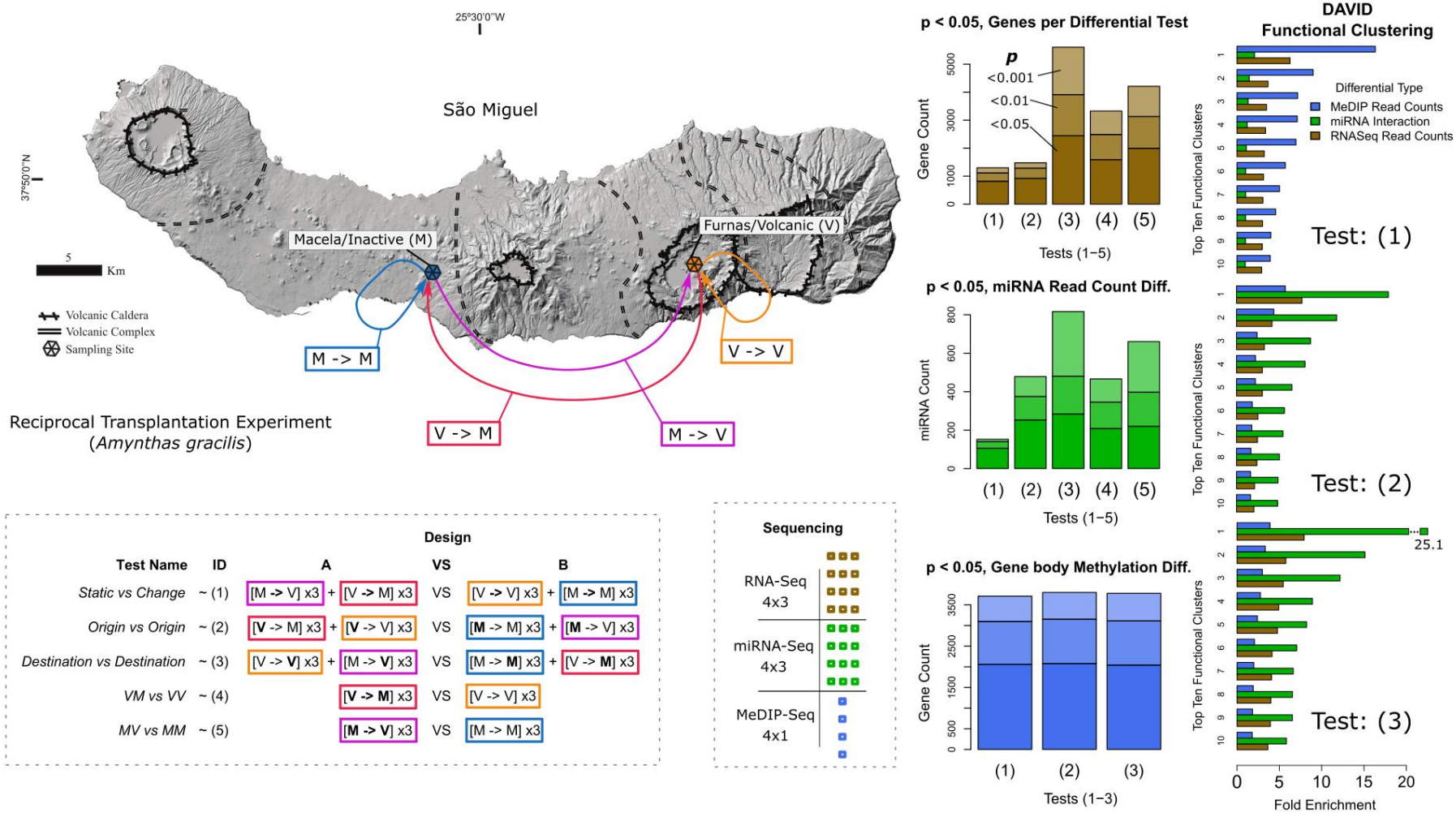


Figure 60. Experimental design and sequencing differential results from left to right (1) Sao Miguel Sampling locations, and differential test design, (2) p-value counts from differential tests described in (1), and (3) Functional enrichment clustering via GO Biological processes: fold enrichment for first three tests.

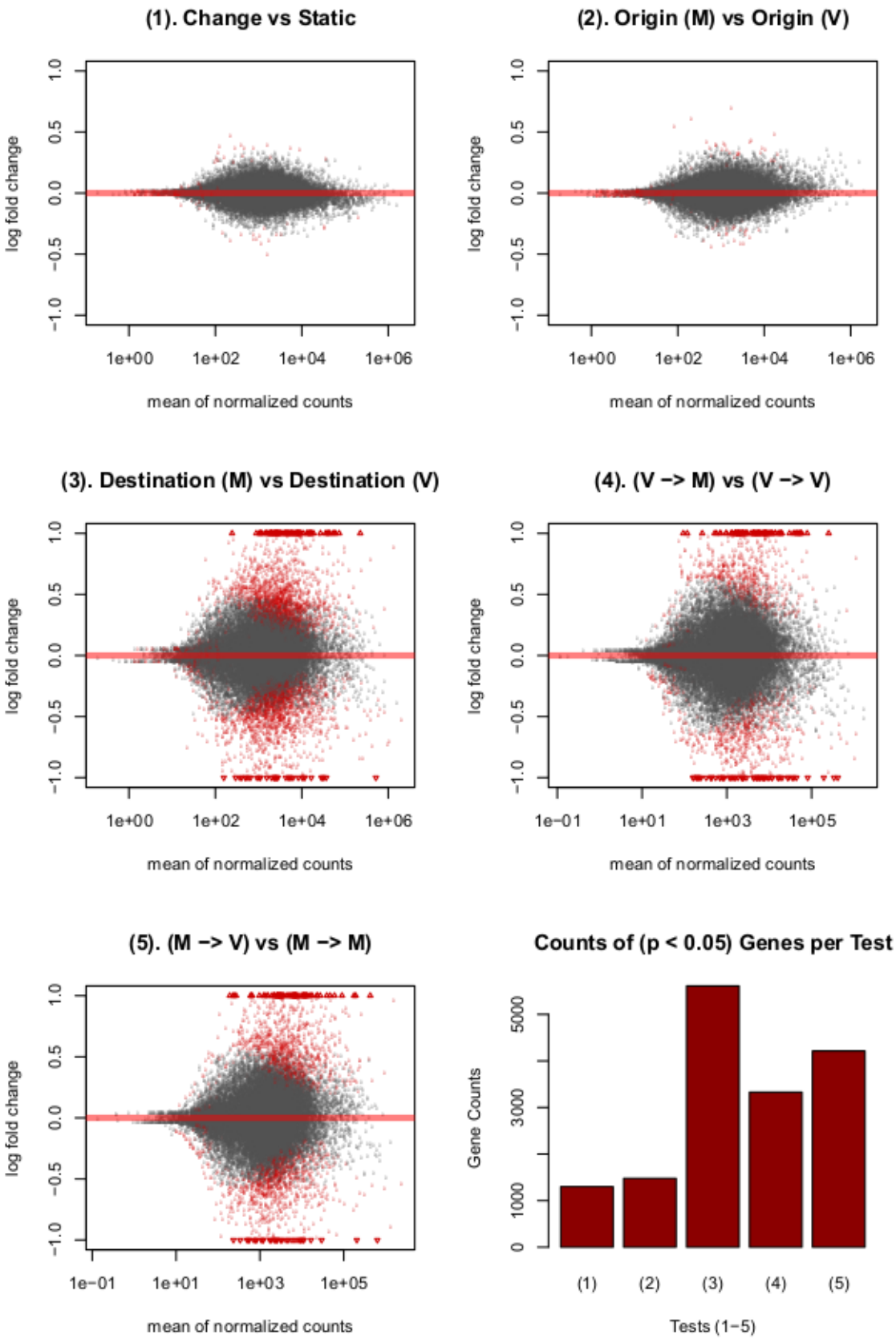


Figure 61. Differential Expression Test Results, Effect-size shrunk Log2 FC against Normalised Sample Means. Expands the RNA-Seq results in Figure 102.

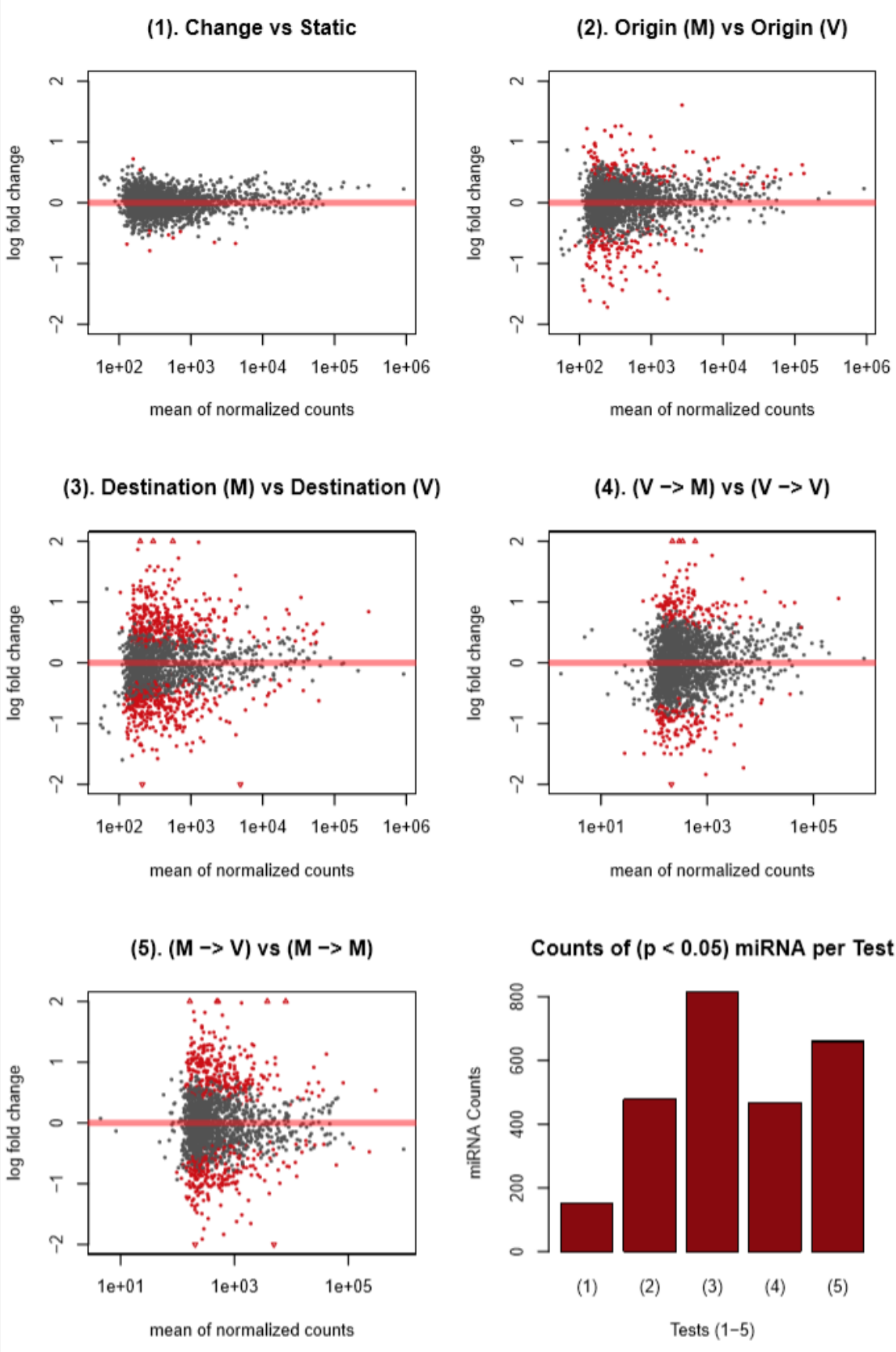


Figure 62. Differential Expression Test Results, Effect-size shrunk Log2 FC against Normalised Sample Means. Expands the miRNA-Seq results in Figure 102.

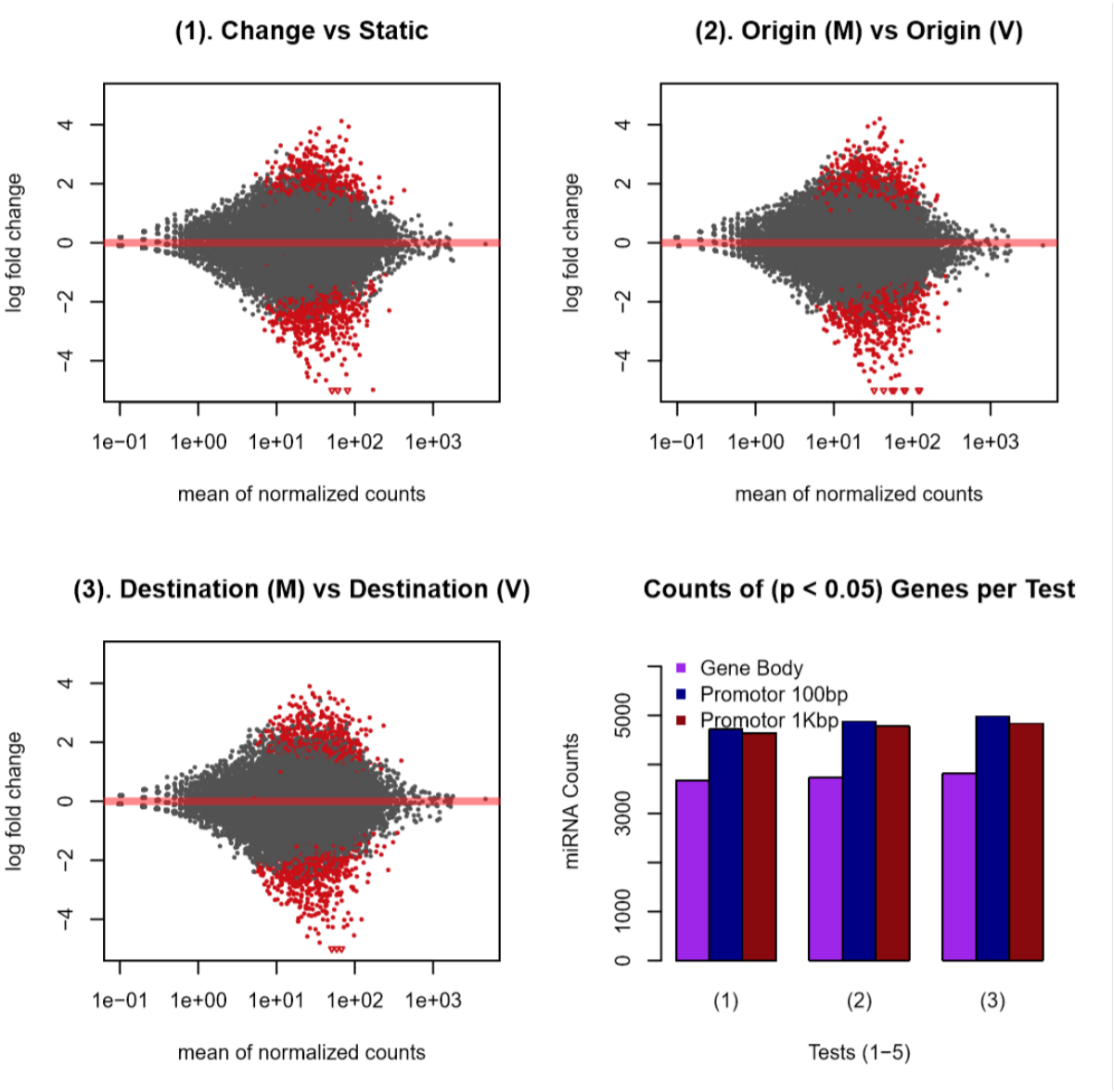


Figure 63. Differential Expression Test Results, Effect-size shrunk Log2 FC against Normalised Sample Means. Expands the MeDIP-Seq results in Figure 102.

3.2.8. Functional Profiles of Plasticity

Enrichment cluster intersection maps in the right halves of Figures 106-108 show the most significant terms per data source. The network view allows direct comparison of source data enrichment effects on identical terms. All the top terms by significance in each graph are members of a cluster which intersects with at least one other data source's enrichment cluster. Some terms and their respective clusters originate from only a single data source; however, these instances are all amongst the least enriched/significant. These clusters intersections are labelled with respect to the lowest *p*-value terms in the clusters. Comparing between Figures 106-108, the same intersections organically re-emerge, however the data sources, enrichments and significances all vary substantially. The top re-emergent cluster intersections are shown in Figure 109 as a trait matrix. The merged clusters summarise the five main functional change categories between the test individuals:

1. Circulatory system Development (GO: 0072359)
2. Epithelial Development (GO: 0060429)
3. Ion Transport (GO:0006811)
4. Neuron Development (GO: 0048666)
5. Signal Transduction (GO: 0007165)

The net up/down regulation of genes annotated with this terms, or associated miRNA up/down regulation, are shown in the trait matrix also. Notably, there is a consistent pattern whereby more genes with significantly up-regulated miRNAs were found in the first four categories for the Destination test. The same categories in the Origin test showed a net down-regulated miRNA-gene component. Reflecting the functional enrichment differences in the MeDIP-Seq tests, significant differential methylation was found to be the most abundant in the Static vs Change test for categories 3-5; in these cases, affecting substantially more genes than any other data source. Methylation significant change rates for categories 1-2 were consistent, suggesting that these level may be more reflective of a certain degree of systemic stochasticity.

Categories 2-5 in the origin test were notable for the miRNA effect size being larger than the RNA-Seq effect size, while category 1 was showed an atypically larger RNA-Seq effect for the Origin test.

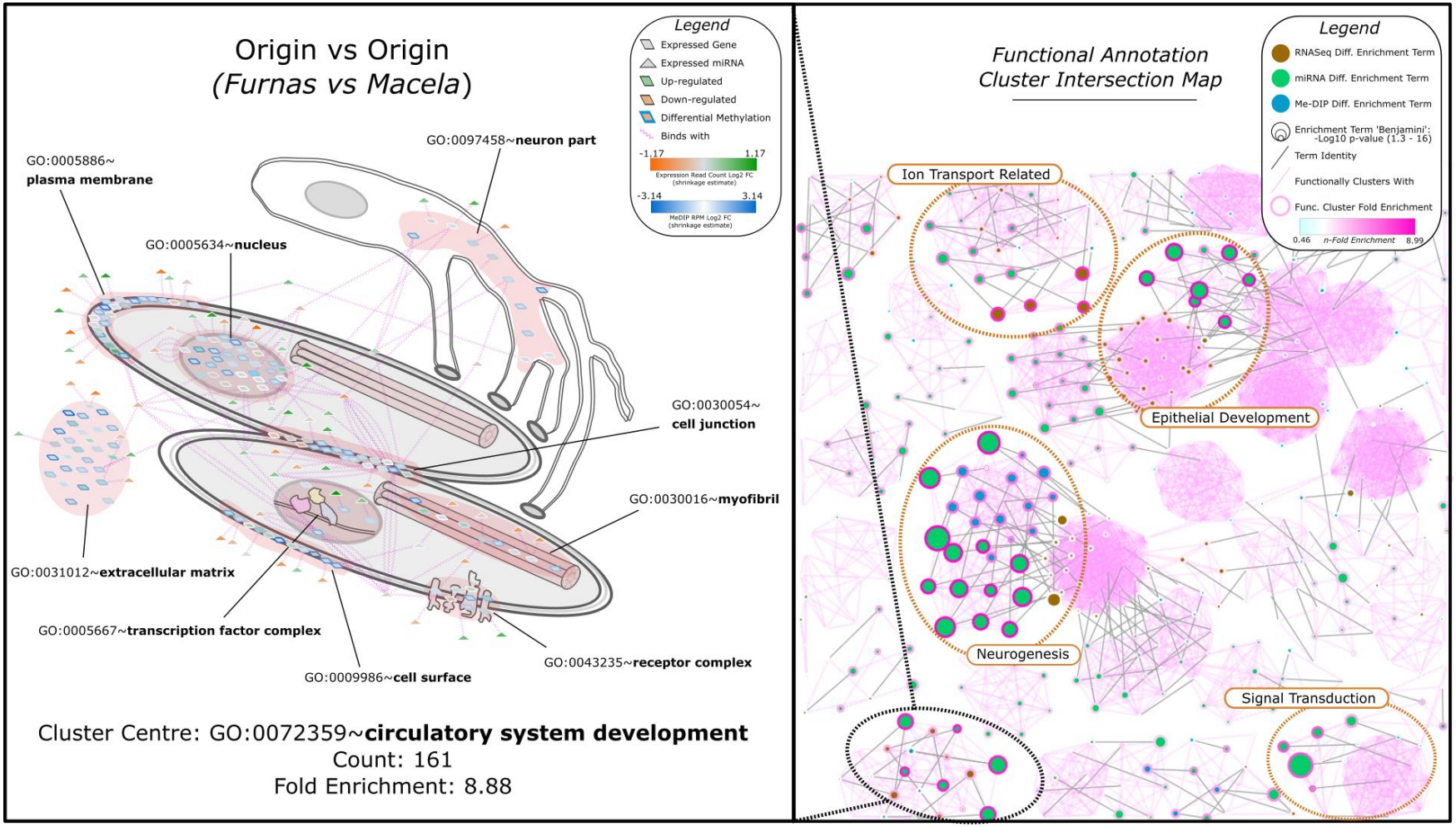


Figure 64. Origin vs Origin, Functional enrichment cluster intersection between data sources (left) and network-view expansion of merged clusters for a single intersection (right). This image shows the miRNA regulatory network for the genes annotation with the term 'GO:0072359~circulatory system development', the network is spatially organised around the GO cellular component terms annotating the cluster, and colours nodes my fold-change in the relevant test.

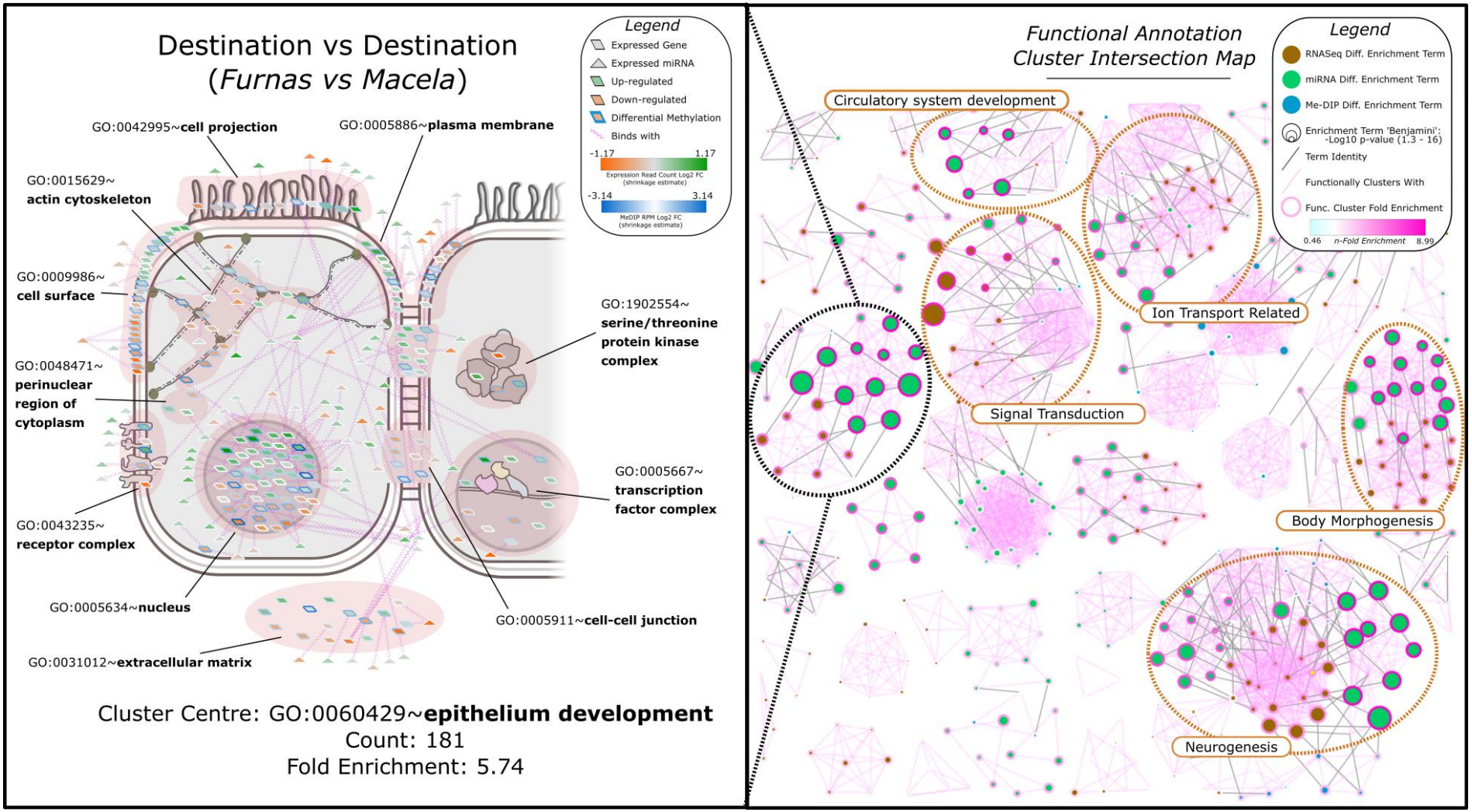


Figure 65. Destination vs Destination, Functional enrichment cluster intersection between data sources (left) and network-view expansion of merged clusters for a single intersection (right). This image shows the miRNA regulatory network for the genes annotated with the term 'GO:0060429~epithelium development', the network is spatially organised around the GO cellular component terms annotating the cluster, and colours nodes my fold-change in the relevant test.

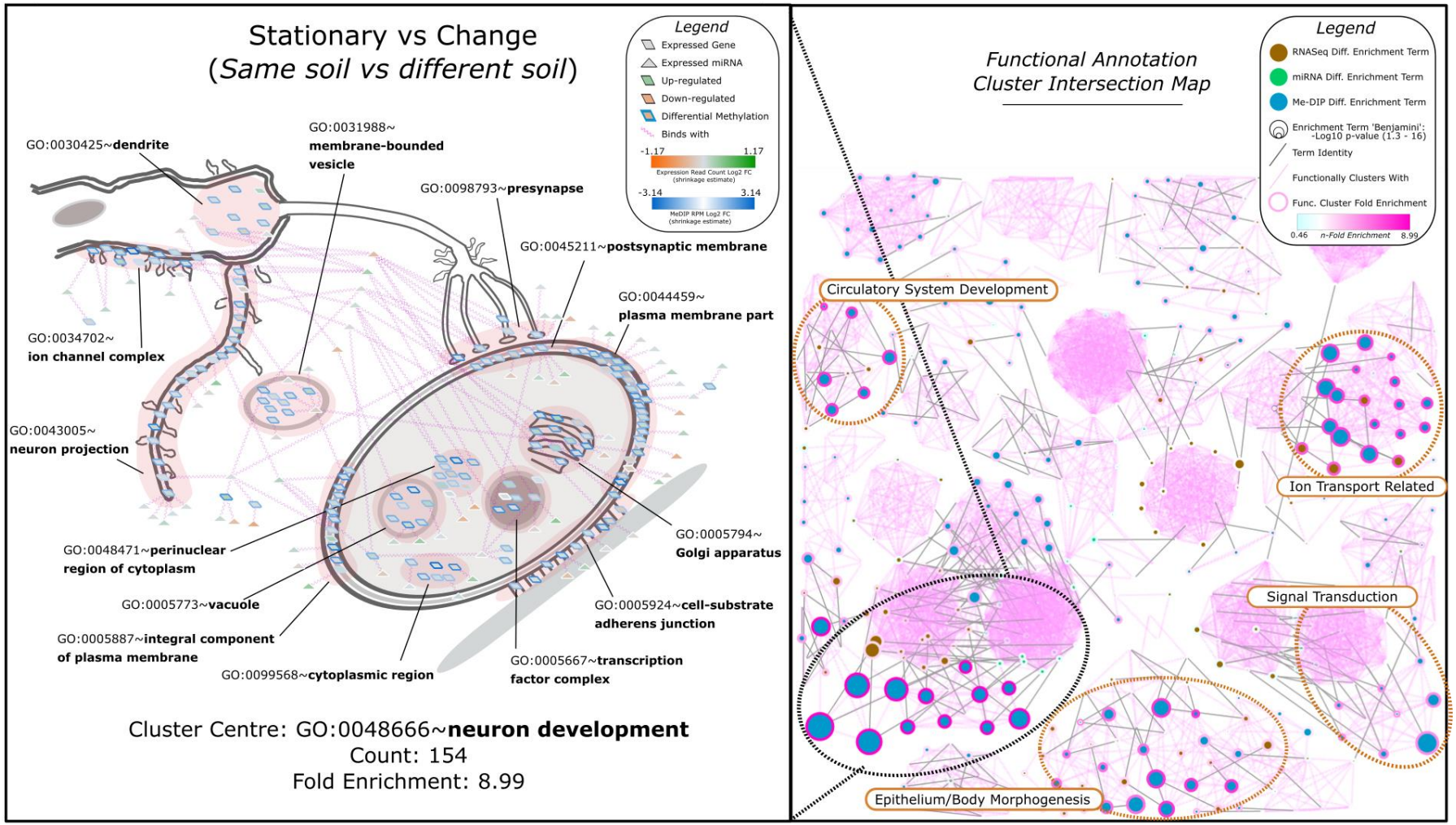


Figure 66. Stationary vs Change, Functional enrichment cluster intersection between data sources (left) and network-view expansion of merged clusters for a single intersection (right). This image shows the miRNA regulatory network for the genes annotated with the term 'GO:0060429~epithelium development', the network is spatially organised around the GO cellular component terms annotating the cluster, and colours nodes my fold-change in the relevant test.

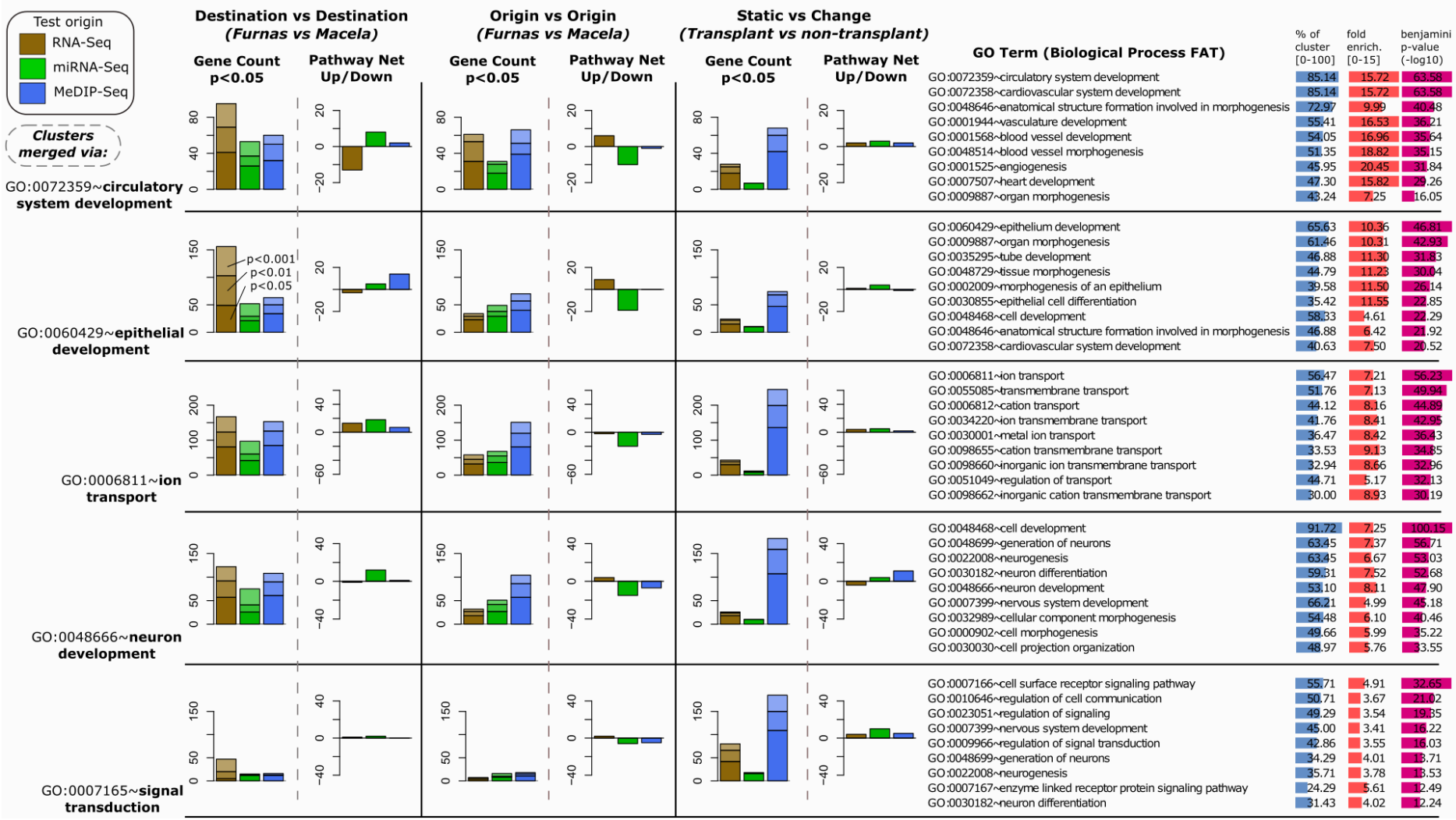


Figure 67. Trait Matrix. (left) Columns describe the largest cluster intersections from Figures 106, 107 and 108. (middle) Bar charts show number of genes affected by p-significant changes per cluster, and the net up/down regulation of that cluster per data source. (right) Re-annotation of cluster sub-terms, top nine terms in cluster by p-significance of enrichment.

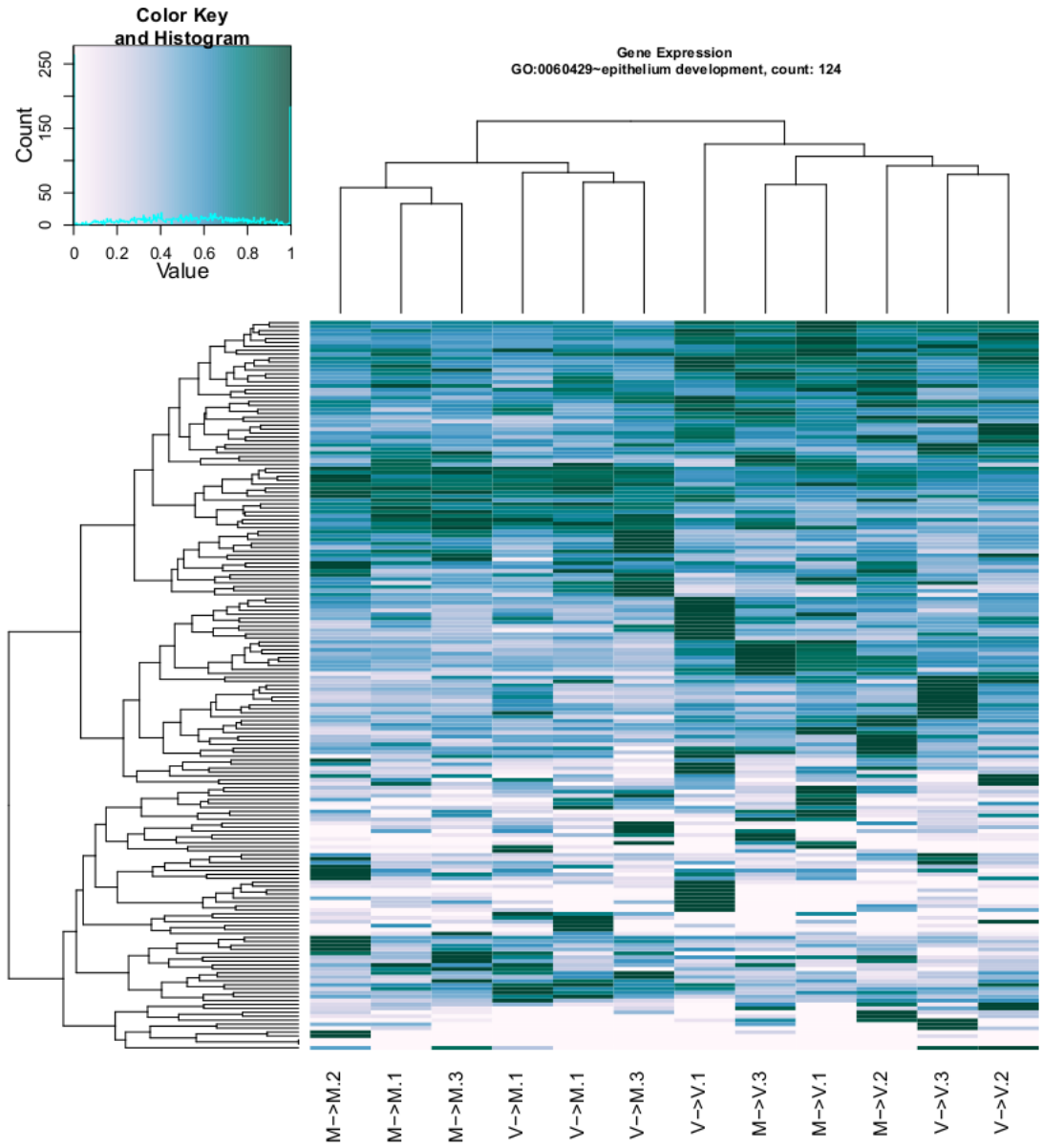


Figure 68. Epithelial Development Heat Map for Transplant Gene Expression.

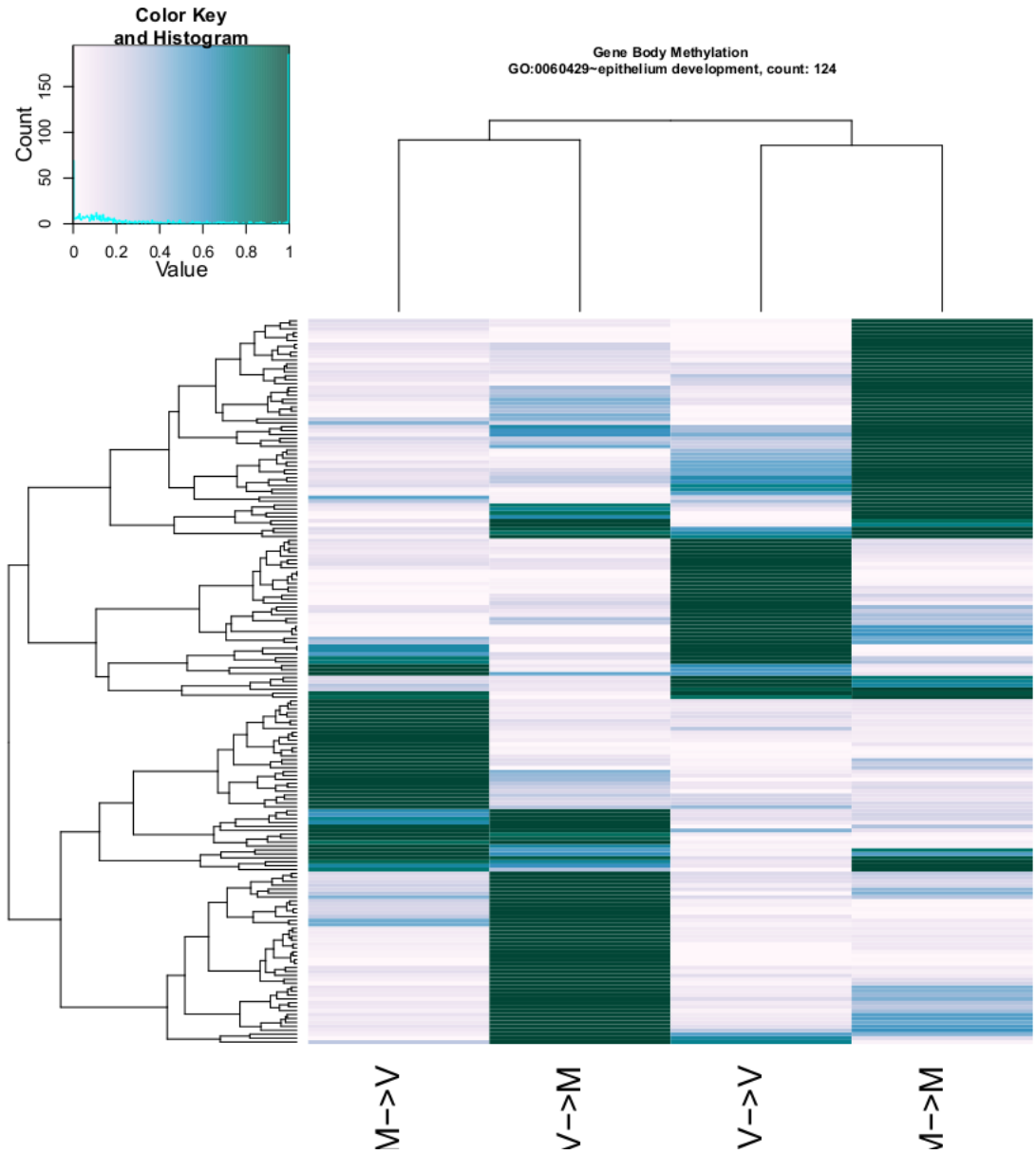


Figure 69. Epithelial Development associated gene heat map for absolute gene body methylation rates.

3.3. Discussion

3.3.1. Genomic Diversity

The genome was found to contain a pervasive and unusual bimodal pattern of allelic divergence. This pattern was not localised to any locus, but was found throughout almost all scaffolds, suggesting that situations such as regional introgression across species barriers are not the cause. There was also no association with gene-density or gene function detected.

The mosaic nature of the genome may be the result of several factors. It appears the multiple introductions of *gracilis* to the Azores could have introduced individuals from distant lineages, which then hybridised. It could also be the case that naturally high diversity was present in the continental

species, given the large effective population size this would not be unprecedented (Small et al. 2007b) and that the local colony dwelling nature of the endogeic worm (Lavelle 1988) regularly introduces low-diversity allelic regions into individual genomes via in-colony in-breeding, however the consistent and evenly distributed presence of variants within even the lower diversity regions of the genome suggests this is not the main cause in this individual. It is also possible that, as the PSMC suggests, the high effective population size of the continental species allowed for the gradual accumulation of a high allelic diversity, which the population bottleneck after introduction to Sao Miguel limited, resulting in a mosaic of allelic regions which reflect the current effective population size (at ~0.5% absolute variation), and the pre-introduction population size (3-4% variation). This final case also requires supposition that insufficient meiotic recombination had occurred since the first introductions of this germ line to homogenize the variant density.

3.3.2. Methylation Spatial Features

Invertebrate methylation studies have indicated, that unlike mammalian biology, DNA methylation regions tend to co-localise with transcribed regions (Suzuki et al. 2007). This is confirmed by these results, which show the majority (95-98%) of genes are associated with methylation to some degree. Although the divergence between invertebrates and vertebrate has been described as a function of the absence of promoter methylation (Keller et al. 2016), the differential models built here suggest that promoter methylation, whilst highly prone to flux, is also amongst the largest relative set-changes in differential tests. Another recent study has suggested that the Pacific Oyster *Crassostrea gigas* may also have a functional role to methylation in promoters given the 5’ bias to its intra-genic methylation (Rivière 2014). Despite the opposite (a 3’ bias) in these models, evidence for promoter methylation activity is still observed in differential tests.

It has also been observed that methylation is associated with splicing via molecular mechanisms which promote exon recognition during transcription (Maunakea et al. 2013), and similar associations have been found in invertebrates (Flores et al. 2012)(Lyko et al. 2010). The overwhelming association of methylation rate abundances with splice junctions in the interval-based gene models strongly suggest that a similar association may exist in the earthworm genome.

Finally, application of the functional sequence signature method developed in Chapter 4 was able to determine than in the case of 3’ and 5’ UTRs, and primary promoters the sites which were methylated also had distinct structural differences with unmethylated sequence. In the case of the UTRs, short motif depletion similar in length to the crucial seed binding region lengths found in the case of miRNA bindings suggest there may be some epigenetic interaction with miRNA regulation.

When comparing the methylation spatial probabilities as in Figure 86, we find that any given 1/20th length division of a 5' UTR is only having a ~0.15 likelihood of being methylated, whilst for 3' UTRs it is closer to ~0.26, despite the summed region probabilities being much higher. However, in Figures 93 and 94, *k*-mer structure fold-changes a between sequence sets groups by methylation, it is actually the lower-rate-group of 5' UTRs by methylation which are almost indistinguishable in sequence structure from the higher-rate group. Whilst the most abundantly methylated 3' UTR only seems to exhibit a sequence correlation amongst the 33% most abundantly methylated subset, with low methylation comparable to no methylation in terms of sequence structural features. This suggest distinct and different regulatory roles played by methylation in these different regions.

3.3.3. Systematic Contributions to Function

The contribution of three large scale regulatory responses will be discussed with respect to each test type in this experiment. Gene expression overall was shown to exhibit the largest response to the Destination soil test versus the origin soil test. This suggests that the response the earthworm produces to these multi-stressor environments is highly specific, and not just a general stress response, this is because worms of the same origin site to the destination must exhibit similar expression patterns to the transplanted worms for the sample-group differences to be discovered as significant.

Methylation was found to be highly variable between samples. The stochasticity in the data suggested that methylation was highly variable in this species. Our results align remarkably with a study titled '*Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease*' (Feinberg & Irizarry 2010), in which the theory that heritable stochastic changes in methylation act as a facilitative mechanism by which members of a population may vary around a mean phenotype. Functional enrichment results found via the accumulation of population VMRs (*variable methylation regions)* in mouse liver tissue revealed generation of Generation of Neurons to by the most enriched term, and Neurite Morphogenesis and Neuron Development were also in the top ten. A similar test performed on human liver tissues found morphogenic and developmental terms highly enriched. Both these results align closely with the functional traits displayed in Figure 108, such as Neuron Development. Morphogenic terms was also substantially enriched in functional clusters. The similarities found here suggest that, no only might variation in methylated between these samples act as an evolutionary function which varies the population phenotypes around a mean in a heritable way, but that the functional traits upon which it acts may be either conserved or convergent across substantial evolutionary distance. That these enrichments only show up in a substantial way in the Static vs Change tests suggests that an environmental stress

response may have the consequence of triggering the individual methylome stochastic flux within/around certain trait-associated loci.

The test significance rate ratios in miRNA tests are interesting for the relatively large impact of the origin soil effect compared to the ratios found in RNA-Seq tests. It is also the case within the miRNA network that some singular miRNAs are capable of binding with 50-100 mRNAs each. Similar situations are reported in mammalian transcriptomics, where most mRNAs are bound by a miRNA (Friedman et al. 2009). This differences suggests that a 'functional memory' of the soil of origin may persist in the transplanted individuals. In the case of the dominant five trait categories this is partially described by a small net loss of miRNA expression in the worms of Furnas soil origin. These molecules continue to act as the post-transcriptional sculptors of gene expression this may serve as a phenotypic 'buffer' which prevents an organism from fully acclimating, and biologically over-committing to what may only be temporary environmental fluctuations. The large scale effect of miRNA regulatory action has been identified to be predominantly repressive (Cai et al. 2009) in many studies (Bartel 2018). This appears to be reflected in at least two trait categories, Epithelial Development and Circulatory system development, where the differences between Destination and Origin tests are an inverse relationship between net miRNA and net mRNA up/down regulation. The general pattern of net upregulation of miRNAs in destination tests, and net down regulation in Origin tests, may suggest that when miRNAs are acting as acute acclimative response intermediaries they are more likely to be actually be acting to suppress more transcriptional targets, whilst their loss of abundance in the Furnas soil suggests a functional memory in the form of a retentive non-repression of adaptive trait-associated gene products.

3.3.4. Physiological acclimation and adaptation

In alignment with previously studies (Cunha et al. 2011b) the epidermal thickness of the earthworm was found to be consistently and substantially thinner in active volcanic soils. This follows observations of the hydrothermal tube worm's morphotype adaptation to life amongst O_2 depleted hydrothermal vents. The change in branchial gas-exchange surface area remains constant between the species' morphotypes, but the diffusion distance across the epithelium shrinks when living amongst the vents (Andersen et al. 2006). The discovery of 181 significantly differentially expressed genes, with a pathway fold-enrichment in the differential set of 5.74, acting upon the epithelium development pathway gives strong evidence that this effect is a specific acclimative response, as opposed to an emergent result of damage/stress. Further re-enforcing the assertion that the thinning is a part of the organism's acclimative toolbox, rather than a heritable population level trait, is the count difference in pathway associated *p*-significant gene expression between the three test types. That the transcript response is incredibly low in the origin test shows that this effect is

deployed is response environmental change, to a far greater degree than it persists in environmental origin. That the transcript and miRNA response significance is incredibly low in the general change test, shows that the pathway-associated genes which either thicken or thin the epidermis (Figure 79) are specific to the environmental change which has occurred, with very little overlap. Differential methylation significance occurs at similar rates between the merged epithelial clusters in Figure 109, however the gene set targeted by methylation in the static test enriches for specific terms far more readily in the network view (Figures 106-108). Despite the stochastic differential methylation effects spread between heterogenous terms within the same merged clusters, the more specific functional profile of the 'static vs change' test methylation changes indicate that it may have a role in the more general change-response mechanism by which both epithelial restructuring pathways are regulated. In crayfish it has been observed that similar gene-body methylation is associated with expression stabilisation of genes limited by chromatin availability (Gatzmann et al. 2018), which may be as a trigger, or as a consequence of chromatin remodelling which takes place (Jeltsch & Jurkowska 2014).

Earthworms have a closed haemoglobin based circulatory system for gas exchange (Monahan-Earley et al. 2013), and an open circulatory hydrostatic skeleton for locomotion called the coelom (Rieger & Purschke 2005) (Reiber & McGaw 2009). Given the annotation of the gene-set was produced via the reference uniport database, it may be that some interference exists given the similar evolutionary origins of these two systems. However, the gas diffusion acclimation performed by the epithelium suggests that the annotation associated a differential morphogenesis of vasculature in the worms was also a compensatory mechanism substantially altered environmental O_2/CO_2 diffusion gradients. The merged cluster annotation shown Figure 109, and Figure 106 suggests both angiogenesis and morphogenic restructuring takes place. This trait is also unique in its test change profile for the abundances of *p*-significant RNA-Seq differential results between the origin and destination tests. This suggests that the circulatory system restructuring occur as an acclimative response, but the expression profile changes either occur on a substantially longer time frame than 31 days or are a result of a population level adaptive variation. Again, methylation acts as a stochastic variation on these genes consistently, but and the functional enrichment is slightly more significant in the static vs change test.

Neuron development in the earthworm is the third morphogenic trait category which is highly enriched in all three test types, although in different ways. This will be discussed in conjunction with signal transduction annotation cluster, which overlaps considerably via its constituent genes. Signal transduction is a very broad category, and is difficult to pick apart, with the exception that neurogenesis associated terms always appear to make a substantial constituent of the signal transduction subset. It is has also been shown than invertebrates such as *D. melanogaster* possess

dedicated O_2 and CO_2 olfactory signalling pathways (Luo et al. 2009), suggesting that similar neural or olfactory chemical signalling may be acting as a triggering mechanism for the morphometric changes discussed above. Hypoxia, an expected stressor given a means of ~15 and 10% O_2 gas composition at depths of 25 and 50cm (Table 12), has also been shown to have substantial effects on neural stem cell differentiation in many organisms. Research has shown neuronal migration defects and axon pathfinding changes in *C. elegans* (Chang & Bargmann 2008), reduction of ion concentrations and consequent hyper-polarization in *D. melanogaster* (Gu & Haddad 1999), and cation co-transporter activity reduction and resultant hyper-polarization in *Lymnaea stagnalis* (Silverman-Gavrila et al. 2009). Most of the literature concerning neural responses to hypoxia in invertebrates specifically concerns the membrane-based changes, with little consideration of differentiation and growth alterations (Mannello et al. 2011). This is reflected in Figure 106 where the abundance of membrane-bound proteins in the miRNA-regulation network is displayed. However, there are also substantial numbers neurogenic genes identified here which may be acting as a more general plasticity response. Signal transduction pathways in general are overwhelmingly differentially methylated in the static vs change test compared to any other test, and the neural development specific pathways are also nearly double the rate of differential methylation in this test compared to the others. This suggests that the 'general' environmental change response exhibited by these earthworms may constitute the epigenetic modification of many of the same signalling pathways regardless of the specific changes encountered. Uniquely, the signal transduction merged cluster also showed more RNA-Seq differential significance in the static vs change test, further indicating that this more general change response, at the epigenetic level has a corresponding gene expression profile too.

Ionic transporters also possess some functional overlap with signalling pathways and are constituted by upwards of 200 genes significantly modified by their expression, miRNA binding or MeDIP read mapping levels. Of the genes in the merged annotation cluster 36.4% were associated with metal ion transport, a primary component of the identified environmental profile differentials (Table 14). Metallothioneins describe a protein family which earthworms are known to utilise to handle heavy metal stress (Höckner et al. 2015). Originating in the Golgi apparatus, these proteins are expected to be found in the soluble fraction of the earthworm's physical mass. By referring to Table 15, the soluble fraction metal abundances show that cadmium, zinc, copper, potassium and magnesium are all likely regulated as soluble protein bound ions in some form. However, many others, including one of the primary environmental differentials, lead, is stored in the cells of the organisms as granules with an extremely positive destination effect. Although the destination effect is found to be more prevalent in the fractionation table, in reflection of the relatively consistent set of miRNA

interactions between the destination and origin tests, there are also origin effects to be found in the fractionation table, for example titanium and strontium accumulations are more dependent on the origin soil, which could be a consequence of their relative non-reactivity with biological systems (Saini 2015) (Pors Nielsen 2004). The higher toxicity chromium (Sivakumar & Subbhuraam 2005) content also exhibits an origin effect, although the abundances were relatively low. Fractionation also indicates that the only metal which the earthworm failed to adequately regulate either by metallothionein solvency or by storage as granules was arsenic, a known toxin, which also had a slight origin effect. Methylation differentials for ion transporters were far higher in the 'static vs change' test also, suggesting a general large scale epigenetic regulatory response type amongst ionic transporter genes in the event of environmental fluctuation.

3.4. Conclusion

A reciprocal transplantation experiment was performed, exchanging groups of earthworms between inactive and active volcanic soils with elevated temperatures, CO_2 degasification, O_2 depletion, and altered chemical and metal abundance profiles. RNA-Seq, miRNA-Seq, and MeDIP-Seq profiles of sample functional changes were created. Three main differential tests were used to assess the acclimative, adaptive and general environmental responses in this earthworm, see Figure 102. *Amynthas gracilis* was found to have highly methylated gene-bodies, with variable gene-component rates, and a clear relationship with between mean expression levels and methylation, see Figures 86, 89 and 99. Independent functional enrichments of significant gene-based differentials generated by the three data sources were intersected to show the pathway contributions to plastic traits via different regulatory mechanisms, see Figures 106-108. Epithelial remodelling was described physically and independently re-discovered as a functional signature in multiple enrichment tests, see Figure 37, its contributory systemic profile suggests this is highly plastic acclimative response. Circulatory system morphogenesis and angiogenesis were repeatedly independently discovered in functional clusters (see Figure 106) and profiles were found to be both acclimative and subject to a soil-origin persistence effect. Neuron development was performed acclimative but was epigenetically modified to a far greater extent in worms experiencing environmental change regardless of the type of change. Signal transduction overall exhibited an even stronger methylation response in the general change test and was also subject to more gene expression profile changes in this case. Metal accumulation in body fractionated body matter was shown to exhibit both a large destination effect and a small origin effect, metal ion transporters were also an independently functionally enriched category by all three data sources, and in all three test types, although the acclimative response was the strongest. Methylation change in the earthworm genome was found to be incredibly noisy, with only the general change test showing large numbers of clear functional

enrichments, suggesting a degree stochasticity to the epigenetic mechanism's relationship with functional plasticity. miRNA networks showed a much higher relative soil origin effect profile, relative to the soil destination effect, than gene expression, suggesting expression sculpting via repression networks may act as a persistent functional memory within individuals exposed to environmental flux.

4. Chapter 4: Towards a high-utility general signature for sequence structure

4.0. Motivation

Detecting the presence of systematic differences in genomic evolution is a difficult problem, particularly for novel organisms. In order to build an understanding of the sequence contained within a genome, a transcriptome, or a proteome – it is necessary to annotate and compare elements contained therein, typically through the lens of prior knowledge. However, an accurate annotation for a novel (to genomics) organism's genes/proteins is almost impossible to achieve, with the available annotation solutions ranging from 26-80% to provide *some* level of annotation for genes of interest (Bolger et al. 2018). Various factors are measurable, such as the ratio between synonymous and nonsynonymous mutations – absolute divergence of alleles – but only if the organism is sufficiently low in allelic diversity that copies of its alleles can be collapsed into a stable reference. If the rate of the divergence between the alleles is highly variable, as was the case in Chapter 2, then the problem of building accurate references becomes harder: at some point there needs to be a separation between haploid and diploid reference sequence, and a unity of measurement between them. Additionally, most variant-calling pipelines require that reads be mappable by short-read aligners from one allele to the other (Poplin et al. 2017). As was the case with *L. anatina*'s genome, this does not always work particularly well for divergent alleles, leaving the variant annotations sparse and unreliable. In order to be able to take the sequence content as a whole and produce singular measurements of the information structures present therein, it was necessary to derive a knowledge-free approach to the problem. Most knowledge free analysis of large-scale sequence data in bioinformatics focuses on the use of k-mers, or measurements of information complexity (Zielezinski et al. 2017). Measurements are made of *k*-mer abundance, unique counts, and frequency distributions over one or more sizes. Measurements can also be made of Kolgomorov complexity or Shannon entropy, and used to compare sequence data. Taking this form of initiative as the starting point, a mathematical and algorithmic approach to describing sequence structure in a knowledge-free manner was developed.

4.1. Introduction

'*K*-mer' is a term typically used to describe the set of fixed length substrings found within a larger string. In recent years *k*-mer based analysis, is used widely to perform QA/QC on NGS data (Andrews 2014), to estimate pre-assembly statistics for genomes (Simpson et al. 2009), to build predictors for sequence associated biological features (Liu et al. 2015), and even to taxonomically classify the content of metagenomes (Ounit et al. 2015). K-mers may also sometimes be referred to as n-grams,

or in the case whereby the length may not be fixed: *l*-mers. For the sake of clarity, in this chapter the following terminological code will be followed: *k*-mer will be used to describe the fixed length substrings which constitute the maximum length inputs to a substring-based method, while *l*-mer will be used to describe instances of substring usage over the range of [1, *k*] within those methods.

A primary problematic issue with *k* or *l*-mer based methods for classification is the high dimensionality of longer DNA sequence. For a *k*-mer of *n* length, the number of available sequence types is 4^n, when predictive sequences reach 15-20 base pairs, it typically becomes necessary for the sake of computing power to develop heuristics which limit the sequence space explored by the classifier. Various programs have also been developed to optimise the containment of high dimensionality in working memory for the sake of *k*-mer counting (Marcais & Kingsford 2012) (a routine operation in various other pipelines, such as the Trinity Transcriptome assembler (Brian J Haas et al. 2013)).

Another issue with k or I-mer based approaches to DNA sequence computing, particularly with respect to machine learning, is the fragility of longer *k*-mers. Most modern machine learning methods rely on inputs of fixed dimensionality and size, which results in DNA classifiers using kernel matrices of *l*-mer frequencies derived from a training set of sequences. Although amongst the most highly predictive sequences, longer *k*-mers are also incredibly sparse entries in kernel matrices, which make models derived from them difficult to train. In response to this shortcoming, work has recently been done to attempt to bandage this issue using a gapped *k*-mer approach to kernel matrix construction for support vector machine (SVM) classifiers (Ghandi et al. 2014).

Ghandi *et al's* algorithmic method involves the construction an efficient tree-like data-structure with additional branching between nodes which differ by N bases, this may allow the aggregation of many similar long *k*-mers into a single entry in a kernel matrix, which can produce a more reliable input to an SVM (Ghandi et al. 2014).

The idea of a gapped *k*-mer tree will be central to the foundations of the method described here. However, there are several other categories of biological sequence processing which inform the development of this method. The first, as mentioned above, are the counting and statistics tools used in the data processing pipelines for many NGS experiments. Work, although limited in scope, has been done to apply these tools to derive an informative bird's eye view of an organism's biology. Most straightforwardly, this has been done by calculating whole-genome *k*-mer frequency histograms as a comparison tool between species (Chor et al. 2009). Another way in which these tools may directly inform us biologically include allelic diversity estimation (Simpson 2014), although a *k*-mer based estimation of heterozygosity will lose sensitivity when the density of genomic SNPs is

regularly greater than *1/k*, or when the overall rate is exceptionally small. A third way might be for the preliminary detection of intracellular parasites or other sources of non-host DNA present in the sequencing experiment without direct classification (Kumar et al. 2013). While useful, these tools also have a relatively low-dimensional output relative to their inputs, often taking the form of a two or three-dimensional distribution of frequencies. The notion that a large-scale sequence set might be described biologically in a knowledge-free manner seems appealing but achieving much depth to the analysis is challenging.

Research not directly related to the use of *k*-mers in the same manner, which yet still attempts to gain a bird's eye view of the DNA's information content comes often from an 'information entropy' perspective. Information Theory developed by Claude Shannon (Shannon 1948) has been the basis for much entropic theory of information and is referred to as Shannon Entropy (Lin 1991). The methods developed around which are principally concerned with the nature of DNA insofar as it diverges from a random noise comprised of the same alphabet (Mantegna et al. 1994). Attempts have been made to describe an information entropic 'signature' of DNA (Schmitt & Herzel 1997). Others have also found novel approaches to the idea of entropy, such as via 'Chaos Game Representation' (CGR) (Oliver et al. 1993). Purely entropic or signature-based descriptions of DNA do not appear to be in frequent use in the age of NGS. There has been some perennial interest in CGR signatures however, efforts have been made to deploy these for the comparison of genomes between species (Karamichalis et al. 2016). Euclidian distances of CGR matrices have also been proposed as a quantified measure of species-distance (Karamichalis et al. 2015). Although the perspective of defining sequences, and even life, by the scale and shape of their entropic properties might capture the signatures of far deeper complexity, the outputs produced by these methods are difficult to translate into stand-alone biological insight in the same way that a whole-genome *k*-mer analysis might be. The objective of this research effort is to determine whether it might be possible to achieve the best of both worlds: deeper complexity signatures containing direct biological insights.

4.2. Methodology Development

4.2.1. Rationale

It is hard to spend much time as a bioinformatician in the modern day without being required to 'choose a value of *k*' for a program. Although some assemblers such as MEGAHIT (D. Li et al. 2015) may by default opt to run multiple *k* values in serial, whether error correction, genome assembly, or read library pre-analysis, typically a single value of *k* is required. This highlights the difficulty of integrating *k*-mer based algorithms across multiple *k* values simultaneously. Consider that, in an

alphabet of size four, ATGC, there may exist one to eight different 9-mers set for every 8-mer. If a given 8-mer's frequency could be explained by the frequency of a single 9-mer, it would be natural to point to the 9-mer as the sequence of interest if it was constitutive of multiple roughly equally frequent different 10-mers. This is quite a simple way of looking at the set of *k*-mers in an entropic manner: If the frequencies of shorter substrings disperse evenly amongst the longer substrings which contain them, it is probably the shorter substrings which carry the biological interest. If the presence of these shorter substrings at an unusually high frequency rate is explained by equivalently high frequency longer substrings which contain them, then perhaps it is the longer that are the more relevant to whatever biological question is being asked. The next step might then be to consider if, for a general methodology which is inclusive of a *range of k (or l)*, rather than selecting *l*-mers by their interestingness or (in the terminology which will be used from here on) distinctness, all *l*-mers over *[1, k]* might be included in the set, but their merit be subject to a 'distinctness weighting'.

To assemble a large set of substring information in such a manner as would allow us to ask this question of an arbitrary *l*-mer, the most basic computational requirement is access to the frequency-containing variable associated with an *l*-mer, and a set of associations between it and the frequency variables of the length *l+1* substrings which may contain it. Fortunately, this condition is satisfied by the widely used efficient *k*-mer tree structure. This is essentially a search-tree with *n* possible children per node, where *n* is the size of the alphabet (In this case, four). From now on the rationale will assume the employment of a *k*-mer tree as its primary data structure. Technically speaking the *k*-mer tree would be defined as a 'trie', rather than a tree, as the actual sequence content of the *k*-mer is not stored in any variable and instead may be inferred from the tree position of a given node. Despite this, since the tree, or trie, is not actually being used for search operations, we will continue to refer to it simply as a '*k*-mer tree'.

4.2.2. Initial Formalisation

Given the parent/child relationship between characters within a set of *k*-mers, and the usage of frequency dispersion to measure distinctness, we can begin to define the formulae employed. Given that child node frequencies are contained by an ascending-value-ordered n-tuple **F =** $(f_1, f_2, \ldots f_n)$.

Formula 1: $$d_{min} = f_p$$

Formula 2: $$d_{max} = \left\lfloor \frac{d_{min}}{n} \right\rfloor n^2 + (d_{min} \bmod n)^2$$

Formula 3: $$d_{child} = \sum_{i=1}^{n-1} (\boldsymbol{F}_i - \boldsymbol{F}_{i+1}) i^2 + \boldsymbol{F}_n{}^2$$

Formula 4: $$D = \frac{d_{child} - d_{min}}{d_{max} - d_{min}}$$

Giving:

Formula 5: $$D \in \mathbb{R}\ (0 \leq D \leq 1)$$

Formulae 1-4 describes the *l*-mer distinctness found for the node described as the parent in this context. Here f_p and f_c refer to the parent and child node frequencies respectively, *n* describes the length of the alphabet, and *d* the various frequency distribution scores. The vector of child node frequencies is also pre-sorted from low to high. The distinctness D thus measures in a linear fashion the distance in frequency distributions between the least distributed state (one child node equals parent node frequency), and the maximally distributed case (child nodes divide the parent node frequency in the most even manner possible given the potential remainder). This linear measurement of distribution equality within a set of values functions as a type of Gini coefficient (Gini 1912), for indivisible integers.

We must first note however an important aspect to the 'distinctness' weight calculation here when using trees over a contiguous range of *l*. Distinct *l*-mers will have evenly distributed child-node frequencies, yet so will even totally indistinct *l*-mers in a tree at a depth shallow enough to be saturated by the input set. This is to say that a null case random 'DNA-noise' input would cause this method to identify many distinct *l*-mers in the tree where $l < log_4(F_r)$, with F_r being the root node frequency (the number of input *k*-mers). To remedy this, we might return to the entropic way of thinking. Simply put, it is not just that structure breaks down at a certain point below an *l*-mer branch, but that it also did *not* do so beforehand. Phrased differently, we could say that a given high frequency branch of the tree ought to have shown some resistance to the expected noise-case dispersion above the depth being considered if its own dispersion of frequency is to be indicative of actual structure. In fact, if the same formulae were applied to both cases, a solution could be to multiply the distinctness of a node at *l* by the inverse distinctness of its immediate parent at *l-1.* To avoid confusion, we might separate distinctness into D_b: 'base' and D_a 'actual'. Such that:

Formula 6: $$D_a = D_b(1 - D_{b_{parent}})$$

To find a sum of all *l*-mers which escape the entropy of noise, it ought to be enough to perform the above on every node over the range *[1, k-1]*. We might also optionally multiply *Da* by the length of the *l*-mer, *l*, to scale the measurement by the sizes of the retained structures, and/or we might multiply by the node frequency. A combination of these terms from now will be referred to as a resistant structure score. Generalising slightly from the range of *all* nodes, we can observe that is would be possible to find the resistant structure score, *S,* of any sub-tree recursively with respect to its root *r,* using a depth-first-search (DFS). See Formula 6. In the case of finding a singular

quantification of the scale of the entropic-resistance in the genome, the root *r* would be the actual head-node in the tree.

Formula 7:

$$S_r = \sum_{v \in Ch(r)} Da_v * f_v * l_v$$

Here *v* represents a node (or vertex), and *Ch* the recursive application of the summation function to its children.

One aspect of the dispersive tree measurements process which has not yet been addressed is the directionality of the tree. As discussed in the Rationale set down in Section 2.1, there are eight, not four, DNA (*l+1)*-mers which may contain any given *l*-mer, however the tree structure accounts for the terminal extension bases. Since the tree expands by powers of four (in the case of DNA), the depth at which the tree becomes less saturated, and more informative will only be increased by an average of 0.25 by doubling the frequency. This means there is perhaps enough wiggle room to merely read all inputs twice: once forwards, once backwards.

However, this issue also intersects with the strandedness of DNA, which contains one forward and one reverse complementry sequence. A simple solution could be to capture the other four base extensions in the form of reverse complemented *k*-mer inputs, this would also have the effect of unifying motifs that have been sequenced on multiple occasions from different strands, thus separating their frequencies, despite their biological identity. For protein sequences however, a simple reversal would suffice.

There is another slightly counter intuitive aspect to this calculation which also requires attention. The statistical means taken of any categorical vector of structure scores will always resolve at their current depth, with respect to the non-dispersed structures of higher values of *l*. This is to say that a high frequency 20-mer which shares a constitutive 8-mer with another low frequency 20-mer will cause a relatively low distinctness score for the 8-mer at the point of separation. This effect lowers the mean for the scores at the 8-mer depth. When the whole tree is summarised however, so long as it is deeper than 20 in this case, the higher distinctness of 20-mers and it's the multiplication by *l*, will yield an overall higher structure score for the entire tree. If, however the tree does not extend to that depth, the unregistered frequencies that have 'escaped' will have the effect of incorrectly lowering the structure-score.

This is a boundary problem – the tree cannot be infinitely deep, in fact computational constraints limit its size quite significantly, and all frequencies cannot be guaranteed to disperse within it. As a

result, spectra which cannot be captured by the tree ought, in the case of aggregation methods *within k*, be negated. This involves simply reducing all leaf node branch frequencies to 1 and propagating those subtractions recursively up the tree to maintain equivalent sum frequencies per depth. If the tree were to be used for non-aggregative methods (i.e. motif discovery) this would not be required, it is also not necessarily the case that this correction be required in the case of signature generation. The boundary frequency correction will thus be applied only to the more compact aggregate matrices.

Next, whilst the above may suffice to inform us of a certain property of the strings in the input set, we also must return to the biological manifestation of the *k*-mer, principally, to return to the classification issue: the biological fragility of long substrings. Not all bases in a string may be constitutive of the active motif. There may also have been duplications of motifs which then experienced mutations, none of these aspects of genomic structure would be detectable by a simple *k*-mer tree as we have described so far. For example, an 8-mer might smoothly distribute its frequencies amongst 9-mers, yet all subsequent substrings up to length 20 may continue identically, yet they will do so in four separate branches of the tree. In this case, the 20-mer with a single flexible base will not be discovered at its true frequency. However, if one were to introduce an extra character 'N' as a child node through which all input strings reaching its parent node are additionally to be passed, the subtree originating from the 'N' child node would describe accurately the full frequency 20-mer sequence. Figure 29 shows a basic example of how the merging of subtrees occurs to create an effective 'N-mask' in a *k*=4 binary tree.

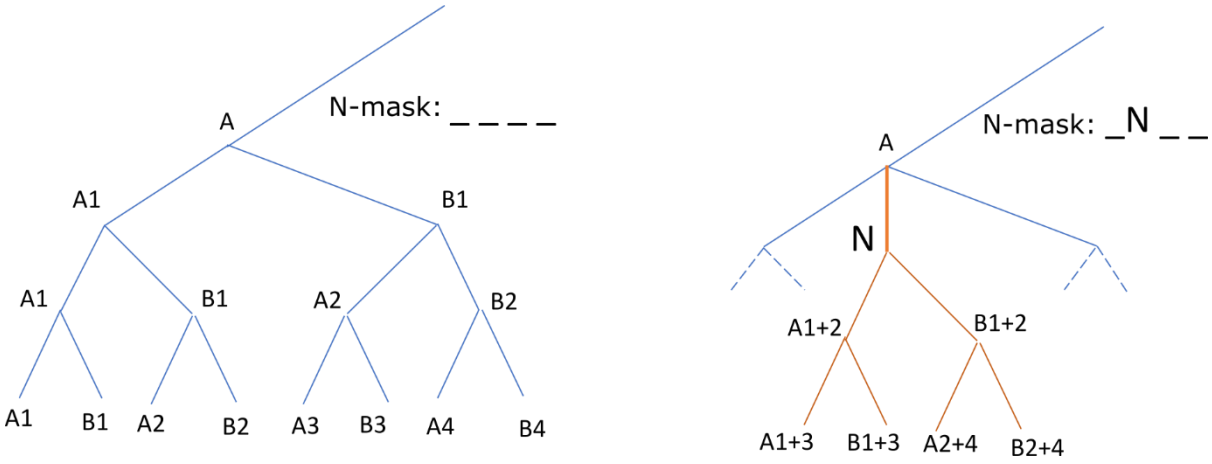


Figure 70. Example of binary tree aggregated for a certain N-mask.

Generalising from this example, we can see than for a single flexible base, an 'N' subtree would then have to be generated for every node in the tree with more than one non-zero frequency child. In the case of multiple 'N'-containing motifs, an 'N' subtree would also have to generated for nodes in the initial 'N'-child subtrees, and so forth. This does however have advantages. Firstly, since each 'N'

subtree is independent from the rest of the tree above its parent, the memory usage can be contained by only generating (and deleting) subtrees as they need to be measured in a single 'depth-first search' (DFS). Secondly, the expansion of computing power and memory usage with additional Ns remains constant when the number of children per node is increased (as we are not at this point investigating transitions vs transversions, or other partially selective evolutionary conditions). As a result, its polynomial efficiency might yet be a suitable trade-off, particularly in the case of larger alphabets, such as with peptide sequences.

4.2.3. The Aggregation Methods for 'N-masked' *l*-mers

Aggregation methods in this case refers to a structured and systematic way that variables can be aggregated from a complex source. The aggregation methods can also be thought of as independent of the variable types gathered. Given that we assess the tree on a per-node basis, an aggregation method could be applied to gather various measurements in the same manner, although at first, we will explore them from the perspective of the development of signatures derived from structure scores.

Let us return temporarily to re-examine what is meant by a 'signature'. It could be said that the signature of an aggregated set of scores is created as much in the process of selective aggregation as it is in the data's original complexity. As in the case of imaging sequence Shannon Entropy (Tenreiro MacHado 2012), or CGR images (Oliver et al. 1993), we can see that the signature is typical displayed as a 2 or 3-dimensional array of points. The case of CGR images used for distance metrics also highlights the importance of comparability between signatures (Karamichalis et al. 2016). This is to say that, a signature ought to retain the same dimensions and size regardless of the input data. When aggregating scores from the tree therefore, we ought to construct the dimensions of the output matrix from sources which can be measured regardless of the sparsity of the tree.

The first dimension seems most suitably to be *l,* over the range of *[1, l-1].* The terminal value of *l* cannot have scores data extracted as the calculation involves the node in question to have children with populated frequencies (*i.e.* it cannot be leaf node). All *k*-mers read into the tree are of length *k* and therefore all depths of the unmasked tree will share an equal sum of frequencies. This means that each category of *l* will always reliably contain measurable structure scores. The most basic output summary of the flat tree will thus be a single vector of structure scores of length *l – 1.*

Formula 8:

$$Sig1D = \begin{bmatrix} S_1 \\ \vdots \\ S_{l-1} \end{bmatrix}$$

When choosing the second dimension, we begin to consider the structure of the N-masked tree also. In this case there are multiple options, and there might also be multiple correct answers. For example, it would be of biological interest perhaps to aggregate all scores which originate from N-masks with equivalent numbers of Ns. This could give us an estimate of the interaction between *k*-mer replication and divergence. This is also quite a straight forward output matrix. It is also worth noting that the first output is a subset of the second: the vector of N=0 scores comprises the first column.

Formula 9:

$$Sig2D = \begin{bmatrix} S_{1,0} & \cdots & S_{1,N} \\ \vdots & \ddots & \vdots \\ S_{l,0} & \cdots & S_{l,N} \end{bmatrix}$$

Whilst the above output matrix is suitably interesting for an expanded *k*-mer spectral summary, and worth including as an informative set of datapoints, it also fails to include much of the inner complexity of the space of N-masked frequencies. One issue with categorising N-masks however is that their categorical dimensionality for deeper tree is very high (at 2^k), and with the sparsity of a DNA tree at *k*=31, the expected sparsity of the individual N-mask categories would disqualify them from direct usage as a means of aggregation for the creation of a signature. Therefore, we might try to find a middle road for the creation of second and third output matrix dimensions. The first pair of dimensional measurements to be investigated here will be the left and right 'seed length'. Seed length refers to the size of the either side of the N-mask (beginning with either the root or leaves of the tree) which contains no Ns. In other words, the length of the fixed seeds pre- or post the variable region of the *l*-mer. Let these index terms be *s,* and *d* (sinistral and dextral). Since not all values of *s* will be valid for all values of *d*, the output matrix will instead be a 3D wedge-shape. This indexing system is further explained by Figure 30.

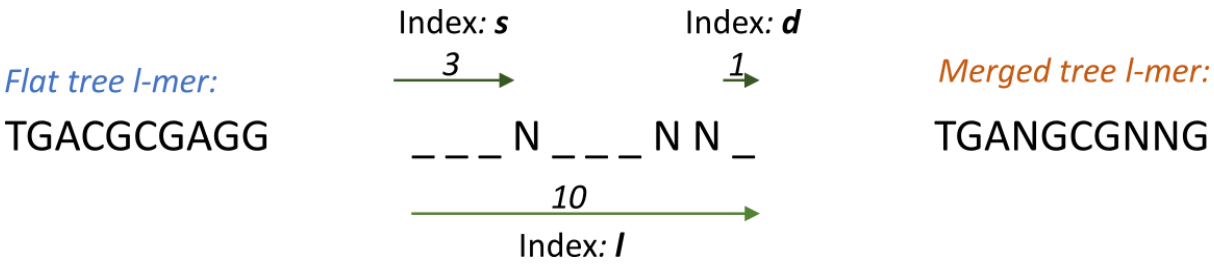


Figure 71. 3-Dimensional Indexing system for N-masks (DNA).

The matrix indices are defined as:

Formula 10:

$$\{ l \in \mathbb{Z} \mid 0 < l < k \}$$

$$\{ s \in \mathbb{Z} \mid 0 \leq s \leq l - d \}$$

$$\{ d \in \mathbb{Z} \mid 0 \leq d < l - s \}$$

Such that:

Formula 11:

$$Sig3D = \begin{bmatrix} \begin{bmatrix} S_{1,0,0} & 0 & 0 \\ \vdots & \ddots & 0 \\ S_{1,s,0} & \cdots & S_{1,s,d} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} S_{l,0,0} & 0 & 0 \\ \vdots & \ddots & 0 \\ S_{l,s,0} & \cdots & S_{l,s,d} \end{bmatrix} \end{bmatrix}$$

This matrix has the property of finding some of the inner complexity in motif flexibility shapes. However, one of its flaws is that the information space from which S is sampled is variable. For example, the lower values of both *s* and *d* present a much larger computational space of sequence flexibility when N is high, than the higher values of *s* and *d*. To create better consistency in the scaling of aggregation categories. We could also re-introduce the number of Ns in the mask as fourth dimension:

Formula 12:

$$Sig4D = \begin{bmatrix} \begin{bmatrix} S_{1,0,0,0} & 0 & 0 \\ \vdots & \ddots & 0 \\ S_{1,s,0,0} & \cdots & S_{1,s,d,0} \end{bmatrix} & \cdots & \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \\ \vdots & \ddots & \vdots \\ \begin{bmatrix} S_{l,0,0} & 0 & 0 \\ \vdots & \ddots & 0 \\ S_{l,s,0,0} & \cdots & S_{l,s,d,0} \end{bmatrix} & \cdots & \begin{bmatrix} S_{l,0,0,n} & 0 & 0 \\ \vdots & \ddots & 0 \\ S_{l,s,0,n} & \cdots & S_{l,s,d,n} \end{bmatrix} \end{bmatrix}$$

However, this version of the aggregation method may not be of much advantage versus the compactness of the 3D version, particularly when the analysis is limited to lower values of N (1-3). For very large sequence input sets (as in the original intended purpose), it may be computationally difficult to increase N substantially, as such the 3D signature aggregation matrix may suffice, however the 4D version perhaps ought to be applied should a version be developed with either smaller input sets, or substantial efficiency improvements in achieving summaries of higher dimensional N masks.

4.2.4. Cases for Aggregation Modes

Although not the principal objective of this research, summarising the total contained structure in the tree is something which might also be useful for large scale projects comparing hundreds or thousands of input sets in an external informative context (i.e. phylogenetics, lifestyle, environmental variables). It might also be useful in the case of segregations made within individual genomes, making experiments between annotation types possible. For example, testing regional information structure between intra-and inter-genic DNA, or between repeat types, or along

physical chromosome maps etc. For this reason, the Structure score summaries will still be included as outputs using the simpler 2D signature matrix (Formula 9).

Formula 7 describes a quantification of *l*-mer structures found in the genome. However, it does not provide us with a metric that is easily comparable between genomes, principally because of the confounding factors of size and ploidy. The simple solution would be to always divide the figure by the head-node frequency (post boundary correction). However, this solution only normalises the un-masked tree, simply because the creation of merged subtrees duplicates and re-measures the same frequencies in a different way. Additionally, in many cases subtrees will not be generated where they are not needed. The solution would be to sum all frequency duplications and add them to the head-node frequency, such that all recorded structure is normalised to the summed scale of the frequencies used in the entire data structure.

Formula 13:

$$S_r = \frac{\left(\sum_{v \in Ch(r)} Da_v * f_v * l_v\right) + \left(\sum_{v \in Ch(r)} \sum_{x \in Merge(Ch(v))} (Da_x * f_x * l_x)\right)}{f_r + \sum_{v \in Ch(r)} f_v}$$

Formula 13 shows the summation of genome-size normalised structure for N=1.

Formula 14:

$$S_r = \frac{\left(\sum_{v \in Ch(r)} Da_v * f_v * l_v\right) + \left(\sum_{v \in Ch(r)} \sum_{x1 \in Merge(Ch(v))} \left(Da_{x1} * f_{x1} * l_{x1} + \sum_{x2 \in Merge(Ch(x1))} Da_{x2} * f_{x2} * l_{x2}\right)\right)}{f_r + \sum_{v \in Ch(r)} \left(f_v + \sum_{x1 \in Merge(Ch(v))} f_{x1}\right)}$$

Formula 14 thus shows the summation function for N=2.

Formula 15:

$$S_r = \frac{\left(\sum_{v \in Ch(r)} Da_v * f_v * l_v\right) + \left(\sum_{v \in Ch(r)} \sum_{x_1 \in Merge(Ch(v))} \left(Da_{x_1} * f_{x_1} * l_{x_1} + \ldots + \sum_{x_n \in Merge(Ch(x_{n-1}))} Da_{x_n} * f_{x_n} * l_{x_n}\right)\right)}{f_r + \sum_{v \in Ch(r)} \left(f_v + \ldots + \sum_{x_{n-1} \in Merge(Ch(v))} f_{x_{n-1}}\right)}$$

Formula 15 shows the generalised extension of the formula for N=*n.* This will be how we assign structure scores to the sequences in the input data.

Although Formula 15 shows the more complete summary of a structure score for a whole tree, its components being frequency, distinctness and size, there are cases in which these components might be more usefully extracted as separate measurements. Indeed, since the signature indexing system does not recursively allocate to the same variable either, a slightly different definition is required.

It is here that we draw a distinction between signatures and summaries. The inter-comparable utility of signatures is maximized not just by equivalent dimensions, but by comparable term regularisation. For example, the mathematics required to compare two sets of variables over *[0, 1]* will inevitably be simpler than in the case of natural range of structure scores, which are essentially unlimited in scale. The variance of the sets of structure scores will also vary wildly with the size of the input sets. Given the range of eukaryotic genome sizes (The 12 MB of *Saccharomyces cerevisiae* to the 149 GB of *Paris japonica*) a signature ought to at least attempt to constrain the distributions of its values to normalised range, even if variance differences will still be inevitable to some degree. For this reason, for the purposes of complex signatures, the aggregation modes described above ought to be applied to gather the *weighted arithmetic mean* of the distinctness of each category. For example, in the case of the 3D signature (Formula 11) matrix:

Formula 16: $DF_{l,s,n} \ni \sum D_a * f$ and $F_{l,s,n} \ni \sum f$, then:

Formula 17: $\overline{D}_{l,s,n} = DF_{l,s,n} \div F_{l,s,n}$

4.2.5. Derived Measurement Types

When considering additional descriptors of the aggregate categories in the signature matrix, it is worth observing that each categories could also be thought of as its own vector of values with its own distribution. Here we propose two additional possible distribution qualities to be measured, formatted as concurrent signature matrices, and the rationale behind them.

The way in which the distribution is qualified will depend on the expected size of the vectors. There are two perspectives considered here. The first is the 'small vector' distribution. This is the case where the signature's input set might be small, for example, a single gene-family, repeat type, or a set of differentially expressed transcripts. Here we might be more concerned about the variance in the distribution, as a single reading may have captured a specific few biologically relevant active motifs. The second perspective is the 'large vector' distribution/large inputs (-omic scale data types). Here we can begin to make safer assumptions about the shapes of the distributions encountered and measure them differently.

Regarding the 'small vector' distributions, as each N-mask category has a given weighted mean *l*-mer of distinctness, this does not tell us the anything about the distribution of that property. Biologically it might be informative to know whether a given N-mask category reliably produces low or high distinctness, or whether its mean is the result of a broad range of inconsistent measurements. For this purpose, we could simply employ a weighted standard deviation (WSD). Like the weighted

mean, the frequencies would be used as weights. This would allow us to produce a parallel signature matrix of deviations.

Formula 18:

$$\sigma_{l,s,n} \ni \sqrt{\frac{\sum_{i=1}^{n} \boldsymbol{F}_i(\boldsymbol{D}_i - \overline{D}^*)^2}{\frac{n-1}{n}\sum_{i=1}^{n} \boldsymbol{F}_i}}$$

The second distribution of interest, regarding the 'large vector' case, relates to the power law. The power law has been observed to be broadly acting property of many natural systems (Newman 2005). Pareto-like distributions of properties have in fact been observed as consistent features of life systems at many scales (West et al. 1999). For example, it has been demonstrated to be a consistently emergent feature of metabolic networks that they be scale-free (Jeong et al. 2000). Additionally long right hand tails on most observed *k*-mer frequency graphs produces of biological sequence also show the Pareto-like distribution of frequency amongst substrings (Chor et al. 2009).

Although it cannot be guaranteed of any given input set that the F * D scores of *l*-mers will follow a pareto distribution, in the case of the largest scale biological data it is an assumption which allows for a more sophisticated measurement. The Pareto distribution formula in its original form is parameterised by two variables, *a* and *m*. The 'shape' parameter, *a,* acts as the variable which may be used to fit the distribution in a real data set, *m* (or minimum) is a simply a translating parameter defined as the minimum value in the data set. We would therefore choose the shape parameter as the most informative component of the distribution to estimate. A maximum likelihood estimated of *a* is quite straightforward (de Zea Bermudez & Kotz 2010):

Formula 19:

$$\hat{a}_{l,s,d} \ni \frac{n}{\sum_{i=1}^{n} \log\left(\frac{\boldsymbol{S}_i}{\widehat{m}}\right)}$$

Where ***S*** is given to be a vector of structure scores, and *m* is the minimum value in that vector. And to avoid confusion, *n* in this case refers to the size of the vector of values. This gives us another parallel signature matrix. This calculation could similarly be applied to any of the defined output matrix types (Formulae 8-12). Given that structure scores below 1 are possible, it could also be a good idea to set a lower bound to the structures included to the calculation (at least > 1).

4.2.6. Null Trees: Local and Absolute

Here we tackle the issues of single base/peptide frequencies, and the limitations imposed by saturation of the data structure.

Saturation in this context refers to the extent to which a random set of strings, present at a high enough frequency, will populate fully the *k*-mer tree data structure up to a certain depth. The relationship between the frequency of the head node, and the absolute null expected saturation is simply $log_n(f_r)$, where *n* is the size of the alphabet, and f_r is the root node frequency. This is the case which assumes all character frequencies are evenly distributed, as are all multi-character combinations. This impacts our measurement of frequency, which is an essential component of most of the measurements used. There are two polarities we must contend with whilst we are measuring frequency: Situations where the depth of the tree is such that the null expectation of *any* given node having a frequency of one or greater is vanishingly small, and situations where the null expectation of frequency may by in the hundreds of thousands. It would be erroneous to attribute low-*l* high frequency nodes the property of possessing an indicator of biological structure particularly when their frequency is comparable to one that might be found in the absolute null tree. Similarly, it would run afoul of multiple-testing error to weigh deep high frequency nodes by their individual improbability. We can also note that frequencies are used in two cases, as in Formulae 1-3 to discover D_b, and as in Formula 16-17, to weight the contribution of D_a to the mean of the given category. The proposed correction to *f* only applies to the weight, rather that the calculation of D_b, as this is not susceptible to the same scaling issues.

Formula 20:

$$f_c = \frac{f_v}{\boldsymbol{max}\left(\frac{f_r}{n^l}, 1\right)}$$

Formula 20 shows the correction of f_v (per vertex/node), by finding the null expectation of frequency saturation at the current node by dividing the root frequency (f_r) by the size of the sequence space of the tree at the current depth (n^l). By providing the lower bound of 1 to the denominator, the effect of the function will only apply at the 'null saturated' depths of the tree. This correction will thus be applied to all cases where *f* is used as a weight.

Since we are expecting some degree of saturation, one complaint we could find against the application of formulae 1-3, is that they range between total conservation and the maximum possible dispersal. Given that most organisms tend to have some bias in their genomic base frequencies, the actual null (i.e. random) dispersal for most of high frequency *l*-mers will rarely reach

the maximum possible. In fact, a 40% GC ratio (as in the human genome) would see many higher structure scores measured in cases where it is absent merely due to the base composition of the input. One seemingly intuitive way this problem could be addressed is by weighting the frequencies of the child nodes based on the input character ratios. However, this creates other unwanted sources of bias due to another aspect of DNA base ratios: they are not evenly distributed. The phenomenon of GC and CpG Islands is quite well established (Aïssani & Bernardi 1991). It refers to regions of the genome which are usually dense in protein coding genes. If a specific mean base frequency were used, it could lead to regions of 50:50 CG:AT having their structure scores weighted higher than they should, and the vice versa for other GC depleted regions.

The solution proposed to 'correct' for the base frequency artefacts is to generate a 'local null' tree prior to the generation of the main tree. The local null is a model of the null distribution of *l*-mers given only the actual uneven distribution of base frequencies as it occurs in the input set. This is created simply by building the main tree in all respects identically, except for a random shuffle performed on all input substrings. This preserves all base frequencies but eliminates their structures. In the case where reverse complements are also input, the random shuffle with occur first. The signature matrices (of *D * f*) generated by the local null tree might then be simply subtracted from the output. Integrating this with the weighted mean calculation would give us Formula 21.

Formula 21: $$\overline{D}_{l,s,d} = \boldsymbol{max}(DF_{l,s,d} - Null_{l,s,d}, 0) \div F_{l,s,d}$$

The subtraction of the local null might also be factored into the calculations of the other derived measurements. We will explore its applications next.

To correct the estimation of a Pareto shape parameter, we could, as mentioned in 2.5. increase the lower bound of the scores processed to the local null mean, however since we know that the local-null effect will apply to all structures, it would only serve to falsely alter the distribution. For the purposes of the single shape parameter which describes in total signature, i.e. as an adjunct to Formula 15, the solution we propose here is to aggregate the total shape parameter in stages, and to weight the contribution of categories based on their null-to-actual structure ratio. To do this, we aggregate the components of the shape MLE separately, *n*, and *log(S/m),* via the 2D aggregation matrix. The categorical actual-to-null ratios then scale each contribution – such that the final shape parameter is largely comprised of the contributions from the tree unaffected by the local null.

Formula 22:

Where: $x = \sum_{i=1}^{n} \log\left(\frac{S_i}{\hat{m}}\right)$, per aggregation category,

$$\hat{a}_r = \sum_{i=1}^{l}\sum_{j=0}^{n} n_{i,j}\left(\frac{\boldsymbol{max}\left(S_{i,j} - Null_{i,j}, 0\right)}{S_{i.j}}\right) \div \sum_{i=1}^{l}\sum_{j=0}^{n} x_{i,j}\left(\frac{\boldsymbol{max}\left(S_{i,j} - Null_{i,j}, 0\right)}{S_{i.j}}\right)$$

The individual category correction for shape parameters is more challenging, as the set of structures generated by both trees will be heterogenous and indirectly comparable in the same way as the output matrix. Currently a correction for single categories will not be deployed, particularly as the shape parameter becomes less stable/informative in lower values of *l* where the saturation is most likely to occur.

In the case of small input sets, we will argue that they ought not be 'local-null' corrected. In larger sets containing multimodal base ratio distributions, and a deeper and uneven saturation, the local-null can mitigate confounding effects to allow the structural content to be inter-comparable despite these factors. However, with small input trees saturation will be minimal and base ratios more typical and descriptive of the specific focus source of sequence, these things could be considered characteristics of the set rather than factors to mitigate. For this reason, the proposed usage of the pair of weighted structure score means and weighted SD for small sets will not be subject to null correction unless the results should provide a compelling reason to do so.

4.3. Implementation

The program was written in C++11 and is only compatible with UNIX-based systems. The program supports multi-threading, although at some memory cost, and at a non-linear performance benefit. The only external library linked is 'pthread'. The maximum depth of the tree in the implementation is currently 32.

Source code is available in Appendix 2.1 'Source code', and on GitHub: https://github.com/OliverCardiff/HighDimensionalSignatures**.**

There are two slightly different versions of the program. 'UGPep' has a slightly altered indexing system optimised for peptide sequence. 'UGLearner' is the original program which works with both DNA and peptide sequences.

4.3.1. Procedure parameterisation

Input Data – a set of 'fasta' formatted strings of alphanumeric characters

K – The depth of the tree

N – The maximum number of 'Ns' to consider in a single 'N-mask'

4.3.2. Core Data Structures

The primary data structure of the *k*-mer tree is simply a search tree derived from a fixed character set. Each node in the tree is of a type representing a single alphanumeric character and contains available memory references to as many potential child node types as the character set defines. Each node also contains one unsigned integer, which describes the frequency with which it has been traversed during the loading phase.

4.3.3. Data Input

This phase of the algorithm comprises the construction of the tree from a set of strings – likely DNA or proteins. *k*-mers of length *D* will be read sequentially from each string in the input set. The *k*-mer substrings define, by their characters, a traversal of the tree. The head node passes the input string to a child node which matches the leading character in the string. The frequency integer within the child node is incremented by one. If no child node has been instantiated yet, then instantiation will occur. Only child nodes which have been traversed will be instantiated in this manner. Following submission of the *k*-mer string to the child node, the leading character is trimmed, and the function is repeated until a tree depth of *D* is reached, and the input string has been depleted of characters.

Once the first *D* characters of the first string in the input set has been read by the tree, the starting *k*-mer index is incremented by one, and in this manner the following *k*-mer is read.

Over the range of *[0, n]*, where *n* is the input string length, indices for *k*-mer substrings are found in the input string: *[(0, k), (n, n-k)].* This process is repeated for every string in the input set. Every *k*-mer read into the tree in this manner will also have its reverse, or in the case of DNA reverse complement, generated, which will be read into the tree in the same manner.

4.3.4. Sub-tree Merge

The implementation of subtree merging, particularly with respect to memory usage, will be covered before the larger DFS algorithm which calls it. When merging sub-trees with respect to a single node, we are theoretically creating an additional tree structure to hold the merged data. However, in many cases the memory allocated to the pre-existing tree structures can be taken advantage of. For this reason, all merged subtrees with respect to a single node store their variables in left-most child's subtree. This is to say that the Node class also implements a 'map' type, which allows it to store additional unsigned integers, paired with an ID which identifies its N-mask ownership.

Mapped IDs are themselves *l*-length binary tree navigation pathways. The formula for navigating the merged sub-tree variable space is as follows.

Given parent ID:

To access leftmost Child's frequency in the same subtree: ChildID = ID * 2

To access other children's frequencies in same subtree: ChildID = ID

To access head-node of new merged sub-tree: ChildID[1] = (ID * 2) + 1

Just so long as all functions follow the above ID manipulation rules with respect to any given node, whereby the initial IDs of all scores in the unmerged tree are zero, the N-mask will always be derivable from the pattern of bits in the integer variable used to identify the score.

This indexing system allows the algorithm to virtualise the retention of merged tree scores within the current tree, without the need to create new Nodes, and the memory overhead that involves.

The tree merging algorithm is a DFS with paired navigation. This is to say that the virtual subtree stored in the leftmost branch of the node of origin is simultaneously traversed alongside an unmerged branch, summing their frequencies into the virtual subtree. This occurs *n* times, once for each of the child-subtrees connected to the node of origin.

4.3.5. Depth-First Search

The aggregation of data within the *k*-mer tree is organised around a recursive DFS function. It is described by Figure 31. The initial values for the parent distinction, depth, and ID arguments are all zero.

```
Function: Search (parentDistinction, Node, ID, depth):
      if(depth < K)
            if (IsMergeable(this))
                  SubtreeMerge(this)

            D_b = FindDistinctness()
            D_a = D_b * (1 – parentDistinction)

            Aggregator.Report(D_a, Frequency, depth, ID)

            if (HasMergedSubtree())
                  Search(D_b, Children[1], (ID * 2) +1, ++depth)

            for i in 1:n
                  if i == 1
                        Search(D_b, Children[i], ID*2, ++depth)
                  else
                        Search(D_b, Children[i], ID, ++depth)
```

Figure 72. DFS tree navigation pseudocode.

4.3.6. Multi-threading

Figure 31 shows that the memory used in the merged trees is created as it is needed. Although not described in the pseudocode, this memory allocation is also deallocated as soon as it is measured and no-longer required for any deeper merges. However, this means that each thread exploring the tree via DFS will have its own substantial memory overhead.

The positive aspect of multi-threading a tree structure is that every child node can point to an independent region of memory. From the root node, *n* threads can be created (one per child), and each thread will not have any memory access conflicts when reading the tree. The aggregator class also implements separate memory buffers per thread, which are periodically read into the output matrices. This mitigates any bottlenecking at that point. The thread allocator can expand tree access for all available threads in this manner, continuing to guarantee independence of memory. The caveat to this approach to threading is that the calling thread also continues to work on one branch of the tree whilst others work on others. Only once all work generated by a node has been completed by the worker threads can the collection of workers be freed up to be reallocated. This has the effect of limiting thread efficiency per spawning node to the performance of the most expensive subtree. Given that character frequencies in biological sequences are rarely equal, this can equate to substantial loss of overall threading efficiency.

The inefficiencies of this threading system are unlikely to be unmitigable. Further performance gains may almost certainly be achieved by optimising the thread allocator. This has not been undertaken yet due to time constraints.

4.3.7. Performance testing

The performance tests were running on a Linux desktop computer with 32Gb of RAM, and a 12-core Intel CPU. Due to the cores available, some of the tests using more than 12 threads may not be indicative of the true efficiency at this scale. However, given the thread availability issue, it can also be beneficial to add more threads to the allocator than can be simultaneously assigned to separate CPU cores.

The following tests are run using subsets of DNA from the NCBI *Escherichia coli* reference genome (Blattner 1997), and subsets of protein sequence from the *Apis mellifera* proteome (Consortium 2006). The depth of all trees, as in the value of *k*, was 30 for all tests.

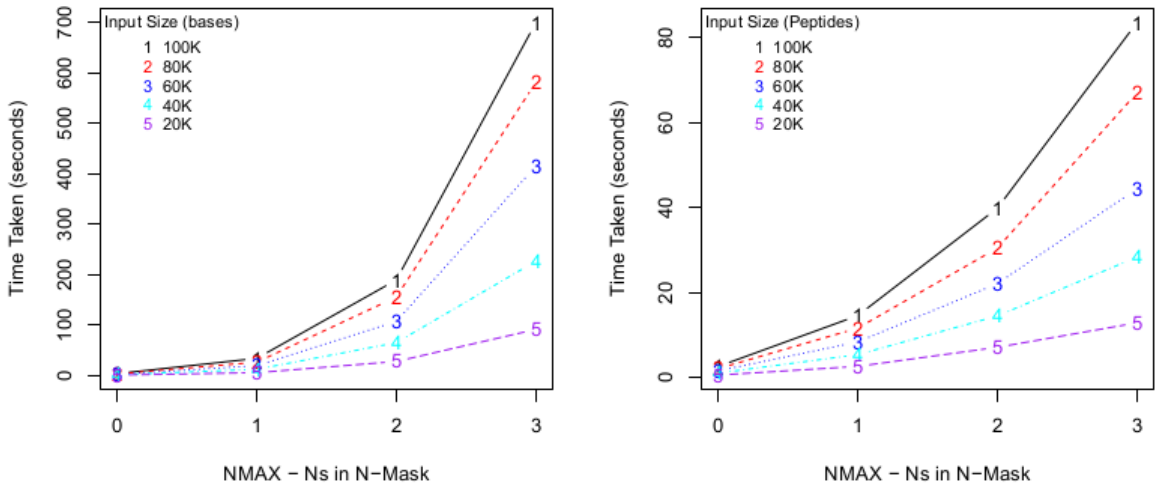


Figure 73. Performance tests for N values over [0,3], one thread. Left: DNA input sets, ranging from 20-100K bases sampled from the E. coli *reference genome. Right: 20-100K peptides sampled from the* Apis mellifera *proteome.*

Insofar as the complexity of the N-mask increases by powers of 2 with every additional N (2^n), the performance of the program reflects this with exponential computation time increments. Interestingly the difference in performance between DNA and peptide input sets are almost a factor of 10. This is likely due to the extreme sparsity of the peptide tree (the space expanding to 20^{30} at the end), resulting in far fewer nodes are meeting the qualifying conditions for the generation of a merged subtree. Additionally, owing to the higher alphabet, the average saturation depth in the peptide tree will also be much shallower (3.94 with reversals in the peptide 100k test set, versus 8.8

in the DNA test).

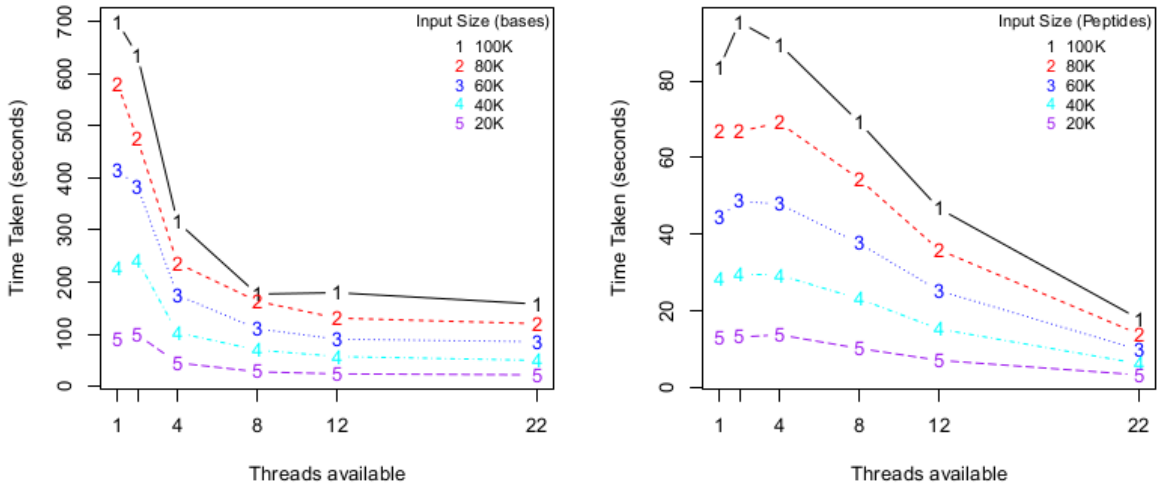


Figure 74. Performance test, multithreaded (N=3), for 1-22 threads executing on a single machine. Left: DNA input sets, ranging from 20-100K bases sampled from the E. coli reference genome. Right: 20-100K peptides sampled from the Apis mellifera proteome.

The results shown in Figure 33 show that the DNA search tree fails to make performance gains above 8 threads. The difference in performance between 2 and 4 threads also suggests that the equal distribution of work between threads from a single originating node of the tree (as discussed in 2.3.6) plays the most significant role in thread efficiency. Figure 34 shows that in both cases, the only time the per-thread efficiency increases following incremental increases from single threaded performance is when the thread number becomes equal to the alphabet size.

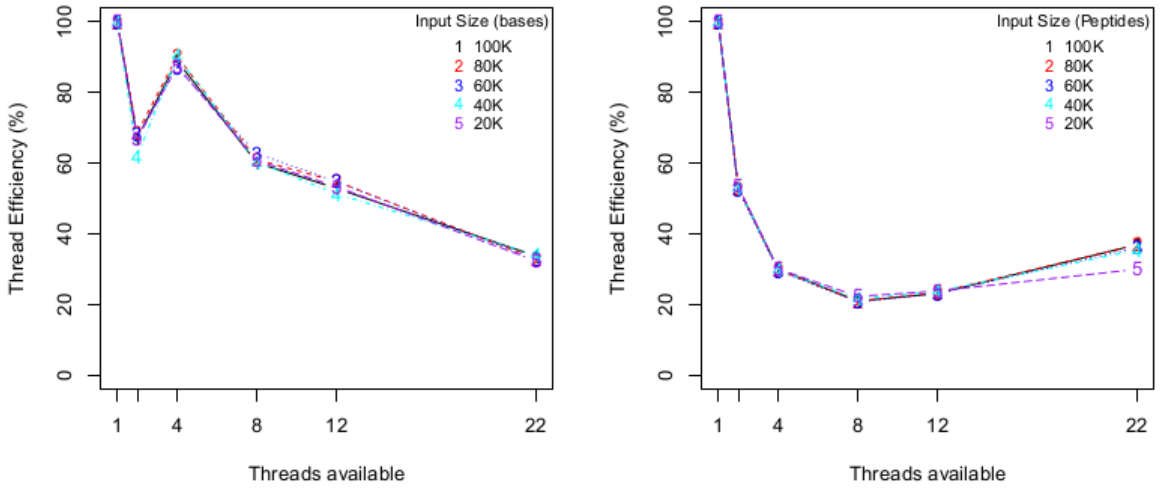


Figure 75. Performance test; Thread Efficiency. Left: DNA input sets, ranging from 20-100K bases sampled from the E. coli reference genome. Right: 20-100K peptides sampled from the Apis mellifera proteome.

The peak RAM usage (see Figure 35), rather than being exponentially related, only increases in a proportional linear manner as the N-mask increases in complexity. This is in part due to the immediate memory deallocation performed on all measured subtrees. Only a single slice of the exponentially increased complexity space need be stored in memory at any given time. Since this test was performed using 12 threads, it would be easy to trade-off performance time for memory usage by decreasing the thread count.

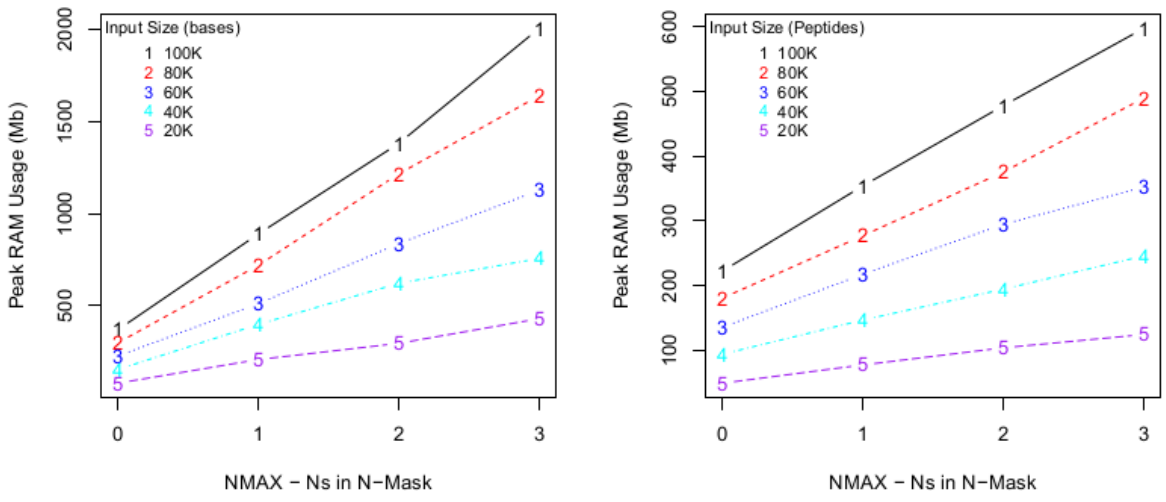


Figure 76. Performance test; Memory Usage (12 threads). Left: DNA input sets, ranging from 20-100K bases sampled from the E. coli reference genome. Right: 20-100K peptides sampled from the Apis mellifera proteome.

The implementation of this method is may still be subject to improvement in terms of computation time and memory usage. Despite this, it is currently enough to calculate signatures for the smaller values of N and has not crashed during testing on several machines.

4.4. Results

Given the exponential time cost of calculating more complex N-masks (as seen in 3.3.7.), the demonstrated application of this program will be limited to values of N at 3 or lower. For the sake of generating inter-comparable signatures, it is also important that all parameters be equal aside from the input set. As in the performance tests, the value of *k* will be 30 in all cases.

4.4.1. Visualisation

Biological information is often only so meaningful as the human eye can comprehend. As the multi-dimensional nature of these signatures does not plot spatially in an intuitive manner in their native dimensions, the visualisations have been flattened into 2D plots, with extra-dimensional information encoded in colour, alpha, and point size. To provide a basic set of interpretive aides for the signatures, the illustrations in Figures 36-38 were created as a reference for users looking at more

complex plots. These can be referred to by the reader whilst viewing later sections. For a quick guide to the colour key see Figure 38.

Another useful visualisation which has been applied to the 3D output matrix is the concept of ‘threads’. As will be shown, the higher dimensional output signatures typically have categories which follow a linear or curved gradient at multiple depths. These categories are usually identical in ‘left seed’ length but increment by one in ‘right seed’ length between depths. For this reason, faint lines have been added to plots which connect all points that observe this single right increment relationship. Figure 8 shows the creation of single thread visualised.

To clarify the meaning of ‘dispersal’ patterns, Figure 36 shows two miniature cases of sequence structures.

The source code written in R for the visualisation functions described here is available in Appendix 2.1.3 ‘Source Code->Visualisation’.

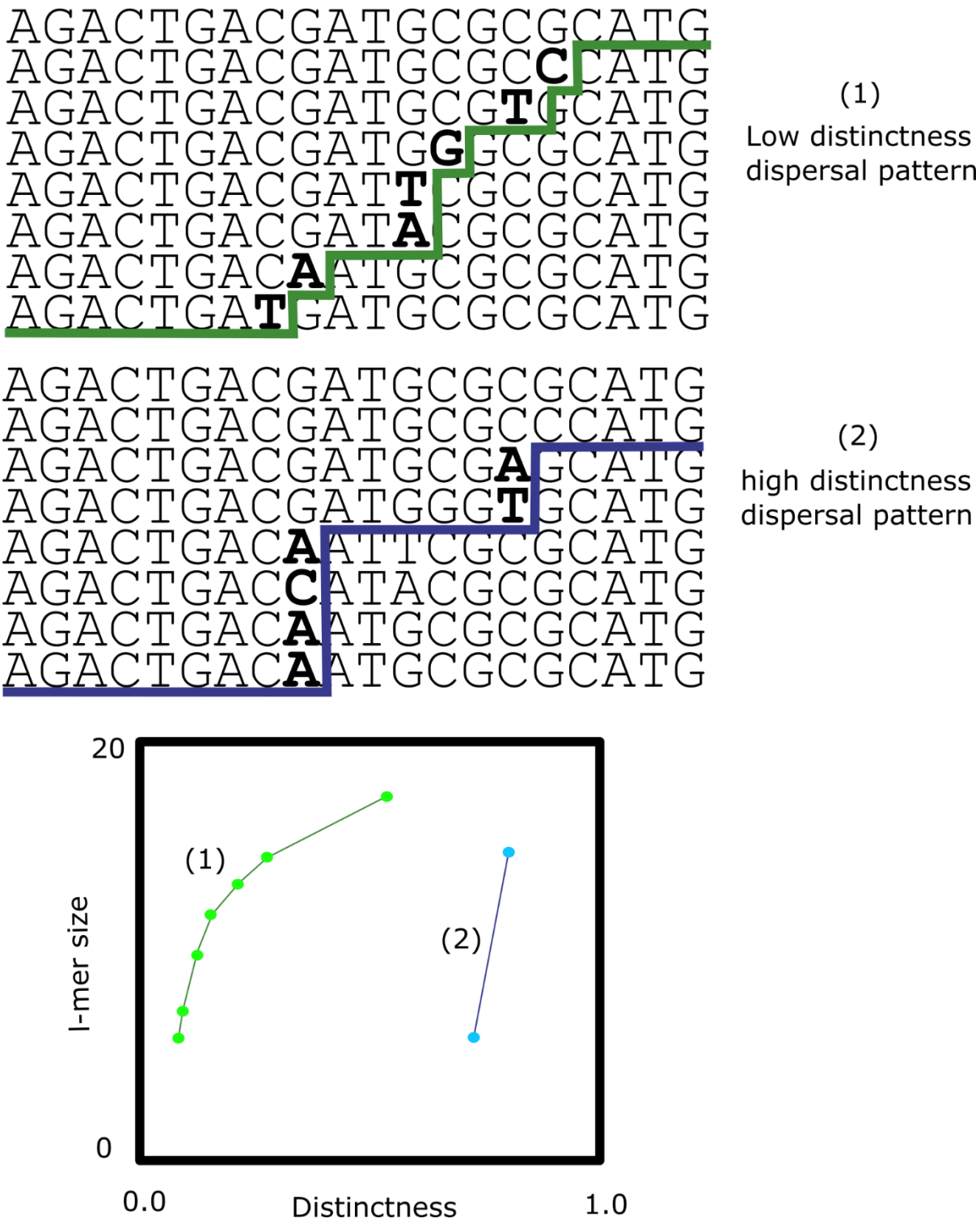


Figure 77. Relationship between unmasked sequence threads and motif variability (not to scale).

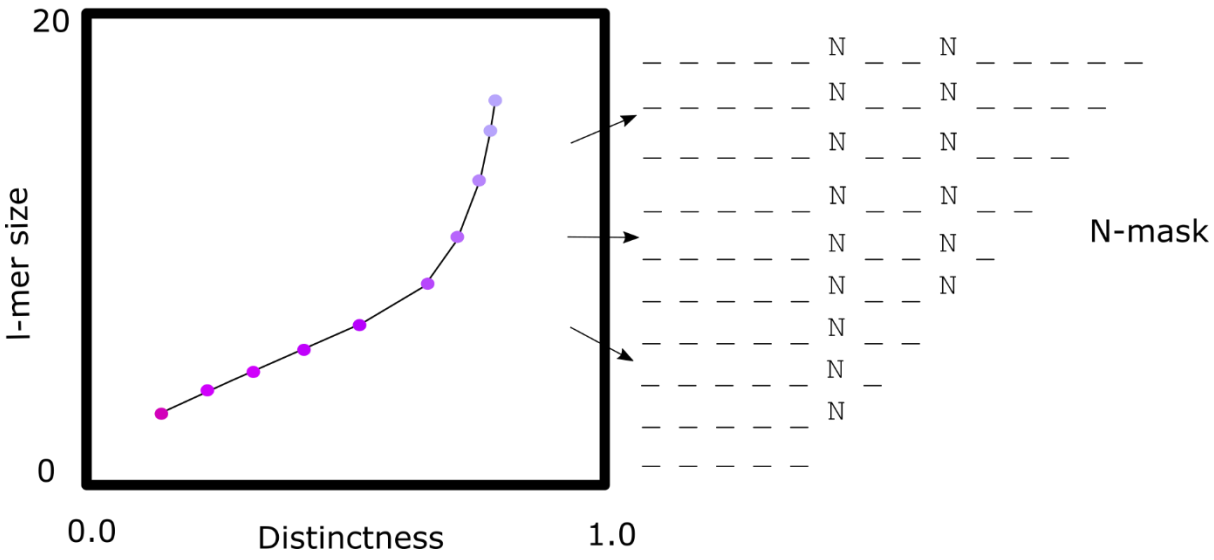


Figure 78. Illustrative relationship between N-masks and threads.

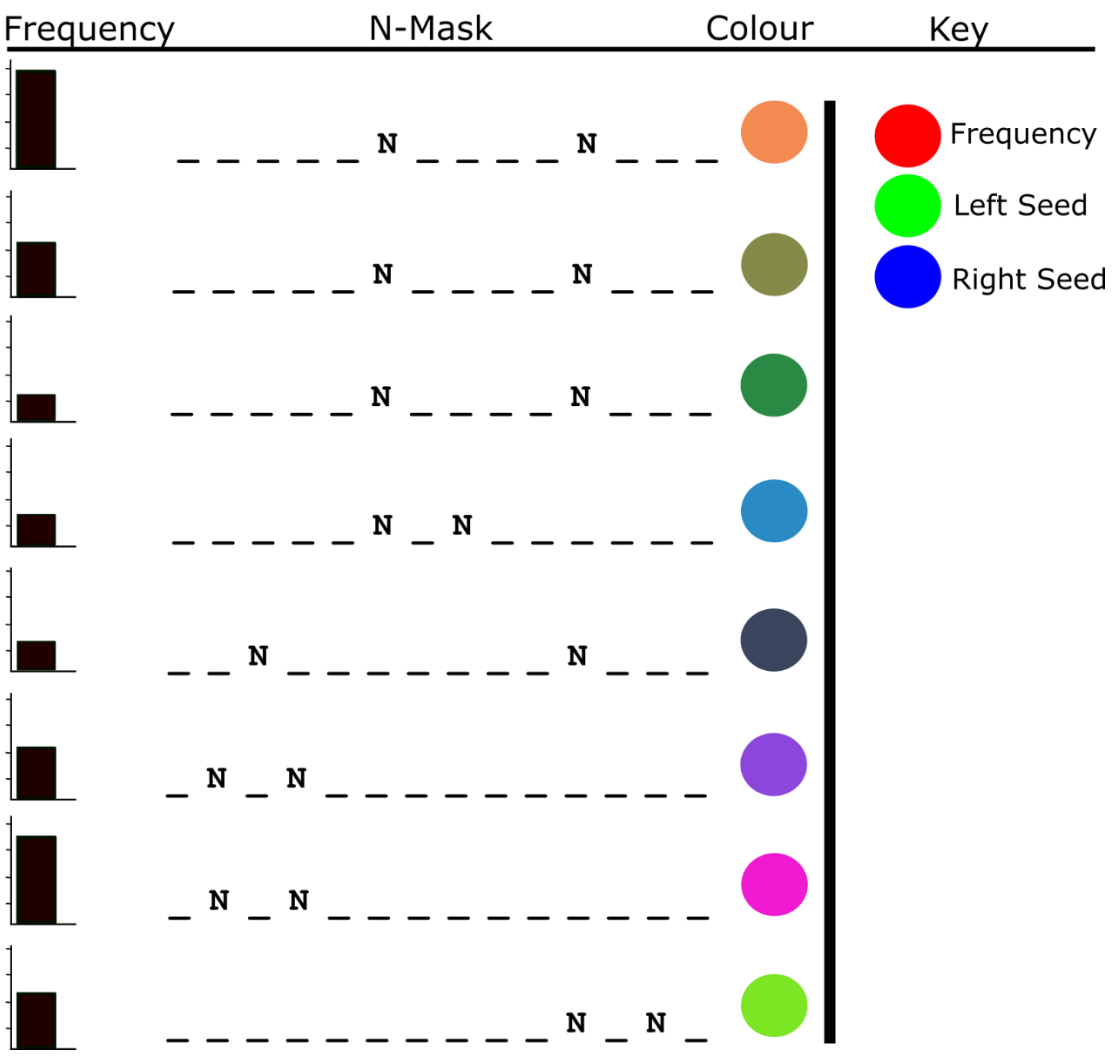


Figure 79. Illustrative relationship between graph colours, seed lengths and l-mer frequency.

4.4.2. Random Case Signatures

To interpret any signature that is produced by this method it is important to understand visually the baseline null case from which all structured sequence inputs will deviate. For this reason, two (DNA and peptide) random sequence noise input sets were tested. Both of 100K characters in length. Since the purpose of the random tests is to establish a null-looking signature, there was one considered difference in generation between DNA and peptides. Random DNA sequences were generated with even base ratios, but random peptide sequences were generated with the average peptide frequencies found in the UniProtKB database (see Table 4).

Table 12. Peptide Frequencies Used in Random Tests (EMBL et al. 2013)

Typical AA Composition of UniProtKB/Swiss-Prot database (%)							
Ala	8.25	GLU	6.75	Met	2.42	Tyr	2.92
Arg	5.53	GLY	7.07	Phe	3.86	Val	6.87
Asn	4.06	His	2.27	Pro	4.7		
Asp	5.54	Ile	5.96	Ser	6.56		
Cys	1.37	Leu	9.66	Thr	5.34		
Gln	3.93	Lys	5.84	Trp	1.08		

The reasoning is simply that the typical peptide ratios vary so greatly between them that even unstructured input sets will universally register higher structures at the top of the tree, unlike most DNA sequence trees, which are expected to be closer to 0.

Figure 39 shows the output of the 2D aggregate matrix (Formula 9) using the local null corrected weighted arithmetic mean distinctness per N, per *l* (as in Formula 21), for the random noise input sets. The typical pattern for distinctness values at *l,* as they ascend beyond saturation depths, is to move swiftly towards 1 (the value found when a frequency 2 branch splits). The return to zero is thus indicative that there are no more structures to be measured for distinctness in the entire tree at this point. Even a very small and improbable number of >1 frequency branch will cause a distinctness mean to be recorded.

Noticeably, the random DNA tree continues to find some measurements of structure even as high as *l*=22 when N=3 (effective minimum sequence space of size 4^{19}). While highly improbable, the frequencies of these small structures may also be due to the slightly inconsistent effects of pseudorandom number generation.

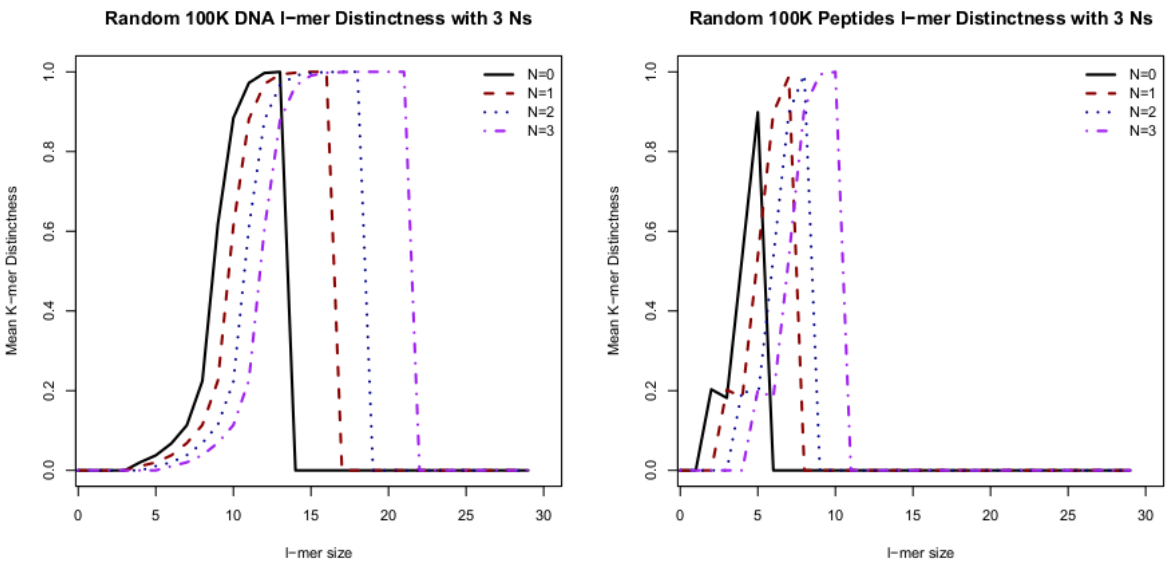


Figure 80. 2D Structure matrix outputs for null cases. Left: 100K DNA bases, Right: 100K protein peptides.

The random peptide tree, with its greater sequence space, loses all structure very quickly in comparison.

The null dispersion of frequency can be seen very clearly in Figure 40, with the value of N only slightly modulating the depths at which the sequences disperse. The relationship between saturation and distinction scores also is clearly displayed. At 200Kb (100Kb input + 100kb reverse complement), the null average saturation depth is approximately 8.8. It is only after this depth that the cohort of means begins to show the results of the dispersal of the set of structures retained by chance. Naturally, as the depth gets lower, the probability of any given structure retaining enough frequencies to disperse amongst the child-nodes decreases exponentially. This also applies to the parent nodes of by-chance dispersals, meaning that the calculation of $(1\text{-}D_{parent})$ component of the calculation of D_a (Formula 6) is far more likely to also be 1. It is this relationship with drives the distinctness curve to 1 in the null/random case.

Looking at the random peptide output, Figure 41, the curve is similar in shape but occurs far more rapidly, as in Figure 39. One positive aspect of the null peptide signature is that we can reasonably expect almost all structures recorded above $2 * log_{20}(f_r)$ to be the result of actual biological effects.

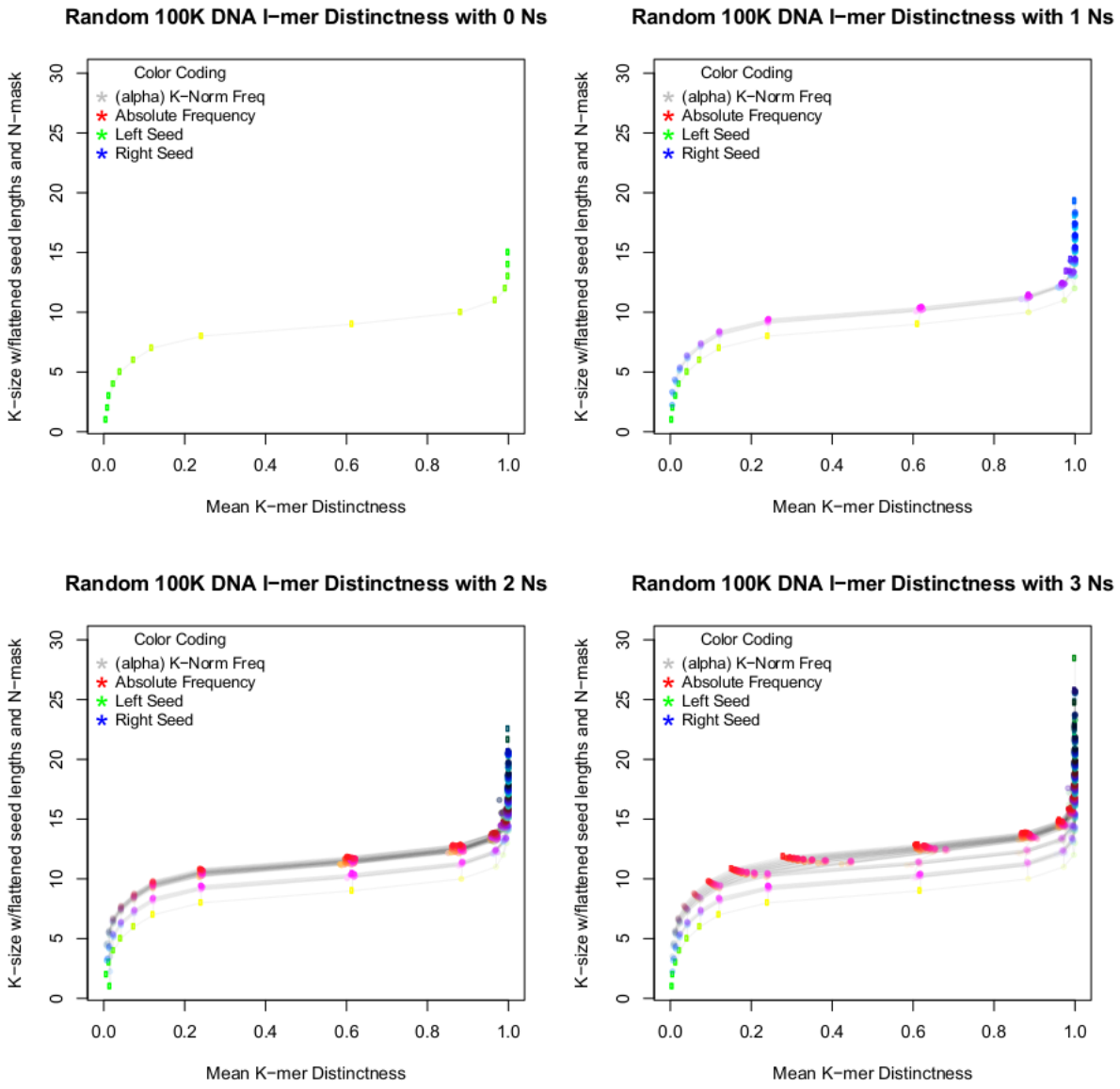


Figure 81. 3-Dimensional output signatures for random DNA. 100Kb random sequence used in each execution, visualisation of four values given for parameter N. Top-left: N=0, Top-right: N=1, Bottom-left: N=2, Bottom-right: N=3.

Insofar as these graphs inform our interpretation of other plots, we should make note of the natural signature of null sequence dispersal and attempt to distinguish it from structured dispersal. For the DNA graphs we observe the steepest part of the 0-1 distinctness curve beginning near the saturation depth, the tendency towards 1 at the top of the signature, and the tendency towards zero near the start. This shape will be referred to as the 'DNA null curve' in discussion of later plots. We should observe therefore the modulations of the null curve as biological signatures. Similarly, the pattern of natural effects which occurs in the peptide graphs, as discussed at the start of this section, varies slightly. We observe the head of the tree commencing at ~0.4 distinctness, moving quite sharply lower, and reversing after the saturation depth to curve back towards 1. Again, this will be referred to in later sections as the 'peptide null curve'.

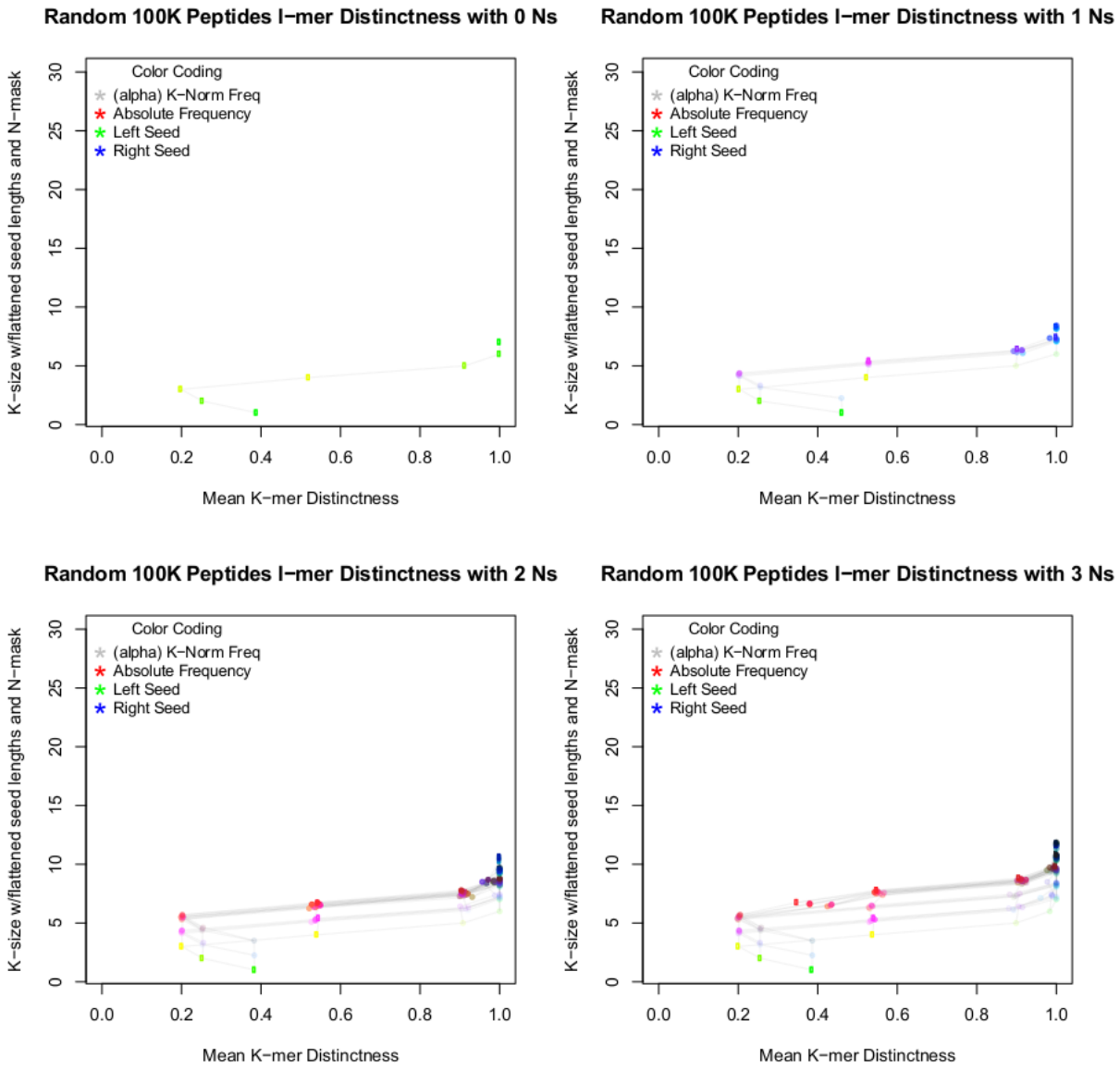


Figure 82. 3-Dimensional output signatures for random peptides. 100K random peptide sequence used in each execution, visualisation of four values given for parameter N. Top-left: N=0, Top-right: N=1, Bottom-left: N=2, Bottom-right: N=3.

4.4.3. Small Subset Signature Tests

The next series of tests involves using subsets of biological sequence at the same scale as the test set (100K characters). There was no additional randomisation of subset, in both cases they were selected under the conditions of being the first 100K characters in the files they were extracted from. The two source material files were as such*: Apis mellifera* proteome retrieved from Uniprot (EMBL et al. 2013), and *Escherichia coli* reference genome retrieved from the NCBI genome database (NCBI 2016).

The objective of these tests is to examine the way in which the null curve begins to change when biological sequences are used, with relatively low structure in the input. In the case of ‘omic scale

datasets, many of the sequence structures that might be found are only discoverable in the context of the entire set. For example, a 30-mer which occurs only five times in the genome is unlikely to be present more than once in a 2% subset. By extension, we can suggest that whilst these small subsets of sequence will be more structured than random, the actual discoverable structure ought to be on a much lower scale than in a typical input set. This makes them a good 'stepping stone' between the random signatures and full input sets. The signatures developed here are purely aggregates of weighted mean distinctness scores and have not been subject to 'local-null' corrections.

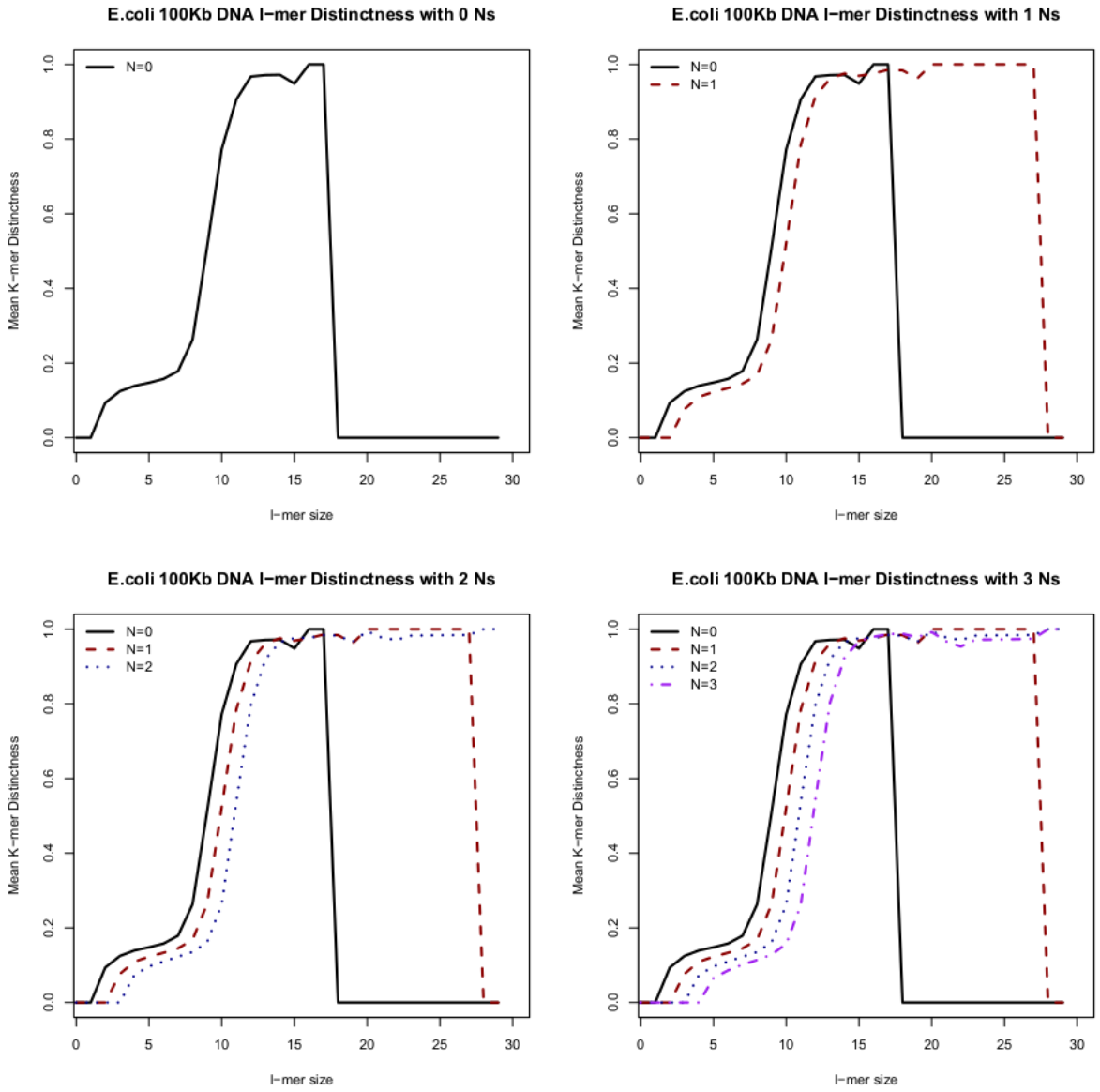


Figure 83. E. coli 100Kb subset 2-Dimensional signature output. Top-left: N=0, Top-right: N=1, Bottom-left: N=2, Bottom-right: N=3.

Figure 42 is directly comparable to the Figure 39 (left), this is to say that the N=0 plot follows a similar pattern with two exceptions, a faster ascent in the saturation depths and a longer reach into

the unsaturated depths (14 vs 18). The effects of introducing Ns has a far more marked effect. N=1 only loses all structure at *l=28*, and the other values of N continue to find low frequency merged long l-mers throughout the set. This speaks to the fragility of long substrings in biological sequence more generally, and would be expected concordance with the development of gk-SVM (Ghandi et al. 2014), as discussed the introduction.

Figures 42-45 are all subsets of larger permutation tests. Their expanded paired images are available in Appendix 2.5, as IMG1-4 respectively.

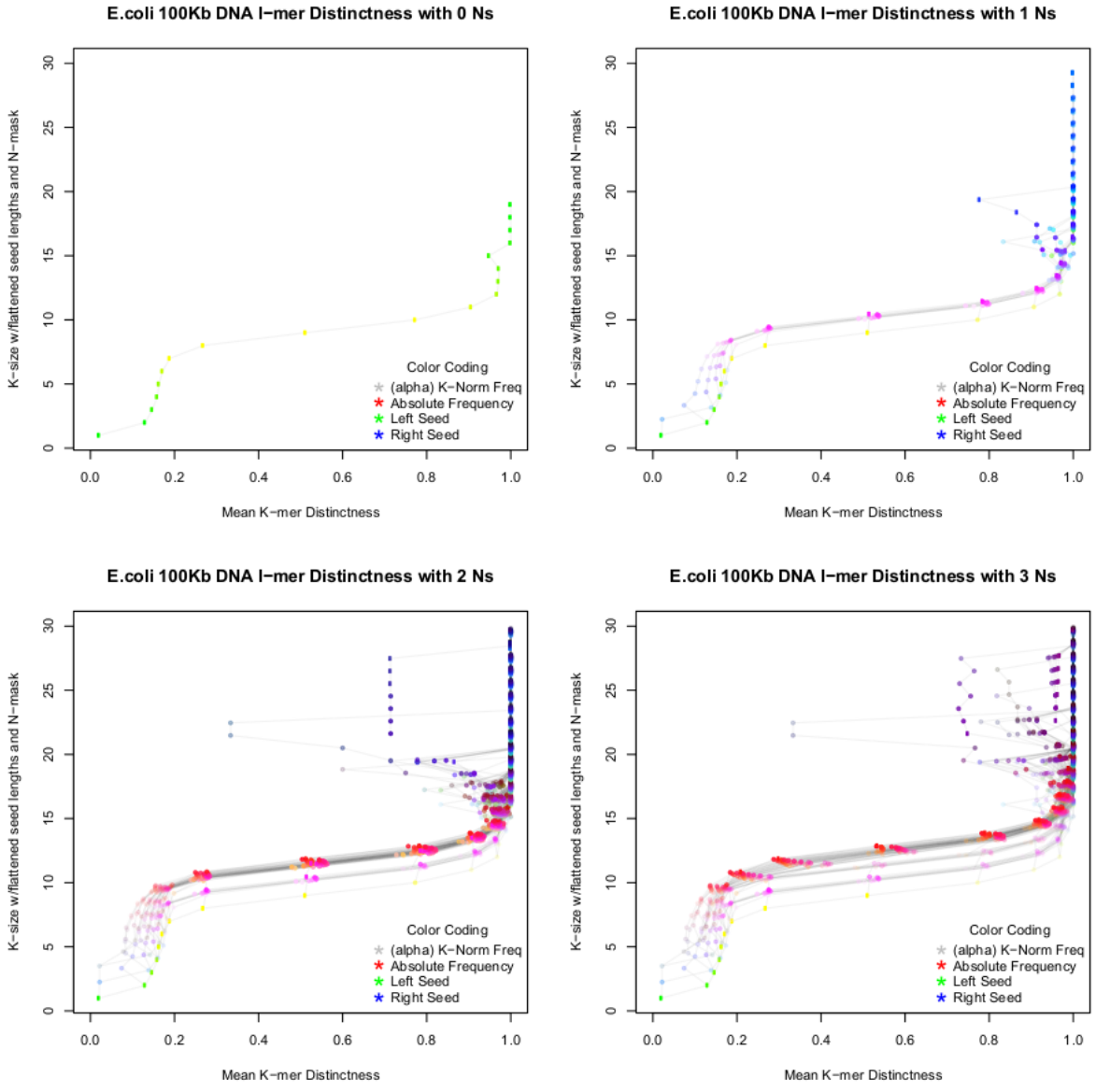


Figure 84. E. coli 100Kb subset 2-Dimensional structure output. Top-left: N=0, Top-right: N=1, Bottom-left: N=2, Bottom-right: N=3.

The 3D matrix outputs (Figure 43) begin to show in more detail the N>0 results found in Figure 42. The braid-like structures forming at the saturated depths show that the detection of uneven 2-8mer

substring frequencies becomes possible at this scale. The masked-index category threads here tend to repeat earlier unmasked, or lesser masked threads at higher depths. The post-saturation gradient is still largely present, however the component which collapsed at 1 in the null curve is showing various relatively distinct medium-to-low frequency N-masked complex structures with long right seeds at the higher depths. This is an example of how the specificity of the signature allows direct description of the type of flexibility found in the reference structures.

The peptide 2D subset test (Figure 44) produces a substantially difference result to the null test. With saturation depths typified by an early spike in distinctness followed by a curve which tends slowly higher. The peptide null curve tendency to return to lower mean distinctness immediately following saturation is repeated here, however the dispersal of frequencies seems to be far more gradual for each structure. A case where a frequency-50 12-mer loses 10% of its frequencies per depth, would be typical of a sequence structure that drags the mean distinctness towards 0.1, as can be seen here.

What this suggests biologically is a set of similar sequences which are each dissimilar from each-other in different ways, suggesting that an N-mask would struggle to reunite them at longer for fragile values of *l.* The opposite case would be a set of sequences which all differ a one or two fixed location, this would disperse over far fewer depths, generation very high distinctness scores.

The rapid spike towards 1 demonstrated by the N>0 should also be considered more the effect of the terminal-k depth backward subtraction process described in 2.2. Figure 44 is also directly comparable to Figure 45 in shape. However, Figure 45 also begins to show another feature related to the single right seed extension per depth relationship discussed in 3.4.1., threading, and a certain banding pattern of threads. A banding pattern can be described as a case where multiple threads cluster into a single channel. To understand banding, consider the opposite cases described above, of high frequency structures which typically disperse either over many depths, or only over one or two, as in Figure 36. Bands represent specific biological prominences in the modes of structure dispersal within that range. This might suggest evolution acting differently on several different types of sequence motifs. Some motifs are flexible in a highly regular manner, these may present as higher distinctness bands, some motifs have the evolutionary flexibility to diverge at almost any base, just so long as most of the sequence remains similar, these types of sequence structure will manifest more as bands towards the lower distinctness range. The number, and complexity of the bands, are thus to be read as indicative of the prominence of sequence structure types exhibiting separate modalities of evolutionary change.

For example, the N=3 graphs in Figure 45 shows in the 17-23 *l*-mer range, an unusually distinct set of structures typified by a relatively small region of flexibility and a long right seed. This structural category decomposes in a slightly more homogenous manner that the other content of the test set.

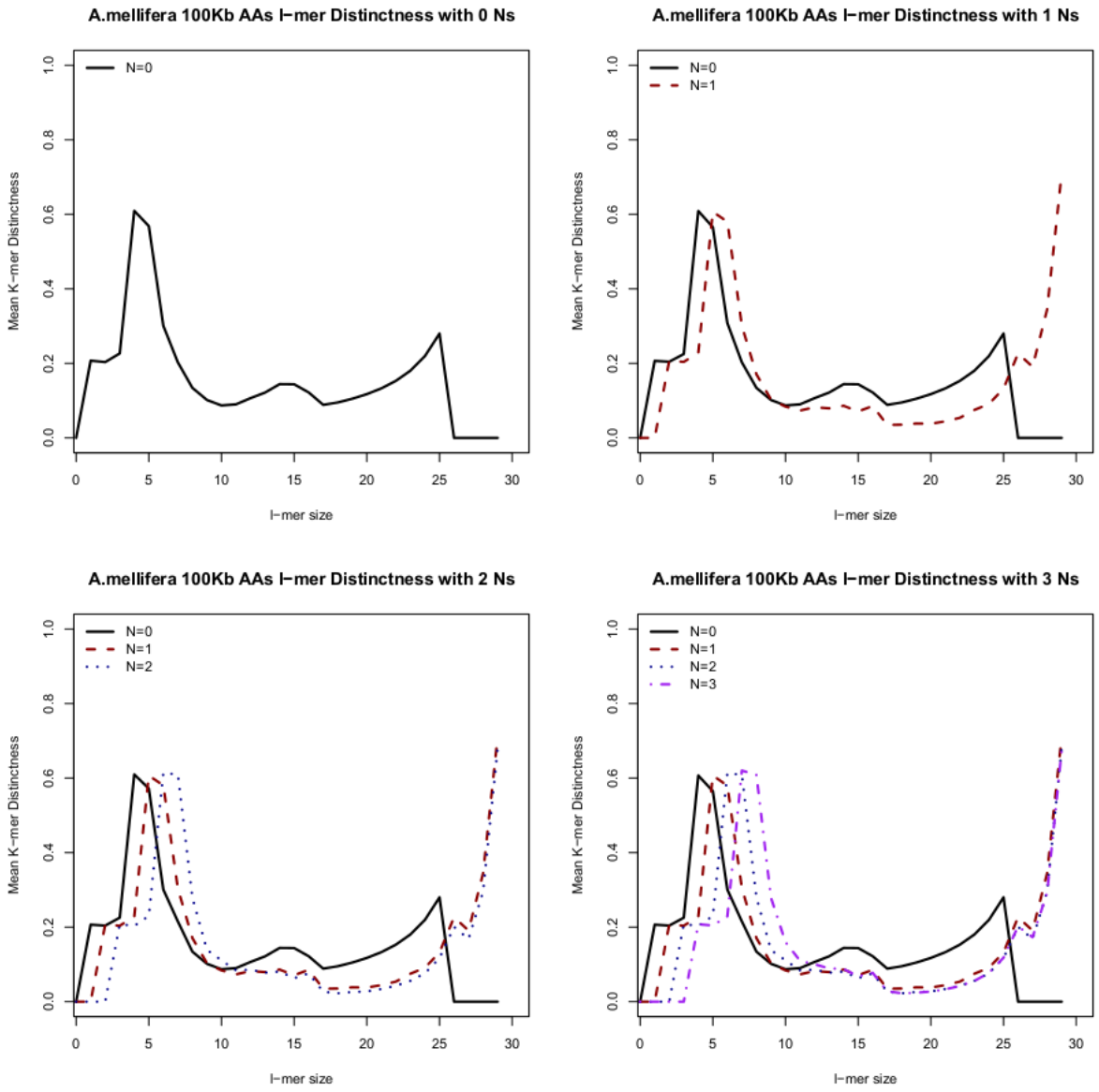


Figure 85. Apis mellifera 100K AA 2-Dimensional structure matrix. Top-left: N=0, Top-right: N=1, Bottom-left: N=2, Bottom-right: N=3.

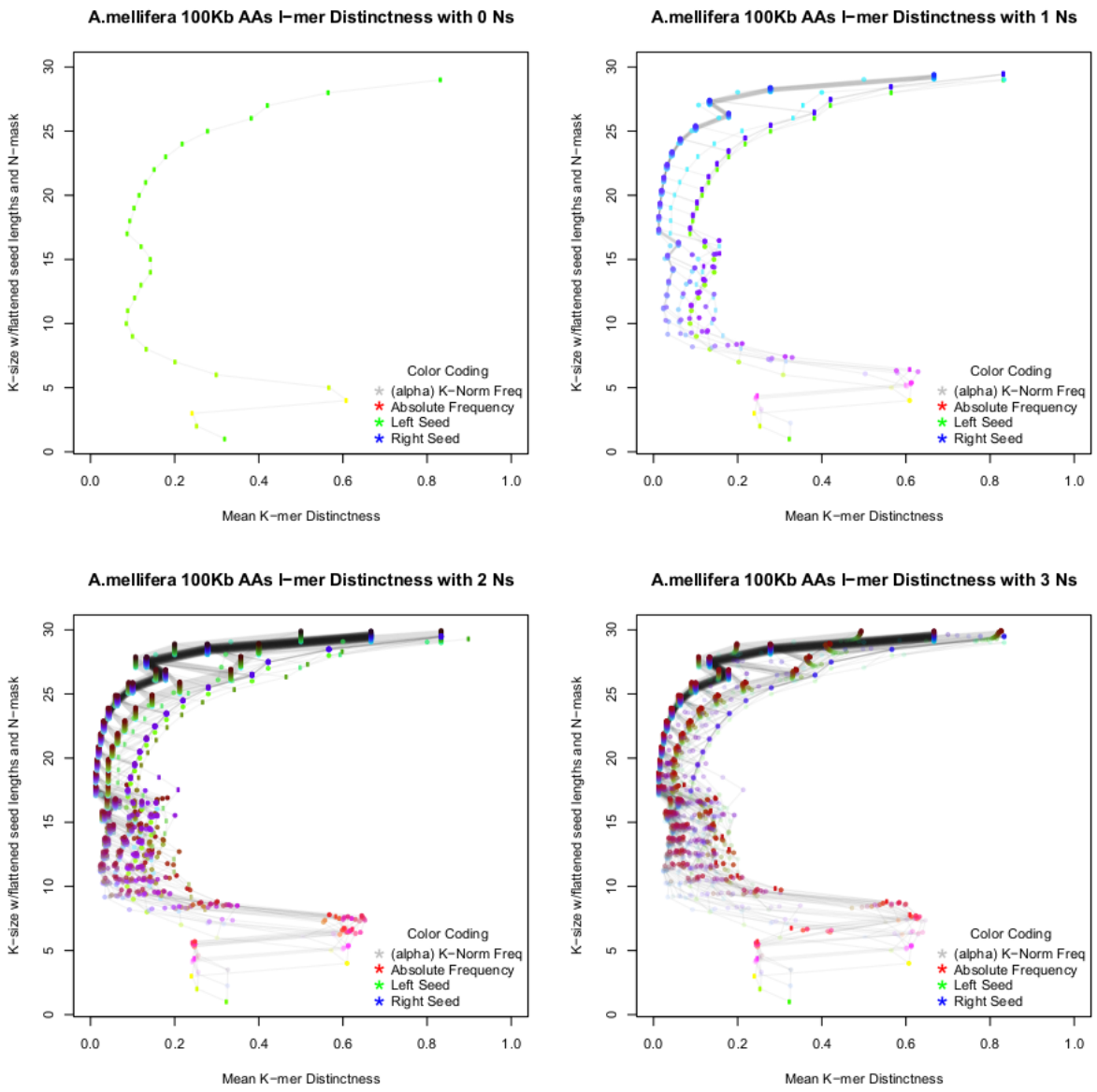


Figure 86. A. mellifera 100K AA 3-Dimensional structure matrix. Top-left: N=0, Top-right: N=1, Bottom-left: N=2, Bottom-right: N=3.

4.4.4. Test Set: Invertebrate Proteome Signatures

This is the first of three test sets designed to explore the ways in the which the signatures can be used to interpret biological data types. This test set also serves to further explore the peculiarity of *L. rubellus* in comparison to relative to three other annelids, and an arthropod. *Capitella teleta, Helobdella robusta* and *Apis mellifera* reference proteomes were all retrieved from Uniprot (EMBL et al. 2013), *L. rubellus* proteome was extracted from a genome assembly discussed in Chapter 2. *A. gracilis* proteome was extracted from the genome assembly performed in Chapter 3.

Here we introduce the test of Pareto shape parameters and full tree summaries as described in 3.2.5. and 3.2.4. respectively. Table 5 shows the summarisation of the total structure in the trees

built out of these proteomes. In addition to the summary scores, it also shows the result of applying the 'local-null' correction/subtraction discussed in 3.2.6.

Table 13. Proteome Tree Summaries

Species	Structure	Struct. - Sub	Shape	Shape - Sub	Size (aa)	>K Freq (%)
Apis mellifera	0.0227	0.0054	1.581	5.082	5,988,832	1.48
Capitella teleta	0.0286	0.0131	1.920	2.095	10,523,041	7.46
Amynthas gracilis	0.0408	0.0237	1.689	1.947	12,106,353	11.61
Helobdella robusta	0.0286	0.0137	1.614	2.671	8,079,707	3.36
Lumbricus rubellus	0.0313	0.0115	1.624	2.684	7,920,655	18.13

The final column in Table 5 also shows the percentage of the sequence structures which did not disperse within the tree and were negated by the terminal subtraction process described in 3.2.2. Interestingly the only non-annelid in the set has the lowest escaped frequency count by far, with *rubellus*, arguably the most difficult genome to analyse conventionally, showing a huge quantity of escape frequency – this suggests a very large number of either recent protein family expansions, or sufficiently divergent allelic copies, which might be right answer given the conclusions drawn in Chapter 3. For example, an >*K* Freq %-score of 50 could be achieved by every sequence being duplicated once identically.

The Pareto shape results are interesting for how consistent they are despite considerable changes across the rest of the scores. This indicative that the distributions of structure in these proteomes has a very steep pareto curve, to see simulated Pareto distributions which demonstrate the meaning of the shape parameters, see 3.4.7., and Figure 70.

Many of the other results in Table 5 are particularly informative when paired with the signature. Rather than describing the rest of Table 5 in depth independently, it will be frequently referred to in the following summary of the five 3D signatures. Each of the signatures also represent the post local-null correction.

The data representation provide in Figure 46, reveals further evidence of how the odd-one out in this set (in evolutionary distance) is also substantially difference in sequence structure. However, first it is also necessary to cover the usage of absolute and relative frequencies in this plot. The individual absolute categorical frequency density is coded to the red component of the pixels. Typically, we would expect the lower *l*-mer categories to be denser in frequency as the sequence space is substantially smaller, and the categories far fewer in number. However, towards the saturation depths the absolute frequencies will be lower for two reasons: 1) local null subtraction 2) absolute null correction, both of which are described in 3.2.6.

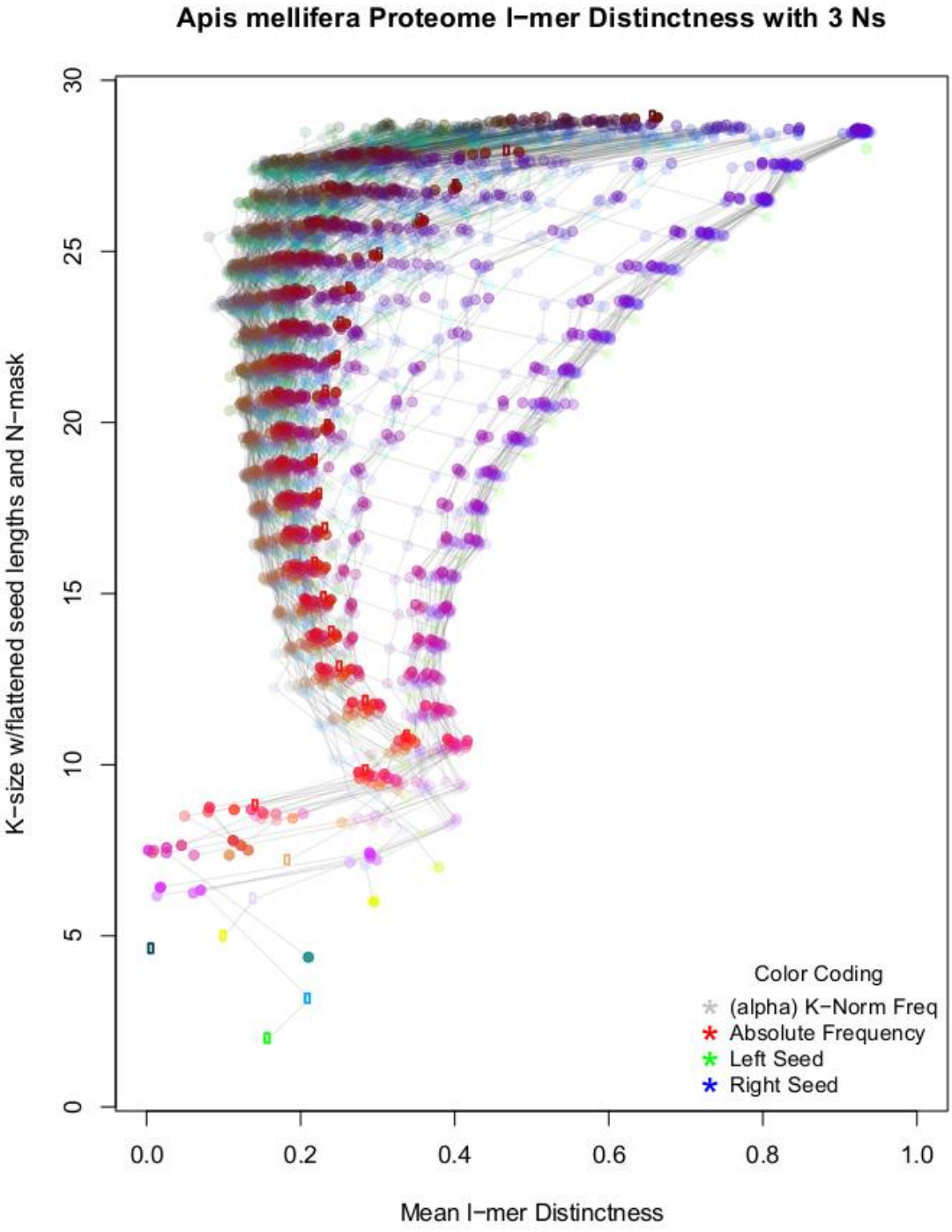


Figure 87. Apis mellifera full proteome 3D-signature.

Secondly, frequency scaling across a single depth has been coded to the alpha-value (transparency) of the point. This allows the user to see the relative quantities composing various features. For example, if we were to observe the faint single-thread proceeding from the left of the rightmost band from depths 11-28. It seems insignificant; however, this is likely the effect of one large, or a small number of similar domain types which have a very specific dispersal profile. Other similar

threads following their own patterns can also be seen further into the signature. One of the flaws of this visualisation methods is that much of the complexity discovered is packed into broad but dense bands of distinctness, making threads impossible differentiate, leaving only colour gradients as informative.

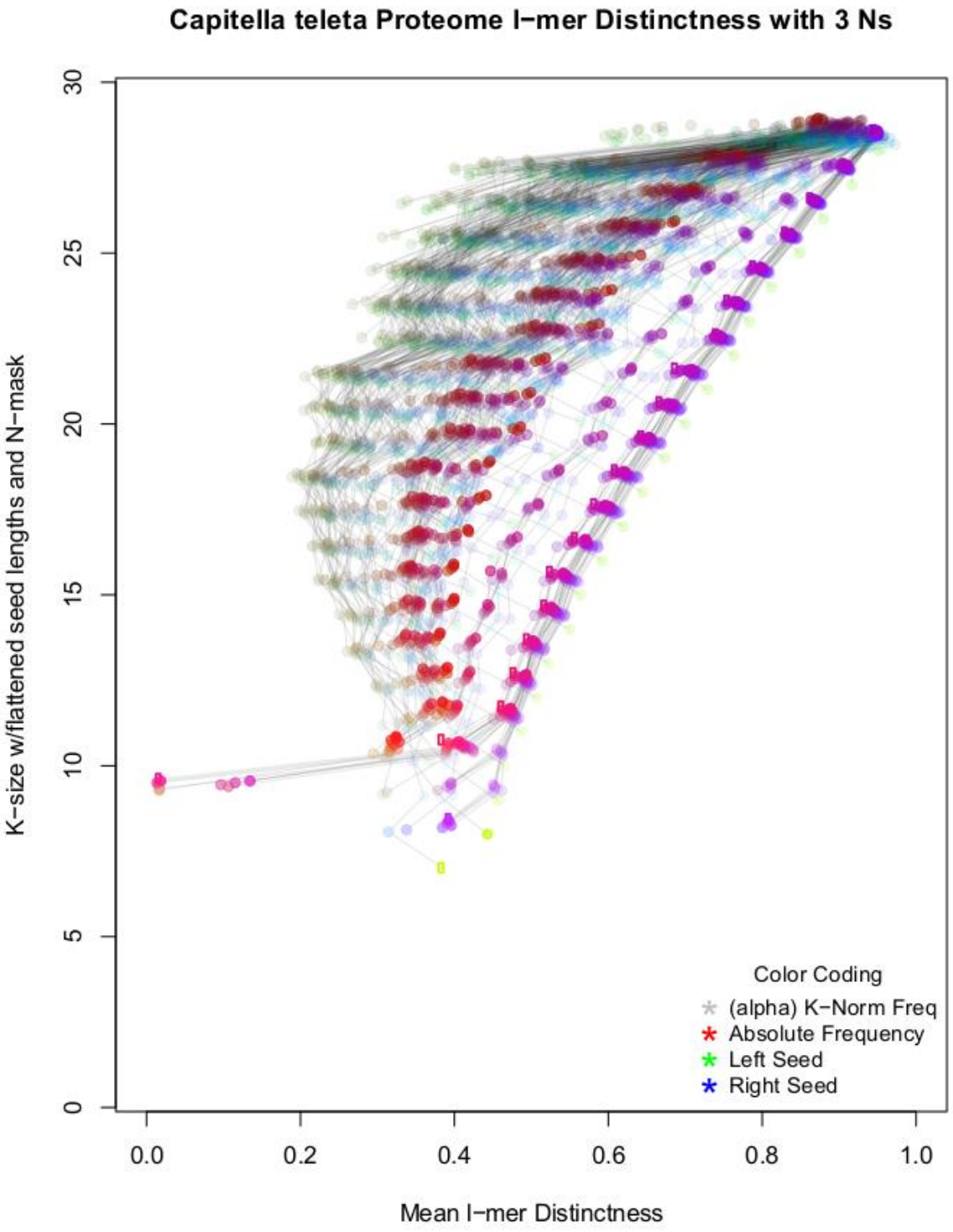


Figure 88. Capitella teleta full proteome 3D-signature.

Comparing Figures 47 and 48 shows the threading complexity which is visually collapsed in *mellifera* and still very hard to discern in *teleta.* Another feature which is particularly prominent in Figure 47 but is present in all proteome signature to some degree is the band separation between a narrower rightmost group, and a much broader leftmost group. Typically, the rightmost group is represented by categories with either a very long left or right seed, and perhaps only a single N in the mask. This can be thought of as the scaled modularity of the motifs which are the most resistant to variation. A rightmost band moving quickly towards 1 suggests a large portion of unique/non-duplicated sequence, or low frequency fragile long motifs. The width of the gap is perhaps more descriptive of distances between similar motifs groups, rather than the conservation patterns within the most uniform.

Returning to Figure 46, there are several datapoints which resonate in the interpretation from Table 5. The signature has the greatest tendency to curve towards 0 of all the proteomes, the tree structure summary is also the lowest of out the set, and when local-null corrected this difference only becomes more pronounced. It has the lowest rate of frequency escape, and the highest post-correction Pareto shape. What this says more broadly about the signature is that it is likely to have many smaller groups of internally homologous sequence motifs, rather than fewer larger groups (proportionally speaking). It is also the case that the most common protein domains are less likely to be disproportionately overabundant in the set. This could be summarised as a 'high complexity, low structure' proteome, insofar as structure is defined in terms of stacked sequence homologies.

Figure 47 and 48 both represent the tied 2nd most unstructured proteomes after *A. mellifera,* although *H. robusta* shows a much greater tendency towards the heterogenous structure dispersal that *C. teleta* this could reflect the higher shape score of *robusta. Capitella* also has a more differentiated set of deeper banding patterns across the categories of medium length left and right seeds, suggesting a wider variety of motif forms.

Looking at the earthworm plots (Figures 49 and 50), we can see a much greater range of banding patterns, particularly in the case of *L. rubellus* which seems to have a combination of a great many diverse smaller structures which can be merged into lower distinctness, higher frequency categories. Most interestingly the type of N-mask applied appears to have a very significant effect on the dispersal patterns, this might be suggestive of large-scale gene family expansions which diverged in several different ways and would also make some sense of the very high rate of frequency escape in Table 5.

One caveat to the 'banding spread equals diversity' argument is that dispersal patterns generating similar levels of distinctness need not originate from the same type of structure, it is only more

apparent when they are more spread out. Additionally, within the narrower, denser bands of many threads there may also be structure which is simply too densely arranged in these plots to be discernible. For this reason, additional visualisations have been generated, using a z-axis to expand these thicker leftmost banding patterns. Figures 51 and 52 show two versions of this additional dimensionality. They have the advantage of separating thick bands when rotated suitably, but the simultaneous disadvantage to obscuring other parts of the plot. To present the data in another form which attempts to make maximum advantage of the 3D plot, a series of animated rotations of the 3D plots for each image have been produced.

Appendix 2.4 contains ANIMATION2-11, which display the five proteome signatures rotating around multiple axes, both with and without 'thread' lines.

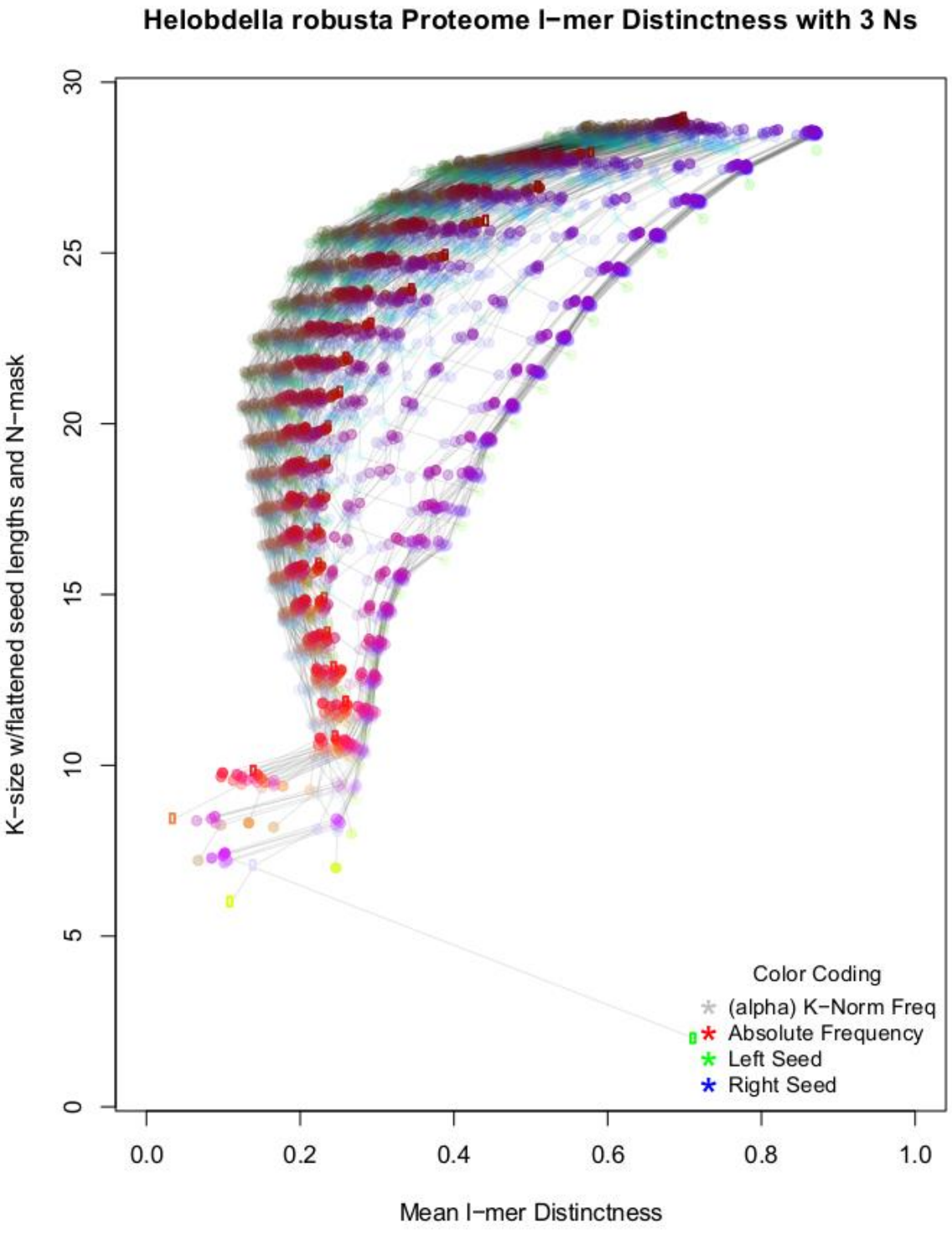


Figure 89. Helobdella robusta full proteome 3D-signature. N=3.

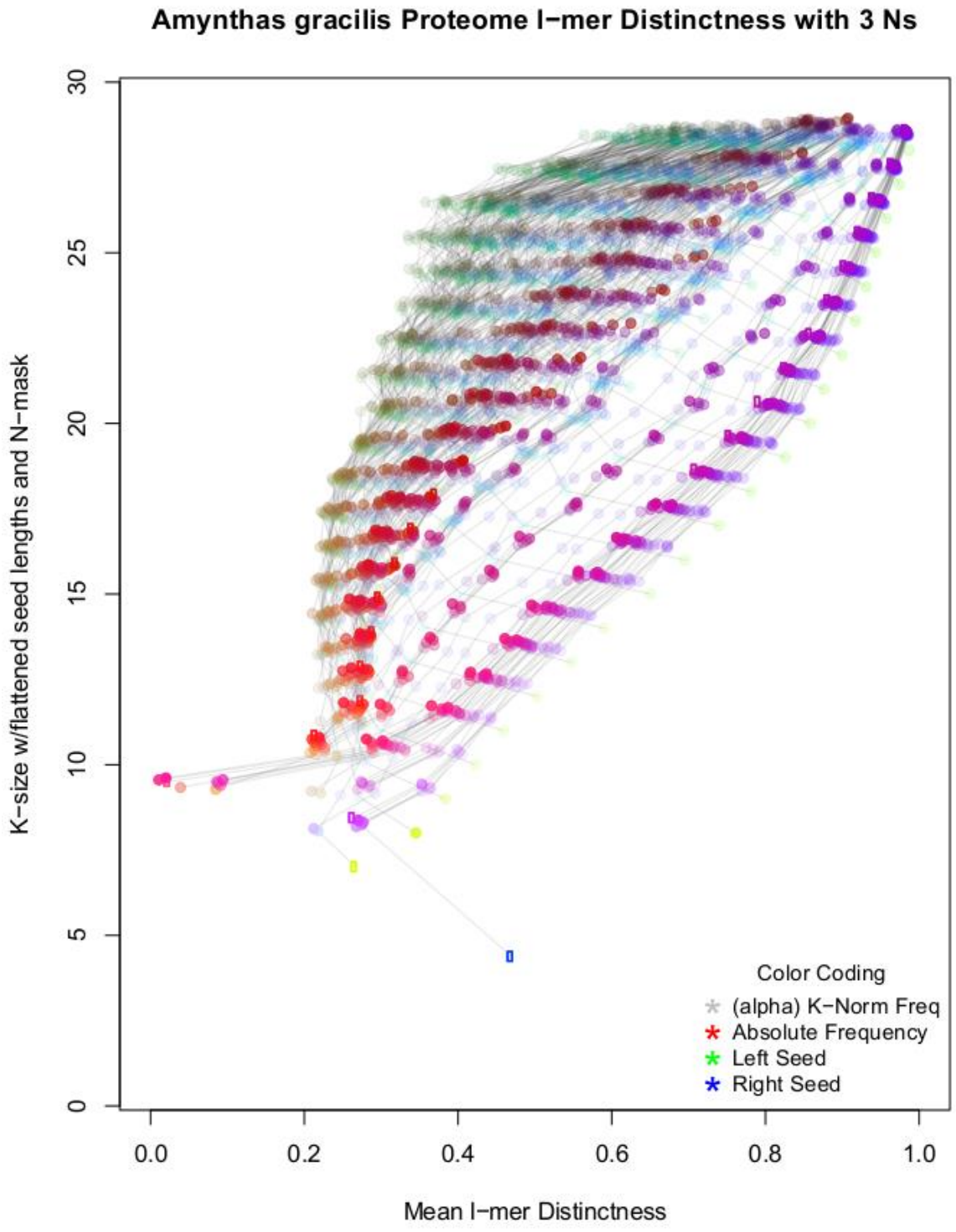


Figure 90. Amynthas gracilis proteome signature. N=3.

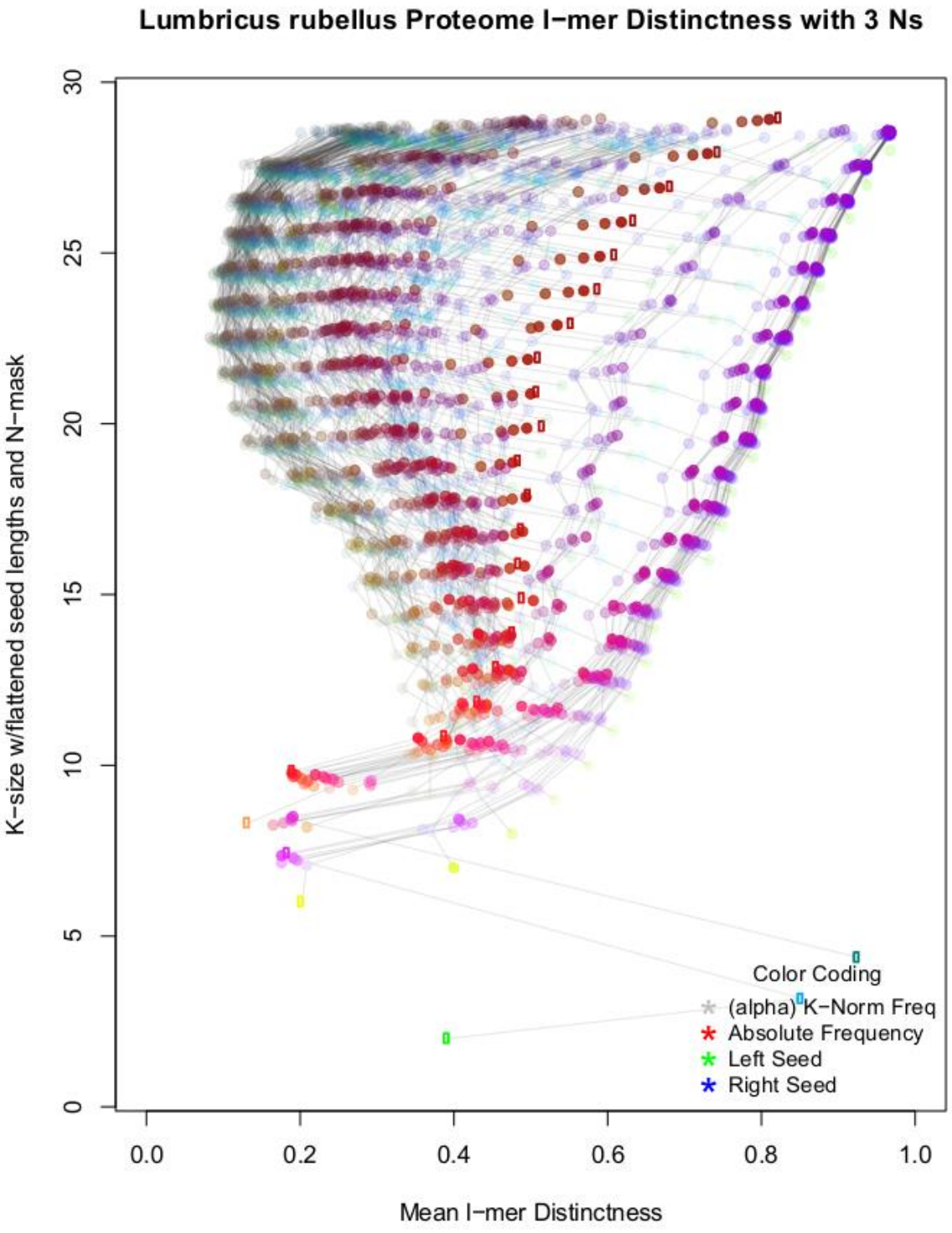


Figure 91. Lumbricus rubellus full proteome signature. N=3.

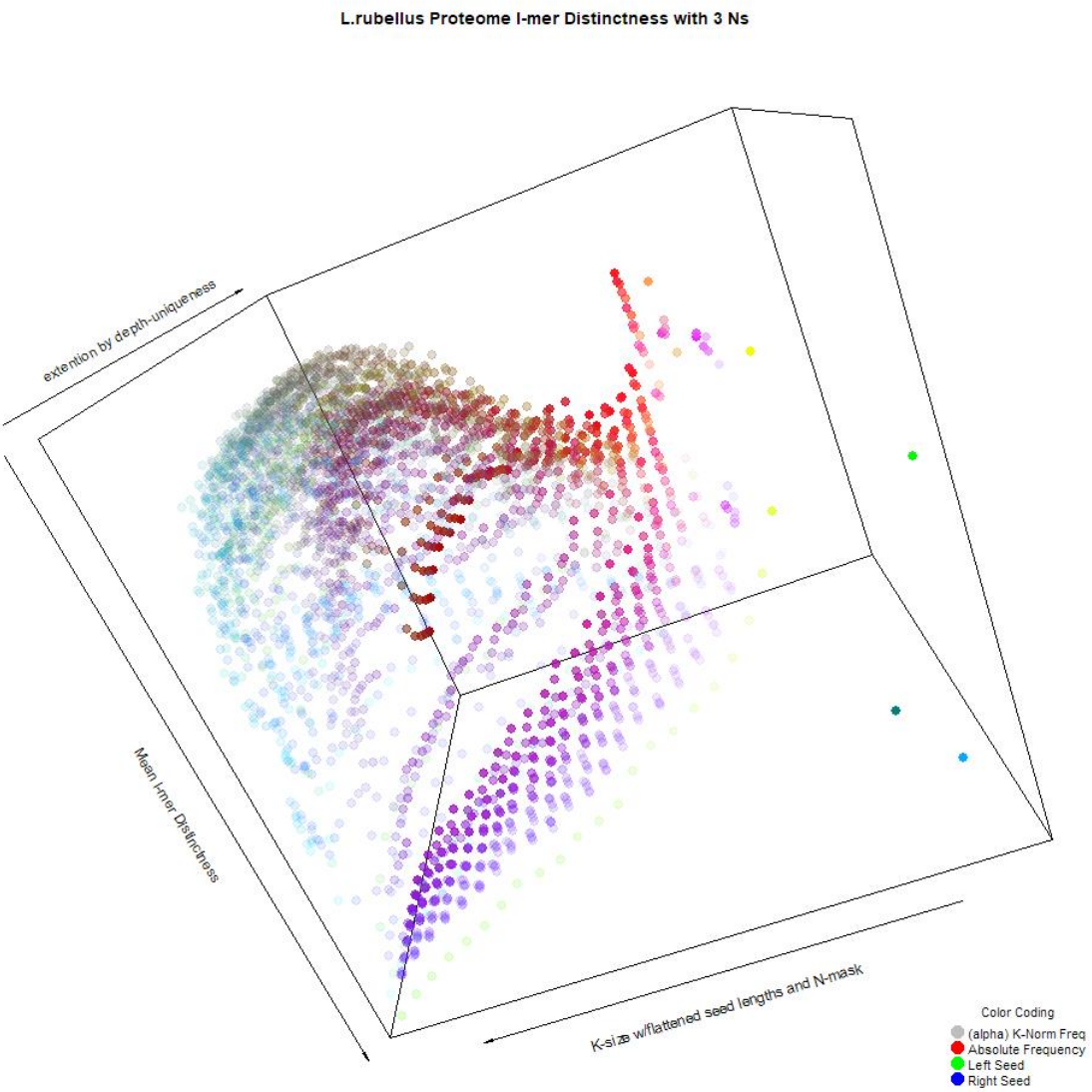


Figure 92. Lumbricus rubellus full proteome signature, alternative visualisation. N=3.

The 3D plots produced of these signatures are primarily illustrative of the depth of complexity discovered by this method. It is relatively difficult to compare between them due to the rotation-occlusion issue. Adding the thread lines, as in Figure 52, makes them particularly dense.

In summary, the proteome test set was able to demonstrate a wide variety of signatures, with key correlates between the singular tree summaries, and the patterns found in the signature graphs. The visualisation density issue is still a limiting factor on the user's interpretation, however there are also higher perspectives in the interpretation which don't always require the discernment of every single category's distinctness, or it's thread pattern.

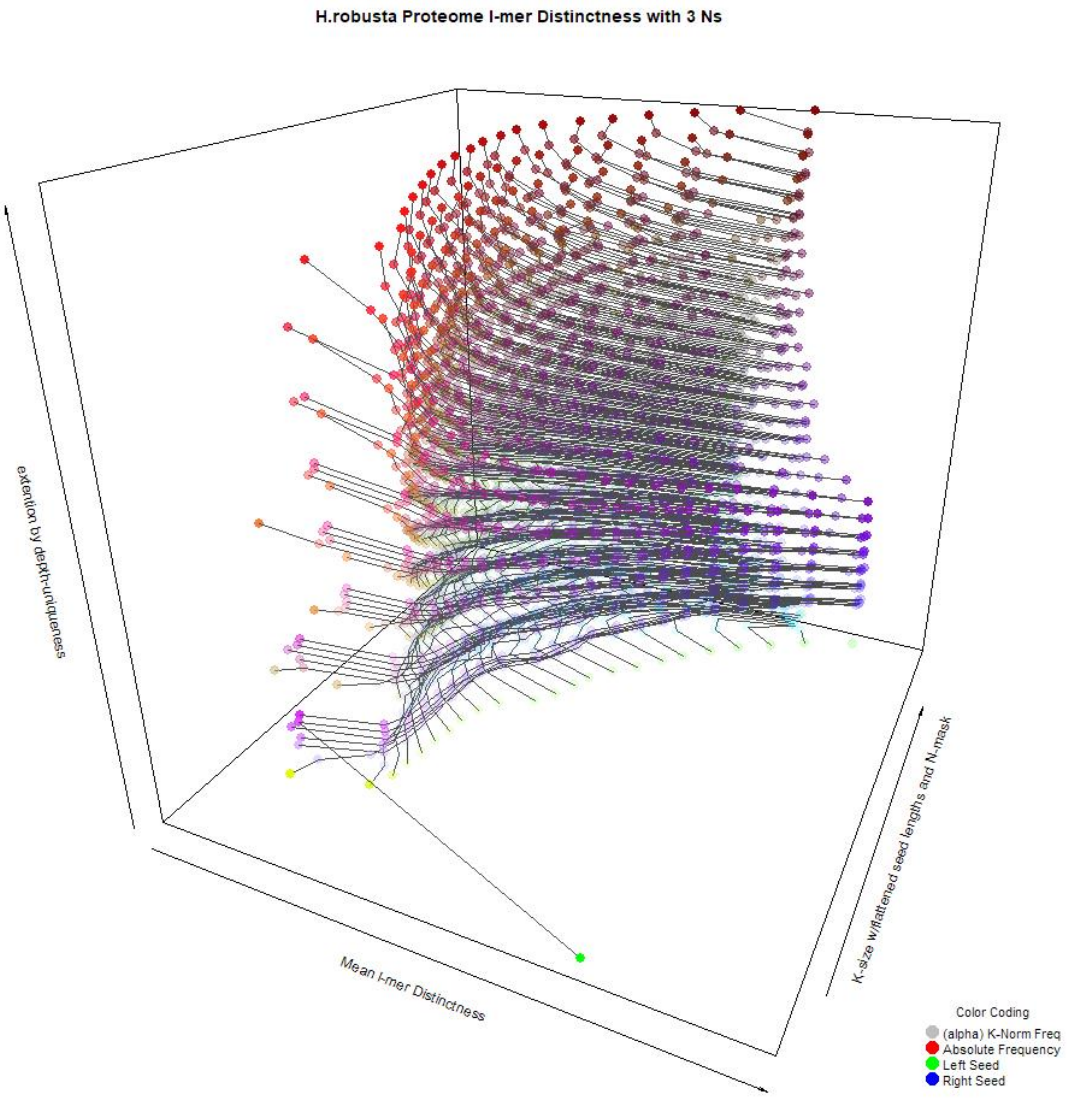


Figure 93. Helobdella robusta full proteome structure, alternative visualisation with threads. N=3.

4.4.5. Test Set: *E. coli* Genome Signatures

The second test set involves the signatures from the DNA of 18 *E. coli* genomes, retrieved from NCBI Genome database (NCBI 2016). The genomes were sampled from six of the seven major phylogroups as defined by the Clermont typing method (Clermont et al. 2013). Group C was only excluded due to data quality/availably issues. The phylogroup selection was applied with the intention of viewing the range of signatures across the broadest range of genomes available within the restriction of a single species. This serves as a counter-point to the previous test, which reached across hundreds of millions of years of evolutionary time. Here we investigate the variability of signatures within a tightly restricted set – to see if it might be informative, and to see the visual differences of with relatively small changes in input.

Table 14. E. coli, 18 genome structure summary scores.

E. coli Test

Strain	Phylogroup	Structure	Structure - Sub	Shape	Shape - Sub	Size (bp)	>K Frequencies (%)
S88	B2	0.1411	0.0378	1.290	3.453	5,166,121	1.85
LF82	B2	0.1400	0.0375	1.289	3.645	4,773,108	0.99
E2348/68	B2	0.1399	0.0369	1.289	3.635	5,069,678	2.67
SMS-3-5	F	0.1409	0.0373	1.286	3.692	5,215,377	1.53
IAI39	F	0.1405	0.0372	1.284	3.607	5,132,068	4.33
B093	F	0.1406	0.0367	1.280	3.804	5,205,351	0.99
TA280	D	0.1412	0.0371	1.284	3.721	5,296,938	1.32
H299	D	0.1417	0.0380	1.287	3.392	5,317,840	1.53
UMN026	D	0.1419	0.0380	1.287	3.491	5,202,090	1.73
ECOR31	E	0.1416	0.0371	1.273	3.661	5,443,045	3.20
B185	E	0.1415	0.0379	1.282	3.556	5,144,306	0.91
E101	E	0.1418	0.0375	1.283	3.499	5,181,904	1.53
_55989	B1	0.1403	0.0372	1.285	3.705	4,989,876	0.45
IAI1	B1	0.1399	0.0379	1.298	3.486	4,700,560	1.81
O111	B1	0.1424	0.0386	1.289	3.301	5,284,381	4.41
HS	A	0.1398	0.0380	1.299	3.434	4,643,538	2.13
ATCC-8739	A	0.1409	0.0389	1.297	3.272	4,746,218	1.99
TA007	A	0.1417	0.0378	1.286	3.364	5,299,319	2.36

Table 6, like Table 5, shows the range of structures and shapes across the test set. Perhaps as expected the structure and shape scores for all entries are highly consistent. More interestingly, the escaped frequency rate remains quite variable from 0.9-4.4%. The phylogroup categories did not have any significant correlation with any of the scores. This could be indicative of the substantial genomic variation present within phylogroups. Additionally, the signature and shape scores are intended as indicators of sequence set structure and complexity rather than evolutionary distance. Any distance between genomes describable by these scores could be thought of more as a biological architecture distance, which is only tangentially related to evolutionary time.

All the signatures generated by this test are available in Appendix 2.2 in the file E.coli_signatures, and additionally presented in series as a short .gif as ANIMATION1, in Appendix 2.4.

To demonstrate the effects of applying the local-null correction to the signature Figures 53 and 54 were created, pre- and post-correction. The main difference is the removal of most of the pre-saturation (~11.6) points. This shows that there was no over-abundance of shorter oligomers which wasn't also emergent in the random-shuffled input.

If there is a trend which Figures 54-57 follow, it is one of similarity to the DNA null-curve. The aspects of the small subset set of *E. coli* DNA which began to emerge as different to the null-curve are exaggerated in scale, but the transformation of signature shape is nowhere near as dramatic as in the proteomes. Additionally, whilst the structure scores in the Table 3 are significantly higher than Table 2, they are not directly comparable, as structure is always measured relative to sequence space occupation only equivalent alphabets may be compared numerically without additional transformation. The coherence to the null curve suggests that these DNA inputs were highly complex, and relatively low in highly duplicated structures. The corrected distribution shapes were also generally higher than most of the proteomes, except *Apis mellifera*, which also bore the most similarity to the small input subset, and to the peptide null curve more generally.

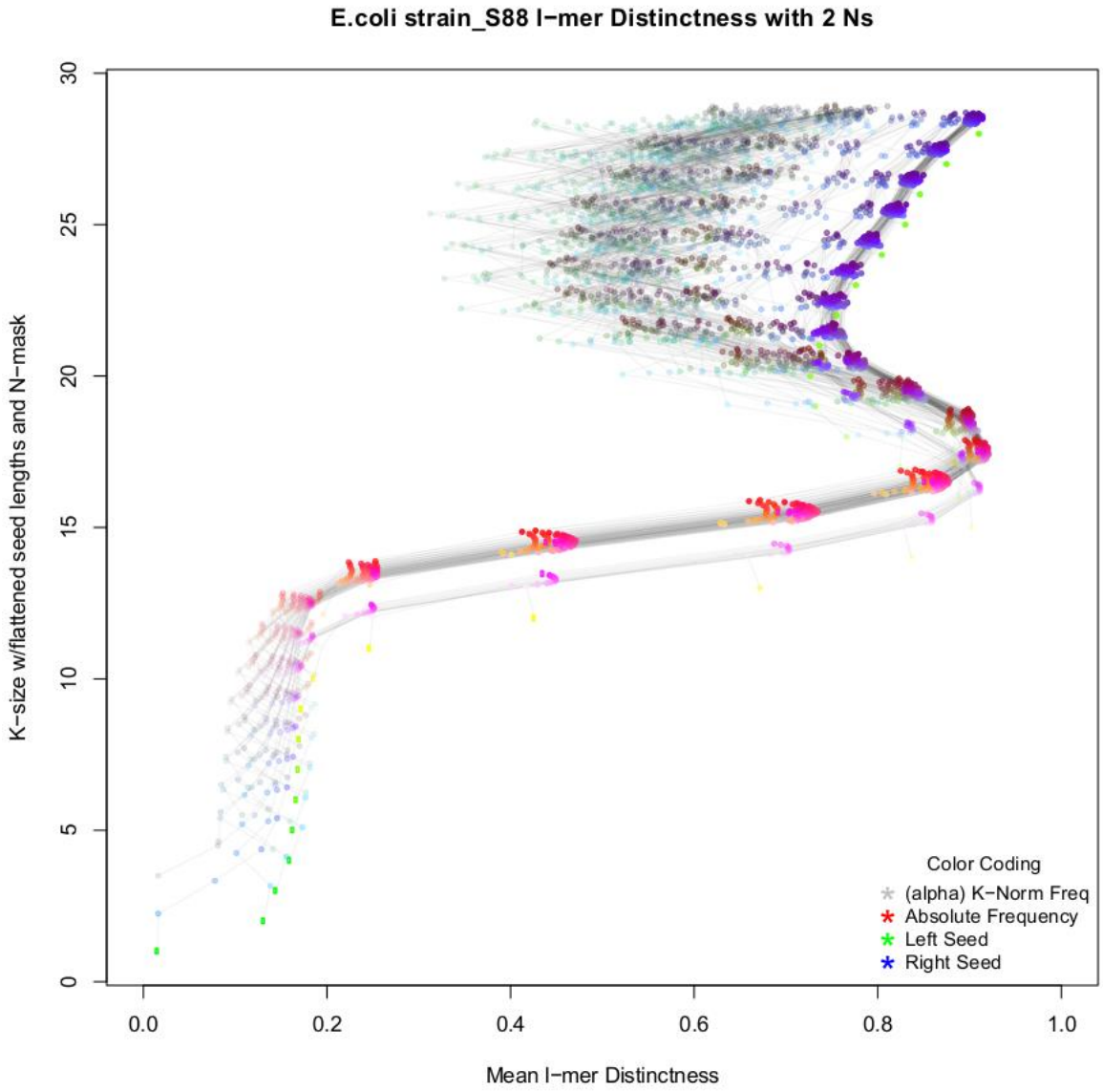


Figure 94. E. coli strain: S88 full genome signature. N=2. Without local-null subtraction.

This relatively low structural scale could be reasonably expected in bacterial genomes that usually have fairly small gene families, with many genes being unique single copies (Pushker et al. 2004). Still there are differences between the strains which may highlight their evolutionary behaviours. For example, despite being very closely related in structure scores, S88 and B185 (Figures 54 and 55) have quite a marked difference in the frequency densities of short left and right seed N-masks across the *l*=20-30 range, suggesting a pattern of motifs with multiple mutations separated by ~15+ bases are far more common in B185.

A similar comparison is possible between strains HS and O111 (Figures 56 and 57), the highest and least structured entries in the table respectively. However, in this case we can see the scale of the structures reflected also in the convexity of the curve of the rightmost band. Interestingly part of the

signature of higher overall structure manifests as less distinct primary sequence threads in these cases.

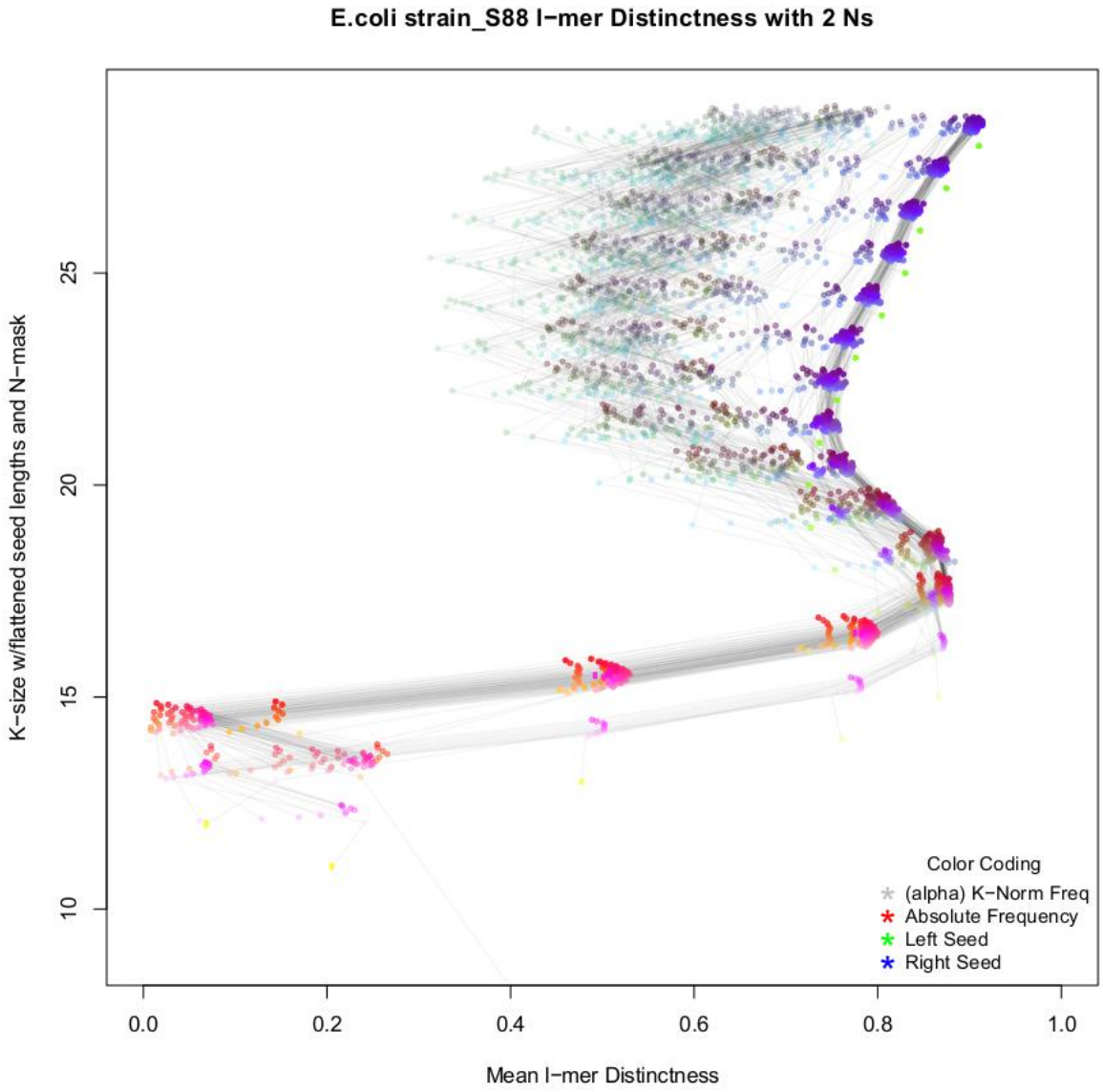


Figure 95. E. coli strain: S88 full genome signature. N=2.

The original conception of the *k*-mer tree signature method was to describe the structures in large and complex genomes, however given the current memory and performance limitations of the software, smaller bacterial genomes were chosen. When the signatures are expanded in 3D plots (see Figure 58), the patterns do not expand to a greater depth of complexity and remain very similar at different depths of the z-axis. Further performance gains must therefore be made before the DNA tree signatures can be suitably refined for the intended 0.5 - 1Gb genome inputs.

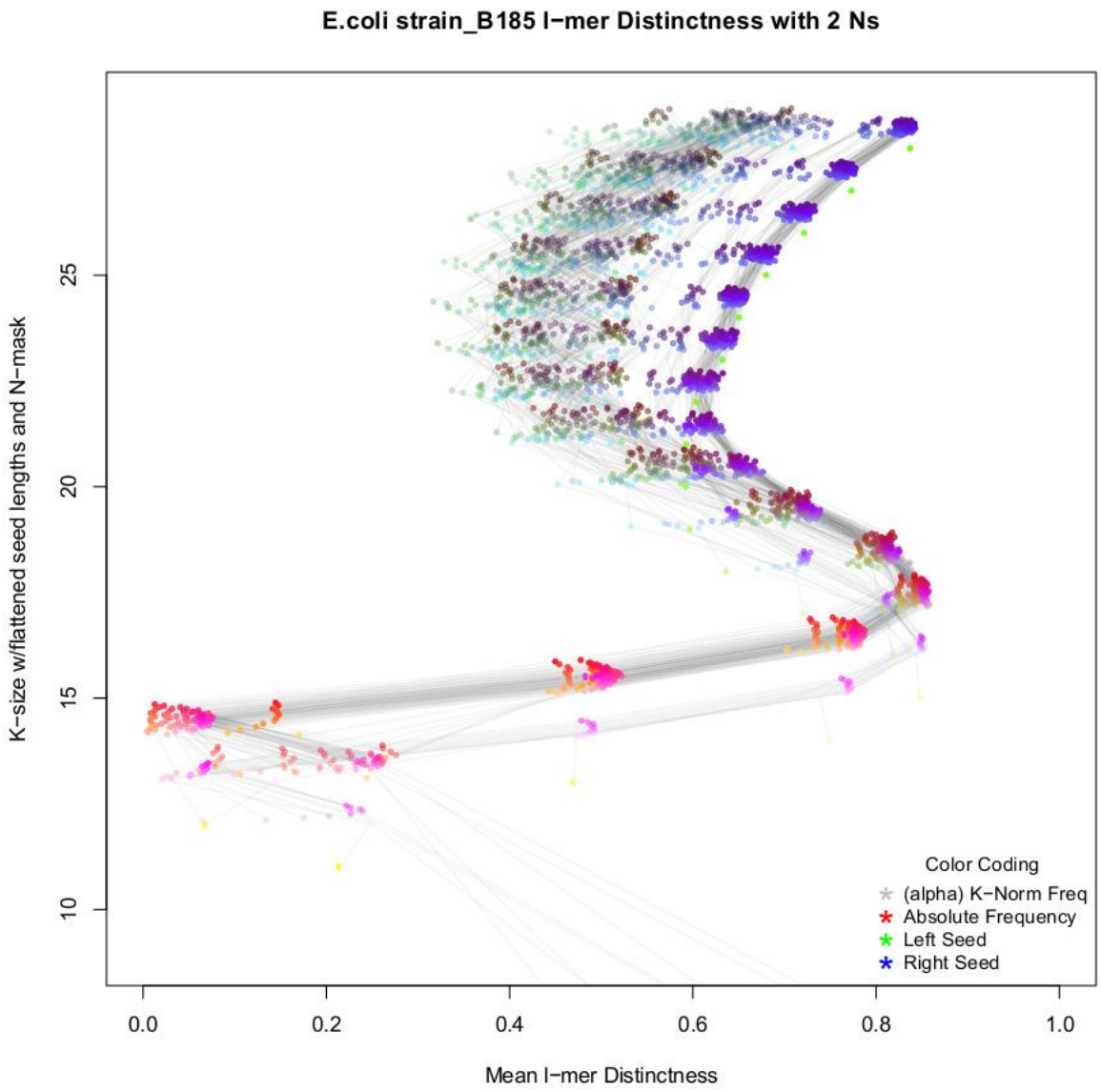


Figure 96. E. coli strain: B185 full genome signature. N=2.

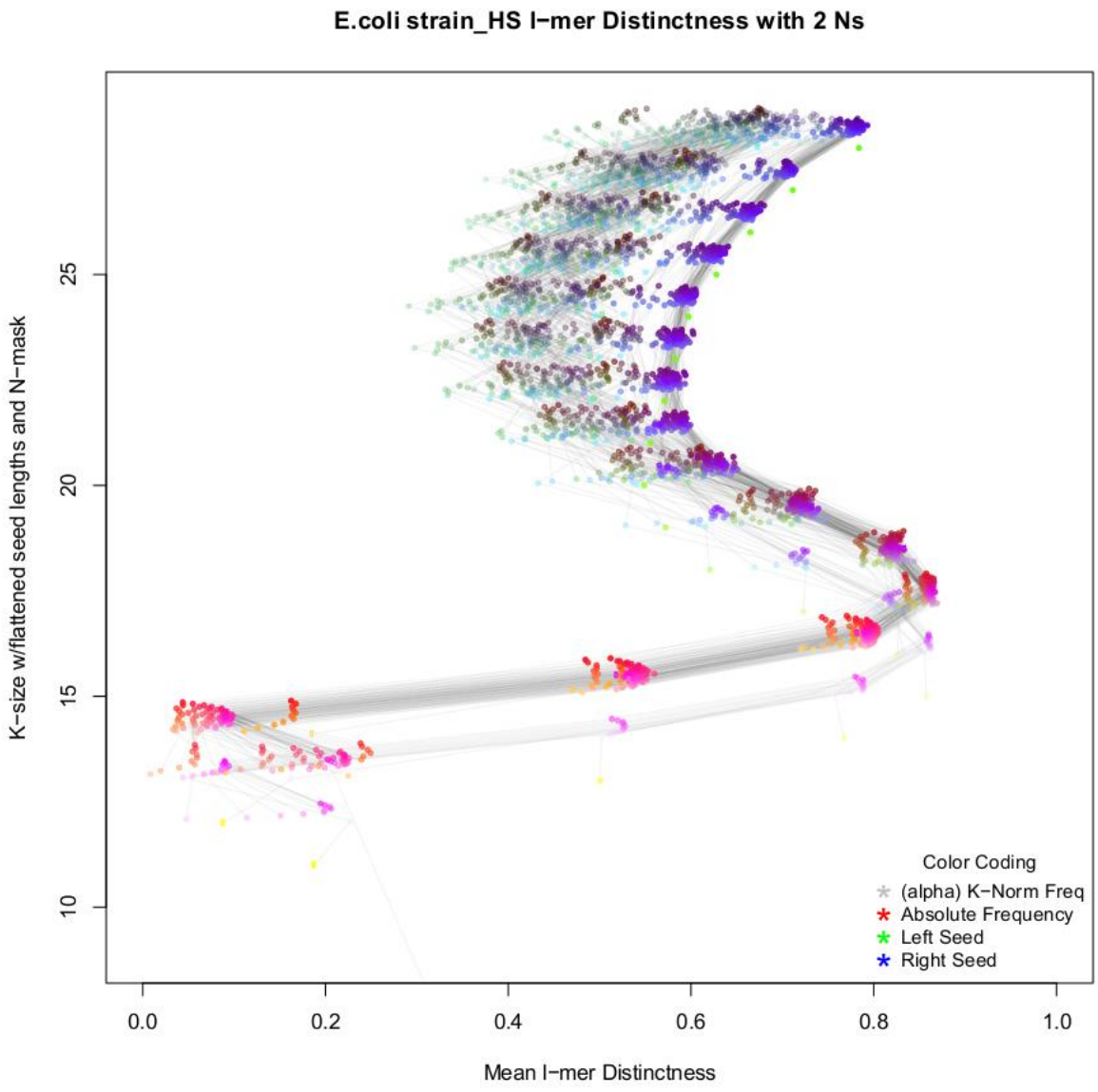


Figure 97. E. coli strain: HS full genome signature. N=2.

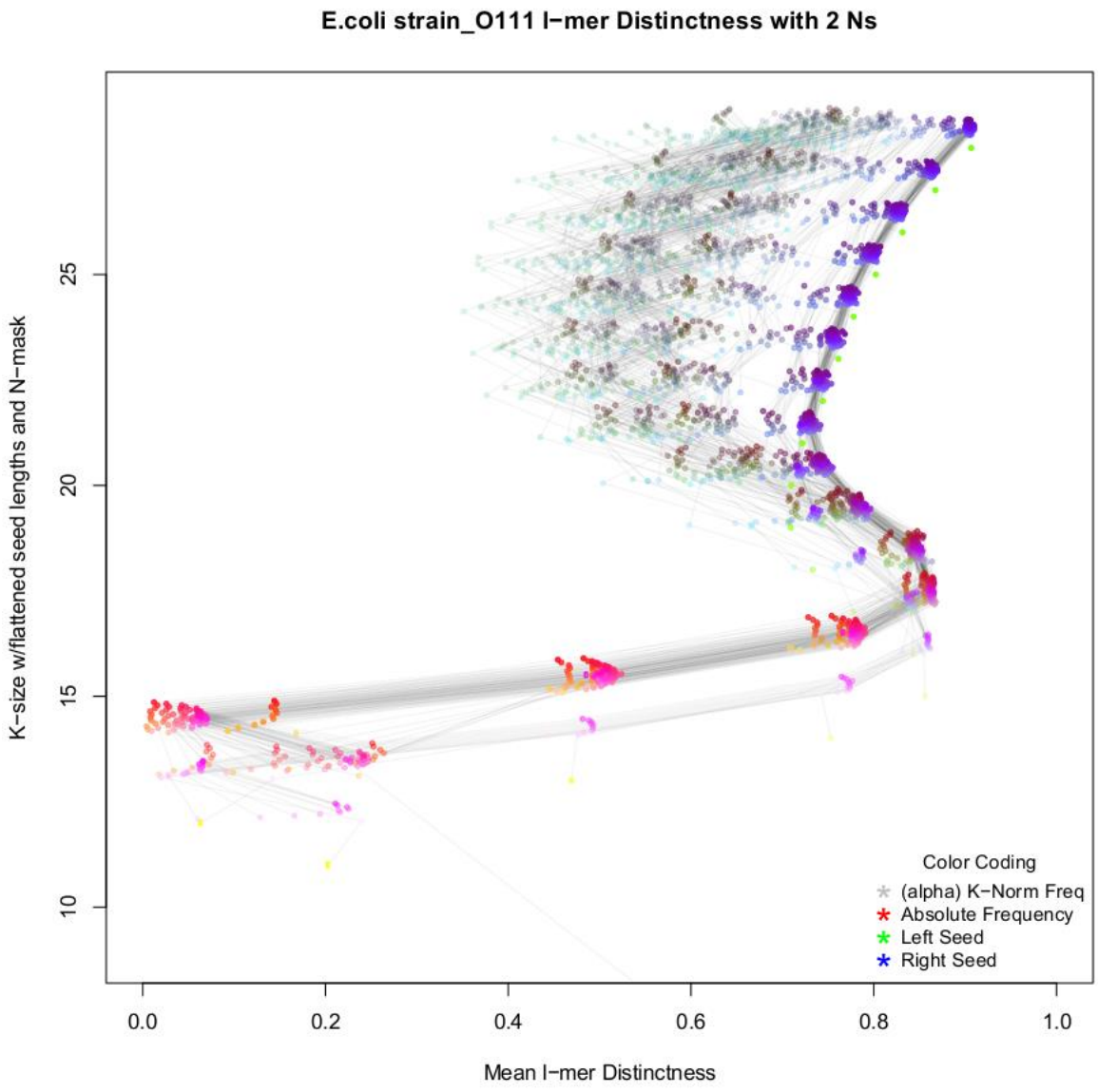


Figure 98. E. coli strain: O111 full genome signature. N=2.

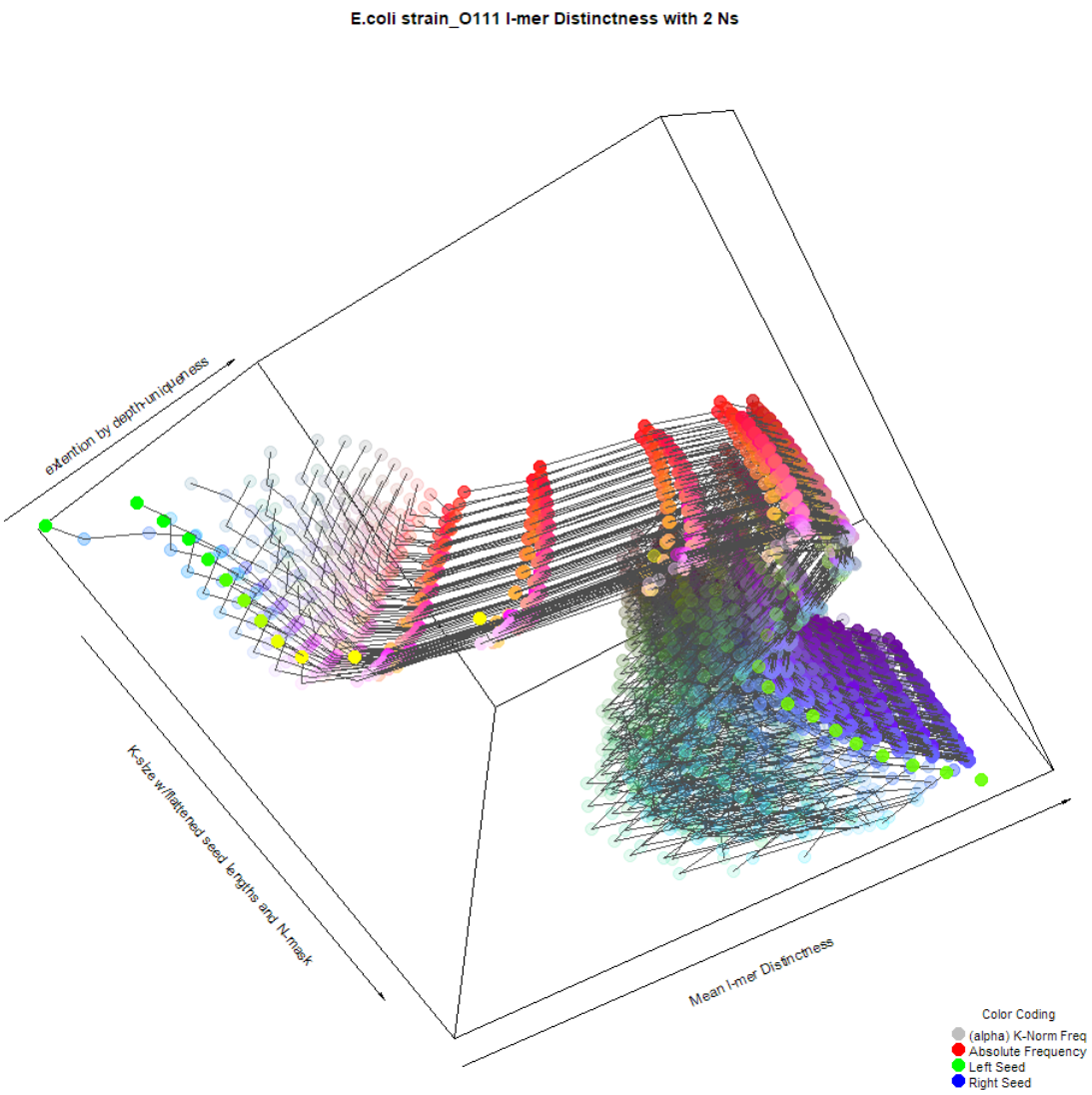


Figure 99. E. coli strain: O111 full genome signature. N=2. 3D visualisation.

4.4.6. Test Set: Protein Families

The third test set was intended to test the program's ability to describe smaller, yet highly structured datasets. Protein families were expected to satisfy this criteria due to the anticipated number of repeated domains in the input set. The input sets were retrieved from PFAM (Lee et al. 2015) ftp server. Due to the extreme size range in the protein family reference sets (100K – 5M), some files were limited to the top 10K lines. To maximise the utility of this test, the protein families selected were identical to the six highly allelically divergent environmentally adaptive families identified in both *Lingula anatina* and *Lumbricus rubellus* in Chapter 2. Of interest is Chapter 2, Figure 21, which shows the variable rates and distributions of allelic divergence amongst them. The hypothesis being that the rates of evolutionary divergence between alleles may have some correlate in the signatures.

Mucin-like glycoproteins were identified as being the more divergent group, followed by ZIP metal transporters. Interestingly, it appears that in Table 7 these two also the highest structure scores, with mucins coming out as the most by far the most structured. The two least divergent families were Glucuronosyltransferase and GPCR Chemoreceptors, and again the extremes align in reverse, with GPCRs achieving the lowest structure. Although this is not a statistically valid proof of allelic divergence and structure correlation more broadly, it does appear to have some interesting intersection in this case.

Table 15. Protein Families Structure Summaries

Protein Family	Structure	Struct. - Sub	Shape	Shape - Sub	Size (aa)	>K Freq (%)
Epithelial Sodium Channels	0.0461	0.0393	1.765	1.858	423,109	7.14
Glucuronosyl-transferase	0.0719	0.0636	1.606	1.754	384,828	2.12
GPCR Chemoreceptors	0.0504	0.0391	1.625	1.888	301,048	6.20
Laminins	0.0494	0.0414	1.630	1.922	259,950	3.01
Mucin-like Glycoprotein	0.1914	0.1774	1.561	1.690	119,725	11.33
ZIP Metal Transporters	0.0886	0.0751	1.625	1.776	380,509	6.38

This test set will also offer the chance to demonstrate the first derived measurement type, the WSD of category distributions (see 3.2.5.). Weighted deviations are more descriptive in the case that a specific set of structures are in question, as the can reveal the extent to which a category represents a singular feature of the protein family. Here the inverse scale of the WSD has been coded to the size of the points used to show each category. The WSD scale has also been normalised to the maximum

per depth, this ought to help combat the effect of WSD always shrinking towards distinctness boundaries, however this effect does persist. To be clear, the smaller the WSD, the narrower the distribution, the larger the point will be drawn on the plots in Figures 59-66, on a linear scale.

The information which can be gleaned simply from the WSD component of the signature is demonstrated by several comparisons within these images. Firstly, looking at the differences between epithelial sodium channels and Glucuronosyltransferase (Figures 59 and 60), there is only particular WSD pattern which stands out. This is the ~0.3 distinctness 'backbone' band between depths 9 and 16 found in Figure 60. Whilst both plots show a typical pattern of high deviation throughout the middle of the plot, suggesting most component signatures found in this range are a diverse structural mix, the 0.3 band feature in Figure 60 suggests a specific consistency to the dispersal patterns within that range, perhaps indicative of a conserved active domain, or conserved motifs within them. Figure 59 by contrast has far more 'distinctness outliers' generated by low-frequency N-masks with relatively large left and right seeds – these being depth-specific points which to not cohere to banding patterns. This suggests the presence of rarer variants present within motifs already typified by more regular variation patterns; smaller groups of domains which break away from the main set in an unusual manner. Given the breadth of the Epithelial Sodium Channel family, and the variety of sub-groups within it, this could be expected (Hanukoglu & Hanukoglu 2016).

All the protein family signatures generated by this test are also available as 3D visualisations in Appendix 2.3.

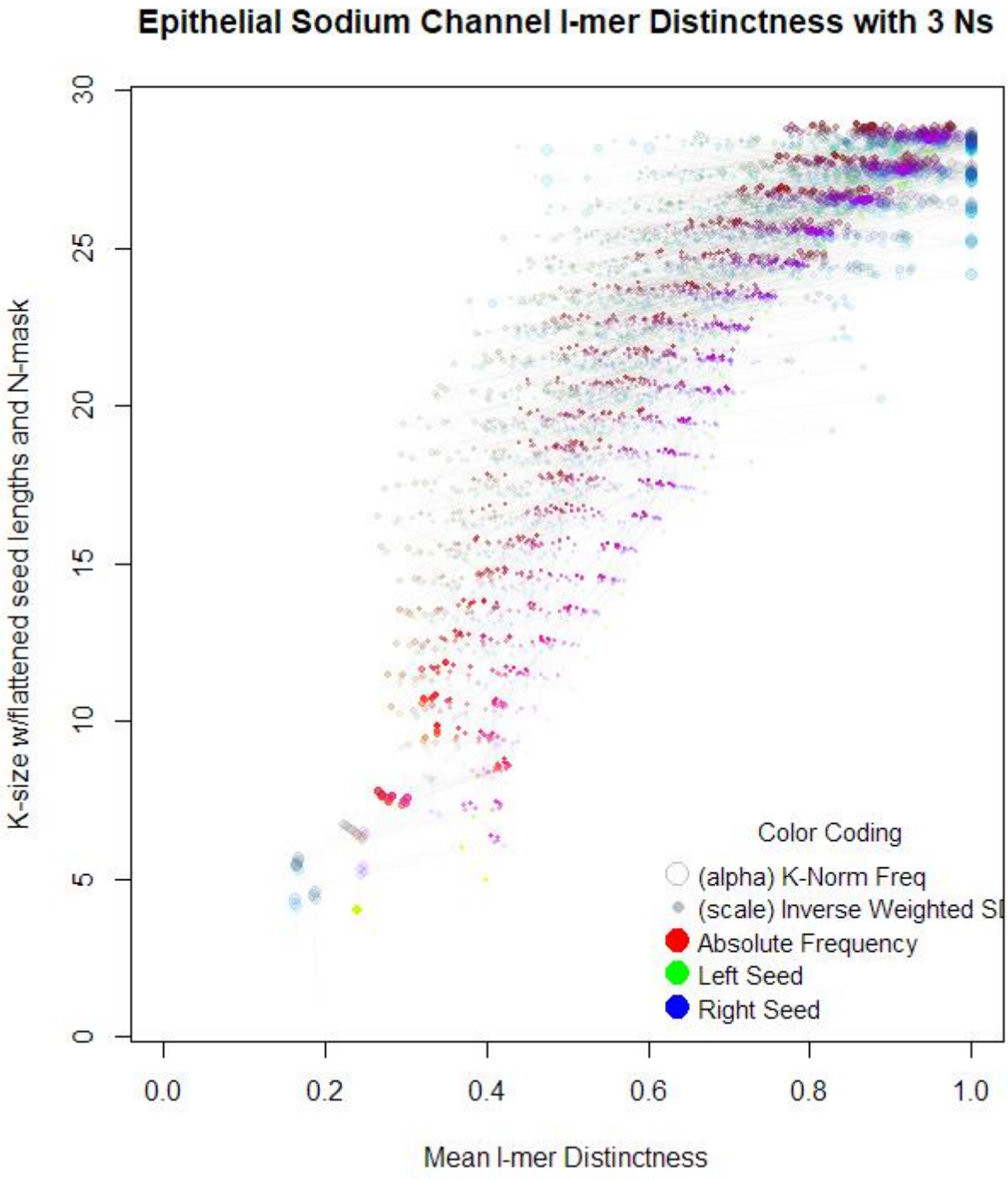


Figure 100. Epithelial Sodium Channel, (PFAM) Protein Family, Signature with WSD. N=3.

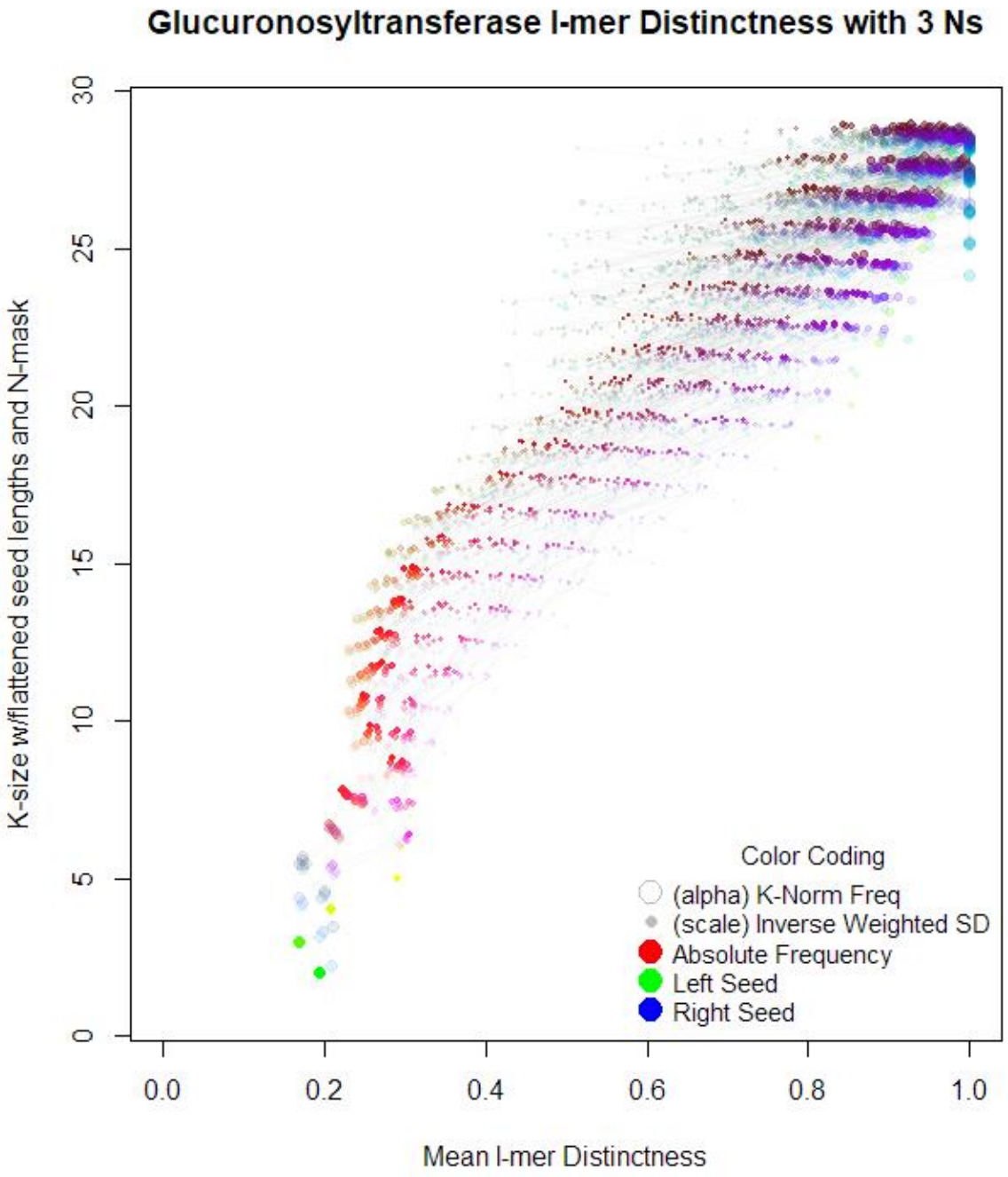


Figure 101. Glucuronosyltransferase, (PFAM) Protein Family, 3D Signature with WSD. N=3.

Comparison between Figures 60 and 61 may also be illustrative of the WSD signatures. It seems that the GPCR Chemoreceptor family, although considerably less structured than Glucuronosyltransferase in Table 7, has far narrower categorical distributions of dispersal type across the entire signature, although as in Figure 59, there are also many distinctness-outliers present. This may also be a commonality of membrane bound proteins with active sites, although it suggests GPCRs as a family have more homogeneity in their predominant variant patterning. A final point of comparison between them is the shape and distinctness position of the lower half of the signature. Although they both trend similarly, Figure 60 shows a more concave shape, whilst 61 is more convex. From this we can also infer than the flexible AA positions in GCPR Chemoreceptors may also be functionally more restricted to a certain set of replacements. Given that GCPRs are known to possess seven membrane spanning a-helices (Hollenstein et al. 2014), this could be a signature of the importance of hydrophilic/lipophilic AA restriction at regular helical sequence positions. Given that the slight convexity is also present in Figure 59, also representing a protein with membrane-spanning domains, this could be a more general signature of that attribute.

Returning to Chapter 2 Figure 21, and the mystery of the hyper divergent Mucins, we can now compare its signature (Figure 63) to the rest of the set. Remarkably, it is incredibly different. In addition to having a less structured tree summary, it also presents a signature far closer to the small subset curve than any of the other family signatures. Additionally, like Figure 60, although to a much greater degree there is a very low deviation dispersal pattern for very low left/right seeds along the leftmost band, reaching all the way up to depth ~25. This suggests a very large and consistently heterogenous variation pattern for most of the sequence content in these proteins. It has been observed that typically only the terminal domains in mucin-like proteins are conserved between species, whilst the central, threonine rich region, is made up of many tandem repeats whose primarily function appears to become highly glycosylated, thus creating the hydrophilic properties required to form gels or mucus (Acosta-Serrano et al. 2001). It seems that the signature of this large highly variable central domain is dominating the protein family signature and is responsible for the huge sequence variation seen in Chapter 2. We can also suggest that the propensity towards many repeats within the protein is the main constituent factor in the higher tree structure scores.

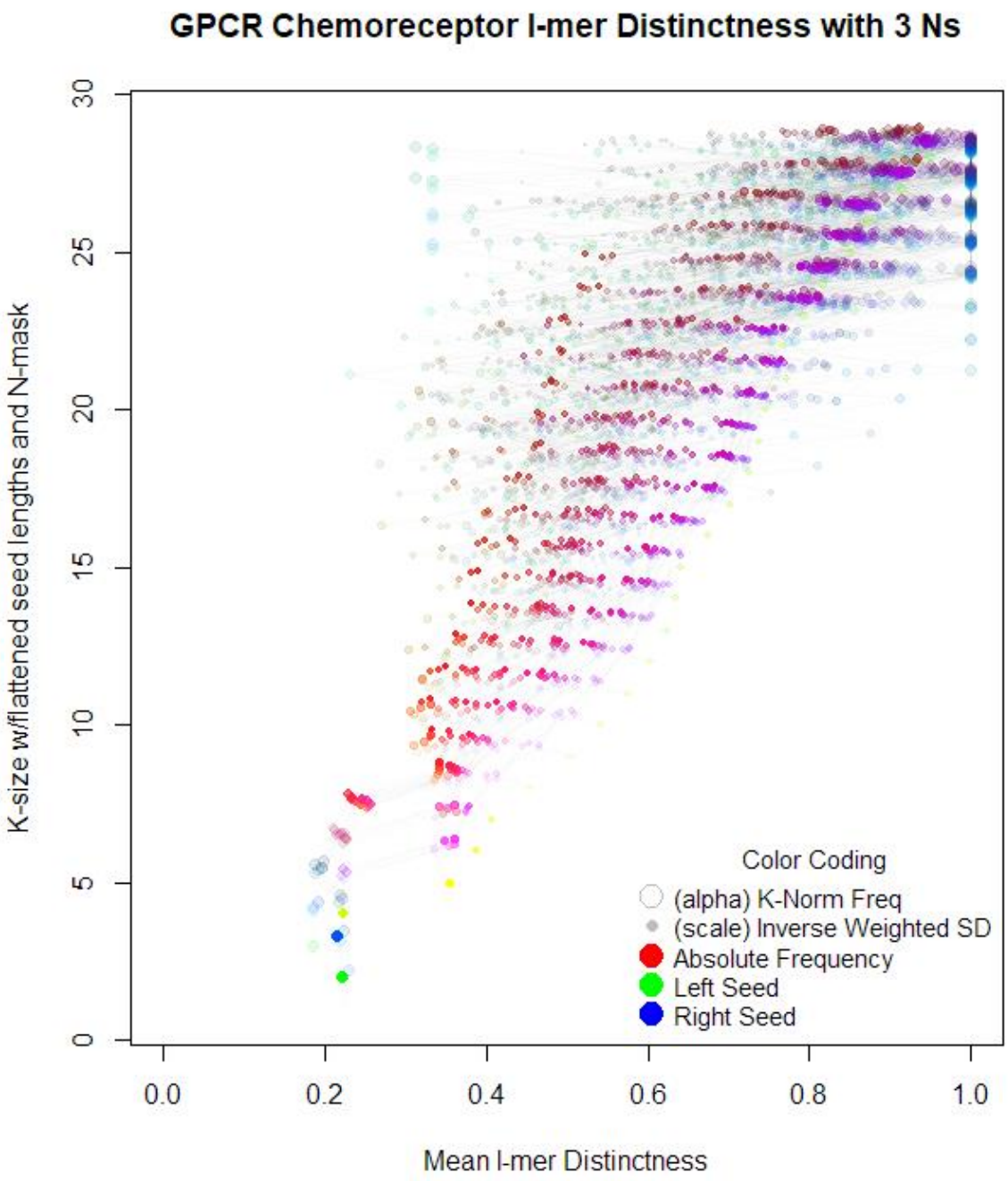


Figure 102. GPCR Chemoreceptors, (PFAM) Protein Family, 3D Signature with WSD. N=3.

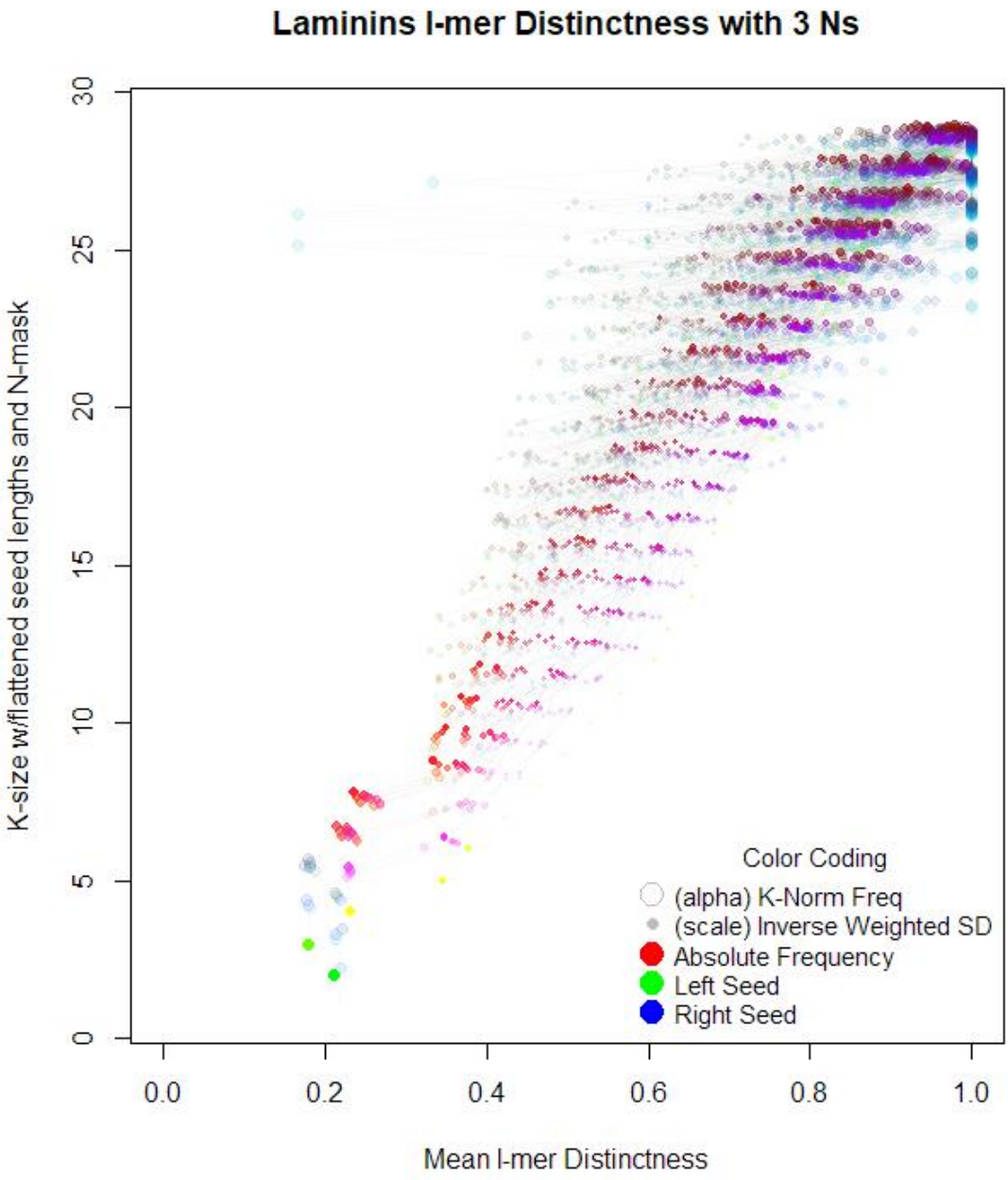


Figure 103. Laminins, (PFAM) Protein Family, 3D Signature with WSD. N=3.

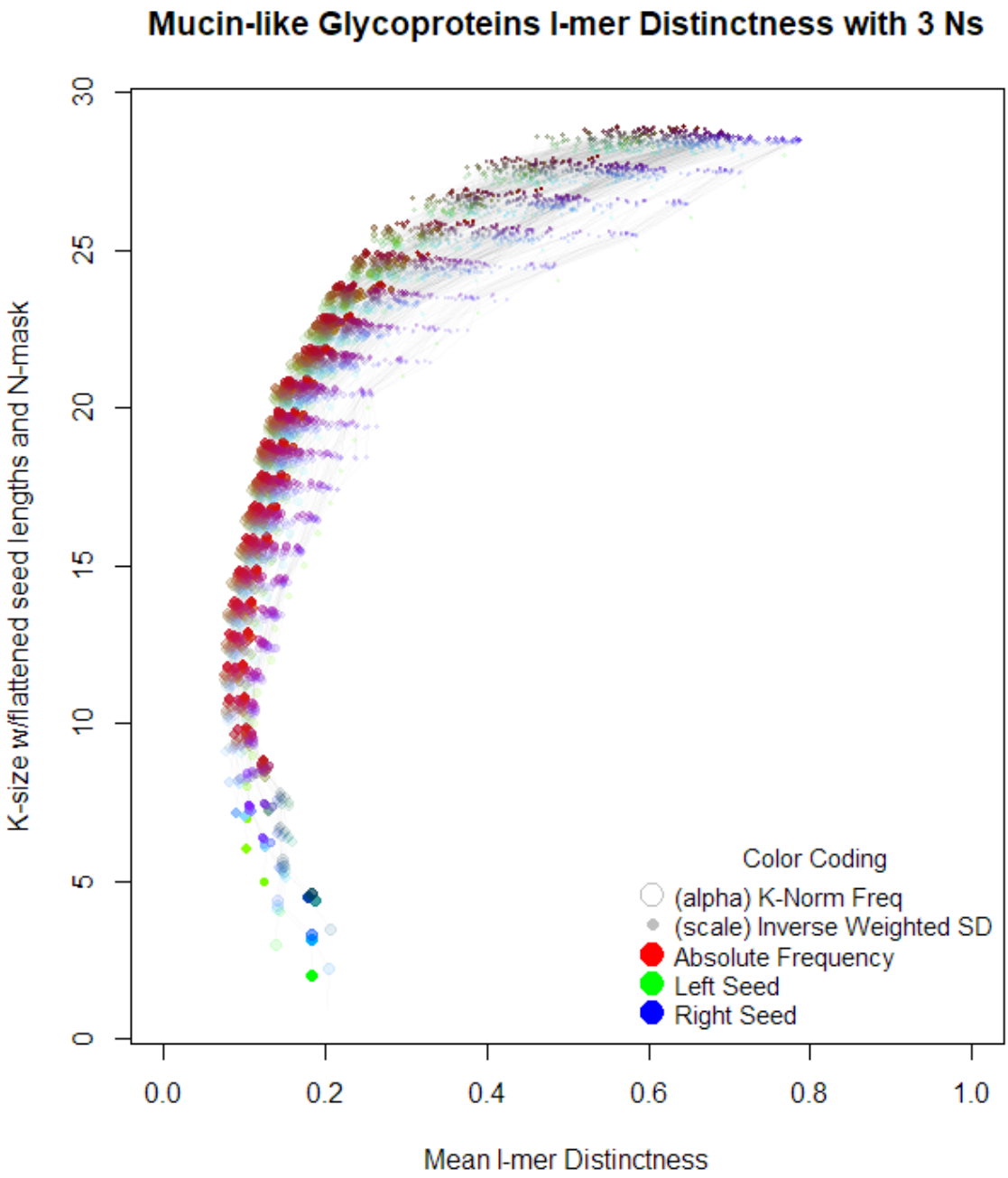


Figure 104. Mucin-like Glycoproteins, (PFAM) Protein Family, 3D Signature with WSD. N=3.

The Mucin-like family, with its high structure summary and low distinctness signature curve, shows us that both measurements must be read together to form a fuller understanding of the *k*-mer tree. From the signature and the summary, we can read that this is an example of abundant highly repetitive, yet highly flexible domains.

As the second most structured entry in Table 7 (although by a considerable margin), ZIP metal transporters. The signatures here show a slight backbone effect, and a concave low to middle distinctness curve, with broad motif distinctness distributions for almost all categories. These

transporter types have been shown to possess eight well conserved transmembrane domains with extra- and intra-cellular loops by contrast being highly divergent (Grotz et al. 1998)(Guerinot 2000). That there is a combination of two variant modalities could be the origin of the high deviation category distributions. The large number of conserved domains increasing overall structure, with the large number of highly variable loops reducing distinctness.

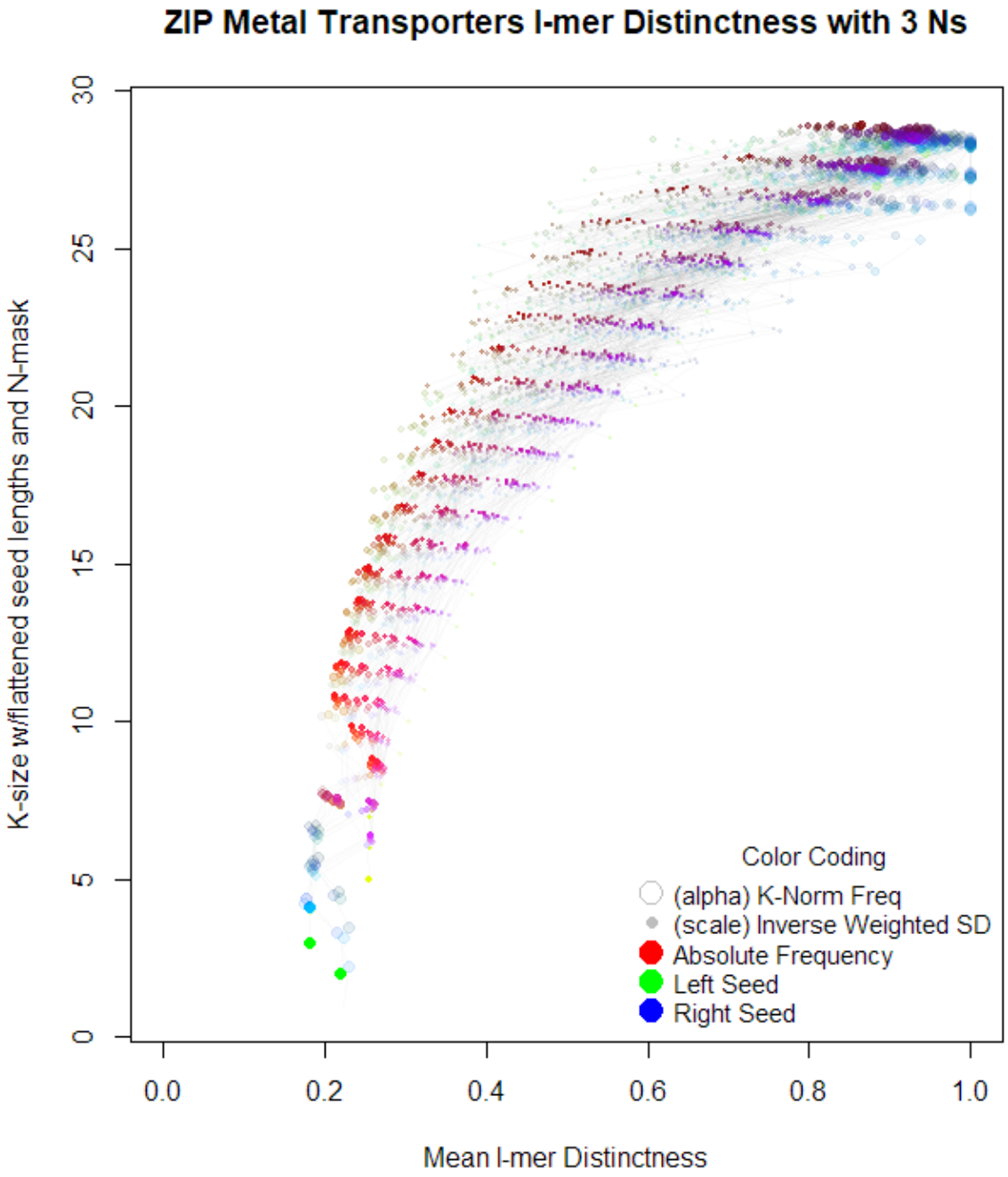


Figure 105. ZIP Metal Transporters, (PFAM) Protein Family, 3D Signature with WSD.

In summary of the family test set. The results combined with those in Chapter 2, Figure 21, suggest that protein families which have less duplicative structures may also be more resistant or sensitive to allelic variation, this could be possibly be described in the high-distinctness, low structure, high complexity category of sequence, and possibly high-fragility. The more duplicative and many-domain proteins typically are more structured in terms of sequence repetition, but with very indistinct variant patterning, and perhaps low overall complexity. These are of course early estimations of the possible set of relationships between signatures and family types. A full study of 1,000+ PFAM families via this interpretive method would be required to yield a stable reference set of guidelines for biological signature interpretation beyond the observations of pure entropic sequence descriptions.

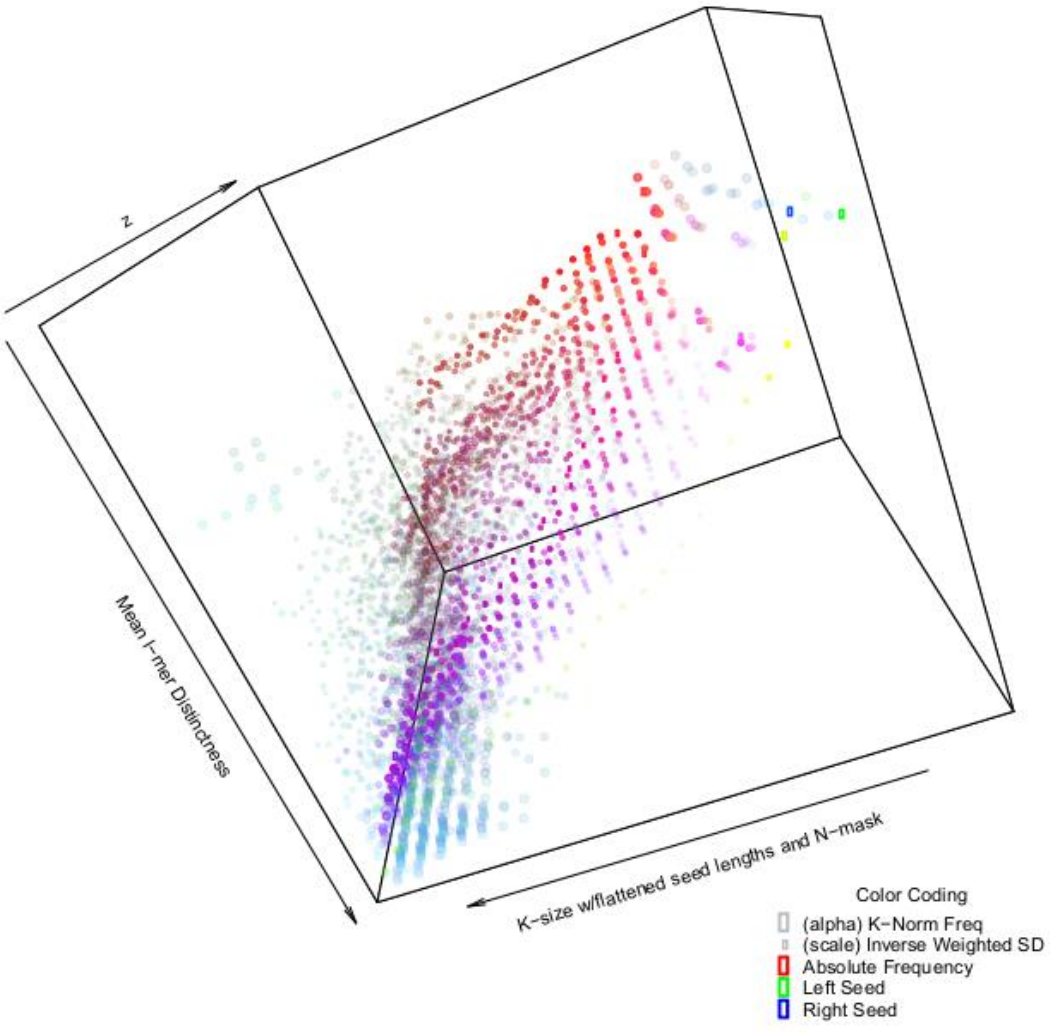


Figure 106. GPCR Chemoreceptors, (PFAM) Protein Family, 3D Signature with WSD, 3D Plot.

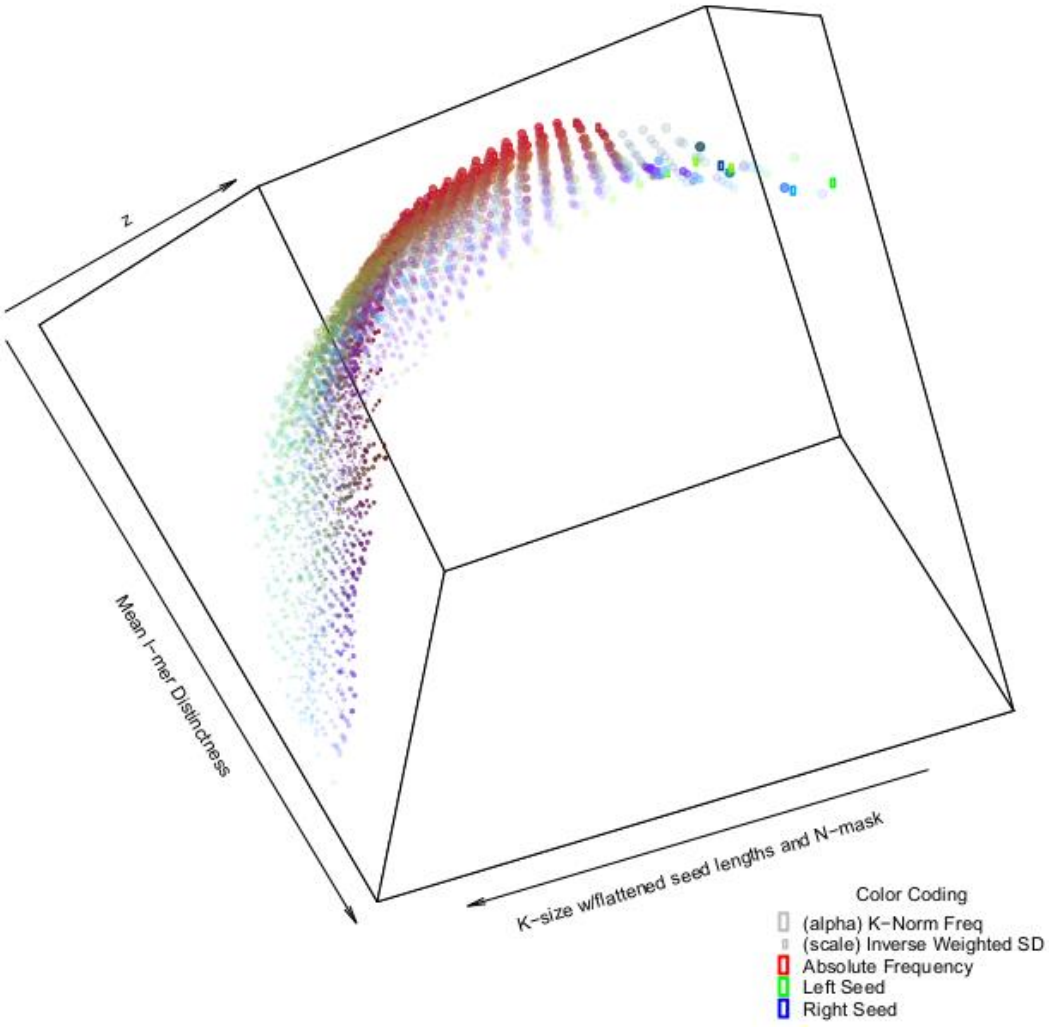


Figure 107. Mucin-like Glycoproteins, (PFAM) Protein Family, 3D Signature with WSD, 3D Plot.

Three-dimensional plots of GPCR chemoreceptors and Mucin-like glycoproteins have both been produced (Figures 65 and 66). These can further indicate visually the inner complexity of the low-structure GPCR family of sequences, and the contrast this presents to the low-complexity, yet highly structured sequence of Mucin domains.

4.4.7. The Pervasiveness of the Power-Law

The shape parameters generated for the uncorrected datasets are remarkably consistent. Figures 67-69 show that between the heterogenous data in the three test sets, the post saturation shape always ends up within the 1.5-2.0 range, with ~1.65 being the most common shape score. The distribution shape is discovered at all post saturation *l* values, and at all counts of *N.* There does not appear to be a correlative relationship (save for the saturation effect) between *l* and *a*, nor does there between *N* and *a*, nor does there between genomic and proteomic sequences.

For reference Figure 70 illustrates the steepness of the Pareto curve for the value of 1.6, in comparison to other values registered by some of the null-corrected summaries. The pervasiveness of this value of *a*, may also suggest to us that perhaps correcting shape parameters by local-nulls could be inadvisable, as it appears that the true biological power-law shape is manifest universally. Instead it could be suggested that the smaller variations within this 1.5-2 range ought to be considered with greater weight. In Table 7, the pre-corrected difference between GPCRs and Mucin-likes is between shapes of 1.625 and 1.561. Likely showing that the higher abundance of the most frequent motifs in Mucins due to the repeat-structured central domain creates the steeper Pareto distribution.

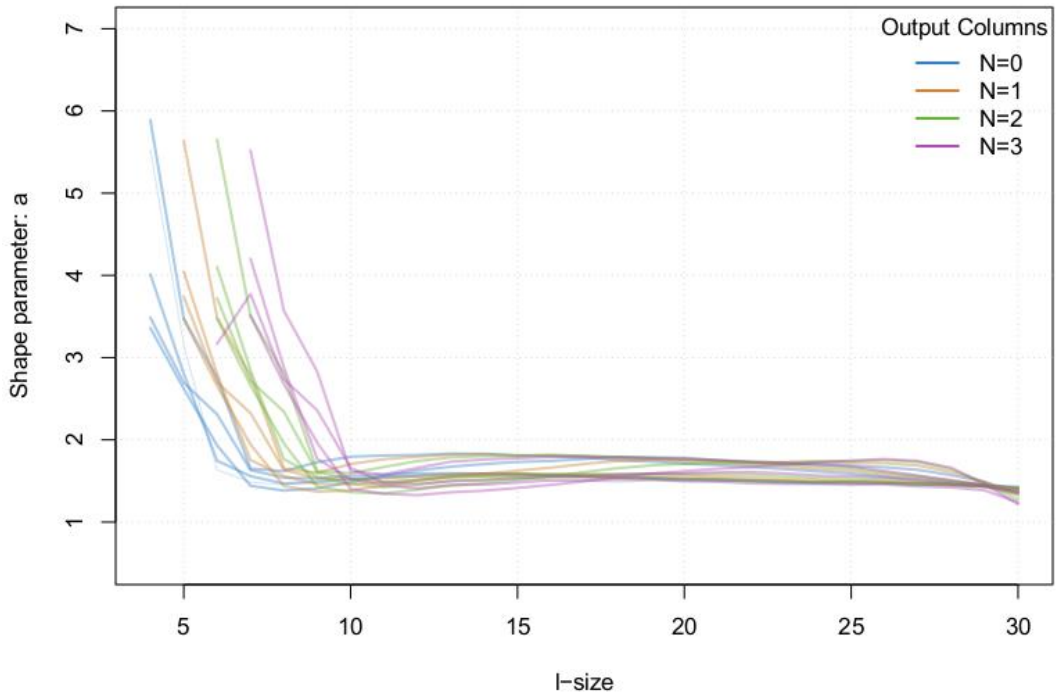


Figure 108. Pareto shape parameters distributed across l and N, Five Proteome Test Set (one line per proteome per N value).

The consistency observed here suggests that further work to establish the impact and value of the range of variation observed in uncorrected shape scores could be useful in maximising the information utility of the signatures.

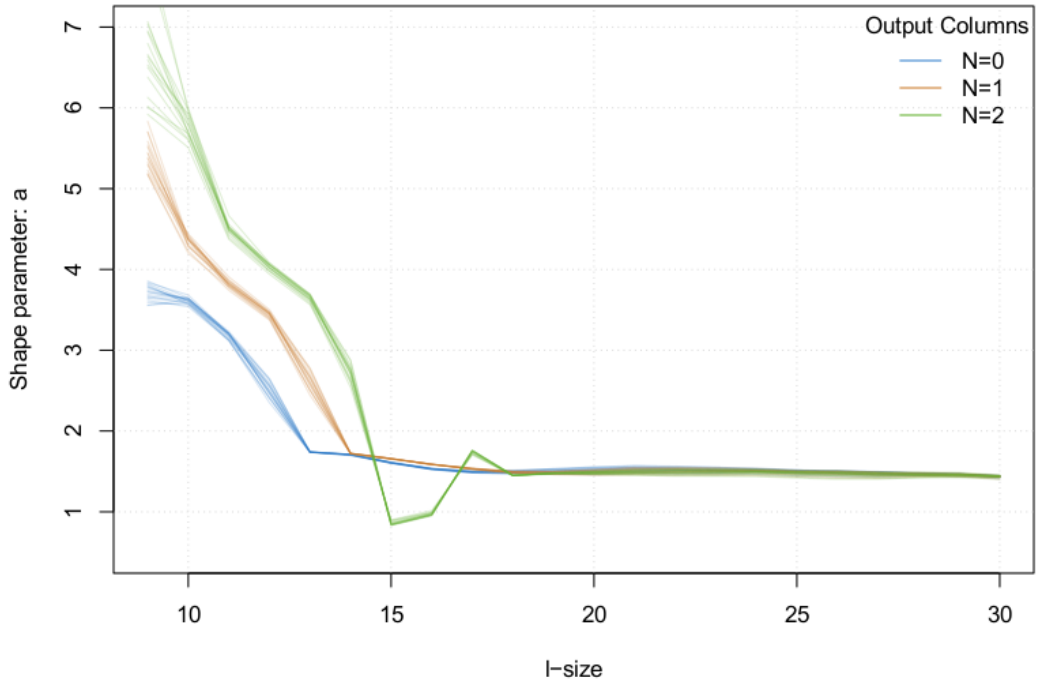


Figure 109. Pareto shape parameters distributed across l and N, E. coli Genome Test Set (one line per genome per N value).

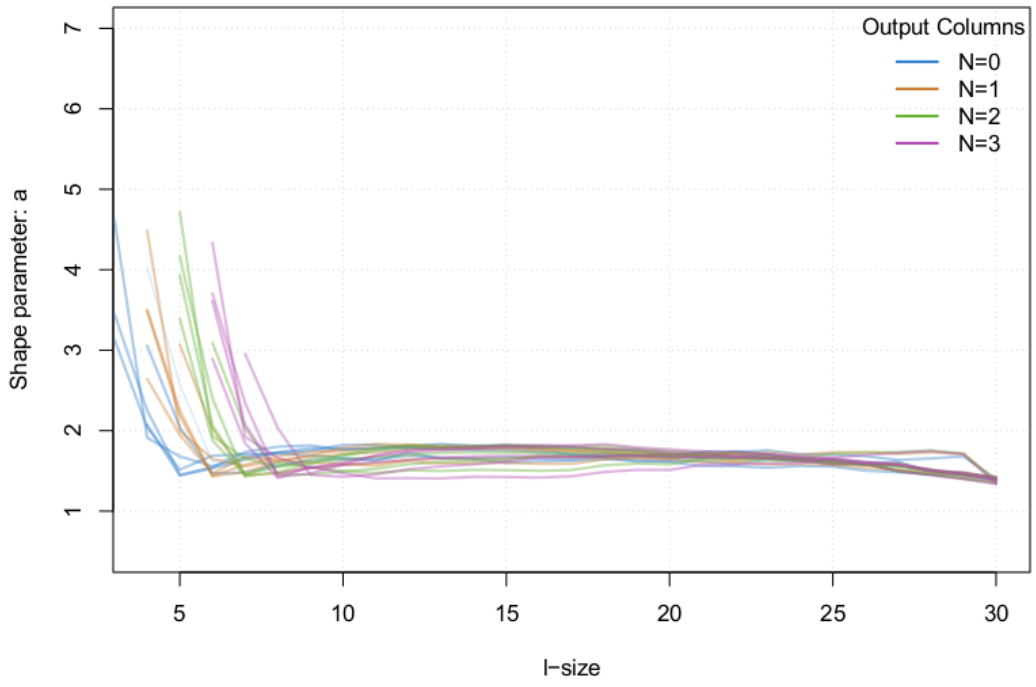


Figure 110. Pareto shape parameters distributed across l and N, Protein Family Test Set, (one line per protein family per N value)

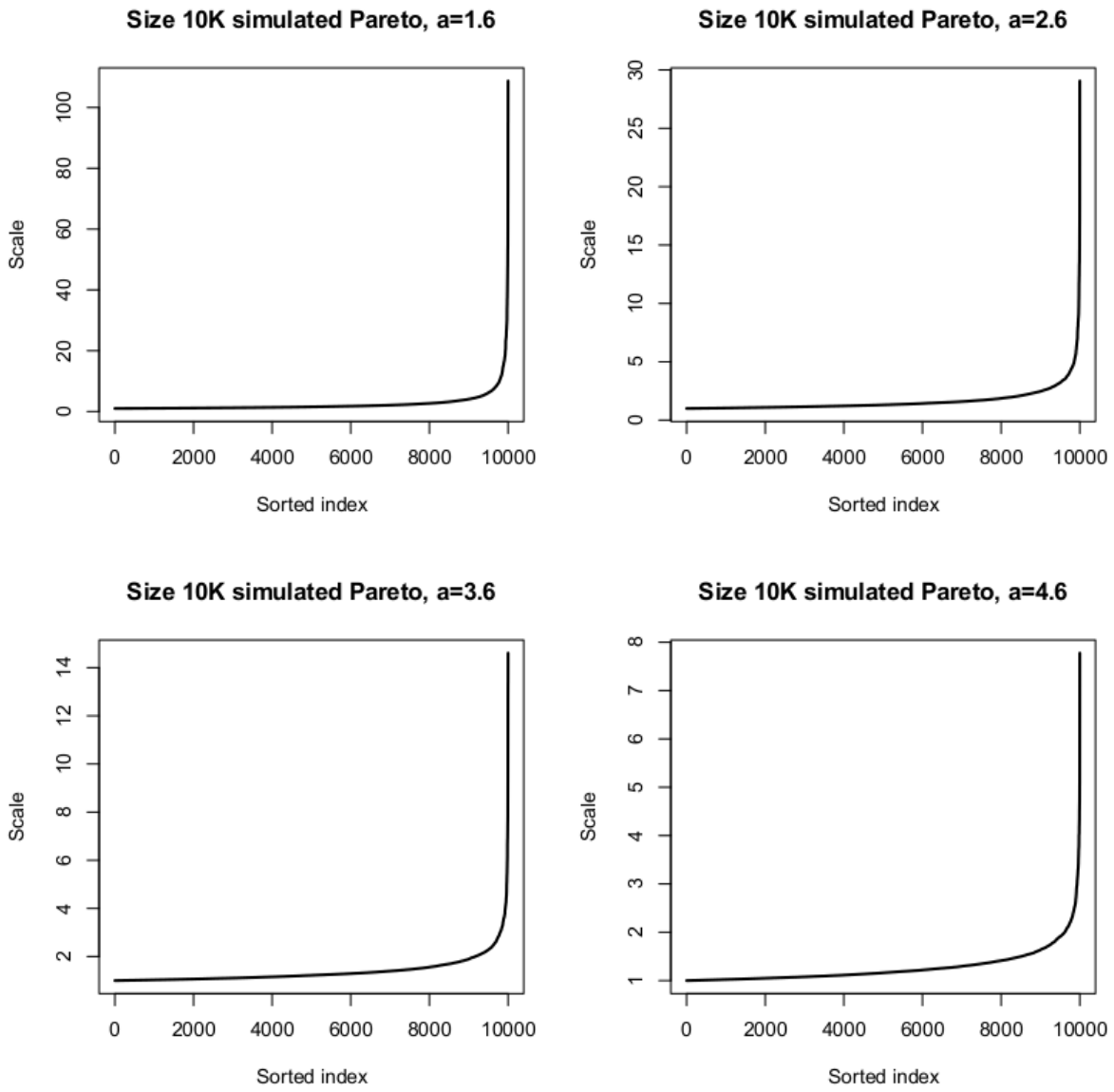


Figure 111. Example Pareto Distributions across shape parameters. Top-left: a=1.6, Top-right: a=2.6, Bottom-left: a=3.6, Bottom-right: a=4.6.

4.5. Discussion

4.5.1. Signature Performance

Having developed these aggregation, summarisation and visualisation methods for N-masked *k*-mer trees, we can begin to review their effectiveness as tools for the description of sequence sets. The objective of this research was to find methods of describing large self-contained sets of biological sequence in a manner which yielded an 'at a glance' impression of the content of that sequence, and that this signature ought to also be composed of datapoints which could be dissected into the points of biological origin that composed such an image.

One of aspects which this work has overlooked to some extent is the potential for further dissection of the signature categories. Any given signature category could, for example, have a paired set of annotation labels from the input set, and could be described in its proportional representation of those categories. This could be secondary structures, or domain types in protein families. It could be functional attributes of entire proteins (cell signalling, transmembrane, DNA binding, enzyme, matrix structure and so forth), in the case of proteomes, or it could be any prominent DNA annotation in the case of genomes (known binding motifs, intra/extra-genic DNA, intronic/exonic, LTRs, etc.).

Although a certain degree of human error might be introduced by adding annotation categories to signatures, this might also further the informativeness of the decomposition of a signature. For example, in the case of *L. rubellus* it could show at a glance which proteins were creating the huge separations in banding patterns, or the in the case of Mucins, the glycosylated repeated domains causing the total distinctness collapse. As every point in the *k*-mer tree has spatial sequence origins it would not be difficult program the propagation annotation categories throughout the tree in the same manner as basic frequencies. Through a hash-map of parallel storage variables in the Node class, an arbitrary dictionary of annotation types could be fit to the tree at run-time. This would however have a larger memory footprint. Another method of programming the annotation overlay with a lower memory footprint, but a higher time-cost, could be to generate the tree as many times as annotation categories are present, using only annotated sequence each subsequent tree. The final set of signatures could then be merged into a categorical ratio overlay.

Many of the observations made of signatures in this work have been post-hoc speculations on the origins of signature sub-formations, each of which would need further investigation to flesh out into suitably dependable theories that might be relied upon in future research. The alternative annotation methods for the tree might thus be a reverse search capacity whereby the user selects signature components of interest, and searches an annotated sequence set for its constitutive spatial information. This might be as simple as re-building the tree, and searching (&DFS subtree

merging) for only a specific set of N-mask patterns, and converting all discovered sequences with any degree of distinctness into a finite state automata, in a similar manner to BLAST (Camacho et al. 2009). This approach might have the advantage, compared to the above paragraph, of allowing the user to search a specific pattern type across a far wider array of annotations without the per-annotation time cost, and with a fixed peak memory footprint. The disadvantage being the signatures remain initially quite abstract.

At present the signature system appears to function quite well at processing entire proteomes, however the DNA processing capacities have limited its application in other ways. A point of interest might have been the *Lumbricus rubellus* and *Lingula anatina* genomic signatures in comparison to other less-divergent genomes in the same clade, or in comparison to the more well researched model species. However, further efficiency gains will need to be made before such a comparison is possible.

4.5.2. Experimental scope for performance gains

In the context of the possibility of major efficiency improvements, there are several parameters that sequences signatures could be expanded. These are 1) the dimensionality of the N-mask, 2) the depth of the tree, and 3) the size of the input sets.

4.5.2.1. N-Mask Dimensionality

Regarding the first parameter, N: It would be ideal to be able to expand the maximum number of Ns in each mask up to $k - log(f_r)$, meaning that all N-masked sequences in the tree would still reach null saturation by their leaves. A more comprehensive summary of the structure, and even more sensitive detection of sparse motifs would be also be achievable if techniques were developed to compute the complete set of all (2^n) N-masks per k-mer, however a more reasonably expectable near-term objective might be a moderate expansion of the N-count. One primary inefficiency of the current method are the excessive heap memory allocation and deallocation during the creation and destruction of subtrees in the merging function. This could be replaced by some fixed size pre-allocated working memory used for tree merges. Another bottleneck in terms of clock cycles is the repetitive paired-DFS function executions required to merge subtrees. It might be possible to by subtraction discover extra merged node combinations for ‘free’ by storing multiple different merged frequency scores in the same node.

4.5.2.2. Tree Depth Limitation

The depth of the tree, unlike the N-mask, does not have a theoretical complete solution. In terms of absolute utility, the depth of the tree could extend to the length of the longest string in the input set, however this is also wildly unfeasible. At present, the frequencies which ‘escape’ the tree ranged

between 1-4% of the total for *E. coli* DNA, and 2-11% for the protein families, and 1.5-18% for the proteomes. It would perhaps be wise to suggest that frequency escape isn't always a bad thing, or that is ought to be considered an effect which renders the analysis compromised. For example, if analysing both alleles of a highly allelically divergent genomes, one might expect 30-40% of frequencies to escape, and be backwards subtracted from the tree. A hyper-conserved protein family could be expected to give a similar reading. The real consideration with depth is whether it captures the breakdown of the high frequency structures which are expected to break down, and which are meaningfully interpretable when they do so. Still, it remains desirable yet to extend the tree's depth, if only to discover the point at which it becomes ineffective.

4.5.2.3. Size of Input Sets

The memory and processing time capacity for larger input sets would be very useful, as mentioned several times, for the sake of full genome signatures. It could also be utilised for other large sequence inputs, such as metagenomes, which can be hard to analyse and whose description might be facilitated by sequence signatures. Another area of large sequence set inputs is transcriptomics. Since frequency is currently coded to 1-per-k-mer, there would be no significant performance penalty for coding *x*-per-*k*-mer, where *x* is the normalised read count spatially resolved for a transcript from a given sample. There are many possible input configurations which could be explored with performance gains, and memory footprint reductions.

4.5.3. Potential Experimental Applications

There are many unexplored potential use-cases for the signatures developed in this research. Those that will be expanded up here include 1) Traits and phylogenetic association, 2) Stress/dose response signatures, 3) Reference signature database development.

4.5.3.1. Intersection with Traits and Phylogeny

Taxonomic classification and the prediction of evolutionary history has become a very powerful tool for understanding evolutionary biology, particularly since the advent of NGS technology. With functionally informative signature tools, there exists the possibility to describe the '-omic scale' architectures many types of lifeform, and to intersect these outputs with the structure of phylogenetic trees of various scales. This might concern a small monophyletic clade of species, or diverse samples separated by 100s of millions of years. This may lead to the discovery of 'typical' signatures for certain clades, or the possible association of signature types with other traits, such as environmental plasticity, k- or r-selection, life history, and other phenotypic qualities.

4.5.3.2. Stress/Dose Response Signatures

Stress responses in transcriptomes are a frequent research objective. Even the earlier transcriptomic studies revealed that organism stress responses can have huge impacts on the entire expression pattern of transcripts, for example, in 2002 *Arabidopsis thaliana* was found to have 30% of its transcripts in some way differentially regulated because of common stressors (Kreps 2002), and currently the results of transcriptome stress experiments have similar results across the board. For example, a 2017 study shows that in human mononuclear blood cells repress two thirds of their genes in response to heat stress, whilst up-regulating many others (Bouchama et al. 2017). Stress responses have the effect of drastically altering gene expression in most organisms. By encoding a k-tree with read-depth scaled frequency scores for all input k-mers, it would be possible to deploy a system of signature differentials for stress response, with the aim of qualifying the extent of a stress response in an organism which annotates very poorly when compared to the available references. There might also be a variety of signature differential types depending on whether stressors are singular or multiple, or based on their severity.

Another avenue of mathematical development which might aide this potential research direction would be the formalisation of distance metrics between signatures within the same set of samples. This would include estimation of null variance of category scores between replicates, and the testing of stress or other variable response samples against them. Distances would also be subject to the signatures form, with the possibility of 'distance signatures' displaying most prominently the range of distances exhibited by the categories or threads which separate the samples the most.

4.5.3.3. Reference Database Development

Although signatures are informative of sequence features by themselves, much of their value can come from comparison. The highly visual aspect of the output allows a researcher to very quickly see if one sequence set 'looks like' another one. By extension, the notion that digital means to quantitatively assess which signatures look-like which seems sensibly forthcoming. In terms of comparing between a small number of pre-calculated outputs this might be trivial, however much of the modern quest for biological insight comes in the form of queries directed at massive data banks via sophisticated search tools. Should enough signatures be generated, it would be useful for a signature-specific distance metric heuristic search function to be developed, such that a user might query a database of signatures with their own outputs to see which other organisms manifest similar sequence variation and structure patterns, in a manner that is liberated from comparisons of homology. The type of sequence data in this use case is not limited to the types of any of the test sets used in this work. However, a relatively low-computational cost first endeavour might be to

generated signatures for all 1000+ PFAM protein families, and to deploy a PFAM-indexed search function, as an adjunct the current protein family knowledge base.

4.6. Conclusion

A biological sequence signature creation tool was developed. The tool was applied to 5 invertebrate proteomes, 6 protein families, and 18 *E. coli* genomes, as test sets. The results showed that the peptide trees are capable generating varied and informative signatures of the sequence inputs, which relate directly to the biology of the sample sources. The DNA trees have yet to achieve the computational performance required to perform large scale analysis on genomic scales. Multiple aggregation methods were proposed for tree aggregation. The research then focused on two aggregation methods that were most suitable for the trees generated, given the performance boundaries of its parameters. Visualisation methods were developed for the outputs, with infographics also created to aide in the interpretation of the 'signature' plots. This work shows that it is possible to create dense and complex signatures of sequence structure, without sacrificing their biological utility, when highly optimised navigations of high dimensional space are deployed instead of generalisations which first seek to reduce it.

5. Chapter 5: Discussion

This thesis aimed to discover more about how mechanistic information sources of evolutionary and environmental flexibility were stored and used in biological systems by examining models which demonstrated an abundance of such information (Chapter 2), by analysing a system in which high tolerance a retained abundance of neutrally selected genetic material could have a large potential selective advantage (Chapter 3), and by further developing information processing methods to characterise the sequence structures in these model systems (Chapter 4). Here follows an initial review of the progress made in this thesis, with the core messages derived from the analysis associated with each specific aim. The rest of the discussion will then focus on several areas of theoretical advancement informed by the results found here, with suggestions for further study.

5.0. Aim 1

To assess the allelic diversity metrics of two highly divergent invasive global species

The genomes of two organisms were analysed: *Lumbricus rubellus* an earthworm, and *Lingula anatina* a marine brachiopod. *Lingula anatina* was found to have a published draft genome (Luo et al. 2015) with substantial mis-assembly as a result of an absolute allelic divergence of ~10%, around the current maximum observed for a natural population (i.e. not an F1 hybrid). *Lumbricus rubellus* was found to have an unprecedented ~33% base sequence divergence between alleles, with even the protein sequences between alleles separating by ~10%. Distributions of the divergences were quite uneven, and both genomes appeared to be mosaic-like, with both a low diversity and a high diversity set of allele fragment pairs in their assemblies. Working towards this aim found that the extent of divergence between alleles in some genomic systems may be almost unlimited and opened the door to new theoretical considerations surrounding re-combination and the Meselson effect (Welch & Meselson 2001).

5.1. Aim 2

To describe the potentially acclimative or adaptive information present in hyper-divergent alleles

Gene families found in the hyper-divergent regions of the genomes analysed in Aim 1 were described. The top six families shared by both genomes were found to be highly environmentally interactive proteins types. This suggested that the mechanistic commonalities between the genomes also led to commonalities in the outcomes for phenotypic diversity, although this diversity would be highly cryptic to a non-molecular taxonomist given the discussed morphic stasis of both species.

5.2. Aim 3

To develop a general theory of redundant information structures

A straightforward summarisation method of k-mer tree data structures was proposed. Other sequence signature evaluation methods were reviewed and found to be either insufficiently detailed, or insufficiently functionally informative. Structure in the total sequence domain (a single genome, proteome etc.) was defined as over-representation of k-mers given their expected incidence rate, and the relationship between the sequential extensions of each *l*-mer within the tree was defined as the product of child node frequency *gini-like* coefficients between a parent node and a child node, with the parent co-efficient being inversed. It was found that the dimensionality to be navigated computationally was extremely high, but that there was an unexplored analytical space available to be discovered, given that most other sequence summaries act by heuristically limiting dimensionality in some way.

5.3. Aim 4

To implement and apply the developed theory to different sequence types

Implementation of the sequence structure scoring algorithm was achieved. Some degree of success was had in implementing multi-core processing, although the threading efficiency could still be dramatically improved. Although the implementation is theoretically capable of processing all k-mers with any number of gaps (N), in practise it was found that limiting N to values of 2 or 3 was required for execution to complete in usable times (<48hrs per input set). A signature visualisation method was developed to explore these scoring systems, and three types of sequence input sets were evaluated (two protein, one DNA). Changes in signature were found to qualitatively describe the prior known characteristics of some of the inputs quite well, whilst in other non-model systems, the results were more informative. *Lumbricus rubellus* was found to have the most diverse set of information structures in its proteome compared to four other invertebrates of lower diversity.

5.4. Aim 5

To assemble the genome of a species with high theoretical need of both adaptive and acclimative mechanisms to cope with its environment

A genome of *Amynthas gracilis* was assembled. The assembly had an N50 of 478kb 4,350 scaffolds, and a size of 589Mb. Several *ad hoc* assembly finishing processes were adapted from work on the similarly allelically divergent *Ciona savygni* genome assembly (Vinson et al. 2005). Allelic divergence was found to be bi-modal, with peaks approximately at 0.5% and 3% (although this varies somewhat with window-size). The transitions between these divergence rates were incredibly sharp, suggesting

that the genome was an allelic mosaic of alleles originating from lineages of several different degrees of evolutionary distance. A full suite of gene prediction programs were run, and their outputs were collapsed with MAKER2 (Holt & Yandell 2011). Overall, 26,951 gene models were created.

5.5. Aim 6

To discover the simultaneous roles of acclimative plasticity and atavistic adaptivity mechanisms in an organism under high environmental stress

A dataset originating from a reciprocal transplant experiment was analysed. This dataset contained next-gen sequencing of RNA-Seq, miRNA-Seq and MEDIP-Seq, for sample individuals which had been transplanted between active and inactive volcanic soils for 31 days. Activation of acclimative responses was most readily detected in RNA-Seq differentials and was also a substantial signature in miRNA expression. The differentials between soil origins in miRNA expression suggested that this system reacted less readily to acclimative plasticity and seemed to preserve some of the long-term molecular expression stability from the original habituated environment. DNA methylation was found to have a highly stochastic abundance differential between samples. However, miniature gene-models of methylation distribution patterning were highly consistent between individuals despite the massive per-gene variations.

5.6. Theory 1: Sexual Dimorphic vs Plastic Multimorphic

Sexual dimorphism is an extremely common feature in many complex lifeforms (Lande 1980). It was only in the 1980s that the idea of sexual phenotypes originating from sex chromosomes began to be formally investigated and accepted (Rice 1984). However, the non-recombinant and partially haploid allele as a sex morphotype determinant is not a single origin evolutionary event, and appears to have independently evolved in many taxa (Ayling & Griffin 2002). Step-wise processes for the evolution of sex-determination have even been identified in fungal genomes (Fraser et al. 2004), suggesting that this pattern of evolutionary genomics has a universality to it which goes far deeper than we presently understand.

The processes by which a collection of linked genes becomes non-recombinant is a destabilisation of classic allelic evolution. Through an inversion, a divergence or some other translational mutation which prevents recombination for a large portion of one chromosome, a collection of linked genes with a combined haplotype effect become a singular morphotype determinant.

Lumbricus rubellus is known to have extremely divergent lineages, as explored in Chapter 2. It is also known that these lineages are reproductively active (Giska et al. 2015b). Annelid earthworms are

also known to be hermaphrodites lacking any sexual dimorphism (Lavelle 1997b), they are also long understood to be facultative parthenogens (Jaenike & Selander 1979). As the regular transmission of alleles between lineages discussed in Chapter 2, the question of recombinant validity allelic compatibility arises. Are there large regions of DNA which no-longer recombine when paired against their ancestral lineage partner in a hybrid individual? Given that most worms surveyed so far appear to be a combination of lineages, is this introduction of non-recombinant regionality a theoretical equivalent to the same process by which sex chromosome evolution converges across the tree of life?

Perhaps it is the case that sex chromosome evolution is in fact merely a large sub-section of a broader trend, of non-recombinant regions which confer contextual function to the phenotype, with dimorphism the result of the naturally emergent phenotypic game between emergent traits governed by the non-recombinant DNA's presence/absence. Alternatively, given that the range of earthworms, and the diversity of environments in which they reside is so vast, perhaps we might think of non-recombinant allele pairings as miniature 'enviro-genders' or an 'enviro-sex'. While not a serious proposition for terminology, this perspective is illustrative of the systematic continuity between these ideas. This theoretical position would also extend to other long-distance hybrid systems in hermaphrodite species, such as the *Mytilus trossulus/edulis* hybrid zone in the Baltic Sea (Strelkov et al. 2017). In this regions two mussels which ordinarily dwell in different substrates hybridise (benthic vs algal adhering) (Katolikova et al. 2016). This might also either lead to or emit from the incidence of aneuploidy which has been widely documented to exist across many species of Oligochaeta (Pavliček et al. 2016).

To further an understanding of the morphotype characteristic of non-recombinant alleles, the first prerequisite would be an improved assembly (500 kb+ N50) of the genomes of both lineages of *rubellus*, and substantial genomic population data which could be obtained in the form of RAD-Seq or low-depth genome sequencing from individuals of various populations. The objective would be to identify large non-recombinant regions of the genome which remain consistently non-recombinant in several populations. The challenge would then be to associate these regions with potential phenotype characteristics (which would more likely be metabolic rather than morphological, given the discussion of morphostasis in Chapter 2).

5.7. Theory 2: Doubling the Distance: Allelic Aivergence and Ploidy

In the Introduction, various evolutionarily active processes by which information was duplicated within the genome were discussed, one of which was genome duplication, and one of which was allelic divergence. Considering the sustained allelic divergence observed in Chapter 2, it might also

be worth considering the potential relationship between allelic divergence and duplication of various scales under the condition of reduced recombination or, as terminology might suit, a preponderance of large and tightly linked genomic islands of divergence.

To consider how these pieces of information machinery might fit together generally, it is worth looking first at the 2R hypothesis. Originally inspired by the discovery of four *hox* gene clusters in the human genome, this hypothesis asks whether the ancestral vertebrate genome was subject to two rounds (2R) of whole genome duplication, and whether this is the source of the size and complexity of the genomes we see today (Spring 1997). Substantial genomic evidence began to amass for this theory in the 2000s (Dehal & Boore 2005), and by now the evidence from next generation sequencing is overwhelming to the point of general consensus around its accuracy (Van de Peer et al. 2010).

One such paper proposing this theory in 1997 asks provocatively, 'are we polyploid?'(Spring 1997). Human genomes are diploid now of course, but the corollary to the consensus around 2R is that researchers must then ask of our alleles, how does four become two? Or more specifically, how did the allele pairs revert to diploid meiosis? To this end *Xenopus* is once again a valuable model. Genetic studies of its evolutionary history suggest that the invasive polyploid *tropicalis* is in fact an allopolyploid, which resulted from a rare event during interspecific hybridisation (Abu-Daya et al. 2012). In this case there will already be paired alleles in the set of four which are more like each-other. This suggestion is that the difference between 'autopolyploidy' and 'allopolyploidy' is one of substantial evolutionary value. Already an often observed feature in plants, such as wheat (Feldman & Levy 2005), newly allopolyploid organisms exhibit dramatic cascades of genomic modifications and gene silencing described as 'genomic shock' (Comai 2000). The capacity of small RNAs to 'defend' one diploid allele pair against the genes of the other pair by gene silencing seems to be deeply involved with an organisms capacity to tolerance these hybrid duplication events (Malone & Hannon 2009). One other vertebrate example is the allopolyploid Iberian cyprinid *Squalius alburnoides, of which there are both diploid and triploid populations (Alves et al. 2001).*

Another attribute of allopolyploidy is the speciation which follows the event. Particularly in plants, (as well as Xenopus tropicalis *(Evans 2008)*), examples of these speciation events have been seen in brassicas *(Widmer & Baltisberger 1999)*, knot-weeds (*Persicaria*) *(Kim et al. 2008),* and cotton (*Gossypium*) (Wendel et al. 1995). The mechanistic model by which speciation occurs with duplication has been proposed as 'divergent resolution' *(Lynch 2000),* by which duplicate genes change function and proffer new reproductively isolating phenotypes. Although other work on

speciation suggests that, as in genomic shock, divergent resolution is more likely to take the form of reciprocal loss or silencing *(Taylor et al. 2001) (McGrath et al. 2014).*

We can summarise this overall view of the process as Hybridisation -> Allopolyploidy -> Genomic Shock -> Divergent Resolution -> Speciation -> Gradual reversion to diploid through karyotype rearrangement, sub/neofunctionalization and subgenome divergence (Zadesenets & Rubtsov 2018).

The problem with this interaction is that it is assumed that it is the hybridisation event which triggers the allopolyploid event. Reasoning this out with respect to the genetics of invasive species appears to reveal something of a blind-spot in evolutionary theory. Consider the following: in the genetics of invasive species there is a paradox. Invasiveness has been widely identified as benefiting from information redundant systems as discussed in the Introduction and in Chapters 2 and 3. However most invasive populations (that are not of deliberate anthropogenic origin) are generally founded by a handful of individuals, or in the case of facultative parthenogens such as the earthworm, even by a single individual (Estoup et al. 2016); severe genetic bottlenecks which can deplete a population of its diversity. The theoretical (and observed) association of hybridisation with invasiveness addresses this to some extent to propose that certain organisms can retain inclusively the combined fitness of multiple lineages or even multiple species, resulting in the type of 'blocky' looking polymorphism pattern in the genome as seen in Chapter 3. For example, a highly fecund R-selected organism with a high percentage of diverse alleles, despite facultative parthenogenic reproduction, may still be able to retain the presence of most of both alleles in the second generation if enough young are successful, although some loss is inevitable.

Organisms already capable of bi-allelic regulation of such extremely divergent alleles may be pre-equipped to deal with the 'genomic shock' of duplication. Regular genome duplication events in populations undergoing the stochastic drift of many hyper-divergent alleles, would eventually coincide with a mosaic genome which harbours a species/lineage selection of allelic diversity that is maximally compatible and retains the essential contextual fitness advantages of both. Since the origins of most genomic duplications are so ancient, it is incredibly difficult to describe the nature of the origin event. It could be that in analysis of the myriad complexities involved in the 'diploidisation' process described following assumed singular allopolyploidy (Zadesenets & Rubtsov 2018), the researcher may overlook the possibility that some apparent 'rearrangement' originates from extreme diploid hybridisation and subsequence intrapopulation drift prior to duplication. Duplications could happen within this population repeatedly and unsuccessfully until an autopolyploidy event occurs within a genome possessed of the type of 'goldilocks' mosaic as described above.

Although this is a speculative theory, studies of earthworm ecology provide various examples of situations in which both extreme hybridisations and regular genome duplication events are detectable – and might provide excellent model systems to further the understanding of ancient duplications in the more 'charismatic' terrestrial taxa. For example, *Amynthas catenus* the Taiwanese mountain earthworm has been shown to exhibit morphotype variants of three different ploidy levels (di, tri-, and tetraploid), with chromosome count and ploidy levels being association with both reproductive isolation and parthenogenic tendency (Shen et al. 2011). The European *Dendrobena* genus contains many species which are both invasive (Pop & Pop 2006) vary in ploidy level (Bakhtadze et al. 2008), *Dendrobaena rubida* alone is known to exhibit di-, tri-, tetra-, hexa-, and octoploid varieties (Cosin et al. 2011), and also acts as an invasive peregrine (Tiunov et al. 2006). The invasive *Dendrobaena octaedra* (Cameron et al. 2008) has been reported to have hexa-, penta- and octoploid variety of the parthenogenic subspecies (Hongell & Terhivuo 1989). *Aporrectodea rosea* is a similar global species present in western Europe, Russia and (more recently) Canadian (Addison 2009) soils, and may be diploid, decaploid, and most duplication levels in-between (Vsevolodova-Perel & Bulatova 2008). This is by no means an exhaustive list and illustrates the potential for these model systems as research tools for evolutionary genomics.

A final piece of the puzzle with respect to allelic divergence and polyploidy might come from a very common correlative observation often made of earthworms: that the polyploids are most often parthenogenic (Jaenike & Selander 1979)(Cosin et al. 2011). As most earthworms are facultative parthenogens already, it is hard to draw a line which indicates that polyploid parthenogenesis becomes obligate, despite the correlation, however it seems to be the preferred reproductive strategy. Some parthenogenic earthworms reproduce meiotically, whilst others appear to have a mitotic process of self-fertilisation (Terhivuo & Saura 2006). The hypothesis here would be that, in the same way that large scale regulatory systems respond to genomic shock, the earthworms' evolved response to the same signal might be to limit sexual breeding. For the sake of the evolutionary question, it would have to be asked, why is this adaptive? Whilst creating large invasive populations from few individuals is clearly an advantage for genetic self-propagation, the case of a sole invader is in the extreme minority. Typically one might expect the function of Muller's ratchet (Haigh 1978) in the absence of non-clonal reproduction, following from the loss of genetic load mitigation (Keightley & Eyre-Walker 2000) to be a substantial long term adaptive cost to the organism's survival. There is a reason why parthenogens appear only to exist at the tips of the leaves of the tree of life. Yet it is not just the chance invaders which participate in this costly exchange. To the question of the benefit we then return to the 'goldilocks' mosaic hypothesis. Diploid hybrids of extreme divergence as seen in Chapter 2 likely must negotiate some trade-off between allelic

incompatibility/misregulation and the potential for inclusive retention of the plasticity or metabolic fitness evolved in multiple lineages or species. If one imagines the continuous range of all possible mosaics by their intrinsic fitness as normally or likewise distributed, the near term (~50 generation) fitness benefit of the right-hand tail might well exceed the cost of no sex, and maintenance of that improbable allelic arrangement could only be sustained by the germline's abstention from meiotic recombination events. Meiotic clonality might limit recombination to between the sub-genomes of the polyploid, whilst mitotic reproduction would prevent unique mosaic alternation altogether. Even though the polyploidisation may occur stochastically (or with increased frequency in hybrids) the fairly frequent (in evolutionary timescales) intersection between 'goldilocks' mosaics and duplication events might be sufficient for nature to selectively retain this function as a survival strategy.

5.8. Theory 3: Fighting the Dimensionality of Entropic Structure

The entropic structure model developed in Chapter 4 was stymied with respect to its application in earthworm genomes by the algorithm's computational memory and time efficiency. Consequently, it was only capable of working with smaller full genome sequences (*E. coli)*. Another limitation of the ID-based N-root tree indexing system (piggybacking off the N=0 tree structure), was that any increase in *K* over 32 would require a doubling of ID variable size from 4 to 8 bytes per index, which would only impose further limitations on the input set size. Application of this algorithm to the full genomes (including full allele copy cohort) of earthworms of various ploidy and mosaic types might have been used to as a knowledge-free metric to assess the information structures in these organisms. The way in which information structure sizes change with genome size or ploidy correlations could also be informative of less recent duplication events, or the scale of the sequence based regulatory mechanics behind them.

Many systems exist to discover disproportionately enriched motifs in DNA sequence, however almost all of these utilise some form of dimensional reduction. For example 'mCUDA-MEME' (Liu et al. 2011) uses hyper-parallelised short alignments based on an identity scoring system as does 'EXTREME' (Quang & Xie 2014), and 'Quick-motif' is built on the basis of pruning the set of subsequence pairs to combat the inefficiency quadratic difficultly (Y. Li et al. 2015). Although these methods are highly advanced and have made significance performance improvements to sequence space navigations where heuristic outcomes are acceptable, it is not possible to adopt most of their optimisations. To stick with the objective of the original mathematical model, it remains necessary not to optimise the search space, but to optimise the methods for navigating it.

In order to implement a measurement system for the mathematical theory of structure scores in the form of manageably efficient algorithm (i.e. requiring less than 1 TB RAM, execution time of no more

than a few days per genome), the central algorithm would need to be re-designed. Another desirable feature for such a redesign would be to alter the efficiency with respect to N, as the present efficiency is unsustainable to discover sparse motifs. Here will follow a discussion of a possible re-design which could accommodate far larger input sets, with the caveat of restriction to a hard-coded alphabet.

The difficulty with fully exploring the k-mer space with flexible bases is that the actual space is far too large. We cannot expect even our most high performing CPUs to do 4^{32} of any simple operation in a reasonable timeframe, and that is only the scale of the N=0 tree. Therefore, as the present program implements, the sparsity of the tree data-structure can be allowed to be reflective of the sparsity of the dataset, and its navigation thus avoids any querying of the unused sequence space. However, because of the extreme sparsity, when 'folding' child-nodes of a single N-root over into merged subtrees, this results in many memory operations to extend the pre-existing data structure, and then deallocate the memory once that operation is complete. However, there is no way around the need for a subtree to hold the merged subtrees, and pre-calculation of the substructure would only yield an unmanageable load in RAM.

The solution proposed to combat this deficiency would be a novel data structure which holds sparse single-depth tree-foldings in sequence space and exploits subtractive redundancy. Firstly, the idea of subtractive redundancy needs exposition. This is simply the idea that, if three values must be stored in memory, *A, B,* and *C*¸ where C = A + B, we can simply store A and B, and calculate C at runtime. This on its own does not seem to be a huge performance gain. However, if it were necessary to store at depth *l*=4, all possible tree-folding combinations of eight terms in a 4-level binary tree:

A

A B

A B C D

A B C D E F G H

Considering that *any* of the nodes above the bottom layer can be turned into an *N* value, and merge their subtrees, what is the set of all values created by the merged subtree?

N=1:

AB CD EF FG

AC BD EG FH

AE BF CG DH

N=2:

ABCD EFGH

AEBF CGDH

ACEG BDFH

N=3:

ABCDEFGH

Above, the letter adjacencies indicate summation. Thinking subtractively, we can see that many of these scores do not need to be stored in memory. If the N=3, ABCDEFGH value is present, only the first or second half of the columns in N=2 are needed, etc. The eventual progress of this line of thinking is the conclusion that if the original nodes were simply stored in a list, an algorithm could calculate all the values for N from 1 to 3 on the fly (and add them to the structure score/save remarkable features etc). But, when the size of the set of all nodes at that depth is 4^{32}, this becomes prohibitive. It would be better to discover merged terms as the sparse tree is folded. But how can we still retain the advantages of subtractive redundancy whilst limiting navigation to the actual sparsity of the tree, and do so in a way which limits the memory allocation churn?

The solution proposed is to fold the *k*-mer tree fully (i.e. such that N=*k*), but *not* to permit any gaps in the N-mask. Effectively this means merging the subtrees of the head-node, and deleting all but the merged tree, then repeating this process on the next increment of *l* (which is now effectively the new head-node) until the tree has been collapsed into a linked list of length *k*. Each merged node however retains the values present in all nodes which were collapsed onto it in a subtractive efficient data structure. The k=4 tree folding to the fourth depth is illustrated in Figure 112.

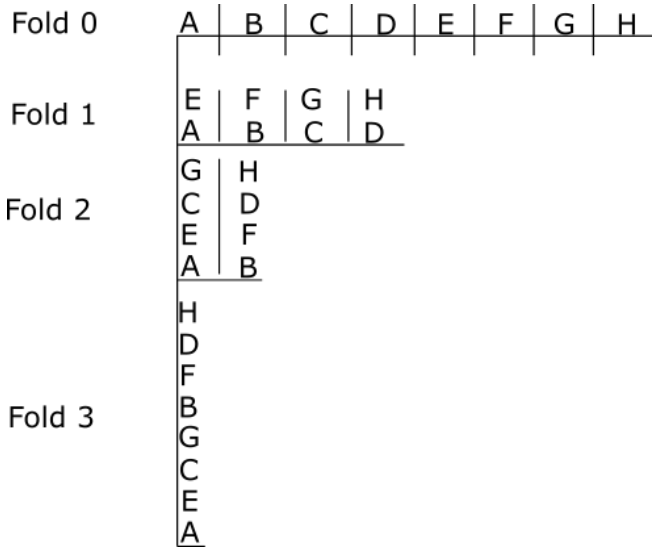


Figure 112. Illustration of successive folding/merging of a binary tree, beginning with the topmost node and terminating with the bottommost. In the original algorithm from Chapter 4, the combined node values are simply summed.

Under the current implementation's rules, AE or BF would simple by the sum of those two nodes' values. However, with the proposed efficiency changes, these would be stored in slightly more advanced structures, see Figure 113.

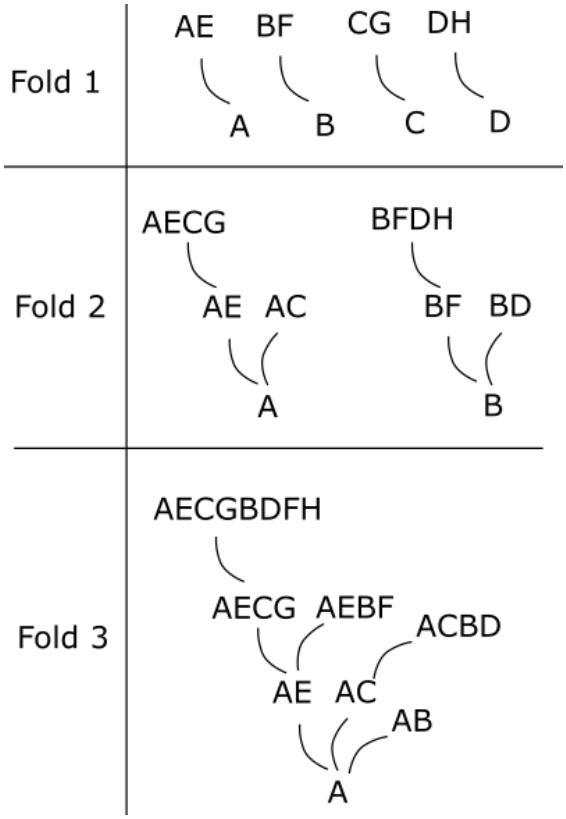


Figure 113. Retentive data structures for storing merged subtree information in single nodes, demonstrating their growth in a binary tree successively merged downwards from the topmost node to the bottommost.

In Figure 113 the curved connections denote the same relationship you might find in a tree (a one way linked memory address connection between node class elements). The final fold now contains 8 elements, the same number as the original number of nodes, however they are structured in a subtractive efficient format, which allows, by traversal, an algorithm to re-calculate on the fly the 27 unique node combinations (19 combined + 8 un-merged) values created by all possible tree folds. It does this by subtracting any node's value from one of its children, and recursively subtracting all children left of the queried child from the queried child's children. For example, see Figure 114.

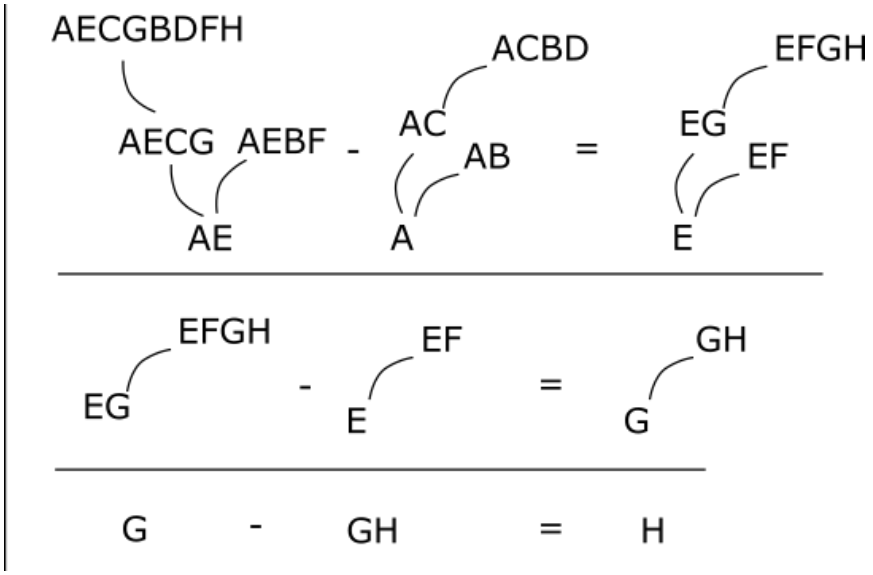


Figure 114. Illustration of algorithmic exploration of subtractive redundancy space within the proposed merged-node data structure.

The advantage of this subtractive system is that by structuring the 8 values as this type of data structure, we can find all required actual folded values on the fly, without needing to store all possible entries in the list or index every list entry by its location in the space. For example, given a sparse version of the demonstration tree, it will still fold successfully (see Figure 115). In this demonstration the absent values are retained as 0s for the point of illustration, but all terminal 0s are simply elements which no longer needs to be allocated in memory. The key advantage of this system is that the tree only needs to be folded once per depth, and the N-mask N_N_ can be derived from subtractive operations on NN__ and NNN_. Eliminating the need for huge numbers of subtree merge operations, whilst retaining the sparsity of the original dataset.

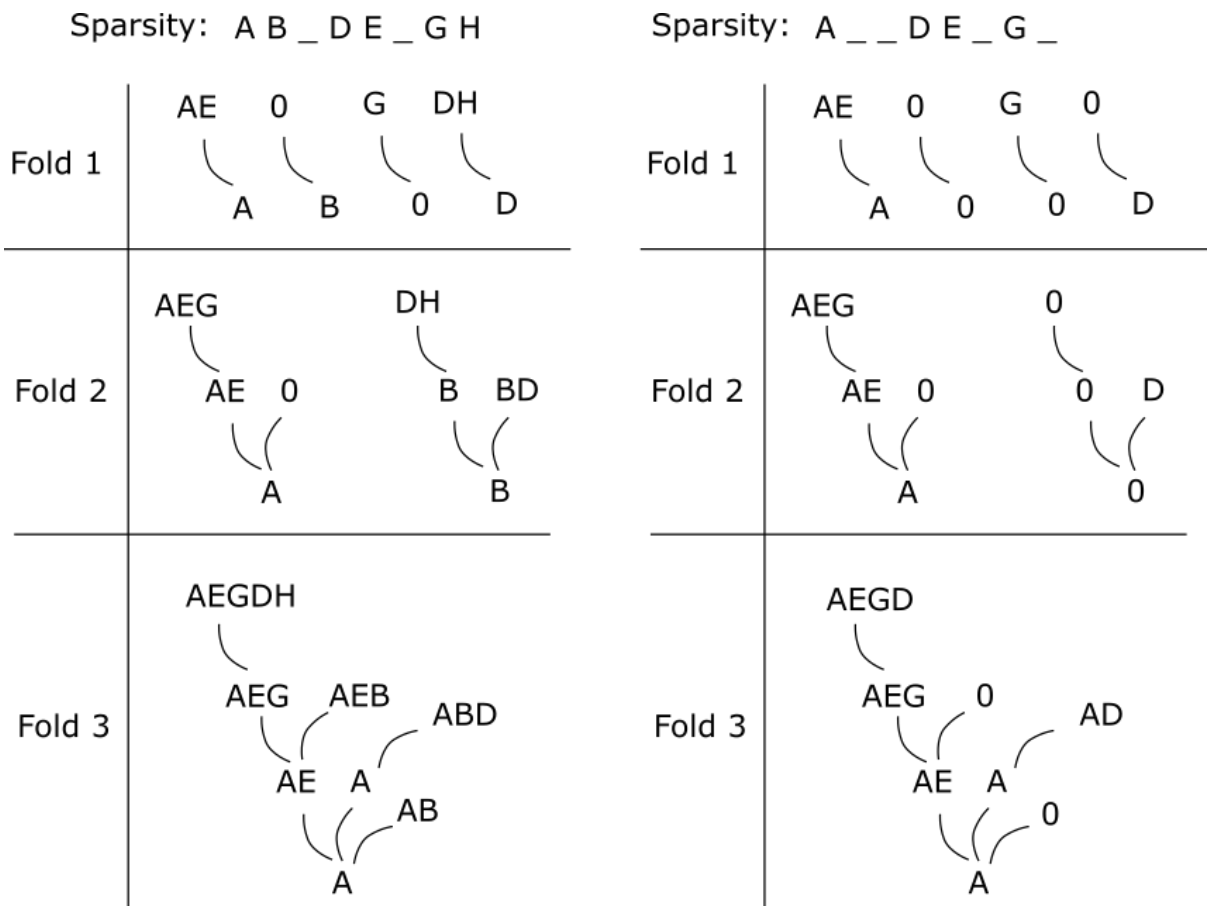


Figure 115. Demonstration of tree-folding data-structure production under conditions of tree sparsity. (Top) underscores indicate absent nodes, (left) two of eight nodes absent, (right) four of eight nodes absent.

This algorithm only applies to the discovery of node merges for a single tree-depth and has been given as an example of a binary tree. However, we can easily convert this system to any alphabet in a power of 2 by adding intermediate depths to the data-structure. Another aspect of this potential solution to be considered is that all structure score calculations are multi-depth operations, so the discovery of the full set of values for one level must occur as part of an algorithm which discovers the connected information for all depths. The fully folded tree would look like Figure 116.

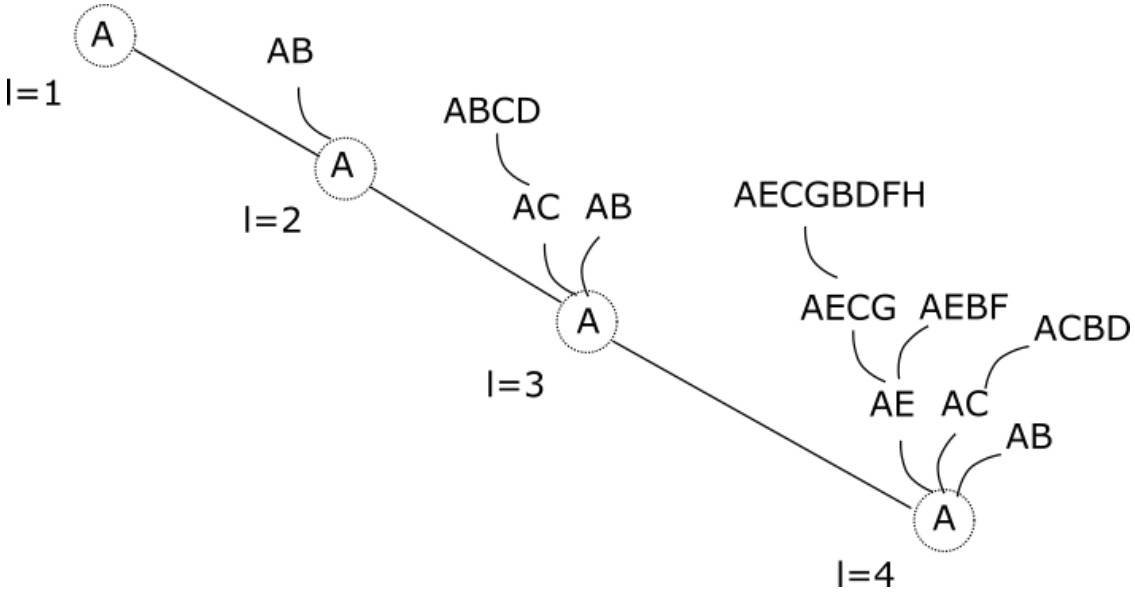


Figure 116. Fully folded/merged four-level binary tree merged via the sparse subtractive redundancy method. This now functions as a linked list between head-nodes of the proposed data structure type.

This is not by any means a thorough formalisation of the algorithm or it's efficiencies but is a proposed starting point from which further work might advance. This would require formal definitions and tests conducted on, **a)** the tree folding/structure building algorithm, **b)** the data structure subtractive navigation algorithm, **c)** the pairing function which associates the results obtained from the iterations of *(b)* over the set of all depths, such that the structure scores could be calculated.

The implementation of this algorithm would likely go a long way towards extending the usefulness of the structure score metric to whole genomes, however this is a task which would require substantial investment and has not therefore yet been completed due to the time constraints of this thesis.

6. Bibliography

Abu-Daya, A., Khokha, M.K. & Zimmerman, L.B., 2012. The Hitchhiker's guide to Xenopus genetics. *Genesis*.

Acosta-Serrano, A., Almeida, I.C., Freitas-Junior, L.H., Yoshida, N. & Schenkman, S., 2001. The mucin-like glycoprotein super-family of Trypanosoma cruzi: Structure and biological roles. *Molecular and Biochemical Parasitology*, 114(2), pp.143–150.

Addison, J.A., 2009. Distribution and impacts of invasive earthworms in Canadian forest ecosystems. In *Ecological Impacts of Non-Native Invertebrates and Fungi on Terrestrial Ecosystems*.

Aguiar, D. & Istrail, S., 2012. HapCompass: A Fast Cycle Basis Algorithm for Accurate Haplotype Assembly of Sequence Data. *Journal of Computational Biology*, 19(6), pp.577–590. Available at: http://online.liebertpub.com/doi/abs/10.1089/cmb.2012.0084.

Aïnouche, M.L., Fortune, P., Salmon, A., Parisod, C., Grandbastien, M.A., Fukunaga, K., Ricou, M. & Misset, M.-T., 2009. Hybridization, polyploidy and invasion: lessons from Spartina (Poaceae). *Biological Invasions*, 11(5), pp.1159–1173. Available at: https://hal.archives-ouvertes.fr/hal-00386234.

Aïssani, B. & Bernardi, G., 1991. CpG islands, genes and isochores in the genomes of vertebrates. *Gene*, 106(2), pp.185–195.

Allen, S.K. & Gaffney, P.M., 1993. Genetic confirmation of hybridization between Crassostrea gigas (Thunberg) and Crassostrea rivularis (Gould). *Aquaculture*, 113(4), pp.291–300.

Alves, M.J., Coelho, M.M. & Collares-Pereira, M.J., 2001. Evolution in action through hybridisation and polyploidy in an Iberian freshwater fish: A genetic review. *Genetica*.

Anders, S., Pyl, P.T. & Huber, W., 2015. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*.

Andersen, A.C., Flores, J.F. & Hourdez, S., 2006. Comparative branchial plume biometry between two extreme ecotypes of the hydrothermal vent tubeworm Ridgeia piscesae. *Canadian Journal of Zoology*.

Anderson, C., Cunha, L., Sechi, P., Kille, P. & Spurgeon, D., 2017. Genetic variation in populations of the earthworm, Lumbricus rubellus, across contaminated mine sites. *BMC Genetics*, 18(1).

Andre, J., King, R.A., Stürzenbaum, S.R., Kille, P., Hodson, M.E. & Morgan, A.J., 2010. Molecular genetic differentiation in earthworms inhabiting a heterogeneous Pb-polluted landscape. *Environmental Pollution*.

Andrews, S., 2014. FastQC. *Barbraham Institute*.

Arnegard, M.E., McGee, M.D., Matthews, B., Marchinko, K.B., Conte, G.L., Kabir, S., Bedford, N., Bergek, S., Chan, Y.F., Jones, F.C., Kingsley, D.M., Peichel, C.L. & Schluter, D., 2014. Genetics of ecological divergence during speciation. *Nature*, 511(7509), pp.307–311. Available at: http://dx.doi.org/10.1038/nature13301.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G., 2000. Gene ontology: Tool for the unification of biology. *Nature Genetics*.

Ayling, L.J. & Griffin, D.K., 2002. The evolution of sex chromosomes. *Cytogenetic and Genome*

Research.

Bakhtadze, N.G., Bakhtadze, G.I. & Kvavadze, E., 2008. The chromosome numbers of Georgian earthworms (Oligochaeta: Lumbricidae). *Comparative Cytogenetics*.

Barros, S.P. & Offenbacher, S., 2009. Epigenetics: Connecting environment and genotype to phenotype and disease. *Journal of Dental Research*.

Bartel, D.P., 2018. Metazoan MicroRNAs. *Cell*.

Te Beest, M., Le Roux, J.J., Richardson, D.M., Brysting, A.K., Suda, J., Kubešová, M. & Pyšek, P., 2012. The more the better? The role of polyploidy in facilitating plant invasions. *Annals of Botany*.

Berger, E., Yorukoglu, D., Peng, J. & Berger, B., 2014. HapTree: A novel bayesian framework for single individual polyplotyping using NGS data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. pp. 18–19.

Blakemore, R.J., 2012. *Cosmopolitan earthworms: an eco-taxonomic guide to the peregrine species of the world*, Robert J. Blakemore.

Blattner, F.R., 1997. The Complete Genome Sequence of Escherichia coli K-12. *Science*, 277(5331), pp.1453–1462. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.277.5331.1453.

Boetzer, M., Henkel, C. V., Jansen, H.J., Butler, D. & Pirovano, W., 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, 27(4), pp.578–579.

Boetzer, M. & Pirovano, W., 2014. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics*, 15(1), p.211. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-211.

Bolger, A.M., Lohse, M. & Usadel, B., 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), pp.2114–2120.

Bolger, M.E., Arsova, B. & Usadel, B., 2018. Plant genome and transcriptome annotations: From misconceptions to simple solutions. *Briefings in Bioinformatics*.

Bouchama, A., Aziz, M.A., Mahri, S. Al, Gabere, M.N., Dlamy, M. Al, Mohammad, S., Abbad, M. Al & Hussein, M., 2017. A Model of Exposure to Extreme Environmental Heat Uncovers the Human Transcriptome to Heat Stress. *Scientific Reports*, 7(1).

Bushnell & Brian, 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner. *Conference: 9th Annual Genomics of Energy & Environment Meeting*.

Cai, Y., Yu, X., Hu, S. & Yu, J., 2009. A Brief Review on the Mechanisms of miRNA Regulation. *Genomics, Proteomics and Bioinformatics*.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L., 2009. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1), p.421. Available at: http://www.biomedcentral.com/1471-2105/10/421.

Cameron, E.K., Bayne, E.M. & Coltman, D.W., 2008. Genetic structure of invasive earthworms Dendrobaena octaedra in the boreal forest of Alberta: Insights into introduction mechanisms. *Molecular Ecology*.

Cavalli-Sforza, L.L. & Feldman, M.W., 1976. Evolution of continuous variation: direct approach through joint distribution of genotypes and phenotypes. *Proceedings of the National Academy of Sciences*.

Cedar, H. & Bergman, Y., 2012. Programming of DNA Methylation Patterns. *Annual Review of Biochemistry*.

Chaisson, M.J. & Tesler, G., 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, 13(1), p.238. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-238.

Chang, A.J. & Bargmann, C.I., 2008. Hypoxia and the HIF-1 transcriptional pathway reorganize a neuronal circuit for oxygen-dependent behavior in Caenorhabditis elegans. *Proceedings of the National Academy of Sciences*.

Chapman, J.A., Kirkness, E.F., Simakov, O., Hampson, S.E., Mitros, T., Weinmaier, T., Rattei, T., Balasubramanian, P.G., Borman, J., Busam, D., Disbennett, K., Pfannkoch, C., Steele, R.E., et al., 2010. The dynamic genome of Hydra. *Nature*, 464(7288), pp.592–596.

Chor, B., Horn, D., Goldman, N., Levy, Y. & Massingham, T., 2009. Genomic DNA k-mer spectra: Models and modalities. *Genome Biology*, 10(10).

Chuang, S.C. & Chen, J.H., 2013. Photooxidation and antioxidant responses in the earthworm Amynthas gracilis exposed to environmental levels of ultraviolet B radiation. *Comp Biochem Physiol A Mol Integr Physiol*, 164(3), pp.429–437. Available at: https://www.ncbi.nlm.nih.gov/pubmed/23164536.

Clermont, O., Christenson, J.K., Denamur, E. & Gordon, D.M., 2013. The Clermont Escherichia coli phylo-typing method revisited: Improvement of specificity and detection of new phylo-groups. *Environmental Microbiology Reports*, 5(1), pp.58–65.

Colacevich, A., Sierra, M.J., Borghini, F., Millán, R. & Sanchez-Hernandez, J.C., 2011. Oxidative stress in earthworms short- and long-term exposed to highly Hg-contaminated soils. *Journal of Hazardous Materials*.

Colbourne, J.K., Pfrender, M.E., Gilbert, D., Thomas, W.K., Tucker, A., Oakley, T.H., Tokishita, S., Aerts, A., Arnold, G.J., Basu, M.K., Bauer, D.J., Cáceres, C.E., Boore, J.L., et al., 2011. The ecoresponsive genome of Daphnia pulex. *Science*.

Cole, P.D., Guest, J.E., Queiroz, G., Wallenstein, N., Pacheco, J.M., Gaspar, J.L., Ferreira, T. & Duncan, A.M., 1999. Styles of volcanism and volcanic hazards on Furnas volcano; Sao Miguel, Azores. *Journal of Volcanology and Geothermal Research*.

Comai, L., 2000. Genetic and epigenetic interactions in allopolyploid plants. In *Plant Gene Silencing*.

Comings, D.E. & MacMurray, J.P., 2000. Molecular heterosis: A review. *Molecular Genetics and Metabolism*, 71(1–2), pp.19–31.

Consortium, T.H.G.S., 2006. Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, 443(7114), pp.931–949. Available at: papers://baa045fb-5c1a-4eda-a33d-dc5addb0683f/Paper/p635.

Cordovado, S.K., Hendrix, M., Greene, C.N., Mochal, S., Earley, M.C., Farrell, P.M., Kharrazi, M., Hannon, W.H. & Mueller, P.W., 2012. CFTR mutation analysis and haplotype associations in CF patients. *Molecular Genetics and Metabolism*.

Cosin, D.J.D., Novo, M. & Fernández, R., 2011. Reproduction of earthworms: sexual selection and parthenogenesis. In *Biology of Earthworms*. Springer, pp. 69–86.

Coyne, J.A. & Allen Orr, H., 1998. The evolutionary genetics of speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 353(1366), pp.287–305. Available at:

http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.1998.0210.

Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E., 2004. WebLogo: A sequence logo generator. *Genome Research*.

Cruickshank, T.E. & Hahn, M.W., 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23(13), pp.3133–3157.

Culley, T.M. & Hardiman, N.A., 2009. The role of intraspecific hybridization in the evolution of invasiveness: a case study of the ornamental pear tree Pyrus calleryana. *Biological Invasions*, 11(5), pp.1107–1119. Available at: http://dx.doi.org/10.1007/s10530-008-9386-z.

Cunha, L., Campos, I., Montiel, R., Rodrigues, A. & Morgan, A.J., 2011a. Morphometry of the epidermis of an invasive megascoelecid earthworm (Amynthas gracilis, Kinberg 1867) inhabiting actively volcanic soils in the Azores archipelago. *Ecotoxicology and Environmental Safety*, 74(1), pp.25–32.

Cunha, L., Campos, I., Montiel, R., Rodrigues, A. & Morgan, A.J., 2011b. Morphometry of the epidermis of an invasive megascoelecid earthworm (Amynthas gracilis, Kinberg 1867) inhabiting actively volcanic soils in the Azores archipelago. *Ecotoxicology and Environmental Safety*.

Davidson, A.M., Jennions, M. & Nicotra, A.B., 2011. Do invasive species show higher phenotypic plasticity than native species and, if so, is it adaptive? A meta-analysis. *Ecology Letters*.

Dawson, M.A. & Kouzarides, T., 2012. Cancer epigenetics: From mechanism to therapy. *Cell*.

Dehal, P. & Boore, J.L., 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology*.

Dennis, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H. & Lempicki, R.A., 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*.

Ding, Q., Hu, Y., Xu, S., Wang, C.C., Li, H., Zhang, R., Yan, S., Wang, J. & Jin, L., 2014. Neanderthal origin of the haplotypes carrying the functional variant Val92Met in the MC1R in modern humans. *Molecular Biology and Evolution*, 31(8), pp.1994–2003.

Domínguez, J., Aira, M., Breinholt, J.W., Stojanovic, M., James, S.W. & Pérez-Losada, M., 2015. Underground evolution: New roots for the old tree of lumbricid earthworms. *Molecular Phylogenetics and Evolution*, 83, pp.7–19.

Durbeej, M., 2010. Laminins. *Cell and Tissue Research*, 339(1), pp.259–268.

Eisenberg, E. & Levanon, E.Y., 2013. Human housekeeping genes, revisited. *Trends in Genetics*.

Ellstrand, N.C., 2009. Evolution of invasiveness in plants following hybridization. *Biological Invasions*, 11(5), pp.1089–1091. Available at: http://dx.doi.org/10.1007/s10530-008-9389-9.

Ellstrand, N.C. & Schierenbeck, K.A., 2000. Hybridization as a stimulus for the evolution of invasiveness in plants? *Proc Natl Acad Sci U S A*, 97(13), pp.7043–7050.

EMBL, SIB Swiss Institute of Bioinformatics & Protein Information Resource (PIR), 2013. UniProt. In *Nucleic acids research*. pp. 41: D43-D47.

Estoup, A., Ravigné, V., Hufbauer, R., Vitalis, R., Gautier, M. & Facon, B., 2016. Is There a Genetic Paradox of Biological Invasion? *Annual Review of Ecology, Evolution, and Systematics*.

Evans, B.J., 2008. Genome evolution and speciation genetics of clawed frogs (Xenopus and Silurana). *Frontiers in Bioscience*.

Facon, B., Jarne, P., Pointier, J.P. & David, P., 2005. Hybridization and invasiveness in the freshwater snail Melanoides tuberculata: Hybrid vigour is more important than increase in genetic variance. *Journal of Evolutionary Biology*.

Feder, J.L. & Nosil, P., 2010. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution; international journal of organic evolution*.

Feinberg, A.P. & Irizarry, R.A., 2010. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proceedings of the National Academy of Sciences*.

Feldman, M. & Levy, A.A., 2005. Allopolyploidy - A shaping force in the evolution of wheat genomes. *Cytogenetic and Genome Research*.

Ferez-Vilar, J. & Hill, R.L., 1999. The structure and assembly of secreted mucins. *Journal of Biological Chemistry*, 274(45), pp.31751–31754.

Flores, K., Wolschin, F., Corneveaux, J.J., Allen, A.N., Huentelman, M.J. & Amdam, G. V., 2012. Genome-wide association between DNA methylation and alternative splicing in an invertebrate. *BMC Genomics*.

Flot, J.F., Hespeels, B., Li, X., Noel, B., Arkhipova, I., Danchin, E.G.J., Hejnol, A., Henrissat, B., Koszul, R., Aury, J.M., Barbe, V., Barthélémy, R.M., Bast, J., Bazykin, G.A., Chabrol, O., Couloux, A., Da Rocha, M., Da Silva, C., Gladyshev, E., Gouret, P., Hallatschek, O., Hecox-Lea, B., Labadie, K., Lejeune, B., Piskurek, O., Poulain, J., Rodriguez, F., Ryan, J.F., Vakhrusheva, O.A., Wajnberg, E., Wirth, B., Yushenova, I., Kellis, M., Kondrashov, A.S., Welch, D.B.M., Pontarotti, P., Weissenbach, J., Wincker, P., Jaillon, O. & Van Doninck, K., 2013. Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga. *Nature*.

Floudas, D., Binder, M., Riley, R., Barry, K., Blanchette, R.A., Henrissat, B., Martínez, A.T., Otillar, R., Spatafora, J.W., Yadav, J.S., Aerts, A., Benoit, I., Hibbett, D.S., et al., 2012. The paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science*.

Frankham, R., Ballou, J.D., Eldridge, M.D.B., Lacy, R.C., Ralls, K., Dudash, M.R. & Fenster, C.B., 2011. Predicting the Probability of Outbreeding Depression. *Conservation Biology*, 25(3), pp.465–475.

Fraser, J.A., Diezmann, S., Subaran, R.L., Allen, A., Lengeler, K.B., Dietrich, F.S. & Heitman, J., 2004. Convergent evolution of chromosomal sex-determining regions in the animal and fungal kingdoms. *PLoS Biology*.

Friedländer, M.R., MacKowiak, S.D., Li, N., Chen, W. & Rajewsky, N., 2012. MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*.

Friedman, R.C., Farh, K.K.H., Burge, C.B. & Bartel, D.P., 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*.

Frouz, J., Elhottová, D., Kuráž, V. & Šourková, M., 2006. Effects of soil macrofauna on other soil biota and soil formation in reclaimed and unreclaimed post mining sites: Results of a field microcosm experiment. *Applied Soil Ecology*, 33(3), pp.308–320.

Garrison, E. & Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. *arXiv preprint*, (arXiv:1207.3907 [q-bio.GN]).

Gatzmann, F., Falckenhayn, C., Gutekunst, J., Hanna, K., Raddatz, G., Carneiro, V.C. & Lyko, F., 2018. The methylome of the marbled crayfish links gene body methylation to stable expression of poorly accessible genes. *Epigenetics & Chromatin*.

Gavrilets, S., 2004. Fitness landscapes and the origin of species. *Monographs in Population Biology*, 41, p.476. Available at: http://press.princeton.edu/titles/7799.html.

Geeleher, P., Huang, S.R., Gamazon, E.R., Golden, A. & Seoighe, C., 2012. The regulatory effect of miRNAs is a heritable genetic trait in humans. *BMC Genomics*.

Geerts, A.N., Vanoverbeke, J., Vanschoenwinkel, B., Van Doorslaer, W., Feuchtmayr, H., Atkinson, D., Moss, B., Davidson, T.A., Sayer, C.D. & De Meester, L., 2015. Rapid evolution of thermal tolerance in the water flea Daphnia. *Nature Climate Change*.

Gendler, S.J. & Spicer, A.P., 1995. Epithelial Mucin Genes. *Annual Review of Physiology*, 57(1), pp.607–634. Available at: http://www.annualreviews.org/doi/10.1146/annurev.ph.57.030195.003135.

Gerdol, M., Venier, P. & Pallavicini, A., 2015. The genome of the Pacific oyster Crassostrea gigas brings new insights on the massive expansion of the C1q gene family in Bivalvia. *Developmental and Comparative Immunology*, 49(1), pp.59–71.

Ghandi, M., Mohammad-Noori, M. & Beer, M.A., 2014. Robust k-mer frequency estimation using gapped k-mers. *Journal of Mathematical Biology*, 69(2), pp.469–500.

Gilad, Y., Rifkin, S.A. & Pritchard, J.K., 2008. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics*.

Gini, C., 1912. *Variabilità e mutabilità*,

Giska, I., Sechi, P. & Babik, W., 2015a. Deeply divergent sympatric mitochondrial lineages of the earthworm Lumbricus rubellus are not reproductively isolated. *BMC Evolutionary Biology*, 15(1).

Giska, I., Sechi, P. & Babik, W., 2015b. Deeply divergent sympatric mitochondrial lineages of the earthworm Lumbricus rubellus are not reproductively isolated. *BMC Evolutionary Biology*.

Gladyshev, E.A. & Arkhipova, I.R., 2010. Genome structure of bdelloid rotifers: Shaped by asexuality or desiccation? In *Journal of Heredity*.

Gluckman, P.D., Hanson, M.A., Spencer, H.G. & Bateson, P., 2005. Environmental influences during development and their later consequences for health and disease: implications for the interpretation of empirical studies. *Proceedings. Biological sciences / The Royal Society*.

Goll, M.G. & Bestor, T.H., 2005. EUKARYOTIC CYTOSINE METHYLTRANSFERASES. *Annual Review of Biochemistry*.

Gregory, T.R., Nicol, J.A., Tamm, H., Kullman, B., Kullman, K., Leitch, I.J., Murray, B.G., Kapraun, D.F., Greilhuber, J. & Bennett, M.D., 2007. Eukaryotic genome size databases. *Nucleic Acids Research*.

Grohme, M.A., Schloissnig, S., Rozanski, A., Pippel, M., Young, G.R., Winkler, S., Brandl, H., Henry, I., Dahl, A., Powell, S., Hiller, M., Myers, E. & Rink, J.C., 2018. The genome of Schmidtea mediterranea and the evolution of core cellular mechanisms. *Nature*, 554, p.56. Available at: http://dx.doi.org/10.1038/nature25473.

Grotz, N., Fox, T., Connolly, E., Park, W., Guerinot, M.L. & Eide, D., 1998. Identification of a family of zinc transporter genes from Arabidopsis that respond to zinc deficiency. *Proceedings of the National Academy of Sciences*, 95(12), pp.7220–7224. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.95.12.7220.

Gu, X.Q. & Haddad, G.G., 1999. Drosophila neurons respond differently to hypoxia and cyanide than

rat neurons. *Brain Research*.

Guerinot, M. Lou, 2000. The ZIP family of metal transporters. *Biochimica et Biophysica Acta - Biomembranes*, 1465(1–2), pp.190–198.

Guo, L., Zhang, S., Rubinstein, B., Ross, E. & Alvarado, A.S., 2016. Widespread maintenance of genome heterozygosity in Schmidtea mediterranea. *Nature Ecology &Amp; Evolution*, 1, p.19. Available at: http://dx.doi.org/10.1038/s41559-016-0019.

Gustafson, K.D., Kensinger, B.J., Bolek, M.G. & Luttbeg, B., 2014. Distinct snail (Physa) morphotypes from different habitats converge in shell shape and size under common garden conditions. *Evolutionary Ecology Research*.

Haas, Brian J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., Macmanes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., Leduc, R.D., Friedman, N. & Regev, A., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8), pp.1494–1512.

Haas, Brian J, Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N. & Regev, A., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8), pp.1494–1512. Available at: http://www.nature.com/doifinder/10.1038/nprot.2013.084.

Haigh, J., 1978. The accumulation of deleterious genes in a population-Muller's Ratchet. *Theoretical Population Biology*, 14(2), pp.251–267.

Hanukoglu, I. & Hanukoglu, A., 2016. Epithelial sodium channel (ENaC) family: Phylogeny, structure-function, tissue distribution, and associated inherited diseases. *Gene*, 579(2), pp.95–132.

Hayes, J.E., Bartoshuk, L.M., Kidd, J.R. & Duffy, V.B., 2008. Supertasting and PROP bitterness depends on more than the TAS2R38 gene. *Chemical Senses*.

He, Y., Jones, C.R., Fujiki, N., Xu, Y., Guo, B., Holder, J.L., Rossner, M.J., Nishino, S. & Fu, Y.H., 2009. The transcriptional repressor DEC2 regulates sleep length in mammals. *Science*.

Heggelund, L.R., Diez-Ortiz, M., Lofts, S., Lahive, E., Jurkschat, K., Wojnarowicz, J., Cedergreen, N., Spurgeon, D. & Svendsen, C., 2014. Soil pH effects on the comparative toxicity of dissolved zinc, non-nano and nano ZnO to the earthworm Eisenia fetida. *Nanotoxicology*.

Höckner, M., Dallinger, R. & Stürzenbaum, S.R., 2015. Metallothionein gene activation in the earthworm (Lumbricus rubellus). *Biochemical and Biophysical Research Communications*.

Hollenstein, K., de Graaf, C., Bortolato, A., Wang, M.-W., Marshall, F.H. & Stevens, R.C., 2014. Insights into the structure of class B GPCRs. *Trends in Pharmacological Sciences*, 35(1), pp.12–22.

Holt, C. & Yandell, M., 2011. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*.

Hongell, K. & Terhivuo, J., 1989. Chromosomal status of the parthenogenetic earthworm Dendrobaena octaedra (Sav.) (Oligochaeta: Lumbricidae) in southern Finland. *Hereditas*.

Huang, S., Chen, Z., Yan, X., Yu, T., Huang, G., Yan, Q., Pontarotti, P.A. ntoin., Zhao, H., Li, J., Yang, P., Wang, R., Li, R., Tao, X., Deng, T., Wang, Y., Li, G., Zhang, Q., Zhou, S., You, L., Yuan, S., Fu, Y., Wu, F., Dong, M., Chen, S. & Xu, A., 2014. Decelerated genome evolution in modern

vertebrates revealed by analysis of multiple lancelet genomes. *Nature communications*, 5, p.5896.

Huang, S., Kang, M. & Xu, A., 2017. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics*, 33(16), pp.2577–2579. Available at: http://dx.doi.org/10.1093/bioinformatics/btx220.

Hughes, A.L., 2012. Evolution of adaptive phenotypic traits without positive Darwinian selection. *Heredity*.

Hurka, H., Bleeker, W. & Neuffer, B., 2003. Evolutionary Processes Associated with Biological Invasions in the Brassicaceae. *Biological Invasions*, 5(4), pp.281–292. Available at: http://dx.doi.org/10.1023/B:BINV.0000005571.19401.81.

Huvet, A., Gérard, A., Ledu, C., Phélipot, P., Heurtebise, S. & Boudry, P., 2002. Is fertility of hybrids enough to conclude that the two oysters Crassostrea gigas and Crassostrea angulata are the same species? *Aquatic Living Resources*, 15(1), pp.45–52.

ISO, 1992. *Determination of water retention characteristic-Laboratory methods*, Geneva.

Jaenike, J. & Selander, R.K., 1979. Evolution and ecology of parthenogenesis in earthworms. *Integrative and Comparative Biology*.

Jänsch, S., Steffens, L., Höfer, H., Horak, F., Roß-Nickoll, M., Russell, D., Toschki, A. & Römbke, J., 2013. State of knowledge of earthworm communities in German soils as a basis for biological soil quality assessment. *Soil Organisms*.

Jeltsch, A. & Jurkowska, R.Z., 2014. New concepts in DNA methylation. *Trends in Biochemical Sciences*.

Jeong, H., Tombor, B., Albert, R., Oltval, Z.N. & Barabásl, A.L., 2000. The large-scale organization of metabolic networks. *Nature*, 407(6804), pp.651–654.

Johnson, R.E., Washington, M.T., Prakash, S. & Prakash, L., 2000. Fidelity of human DNA polymerase η. *Journal of Biological Chemistry*.

Jones, G.L., Wills, A., Morgan, A.J., Thomas, R.J., Kille, P. & Novo, M., 2016. The worm has turned: Behavioural drivers of reproductive isolation between cryptic lineages. *Soil Biology and Biochemistry*, 98, pp.11–17.

Jones, P. a & Laird, P.W., 1999. Cancer epigenetics comes of age. *Nature Genetics*.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., Kohara, Y., Fujiyama, A., Hayashi, T. & Itoh, T., 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, 24(8), pp.1384–1395.

Karamichalis, R., Kari, L., Konstantinidis, S. & Kopecki, S., 2015. An investigation into inter- and intragenomic variations of graphic genomic signatures. *BMC Bioinformatics*, 16(1).

Karamichalis, R., Kari, L., Konstantinidis, S., Kopecki, S. & Solis-Reyes, S., 2016. Additive methods for genomic signatures. *BMC Bioinformatics*, 17(1).

Karlsen, B.O., Klingan, K., Emblem, Å., Jørgensen, T.E., Jueterbock, A., Furmanek, T., Hoarau, G., Johansen, S.D., Nordeide, J.T. & Moum, T., 2013. Genomic divergence between the migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology*.

Katolikova, M., Khaitov, V., Väinölä, R., Gantsevich, M. & Strelkov, P., 2016. Genetic, ecological and

morphological distinctness of the blue mussels Mytilus trossulus gould and M. edulis l. in the White Sea. *PLoS ONE*, 11(4).

Kauppi, L., Barchi, M., Baudat, F., Romanienko, P.J., Keeney, S. & Jasin, M., 2011. Distinct properties of the XY pseudoautosomal region crucial for male meiosis. *Science*.

Kawakami, T., Backström, N., Burri, R., Husby, A., Olason, P., Rice, A.M., Ålund, M., Qvarnström, A. & Ellegren, H., 2014. Estimation of linkage disequilibrium and interspecific gene flow in Ficedula flycatchers by a newly developed 50k single-nucleotide polymorphism array. *Molecular Ecology Resources*.

Keightley, P.D. & Eyre-Walker, A., 2000. Deleterious mutations and the evolution of sex. *Science*.

Keith, N., Tucker, A.E., Jackson, C.E., Sung, W., Lledó, J.I.L., Schrider, D.R., Schaack, S., Dudycha, J.L., Ackerman, M., Younge, A.J., Shaw, J.R. & Lynch, M., 2016. High mutational rates of large-scale duplication and deletion in Daphnia pulex. *Genome Research*.

Kellenberger, S. & Schild, L., 2002. Epithelial Sodium Channel/Degenerin Family of Ion Channels: A Variety of Functions for a Shared Structure. *Physiological Reviews*, 82(3), pp.735–767. Available at: http://physrev.physiology.org/lookup/doi/10.1152/physrev.00007.2002.

Keller, O., Kollmar, M., Stanke, M. & Waack, S., 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*.

Keller, S.R. & Taylor, D.R., 2010. Genomic admixture increases fitness during a biological invasion. *Journal of Evolutionary Biology*.

Keller, T.E., Han, P. & Yi, S. V., 2016. Evolutionary transition of promoter and gene body DNA methylation across invertebrate-vertebrate boundary. *Molecular Biology and Evolution*.

Kim, S.-T., Sultan, S.E. & Donoghue, M.J., 2008. Allopolyploid speciation in Persicaria (Polygonaceae): Insights from a low-copy nuclear region. *Proceedings of the National Academy of Sciences*.

King, R.A., Tibble, A.L. & Symondson, W.O.C., 2008. Opening a can of worms: Unprecedented sympatric cryptic diversity within British lumbricid earthworms. *Molecular Ecology*, 17(21), pp.4684–4698.

Korf, I., 2004. Gene finding in novel genomes. *BMC Bioinformatics*.

Kozomara, A. & Griffiths-Jones, S., 2014. miRBase. *Nucleic acids research*.

Kreps, J.A., 2002. Transcriptome Changes for Arabidopsis in Response to Salt, Osmotic, and Cold Stress. *PLANT PHYSIOLOGY*, 130(4), pp.2129–2141. Available at: http://www.plantphysiol.org/cgi/doi/10.1104/pp.008532.

Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M. & Blaxter, M., 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Frontiers in Genetics*, 4. Available at: http://journal.frontiersin.org/article/10.3389/fgene.2013.00237/abstract.

Kutscher, L.M., 2014. Forward and reverse mutagenesis in C. elegans. *WormBook*.

Lamichhaney, S., Han, F., Webster, M.T., Andersson, L., Grant, B.R. & Grant, P.R., 2018. Rapid hybrid speciation in Darwin's Finches. *Science*, 359(6372), pp.224–228.

Lande, R., 1980. Sexual Dimorphism, Sexual Selection, and Adaptation in Polygenic Characters. *Evolution*.

Lane, N. & Martin, W., 2010. The energetics of genome complexity. *Nature*.

Langmead, B., Salzberg, S.L. & Langmead, 2013. Bowtie2. *Nature methods*, 9(4), pp.357–359. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22388286%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3322381%0Ahttp://www.nature.com/doifinder/10.1038/nmeth.1923.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*.

Lavelle, P., 1997a. Biology and ecology of earthworms. *Agriculture, Ecosystems & Environment*, 64, pp.78–79.

Lavelle, P., 1997b. Biology and ecology of earthworms. *Agriculture, Ecosystems & Environment*.

Lavelle, P., 1988. Earthworm activities and the soil system. *Biology and Fertility of Soils*.

Lee, C.E. & Gelembiuk, G.W., 2008. Evolutionary origins of invasive populations. *Evolutionary Applications*.

Lee, J.C., Biasci, D., Roberts, R., Gearry, R.B., Mansfield, J.C., Ahmad, T., Prescott, N.J., Satsangi, J., Wilson, D.C., Jostins, L., Anderson, C.A., Traherne, J.A., Lyons, P.A., Parkes, M. & Smith, K.G.C., 2017. Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. *Nature Genetics*.

Lee, T.H.F., Zhai, J.X., Meyers, B.C., Gioia, L.C., Kate, M.P., McCourt, R., Gould, B., Coutts, S.B., Dowlatshahi, D., Asdaghi, N., Jeerakathil, T., Hill, M.D., Ullman, L., et al., 2015. The Pfam protein families database. *Bioinformatics*, 8(1), pp.9–15. Available at: http://hsrl.rutgers.edu/research/shellfishdiseasestudies/eid/dermo.htm%5Cn%3CGo to ISI%3E://WOS:A1989T194400010%5Cn%3CGo to ISI%3E://WOS:A1994PV82000001%5Cnhttp://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=889890.

Leitch, A.R. & Leitch, I.J., 2008. Genomic plasticity and the diversity of polyploid plants. *Science*.

Lewis, B.P., Burge, C.B. & Bartel, D.P., 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*.

Li, D., Liu, C.M., Luo, R., Sadakane, K. & Lam, T.W., 2015. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10), pp.1674–1676.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*.

Li, Y., Leong Hou, U., Yiu, M.L. & Gong, Z., 2015. Quick-motif: An efficient and scalable framework for exact motif discovery. In *Proceedings - International Conference on Data Engineering*.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B. & Fan, W., 2012. Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph. *Briefings in Functional Genomics*.

Liebers, R., Rassoulzadegan, M. & Lyko, F., 2014. Epigenetic Regulation by Heritable RNA. *PLoS Genetics*.

Lin, J., 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1), pp.145–151.

Liscovitch-Brauer, N., Alon, S., Porath, H.T., Elstein, B., Unger, R., Ziv, T., Admon, A., Levanon, E.Y., Rosenthal, J.J.C. & Eisenberg, E., 2017. Trade-off between Transcriptome Plasticity and Genome

Evolution in Cephalopods. *Cell*.

Liu, B., Fang, L., Wang, S., Wang, X., Li, H. & Chou, K.C., 2015. Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. *Journal of Theoretical Biology*.

Liu, Y., Schmidt, B. & Maskell, D.L., 2011. An ultrafast scalable many-core motif discovery algorithm for multiple GPUs. In *IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum*.

Liu, Y., Schröder, J. & Schmidt, B., 2013. Musket: A multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics*, 29(3), pp.308–315.

Lokk, K., Modhukur, V., Rajashekar, B., Märtens, K., Mägi, R., Kolde, R., Koltšina, M., Nilsson, T.K., Vilo, J., Salumets, A. & Tõnisson, N., 2014. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biology*.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O. & Borodovsky, M., 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*.

Love, M.I., Huber, W. & Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*.

Lu, Y., Zhang, S., Shah, T., Xie, C., Hao, Z., Li, X., Farkhari, M., Ribaut, J.-M., Cao, M., Rong, T. & Xu, Y., 2010. Joint linkage-linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. *Proceedings of the National Academy of Sciences*.

Luo, M., Sun, L. & Hu, J., 2009. Neural detection of gases - carbon dioxide, oxygen - in vertebrates and invertebrates. *Current Opinion in Neurobiology*.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J. & Wang, J., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1(1), p.18.

Luo, Y.-J., Takeuchi, T., Koyanagi, R., Yamada, L., Kanda, M., Khalturina, M., Fujie, M., Yamasaki, S., Endo, K. & Satoh, N., 2015. The Lingula genome provides insights into brachiopod evolution and the origin of phosphate biomineralization. *Nature Communications*, 6, p.8301. Available at: http://www.nature.com/doifinder/10.1038/ncomms9301.

Lyko, F., Foret, S., Kucharski, R., Wolf, S., Falckenhayn, C. & Maleszka, R., 2010. The honey bee epigenomes: Differential methylation of brain DNA in queens and workers. *PLoS Biology*.

Lynch, M., 2000. The Evolutionary Fate and Consequences of Duplicate Genes. *Science*.

Maegawa, S., Hinkal, G., Kim, H.S., Shen, L., Zhang, L., Zhang, J., Zhang, N., Liang, S., Donehower, L.A. & Issa, J.P.J., 2010. Widespread and tissue specific age-related DNA methylation changes in mice. *Genome Research*.

Malone, C.D. & Hannon, G.J., 2009. Small RNAs as Guardians of the Genome. *Cell*.

Mannello, F., Medda, V. & Tonti, G.A., 2011. Hypoxia and neural stem cells: From invertebrates to brain cancer stem cells. *International Journal of Developmental Biology*.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., MacKay, T.F.C., McCarroll, S.A. & Visscher, P.M., 2009. Finding the missing heritability of complex diseases. *Nature*.

Mantegna, R.N., Buldyrev, S. V., Goldberger, A.L., Havlin, S., Peng, C.K., Simons, M. & Stanley, H.E., 1994. Linguistic features of noncoding DNA sequences. *Physical Review Letters*, 73(23), pp.3169–3172.

Marcais, G. & Kingsford, C., 2012. Jellyfish : A fast k-mer counter. *Tutorialis e Manuais*, (1), pp.1–8.

Martin, S.H., Dasmahapatra, K.K., Nadeau, N.J., Salazar, C., Walters, J.R., Simpson, F., Blaxter, M., Manica, A., Mallet, J. & Jiggins, C.D., 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Research*, 23(11), pp.1817–1828.

Martindale, J.L. & Holbrook, N.J., 2002. Cellular response to oxidative stress: Signaling for suicide and survival. *Journal of Cellular Physiology*.

Martoja, R. & Martoja-Pierson, M., 1970. *Técnicas de histolog{\'\i}a animal*,

Maunakea, A.K., Chepelev, I., Cui, K. & Zhao, K., 2013. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Research*.

McDaniell, R., Lee, B.K., Song, L., Liu, Z., Boyle, A.P., Erdos, M.R., Scott, L.J., Morken, M.A., Kučera, K.S., Battenhouse, A., Keefe, D., Collins, F.S., Willard, H.F., Lieb, J.D., Furey, T.S., Crawford, G.E., Iyer, V.R. & Birney, E., 2010. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*.

McGrath, C.L., Gout, J.F., Johri, P., Doak, T.G. & Lynch, M., 2014. Differential retention and divergent resolution of duplicate genes following whole-genome duplication. *Genome Research*.

Mckinnell, Z. & Wessel, G., 2012. Ligers and tigons and .....what? ....oh my! *Molecular Reproduction and Development*.

Measey, G.J., Rödder, D., Green, S.L., Kobayashi, R., Lillo, F., Lobos, G., Rebelo, R. & Thirion, J.M., 2012. Ongoing invasions of the African clawed frog, Xenopus laevis: A global review. *Biological Invasions*.

Michel, A.P., Sim, S., Powell, T.H.Q., Taylor, M.S., Nosil, P. & Feder, J.L., 2010. Widespread genomic divergence during sympatric speciation. *Proceedings of the National Academy of Sciences*, 107(21), pp.9724–9729. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.1000939107.

Moehler, J., Wegner, K.M., Reise, K. & Jacobsen, S., 2011. Invasion genetics of Pacific oyster Crassostrea gigas shaped by aquaculture stocking practices. *Journal of Sea Research*, 66(3), pp.256–262. Available at: http://www.sciencedirect.com/science/article/pii/S1385110111001201.

Monahan-Earley, R., Dvorak, A.M. & Aird, W.C., 2013. Evolutionary origins of the blood vascular system and endothelium. *Journal of Thrombosis and Haemostasis*.

NCBI, G., 2016. The Genome Database. *NCBI Handout Series*. Available at: https://www.ncbi.nlm.nih.gov/genome.

Newman, M.E.J., 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5), pp.323–351.

Nosil, P. & Schluter, D., 2011. The genes underlying the process of speciation. *Trends in Ecology and Evolution*, 26(4), pp.160–167.

Novo, M., Almodóvar, A., Fernández, R., Trigo, D., Dáaz-Cosán, D.J. & Giribet, G., 2012. Appearances can be deceptive: Different diversification patterns within a group of mediterranean earthworms (Oligochaeta, Hormogastridae). *Molecular Ecology*, 21(15), pp.3776–3793.

Novo, M., Almodóvar, A., Fernández, R., Trigo, D. & Díaz Cosín, D.J., 2010. Cryptic speciation of hormogastrid earthworms revealed by mitochondrial and nuclear data. *Molecular Phylogenetics and Evolution*, 56(1), pp.507–512.

Novo, M., Cunha, L., Maceda-Veiga, A., Talavera, J.A., Hodson, M.E., Spurgeon, D., Bruford, M.W., Morgan, A.J. & Kille, P., 2015. Multiple introductions and environmental factors affecting theestablishment of invasive species on a volcanic island. *Soil Biology and Biochemistry*.

Oakeshott, J.G., Horne, I., Sutherland, T.D. & Russell, R.J., 2003. The genomics of insecticide resistance. *Genome Biology*.

Oda, S., Fukami, T., Yokoi, T. & Nakajima, M., 2015. A comprehensive review of UDP-glucuronosyltransferase and esterases for drug development. *Drug Metabolism and Pharmacokinetics*, 30(1), pp.30–51.

Oliver, J.L., Bernaola-Galván, P., Guerrero-Garcia, J. & Román-Roldán, R., 1993. Entropic profiles of DNA sequences through chaos-game-derived images. *Journal of Theoretical Biology*, 160(4), pp.457–470.

Ounit, R., Wanamaker, S., Close, T.J. & Lonardi, S., 2015. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, 16(1).

Padilla, D.K. & Savedo, M.M., 2013. A systematic review of phenotypic plasticity in marine invertebrate and plant systems. *Advances in Marine Biology*.

Pardo-Diaz, C., Salazar, C., Baxter, S.W., Merot, C., Figueiredo-Ready, W., Joron, M., McMillan, W.O. & Jiggins, C.D., 2012. Adaptive Introgression across Species Boundaries in Heliconius Butterflies. *PLOS Genetics*, 8(6), p.e1002752. Available at: https://doi.org/10.1371/journal.pgen.1002752.

Parelho, C., Rodrigues, A. dos santos, Bernardo, F., do Carmo Barreto, M., Cunha, L., Poeta, P. & Garcia, P., 2017. Biological endpoints in earthworms (Amynthas gracilis) as tools for the ecotoxicity assessment of soils from livestock production systems. *Ecological Indicators*.

Pavliček, T., Cohen, T., Yadav, S., Glasstetter, M., Král, P. & Pearlson, O., 2016. Aneuploidy occurrence in Oligochaeta. *Ecol. Evol. Biol*, 1, pp.57–63.

Peer, K. & Taborsky, M., 2005. Outbreeding depression, but no inbreeding depression in haplodiploid ambrosia beetles with regular sibling mating. *Evolution*.

Van de Peer, Y., Maere, S. & Meyer, A., 2010. 2R or not 2R is not the question anymore. *Nature Reviews Genetics*.

Van de Peer, Y., Maere, S. & Meyer, A., 2009. The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*.

Pelosi, C., Barot, S., Capowiez, Y., Hedde, M. & Vandenbulcke, F., 2014. Pesticides and earthworms. A review. *Agronomy for Sustainable Development*.

Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R., Carter, N.P., Lee, C. & Stone, A.C., 2007. Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*.

Perry, G.H., Yang, F., Marques-Bonet, T., Murphy, C., Fitzgerald, T., Lee, A.S., Hyland, C., Stone, A.C., Hurles, M.E., Tyler-Smith, C., Eichler, E.E., Carter, N.P., Lee, C. & Redon, R., 2008. Copy number variation and evolution in humans and chimpanzees. *Genome Research*.

Pfennig, K.S., Kelly, A.L. & Pierce, A.A., 2016. Hybridization as a facilitator of species range expansion.

Proceedings of the Royal Society B: Biological Sciences.

Pigliucci, M., Murren, C.J. & Schlichting, C.D., 2006. Phenotypic plasticity and evolution by genetic assimilation. *The Journal of Experimental Biology*.

Pinel, N., Davidson, S.K. & Stahl, D.A., 2008. Verminephrobacter eiseniae gen. nov., sp. nov., a nephridial symbiont of the earthworm Eisenia foetida (Savigny). *International Journal of Systematic and Evolutionary Microbiology*.

Plytycz, B., Bigaj, J., Osikowski, A., Hofman, S., Falniowski, A., Panz, T., Grzmil, P. & Vandenbulcke, F., 2018. The existence of fertile hybrids of closely related model earthworm species, Eisenia andrei and E. fetida. *PLoS ONE*, 13(1).

Pogson, G.H., 2016. Studying the genetic basis of speciation in high gene flow marine invertebrates. *Current Zoology*, 62(6), pp.643–653.

Pop, V. V. & Pop, A.A., 2006. Lumbricid earthworm invasion in the Carpathian Mountains and some other sites in Romania. In *Biological Invasions Belowground: Earthworms as Invasive Species*.

Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Auwera, G.A. Van der, Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M.J., Neale, B., MacArthur, D.G. & Banks, E., 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*.

Pors Nielsen, S., 2004. The biological role of strontium. *Bone*.

Prentis, P.J., Wilson, J.R.U., Dormontt, E.E., Richardson, D.M. & Lowe, A.J., 2008. Adaptive evolution in invasive species. *Trends in Plant Science*.

Pushker, R., Mira, A. & Rodríguez-Valera, F., 2004. Comparative genomics of gene-family size in closely related bacteria. *Genome biology*, 5(4), p.R27.

Quang, D. & Xie, X., 2014. EXTREME: An online em algorithm for motif discovery. *Bioinformatics*.

R Development Core Team, 2016. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing Vienna Austria*, 0, p.{ISBN} 3-900051-07-0. Available at: http://www.r-project.org/.

Ramadass, K., Megharaj, M., Venkateswarlu, K. & Naidu, R., 2015. Ecological implications of motor oil pollution: Earthworm survival and soil health. *Soil Biology and Biochemistry*.

Reiber, C.L. & McGaw, I.J., 2009. A review of the open and closed circulatory systems: New terminology for complex invertebrate circulatory systems in light of current findings. *International Journal of Zoology*.

Renaut, S., Grassa, C.J., Yeaman, S., Moyers, B.T., Lai, Z., Kane, N.C., Bowers, J.E., Burke, J.M. & Rieseberg, L.H., 2013. Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, 4.

Rice, W.R., 1984. SEX CHROMOSOMES AND THE EVOLUTION OF SEXUAL DIMORPHISM. *Evolution*.

Richards, C.L., Bossdorf, O., Muth, N.Z., Gurevitch, J. & Pigliucci, M., 2006. Jack of all trades, master of some? On the role of phenotypic plasticity in plant invasions. *Ecology Letters*.

Rieger, R.M. & Purschke, G., 2005. The coelom and the origin of the annelid body plan. *Hydrobiologia*.

Rieseberg, L.H., Archer, M. a & Wayne, R.K., 1999. Transgressive segregation, adaptation and speciation. *Heredity*, 83 (Pt 4)(July), pp.363–372.

Rivière, G., 2014. Epigenetic features in the oyster Crassostrea gigas suggestive of functionally relevant promoter DNA methylation in invertebrates. *Frontiers in Physiology*.

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N., Loire, E., Lourenco, J.M., Nabholz, B., Roux, C., Tsagkogeorga, G., Weber, A.A.T., Weinert, L.A., Belkhir, K., Bierne, N., Glémin, S. & Galtier, N., 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*.

Roux, C., Tsagkogeorga, G., Bierne, N. & Galtier, N., 2013. Crossing the species barrier: Genomic hotspots of introgression between two highly divergent ciona intestinalis species. *Molecular Biology and Evolution*, 30(7).

Rowell, D.L., 1994. Soil science: methods and applications. *Soil science: methods and applications.*

Ru, D., Mao, K., Zhang, L., Wang, X., Lu, Z. & Sun, Y., 2016. Genomic evidence for polyphyletic origins and interlineage gene flow within complex taxa: a case study of Picea brachytyla in the Qinghai-Tibet Plateau. *Molecular ecology*, 25(11), pp.2373–2386.

Safonova, Y., Bankevich, A. & Pevzner, P.A., 2014. DIPSPADES: Assembler for highly polymorphic diploid genomes. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. pp. 265–279.

Saini, M., 2015. Implant biomaterials: A comprehensive review. *World Journal of Clinical Cases*.

Salmon, A., Clotault, J., Jenczewski, E., Chable, V. & Manzanares-Dauleux, M.J., 2008. Brassica oleracea displays a high level of DNA methylation polymorphism. *Plant Science*.

Sánchez, J.A., Aguilar, C., Dorado, D. & Manrique, N., 2007. Phenotypic plasticity and morphological integration in a marine modular invertebrate. *BMC Evolutionary Biology*.

Sarich, V.M. & Wilson, A.C., 1973. Generation time and genomic evolution in primates. *Science*.

Savazzi, E., 1991. Burrowing in the inarticulate brachiopod Lingula anatina. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 85(1–2), pp.101–106.

Schachtman, D.P. & Goodger, J.Q.D., 2008. Chemical root to shoot signaling under drought. *Trends in Plant Science*.

Schierenbeck, K.A. & Ellstrand, N.C., 2008. Hybridization and the evolution of invasiveness in plants and other organisms. *Biological Invasions*, 11(5), p.1093. Available at: http://dx.doi.org/10.1007/s10530-008-9388-x.

Schmitt, A.O. & Herzel, H., 1997. Estimating the entropy of DNA sequences. *Journal of Theoretical Biology*, 188(3), pp.369–377.

Schrader, S., 1994. Influence of Earthworms on the pH conditions of their environment by cutaneous mucus secretion. *Zoologischer Anzeiger*, 233, pp.211–219.

Schübeler, D., 2015. Function and information content of DNA methylation. *Nature*.

Schwarz, E.M., Hu, Y., Antoshechkin, I., Miller, M.M., Sternberg, P.W. & Aroian, R. V., 2015. The genome and transcriptome of the zoonotic hookworm Ancylostoma ceylanicum identify infection-specific gene families. *Nature Genetics*, 47(4), pp.416–422.

Schwarzenberger, A., Keith, N.R., Jackson, C.E. & Elert, E., 2017. Copy number variation of a protease gene of Daphnia: Its role in population tolerance. *Journal of Experimental Zoology Part A: Ecological and Integrative Physiology*, 327(2–3), pp.119–126. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/jez.2077.

Seehausen, O., 2004. Hybridization and adaptive radiation. *Trends in Ecology & Evolution*, 19(4), pp.198–207. Available at: http://www.sciencedirect.com/science/article/pii/S0169534704000047.

Shannon, C.E., 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(4), pp.379–423. Available at: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=%7B&%7Darnumber=6773067%7B&%7DmatchBoolean%7B%25%7D3Dtrue%7B%25%7D26rowsPerPage%7B%25%7D3D30%7B%25%7D26searchField%7B%25%7D3DSearch%7B_%7DAll%7B%25%7D26queryText%7B%25%7D3D%7B%25%7D28p%7B_%7DTitle%7B%25.

Shen, H.P., Tsai, C.F., Fang, Y.P. & Chen, J.H., 2011. Parthenogenesis, polyploidy and reproductive seasonality in the Taiwanese mountain earthworm Amynthas catenus Tsai et al., 2001 (Oligochaeta, Megascolecidae). *Pedobiologia*.

Shields, J.L., Heath, J.W. & Heath, D.D., 2010. Marine landscape shapes hybrid zone in a broadcast spawning bivalve: Introgression and genetic structure in Canadian west coast Mytilus. *Marine Ecology Progress Series*, 399, pp.211–223.

Sievers, F. & Higgins, D.G., 2014. Clustal Omega. *Current Protocols in Bioinformatics*, 2014, pp.3.13.1-3.13.16.

Silverman-Gavrila, L.B., Lu, T.Z., Prashad, R.C., Nejatbakhsh, N., Charlton, M.P. & Feng, Z.P., 2009. Neural phosphoproteomics of a chronic hypoxia model-Lymnaea stagnalis. *Neuroscience*.

Simakov, O., Marletaz, F., Cho, S.-J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D.-H., Larsson, T., Lv, J., Arendt, D., Savage, R., Osoegawa, K., de Jong, P., Grimwood, J., Chapman, J.A., Shapiro, H., Aerts, A., Otillar, R.P., Terry, A.Y., Boore, J.L., Grigoriev, I. V., Lindberg, D.R., Seaver, E.C., Weisblat, D.A., Putnam, N.H. & Rokhsar, D.S., 2012. Insights into bilaterian evolution from three spiralian genomes. *Nature*, 493(7433), pp.526–531. Available at: http://www.nature.com/doifinder/10.1038/nature11696.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E.M., 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), pp.3210–3212.

Simpson, J.T., 2014. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics*, 30(9), pp.1228–1235.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M. & Birol, I., 2009. ABySS: A parallel assembler for short read sequence data. *Genome Research*, 19(6), pp.1117–1123.

Simpson, L.M., Wall, I.D., Blaney, F.E. & Reynolds, C.A., 2011. Modeling GPCR active state conformations: The β 2-adrenergic receptor. *Proteins: Structure, Function and Bioinformatics*, 79(5), pp.1441–1457.

Sivakumar, S. & Subbhuraam, C. V., 2005. Toxicity of chromium(III) and chromium(VI) to the earthworm Eisenia fetida. *Ecotoxicology and Environmental Safety*.

Sizmur, T., Watts, M.J., Brown, G.D., Palumbo-Roe, B. & Hodson, M.E., 2011. Impact of gut passage and mucus secretion by the earthworm Lumbricus terrestris on mobility and speciation of arsenic in contaminated soil. *Journal of Hazardous Materials*, 197, pp.169–175.

Skoufos, E., Marenco, L., Nadkarni, P.M., Miller, P.L. & Shepherd, G.M., 2000. Olfactory receptor database: a sensory chemoreceptor resource. *Nucleic acids research*.

Slomko, H., Heo, H.J. & Einstein, F.H., 2012. Minireview: Epigenetics of obesity and diabetes in

humans. *Endocrinology*.

Small, K.S., Brudno, M., Hill, M.M. & Sidow, A., 2007a. A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome. *Genome Biology*, 8(3), p.R41. Available at: http://dx.doi.org/10.1186/gb-2007-8-3-r41.

Small, K.S., Brudno, M., Hill, M.M. & Sidow, A., 2007b. Extreme genomic variation in a natural population. *Proceedings of the National Academy of Sciences of the United States of America*, 104(13), pp.5698–703. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17372217%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1838466.

Smit, A., Hubley, R. & Green, P., 1996. RepeatMasker Open-3.0. *RepeatMasker Open-3.0*, p.www.repeatmasker.org.

Smit, A.F.A. & Hubley, R., RepeatModeler Open-1.0. 2008-2010.

Sodeland, M., Jorde, P.E., Lien, S., Jentoft, S., Berg, P.R., Grove, H., Kent, M.P., Arnyasi, M., Olsen, E.M. & Knutsen, H., 2016. "Islands of Divergence" in the Atlantic Cod Genome Represent Polymorphic Chromosomal Rearrangements. *Genome biology and evolution*, 8(4), pp.1012–1022.

Sodergren, E., Weinstock, G.M., Davidson, E.H., Cameron, R.A., Gibbs, R. a, Angerer, R.C., Angerer, L.M., Arnone, M.I., Burgess, D.R., Burke, R.D., Coffman, J. a, Dean, M., Wright, R., et al., 2006. The genome of the sea urchin Strongylocentrotus purpuratus. *Science (New York, N.Y.)*, 314(5801), pp.941–52. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3159423&tool=pmcentrez&rendertype=abstract.

Song, Y., Endepols, S., Klemann, N., Richter, D., Matuschka, F.R., Shih, C.H., Nachman, M.W. & Kohn, M.H., 2011. Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Current Biology*, 21(15), pp.1296–1301.

Spring, J., 1997. Vertebrate evolution by interspecific hybridisation - Are we polyploid? *FEBS Letters*.

Springer, N.M., Ying, K., Fu, Y., Ji, T., Yeh, C.T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., Iniguez, A.L., Barbazuk, W.B., Jeddeloh, J.A., Nettleton, D. & Schnable, P.S., 2009. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genetics*.

Spurgeon, D.J. & Hopkin, S.P., 1996. Effects of variations of the organic matter content and pH of soils on the availability and toxicity of zinc to the earthworm Eisenia fetida. *Pedobiologia*.

Spurgeon, D.J., Liebeke, M., Anderson, C., Kille, P., Lawlor, A., Bundy, J.G. & Lahive, E., 2016. Ecological drivers influence the distributions of two cryptic lineages in an earthworm morphospecies. *Applied Soil Ecology*, 108, pp.8–15.

Spurgeon, D.J., Ricketts, H., Svendsen, C., Morgan, A.J. & Kille, P., 2005. Hierarchical Responses of Soil Invertebrates (Earthworms) to Toxic Metal Stress. *Environmental Science & Technology*.

Strelkov, P., Katolikova, M. & Väinolä, R., 2017. Temporal change of the Baltic Sea–North Sea blue mussel hybrid zone over two decades. *Marine Biology*.

Sudmant, P.H., Kitzman, J.O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J. & Eichler, E.E., 2010. Diversity of human copy number variation and multicopy genes. *Science*.

Suzuki, M.M., Kerr, A.R.W., De Sousa, D. & Bird, A., 2007. CpG methylation is targeted to

transcription units in an invertebrate genome. *Genome Research*.

Taylor, J.S., Van de Peer, Y. & Meyer, A., 2001. Revisiting recent challenges to the ancient fish-specific genome duplication hypothesis (multiple letters). *Current Biology*.

Tenaillon, O., Barrick, J., Ribeck, N., Deatherage, D., Blanchard, J., Dasgupta, A., Wu, G., Wielgoss, S., Cruveiller, S., Medigue, C., Schneider, D. & Lenski, R., 2016. Tempo and mode of genome evolution in a 50,000-generation experiment. *bioRxiv*.

Tenreiro MacHado, J.A., 2012. Shannon entropy analysis of the genome code. *Mathematical Problems in Engineering*, 2012.

Terhivuo, J. & Saura, A., 2006. Dispersal and clonal diversity of north-european parthenogenetic earthworms. In *Biological Invasions Belowground: Earthworms as Invasive Species*.

Tiunov, A. V., Hale, C.M., Holdsworth, A.R. & Vsevolodova-Perel, T.S., 2006. Invasion patterns of Lumbricidae into the previously earthworm-free areas of northeastern Europe and the western Great Lakes region of North America. In *Biological Invasions Belowground: Earthworms as Invasive Species*.

Tourasse, N.J. & Li, W.H., 2000. Selective constraints, amino acid composition, and the rate of protein evolution. *Molecular Biology and Evolution*.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. & Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*.

Turner, T.L., Hahn, M.W. & Nuzhdin, S. V., 2005. Genomic islands of speciation in Anopheles gambiae. *PLoS Biology*, 3(9), pp.1572–1578.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D.A., Cestaro, A., Pruss, D., Pindo, M., FitzGerald, L.M., Vezzulli, S., Reid, J., Malacarne, G., Iliev, D., Viola, R., et al., 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE*, 2(12).

Via, S., 2012. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society B: Biological Sciences*.

Vinson, J.P., Jaffe, D.B., O'Neill, K., Karlsson, E.K., Stange-Thomann, N., Anderson, S., Mesirov, J.P., Satoh, N., Satou, Y., Nusbaum, C., Birren, B., Galagan, J.E. & Lander, E.S., 2005. Assembly of polymorphic genomes: algorithms and application to Ciona savignyi. *Genome Res*, 15(8), pp.1127–1135.

Viveiros, F., Ferreira, T., Cabral Vieira, J., Silva, C. & Gaspar, J.L., 2008. Environmental influences on soil CO2 degassing at Furnas and Fogo volcanoes (Sao Miguel Island, Azores archipelago). *Journal of Volcanology and Geothermal Research*.

Vsevolodova-Perel, T.S. & Bulatova, N.S., 2008. Polyploid races of earthworms (Lumbricidae, Oligochaeta) in the East European plain and Siberia. *Biology Bulletin*.

Way, S., Williams, P.H., Gaston, K.J., Humphries, C.J., GASTONt, K.J. & HUMPHRIES Biogeography, C.J., 1994. Do Conservationists and Molecular Biologists Value Differences between Organisms in the Do conservationists and molecular biologists value differences between organisms in the same way? *Source: Biodiversity Letters Biodiversity Letters*, 2(2), pp.67–7867. Available at: http://www.jstor.org/stable/2999760%5Cnhttp://about.jstor.org/terms.

Weber, E. & D'Antonio, C.M., 2000. Phenotypic plasticity in hybridizing *Carpobrotus* spp. (Aizoaceae) from coastal California and its role in plant invasion. *Canadian Journal of Botany*.

Welch, D.B.M. & Meselson, M.S., 2001. Rates of nucleotide substitution in sexual and anciently asexual rotifers. *Proceedings of the National Academy of Sciences*.

Welch, D.M., 2000. Evidence for the Evolution of Bdelloid Rotifers Without Sexual Reproduction or Genetic Exchange. *Science*, 288(5469), pp.1211–1215. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.288.5469.1211.

Wen, B., Hu, X.Y., Liu, Y., Wang, W.S., Feng, M.H. & Shan, X.Q., 2004. The role of earthworms (Eisenia fetida) in influencing bioavailability of heavy metals in soils. *Biology and Fertility of Soils*.

Wendel, J.F., Schnabel, A. & Seelanan, T., 1995. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (Gossypium). *Proceedings of the National Academy of Sciences*.

West, G.B., Brown, J.H. & Enquist, B.J., 1999. The fourth dimension of life: Fractal geometry and allometric scaling of organisms. *Science*, 284(5420), pp.1677–1679.

West, M.A.L., Kim, K., Kliebenstein, D.J., Van Leeuwen, H., Michelmore, R.W., Doerge, R.W. & St. Clair, D.A., 2007. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics*.

Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., Zhernakova, A., Zhernakova, D. V., Franke, L., et al., 2013. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*.

White, B.J., Cheng, C., Simard, F., Costantini, C. & Besansky, N.J., 2010. Genetic association of physically unlinked islands of genomic divergence in incipient species of Anopheles gambiae. *Molecular Ecology*, 19(5), pp.925–939.

Widmer, A. & Baltisberger, M., 1999. Molecular evidence for allopolyploid speciation and a single origin of the narrow endemic Draba ladina (Brassicaceae). *American Journal of Botany*.

Williams, A., Carlson, S.J., Brunton, C.H.C., Holmer, L.E. & Popov, L.E., 1996. A supra-ordinal classification of the Brachiopoda. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 351(1344), pp.1171–1193.

Williams, A., Cusack, M. & Mackay, S., 1994. Collagenous Chitinophosphatic Shell of the Brachiopod Lingula. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 346(1316), pp.223–266. Available at: http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.1994.0143.

Wolf, J.B.W. & Ellegren, H., 2017. Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*.

Wood, E., 1983. Molecular Cloning. A Laboratory Manual. *Biochemical Education*.

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., Wang, J., Orjeda, G., Guzman, F., Visser, R.G.F., et al., 2011. Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355), pp.189–195.

Xue, Y., Wang, Q., Long, Q., Ng, B.L., Swerdlow, H., Burton, J., Skuce, C., Taylor, R., Abdellah, Z., Zhao, Y., MacArthur, D.G., Quail, M.A., Carter, N.P., Yang, H. & Tyler-Smith, C., 2009. Human Y Chromosome Base-Substitution Mutation Rate Measured by Direct Sequencing in a Deep-Rooting Pedigree. *Current Biology*.

Zadesenets, K.S. & Rubtsov, N.B., 2018. Genome Duplication in Animal Evolution. *Russian Journal of Genetics*, 54(10), pp.1125–1136. Available at: https://doi.org/10.1134/S1022795418090168.

de Zea Bermudez, P. & Kotz, S., 2010. Parameter estimation of the generalized Pareto distribution-Part I. *Journal of Statistical Planning and Inference*, 140(6), pp.1353–1373.

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H., Xiong, Z., Que, H., Wang, J., et al., 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, 490(7418), pp.49–54. Available at: http://dx.doi.org/10.1038/nature11413.

Zhang, W., Hendrix, P.F., Snyder, B.A., Molina, M., Li, J., Rao, X., Siemann, E. & Fu, S., 2010. Dietary flexibility aids Asian earthworm invasion in North American forests. *Ecology*.

Zhang, Y., Wang, Z., Yan, X., Yu, R., Kong, J., Liu, J., Li, X., Li, Y. & Guo, X., 2012. Laboratory Hybridization between Two Oysters: Crassostrea gigas and Crassostrea hongkongensis. *Journal of Shellfish Research*, 31(3), pp.619–625. Available at: http://www.bioone.org/doi/abs/10.2983/035.031.0304.

Zheng, G.X.Y., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., Terry, J.M., Mudivarti, P.A., Wyatt, P.W., Ji, H.P., et al., 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, 34(3), pp.303–311.

Zielezinski, A., Vinga, S., Almeida, J. & Karlowski, W.M., 2017. Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biology*.