The Development and Application of SNP Discovery Technologies from 2005 to 2012

Submitted for the degree of Doctor of Philosophy at Cardiff University

James P M Colley

2012

DECLARATION

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed ……(candidate) Date ……

STATEMENT 1

This thesis is being submitted in partial fulfilment 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.

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 Graduate Development Committee.**

Signed ……(candidate) Date ……

Summary

Single Nucleotide Polymorphisms (SNPs) are used as markers in association studies and may contribute directly to inherited disease. Here, we investigated high throughput *in silico* and *in vitro* methods for identifying SNPs and applied these to large scale genomic projects. We evaluated the utility of the Transgenomic Navigator™ software to facilitate automated detection of aberrant denaturing high performance liquid chromatography (dHPLC) elution profiles. 3,747 dHPLC profiles were analysed with this software and 98.3% of products with profiles distinct from wild type harboured novel variants (Chapter 3). We applied this software to investigate whether rare inherited variants in genes that play a role in oxidative DNA damage repair (OxDR) predispose to colorectal adenomas (CRA) and found that unlike MUTYH, inherited variants in other OxDR genes are unlikely to be a frequent cause of CRA (Chapter 4). We evaluated the sequence analysis packages Sequencher, InSNP, Mutation Surveyor and Staden to identify variants in patients with CRAs (using >4,000 chromatograms). Staden and Mutation Surveyor correctly identified 76/77 (98.7%) SNPs and 96.7% and 99.3% of the genotypes respectively (Chapter 5). We compared an optimised version of Staden against Sequencher for variant detection over a 2.5kb region of the adenomatous polyposis coli (*APC*) gene in 969 healthy controls. We found 100% concordance between these packages and found that rare nonsynonymous variants in *APC* were significantly over-represented in patients with CRAs as compared to healthy controls (32/480 vs. 37/969, P=0.0166) (Chapter 5). We evaluated data held in dbSNP (build 129) for common variants in the Caucasian population by examining ten DNA repair genes and subsequently developed software for automatically extracting selected information from dbSNP (Chapter 6). This program was used to rapidly identify 221 common nonsynonymous SNPs in every DNA repair gene in the human genome; these were subsequently typed in 480 publically-available human lymphoblastoid cell lines generating a resource for functional analyses (Chapter 6). We also assessed the role of these SNPs in susceptibility to CRC and response to therapy by exploiting data and samples for the randomised controlled trial COIN (Chapter 7). Finally, we used Next Generation Sequencing (NGS) to discover SNPs in the exomes of 10 patients that exhibited peripheral neuropathy in response to chemotherapy and discovered that *ERCC4* nonsense mutations may contribute to this toxicity (Chapter 8). NGS is likely to become the key SNP discovery technique over the next decade due to its potential to comprehensively search an entire genome at a comparatively low cost.

Acknowledgments

I should like to express my thanks to everyone at the Institute of Medical Genetics who has helped and supported me during this study, in particular my supervisor Jeremy Cheadle for helping me throughout. Julie Maynard, Vikki Humphreys and Rebecca Harris for technical support, Chris Smith and Hannah West for helping me with writing up, Duncan Azzopardi and Anthony Dallosso for accepting my help on their projects and Shelley, Peter, Sherrie, Hannah and Linda for everything in between.

I would also like to thanks the Michael Wagner and his group at the University of North Carolina, for assisting me during my stay in Chapel Hill and finally, Julian Sampson for allowing me to undertake my studies whilst working at the Institute of Medical Genetics.

For Megan Olwen, Fy Nghalon Melys.

Abbreviations

5FU	- Fluorouracil
ACN	- Acetonitrile
ADR	- Adverse Drug Reaction
AFAP	- Attenuated Familial Adenomatous Polyposis
AGBT	- Advances in Genome Biology and Technology
AE	- Acridinium Ester
ARMS	- Amplification Refractory Mutation System
ASA	- Allele Specific Amplification
ASP	- Allele Specific PCR
BER	- Base Excision Repair
BLAST	- Basic Local Alignment Search Tool
BSC	- Best Supportive Care
bp	- Base Pair
C	- Cytosine
CCD	- Charge-Coupled Device
CCMM	- Chemical Cleavage of Mismatch
CEA	- Carcinoembryonic Antigen
CHRPE	- Congenital Hypertrophy of the Retinal Pigment Epithelium
cPAL	- Combinatorial Probe-Anchor Ligation
CR	- Complete Response
CRA	- Colorectal Adenoma
CRC	- Colorectal Cancer
DASH	- Dynamic Allele Specific Hybridisation
DAVID	- Database for Annotation, Visualization and Integrated Discovery
ddNTP	- Di-deoxynucleotide
DGGE	- Denaturing Gradient Gel electrophoresis
dHPLC	- Denaturing High Performance Liquid Chromatography
DMD	- Duchenne Muscular Dystrophy
DMSO	- Dimethyl Sulfoxide
dNTPs	- Deoxynucleotides
DSB	- Double Strand Break
dsDNA	- Double Stranded DNA
DVB	- Divinylbenze
ECACC	- European Collection of Cell Cultures
EDTA	- Ethylenediaminetetraacetic Acid,
EM	- Electron Microscopy
EMBL	- European Molecular Biology Laboratory
emPCR	- Emulsion PCR
FAP	- Familial Adenomatous Polyposis
FEN	- Flap Endonuclease
FP	- Fluoropyrimidine

FRET	- Fluorescence Resonance Energy Transfer
G	- Guanine
GATK	- Genome Analysis Toolkit
GI	- Gastro Intestinal
GWA	- Genome-Wide Association
GWAS	- Genome-Wide Association Studies
HGMD	- Human Genome Mutation Database
HIV-1	- Human Immunodeficiency virus 1
HRM	- High Resolution Melt
HWE	- Hardy Weinberg Equilibrium
KASPar	- KBiosciences Competitive Allele-Specific PCR Array
LD	- Linkage Disequilibrium
LOH	- Loss of Heterozygosity
MAF	- Minor Allele Frequency
MALDI	- Matrix-assisted Laser Desorption/Ionization
MAP	- MUTYH Associated Polyposis
Mass-spec	- Mass Spectrometry
MCR	- Mutation Cluster Region
mCRC	- Metastatic CRC
MI	- Microsatellite Instability
MIM	- Mendelian Inheritance in Man
MMR	- Mismatch Repair
NC	- Non-Conserved
NCBI	- National Centre for Biotechnology Information
NER	- Nucleotide Excision Repair
NGS	- Next Generation Sequencing
NHGRI	- National Human Genome Research Institute
NIH	- National Institutes of Health
OLA	- Oligo Ligation Assay
OMIM	- Online Mendelian Inheritance in Man
ONT	- Oxford Nanopore Technologies
OR	- Odds Ratio
ORF	- Open Reading Frame
OS	- Overall Survival
OxDR	- Oxidative DNA damage repair
OxFU	- Oxaliplatin-5-fluorouracil
PAC	- Polyacrylamide
PASA	- PCR Allele-Specific Amplification
PCR	- Polymerase Chain Reaction
PD	- Progressive Disease
PFS	- Progression-Free Survival
PN	- Peripheral Neuropathy
PNAO	- Peripheral Neuropathy Associated with Oxaliplatin
RCA	- Rolling Circle Amplification

RCT	- Randomised Controlled Trial
RFLP	- Restriction Fragment Length Polymorphism
RNA	- Ribonucleic Acid
RR	- Response Rates
rs	- Reference SNP
SC	- Semi-Conserved
SD	- Stable Disease
SMRT	- Single Molecule Real Time
SNP	- Single Nucleotide Polymorphism
SOLiD	- Sequencing by Oligonucleotide Ligation and Detection
SRA	- Short Read Archive
ss	- Submitted Sequence
SSCP	- Single Stranded Conformational Polymorphism
T	- Thymine
TEAA	- Triethylammonium Acetate
TGGE	- Temperature Gradient Gel electrophoresis
TOF	- Time-of–Flight
U	- Units
UV	- Ultraviolet
WAG	- Welsh Assembly Government
WHO-PS	- World Health Organisation – Performance Status
XFE	- XPF-ERCC1
XP	- Xeroderma Pigmentosum
ZWM	- Zero Mode Waveguides

Table of Contents

List of Tables

List of Figures

1 Chapter One

General Introduction

1.1 Single Nucleotide Polymorphisms

When a nucleotide locus within a genome deviates from the expected sequence the feature is described as a Single Nucleotide Polymorphism (SNP). Technically the term polymorphism only describes such variations that occur within a population at a frequency of at least 1%; however, the interpretation seems to have broadened recently to include less frequent variation, perhaps as a result of these rare loci becoming of greater interest to the research community.

SNPs account for the majority of sequence variation observed in the human genome, their distribution is fairly random but they occur frequently (on average every 1,200 bases). SNPs exist in coding and non-coding DNA and there are numerous techniques for determining their distribution in sample cohorts (Cotton et al., 1998, Taylor and Day, 2005). These factors make SNPs ideal as candidates to study the relationship between an individual's genotype and phenotype.

The impact of a SNP can vary from being completely neutral to completely pathogenic and they can have a dramatic effect on drug response, if genetics is to achieve the goal of delivering 'Personalised Medicine' it is crucial to be able to identify and understand the SNPs of individual genomes, recently termed the 'Variome'.

The recent data explosion in genetics has not come without problems, the NCBI short read archive (SRA) is set to be phased out though 2012[1] as sequencing

[1] Accurate at the time of writing though discussions to retain the database are on going

genomes has become cheaper and faster to such an extent that the SRA cannot cope with the submissions. One suggested solution to this problem is not to store individual genomes but rather to store 'diff' files. diff is a computer programme that outputs the difference between two files, in the case of genetics this can be the difference between an individual's genome and a reference genome, naturally the files are smaller and more manageable but this also demonstrates that it is an individual's differences that have relevance.

1.2 SNP Discovery

Genetic variations have always been sought in order to identify the underlying cause of hereditary disease but they are also significant in identifying disease susceptibility and latterly as bio-markers for drug response/suitability. Early research focussed on simple Mendelian disorders and small numbers of genes but current studies deal with entire genomes and millions of SNPs. The expansion of research projects has driven development of many techniques which can identify samples that contain variation without incurring the costs of large scale sequencing projects, sensitivity and specificity of each technique is widely debated (Gross et al., 1999, Lin et al., 2008, Schwaab et al., 1997). The chosen method for a project is not necessarily the most sensitive as compromises must be made to achieve a reasonable rate of discovery within a budget and consideration must be given to the available skills and technology of each research group.

Discriminating between samples with different genotypes at a SNP locus is achieved by one of two methods; demonstrating a conformational difference between molecules or identifying heteroduplexes.

1.2.1 SNP Discovery Techniques - Conformational Discrimination

1.2.1.1 Denaturing and Temperature Gradient Gel Electrophoresis

Denaturing Gradient Gel electrophoresis (DGGE) was one of the first described methods for detecting SNPS within a PCR fragment (Fischer and Lerman, 1979) and Temperature Gradient Gel electrophoresis (TGGE) was a development of the same principal (Riesner et al., 1989). PCR fragments between 200 and 700bp in length are designed with terminating GC clamps to reduce the risk of complete denaturation. Fragments ideally have a simple melt profile which could be determined by programmes such as MELT94 (http://web.mit.edu/osp/www/melt.html) and latterly uMelt (http://www.dna.utah.edu/umelt/um.php) which predict the likely melt profiles of fragments based on sequence composition.

A polyacrylamide (PAC) gel is created which contains a denaturing agent that increases in concentration through the gel such that the highest concentration is nearest to the positive electrode. The negative charge of the DNA is utilised to draw the double stranded DNA (dsDNA) fragment through the gel from the negative towards the positive electrode there by exposing it to the increasing concentration of denaturant as it migrates. Sequence differences cause fragments to denature at different concentrations and denatured fragments travel through PAC gel at a significantly reduced rate therefore when running samples in parallel a difference in an individual band's position is indicative of a sequence difference.

DGGE is a non-destructive technique, individual fragments can be extracted from the gel for further analysis and only basic laboratory equipment is required. The gradients themselves cannot be precisely replicated which means different sample sets cannot be tested under identical conditions also if the denaturant becomes concentrated enough to completely denature a fragment all resolution is lost. TGGE over comes these problem as it possible to control temperature (the denaturant) with greater accuracy therefore generating the same conditions between runs.

In theory these techniques are capable of detecting all SNPs within a sequence but the technique has largely been replaced by higher throughput alternatives, more recent publications that do report the technique are largely microbiology based but there are exceptions (da Costa Aguiar et al., 2011) who are investigating human variation.

1.2.1.2 Single Stranded Conformational Polymorphism

Single Stranded Conformational Polymorphism (SSCP) analysis (Orita et al., 1989) is a technique that works on the principal that two strands of DNA which differ by a single nucleotide will form distinct 3D structures that migrate at different rates through a non-denaturing matrix. Sensitivity has been reported over a wide range (3 to 90%, (Sheffield et al., 1993)) with 150bp fragments giving the optimal results. With traditional 'radioisotopic label/slab gel' approach it was found that environmental changes such as apparatus temperature were key to obtaining a good rate of detection and so multiple temperatures would be run to achieve maximum sensitivity (it is thought that these 'environmental' conditions altered the molecular interactions of the single strands (Cotton et al., 1998). As new technologies became available the isotopic method was replaced by fluorescent detection (Makino et al., 1992) which was carried out on DNA sequencing instruments that had effective temperature control built in (Iwahana et al., 1996). This technology made it easier to perform multiple analysis at the range of temperatures required and improved the sensitivity of the process. The CCD camera system within the DNA Sequencers meant that multiple fragments could be run in the same lane by using different fluorophores to label the fragments, increasing the potential throughput.

This technique does not exploit the existence of heteroduplexes directly as they are a feature of double stranded DNA; however, as samples are loaded onto non denaturing gels dsDNA reforms and homoduplexes and heteroduplexes may resolve at the bottom of the gel. Despite the greater molecular weight compared to

single stranded DNA (ssDNA) dsDNA migrates faster due to the simplified 3D structure. Though technically heteroduplexes are artefacts of the technique they can be informative.

1.2.2 SNP Discovery Techniques - Heteroduplex Identification

Heteroduplex fragments occur naturally during PCR when a sample is heterozygous at a given locus (Figure 1.1). When an amplicon from a paternal chromosome anneals to a complimentary amplicon from a maternal chromosome at the point at which that sample exhibits heterozygosity there will be a mismatch between bases and hydrogen bonds will not form. This feature is known as a heteroduplex and it reduces the stability of the double strand compared to those with complete matches.

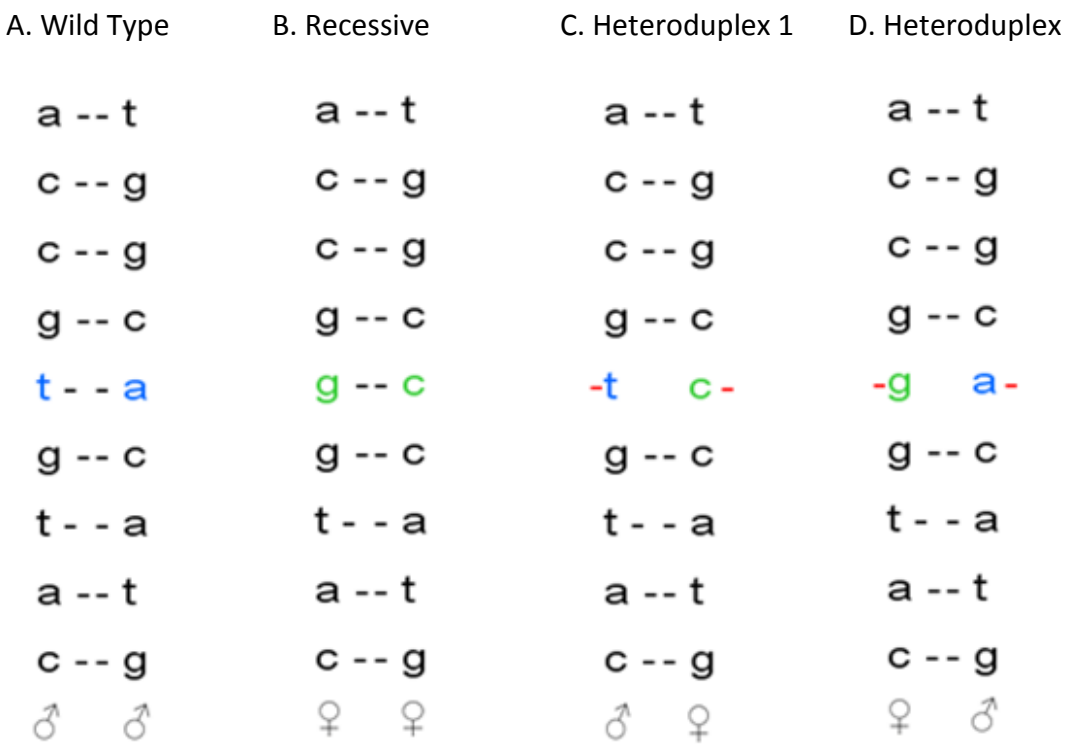


Figure 1.1: Homo and Heteroduplex molecules.

A PCR product generated from a heterozygous sample contains two alleles at a SNP locus. During PCR the amplicon is heated and cooled cyclically, as the reaction cools double stranded products are re-formed. Complimentary strands may form homoduplexes (A and B) or heteroduplex containing mismatched bases (C and D) where the strand pairs originate from different parents. All four species occur naturally during PCR.

1.2.2.1 Chemical Cleavage

By cleaving a heteroduplex fragment at the point where the mis-match occurs the heterozygosity of a sample can be evidenced by the presence of 2 bands on a gel, this is known as Chemical Cleavage of Mismatch (CCMM). Cotton (1988) originally described this process as a modification of the Maxam and Gilbert sequencing technique (Maxam and Gilbert, 1977). Hydroxylamine and Osmium tetroxide are used to modify C or T mismatches respectively, the modified mismatches are cleaved by piperidine and the resulting fragments are separated by gel electrophoresis to determine their sizes.

With the original isotopic technique the sizes of the fragments would provide an indication of where the SNP occurs within the fragment but before end labelling was possible this would not be definitive. The original PCR is cleaved at the SNP locus so when two bands are observed on a gel the SNP locus it is possible to that the locus is 'n' bases from one end of the original PCR fragment but from which end would be unknown.

The technique is highly sensitive (Whittock, 2005) but the methodology is complicated by the toxicity of the chemicals used (requiring multiple ethanol precipitations to be carried out in a fume hood) and by the fact that every fragment requires separate tests to identify C and T mismatches. Developments of the original technique successfully removed some of the more toxic compounds (Potassium permanganate and tetraethylammonium chloride replaced Osmium tetroxide (Roberts et al., 1997)) but with the development of the enzyme cleavage technique the toxic compounds were no longer required.

1.2.2.2 Enzymatic Cleavage

Enzymatic cleavage (Youil et al., 1995) works on the same principal as CCMM but is more readily adapted to fluorescent analysis and therefore offers improved throughput and a simplified, less hazardous procedure when compared to the

chemical method. The combination of fluorescent end labelling and fluorescent ladders improves the accuracy of determining SNP location as a locus can be determined as being 'n' bases from the fluorescently labelled end of the fragments. The technique requires equipment that is widely available in genetics laboratories. To improve the throughput seen with capillary electrophoresis Schmalzing (2000) applied microchip technology to the analysis of the digested products, where micro-channel structures etched into glass replaced polymer filled capillaries and allow for a faster electrophoresis.

1.2.2.3 Denaturing High Performance Liquid Chromatography (dHPLC)

dHPLC uses the principals of chromatography to separate different molecules based on the length of time they are retained on the 'stationary phase' (Oefner and Underhill, 1998). The Transgenomic WAVE ™ DNA Fragment Analysis System uses a DNASep® column which contains a polystyrene-divinylbenzene (DVB) copolymer creating a stationary phase for chromatography. The mobile phase is comprised of the ion-pairing agent triethylammonium acetate (TEAA) which mediates binding of DNA to the stationary phase and the organic solvent Acetonitrile (ACN). Under specific conditions heteroduplex fragments elute at a lower concentration of solvent than homoduplexes.

Sensitivity of the system is determined by selection of a column temperature which creates partially denaturing conditions for the PCR that is bound by the TEAA to the solid phase. Several programmes, are available for predicting the melting profile of a particular PCR fragment (MELT95, uMELT, DNAmelt) which may also be empirically determined (Rudolph et al., 2002) by processing an aliquot of PCR product at a series of sequential temperature increments and plotting the resulting retention time against temperature. Both the computer based and empirical approach produce a 'melt profile' for the fragments from which a temperature (or

temperatures) is selected at which the dsDNA is partially denatured so that the effect of the mismatch on the retention time can be maximised.

Within the temperature controlled column the hydrophobic groups of the TEAA (from the TEAA-DNA complex) interact with the hydrophobic C-18 chains on the DVB beads immobilising the DNA, the ACN concentration increases through a range specific to each fragment. As the ANC concentration increases the first fragments to elute will be those where the binding between the DNA and TEAA is weakest. If there are no heteroduplexes elution will occur at the same point for all amplicons but if heteroduplexes are present the mismatches will weaken the bond between the negative phosphate group of the DNA backbone and the positive ammonium group of the TEAA. More weakly bonding heteroduplexes will elute before the homoduplexes which will require a higher concentration of solvent to overcome the bonds with the TEAA molecules (Figure 1.2).

The eluting PCR products are detected by measuring the UV absorbance of the eluting solution and because the PCR is completely un-modified it is possible to collect this solution, purify it and use the DNA for downstream analyses such as cloning or sequencing.

When the dHPLC temperature is optimal, all four double stranded products may be separated (two homoduplexes and two heteroduplexes (Figure 1.3 B) but this level of sensitivity is not necessary for SNP detection. Often the only difference between a homozygous and heterozygous product is to observe two peaks or a single ‘split’ peak (Figure 1.3 A) heterozygotes can even be identified from a small change in the leading edge of a single peak.

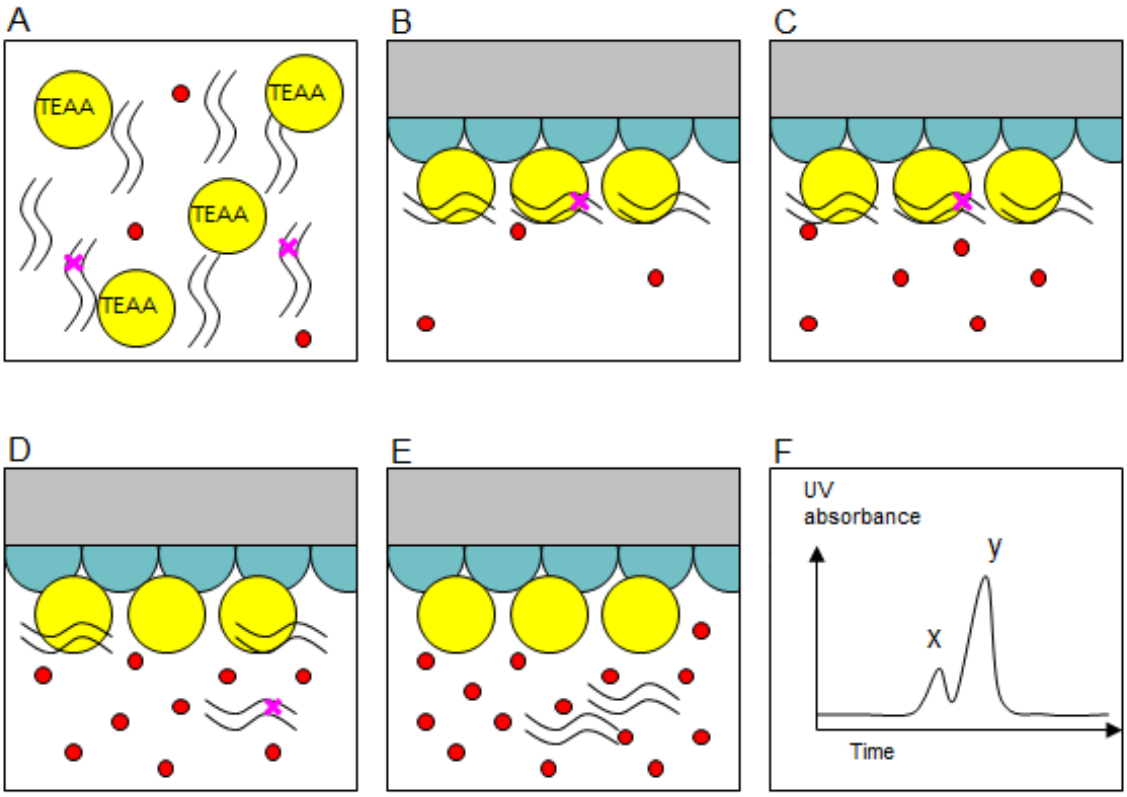


Figure 1.2: dHPLC.

(A) Double stranded PCR is mixed in solution with the solvent Acetonitrile (ACN, red) and the ion pairing agent Triethylammonium Acetate (TEAA, yellow) which binds the DNA. (B) TEAA also binds to the solid phase divinybenzine beads (DVB, blue) immobilising the DNA. (C) The ACN concentration increases. (D) ACN reaches a particular concentration where heteroduplexes elute as their binding to the TEAA is weakened by the mismatch. (E) The ACN concentration increases to a particular point where the homoduplexes are eluted. (F) The eluent passes a UV detector and absorbance is measured and related to retention time.

1.2.2.4 Spiking

Where SNP detection is performed on the basis of identifying heteroduplexes samples homozygous for the minor allele are not necessarily identified as they may elute with wildtype homozygotes under sub-optimal conditions. By introducing a small quantity of PCR that has come from a known reference sequence to every sample we enable heteroduplex molecules to form in samples that are homozygous for a minor allele without affecting samples with other genotypes (Kaler et al., 2000, Pirulli et al., 2000, Schaeffeler et al., 2001). Our own unpublished data has shown that a ratio of 4:1, sample to reference is optimal to identify homozygotes without affecting homozygous calls when using dHPLC to screen sample sets for SNPs (Figure 1.4).

1.2.2.5 High Resolution Melt Profile Analysis

Melt profile analysis provides a 'real time' sample analysis with parallel processing of 96 or 384 well plates analysis time is a matter of minutes making this new technology one of the highest throughput formats (Gundry et al., 2003). PCR fragments are generated in the presence of the intercalating Dye 'LC Green' with which they become saturated. Analysis occurs by measuring the fluorescence given off by each well as the sample plate is gradually heated from 75^0C to 95^0C over a period of 5 minutes during which time the PCR fragments denature releasing the dye. Fragments with heteroduplexes will release their dye more quickly on reaching a certain temperature due to the instability caused by the mismatch and this change in the rate of 'fluorescent decay' can be used to separate samples that carry heterozygous SNPs from those which do not. Unlike dHPLC all samples are run with the same conditions and the optimum temperature does not need to be calculated as it will be covered by the dynamic temperature range of the system. Though high resolution Melt (HRM) has a lower sensitivity compared to dHPLC (Lin et al., 2008) its throughput is considerably greater due to its parallel processing, in certain cases

a researcher may accept a proportion of false negative results in return for the ability to analyse greater sample sets and more amplicons.

1.2.2.6 Sequencing and SNP Calling Software

DNA sequencing (Section 1.4) is the most comprehensive method for identifying and characterising new mutations. Until recently its cost has necessitated the use of the above mentioned techniques to screen samples for those which contain variation; however, where cost is not an issue modern capillary sequencing combined with SNP calling software provides a fast and comprehensive method for screening samples.

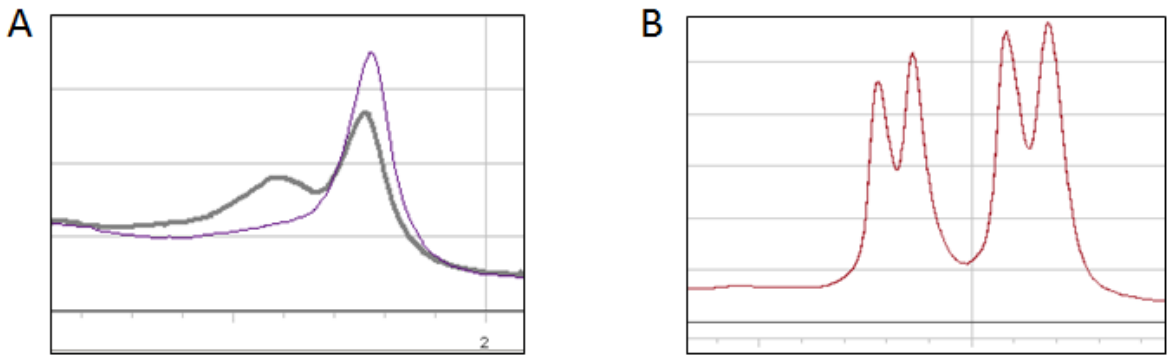


Figure 1.3: Elution of homo- and hetero- duplexes.

(A) The Heterozygous sample (grey) differs in shape from the wildtype sample (purple) the smaller 'peak' on the left is indicative of heterozygosity. (B) All four species of PCR fragment may be resolved with precise conditions, this image shows the 2 heteroduplex and 2 homoduplex peaks produced during PCR of a heterozygous sample.

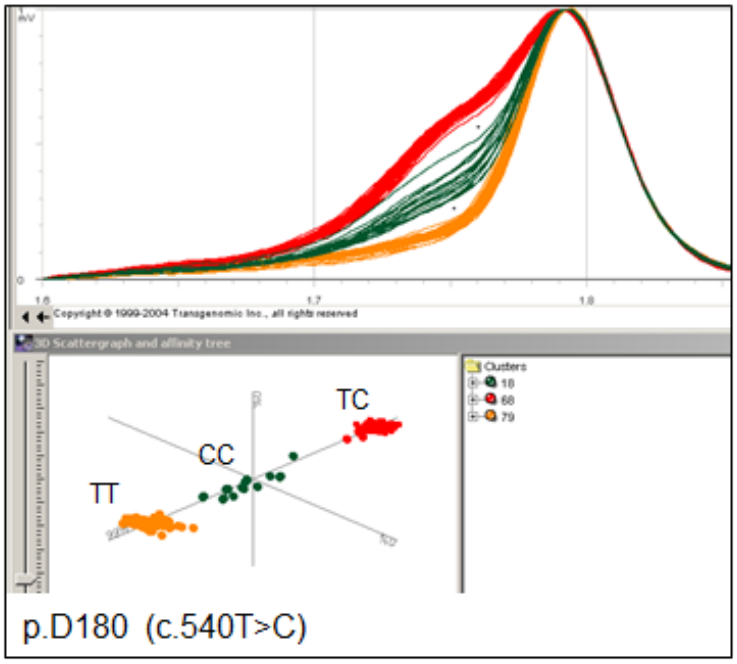


Figure 1.4: Spiking.

When wild type PCR is used to 'spike' PCR products then samples heterozygous for a recessive allele will be able to produce heteroduplexes with the introduced PCR product making the recessive allele detectable by dHPLC. The scatter plot shows that the homozygous recessive alleles (green) appear with a similar but less pronounced shape as the genuine heterozygous samples (red).

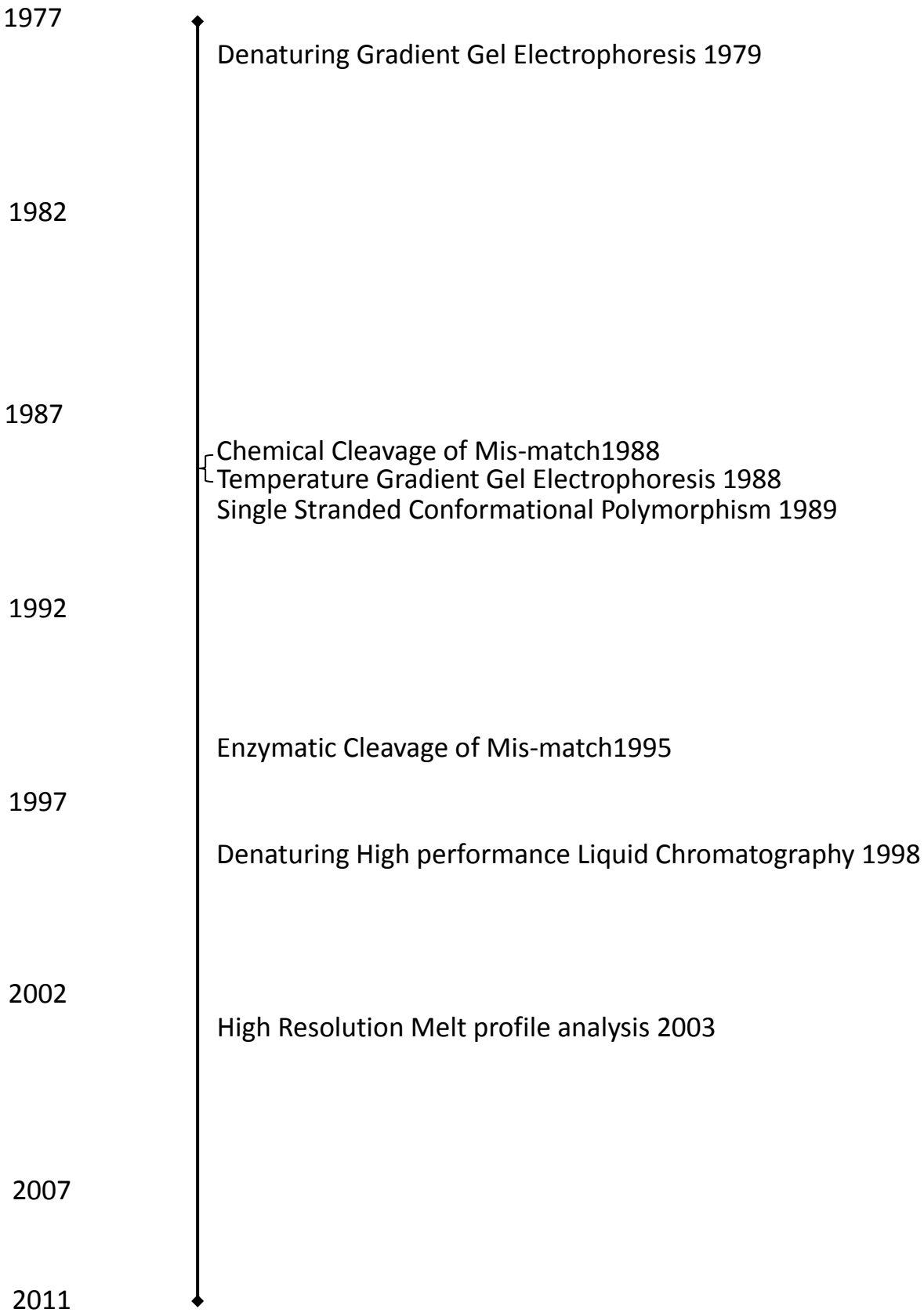

1977
Denaturing Gradient Gel Electrophoresis 1979
1982
1987
Chemical Cleavage of Mis-match1988
Temperature Gradient Gel Electrophoresis 1988
Single Stranded Conformational Polymorphism 1989
1992
Enzymatic Cleavage of Mis-match1995
1997
Denaturing High performance Liquid Chromatography 1998
2002
High Resolution Melt profile analysis 2003
2007
2011

Figure 1.5: SNP Discovery Technique Time Line.

1.3 Genotyping

SNP discovery is still an essential part of genetic research and identifying personal SNPs and haplotypes will always be necessary; however, for projects such as association studies sufficient information regarding common variation is likely to exist in resources such as the NCBI dbSNP database. Where the necessary SNP information is available genotyping assays are used to test for allele distribution within sample cohorts in a fast and cost effective manner.

1.3.1 Specific Oligo Hybridisation

Under sequence specific conditions an oligo probe will anneal to an exactly complimentary sequence, even a single mismatch can be sufficient to prevent this. By testing samples in duplicate with separate probes for the reference and alternative alleles it is possible to determine genotypes. The process of discriminating between alleles in this way is termed 'specific hybridisation' and it is common to several genotyping techniques.

In the original technique (Wallace et al., 1979) probes were designed to span a SNP loci, ideally with the SNP site towards the centre of an 18 to 25mer to maximise the effect of a mismatch, ideally probe ends with high GC content are avoided as they may stabilise a hybridisation even in the presence of a mismatch. G:A and G:T mismatches are the least disruptive heteroduplexes but can be avoided by designing probes on the assay on the alternative strand which will exhibit C:T and C:A mismatches respectively.

The assay was commonly performed by immobilising sample PCRs on a membrane (hybridisation dot blot) and probing with a radio-labelled oligo though it is also possible to immobilise the probes instead and test them with labelled PCR products. The former process is used to test multiple samples for a single SNP the

later tests one sample for multiple SNPs. As with all assays where annealing complimentary sequences is the mechanism of discrimination, annealing conditions must be titrated exactly which necessitates a lot of preparatory work before an assay can be run and numerous sets of different conditions where multiple loci are being tested.

In 1996 Gen-Probe described a variation of the process (Nelson et al., 1996) where probes were labelled with a highly chemiluminescent acridinium ester (AE). Complete annealing of a probe would prevent it from being hydrolysed, the hydrolysis step destroyed the chemiluminescence capability of probes that were not bound hence identifying alleles. Nelson was able to demonstrate the detection of 12 different mismatches in HIV-1 but the technique was not widely used, possibly because of the popularity of other assays available at that time with more simplified methodology.

1.3.2 Restriction Fragment Length Polymorphism (RFLP)

Restriction enzymes recognise and cut specific DNA sequences (Figure 1.6 A). If a locus of interest happens to create or disrupt a naturally occurring restriction site the corresponding enzyme can be used to test for the presence or absence of alleles (Figure 1.6 B).

Originally RFLP analysis was carried out on genomic DNA which was hybridised to membranes and probed with radio-labelled oligos, this technique is described in the landmark DMD paper which investigated 11 X chromosome RFLP markers (Bakker et al., 1985). As PCR was developed it became possible to digest amplicons measuring the sizes of resulting fragments with agarose gel electrophoresis, this made the assay more accessible as more commonly cutting enzymes could be utilised which were not practical to use when digesting very large fragments (due to the number of restriction sites that would occur) with PCR fragments ubiquitously occurring sites can be avoided or accounted for in the interpretation of the results.

RFLP is arguably the most simple genotyping technique, digestion conditions relate to the chosen enzyme reducing the need for comprehensive assay titration; however, it is reliant on the locus of interest creating or destroying a restriction site limiting the number of loci it can be applied to. It is possible to use primers which will change or insert a nucleotide within the primer portion of a PCR therein creating a new restriction site (Figure 1.6, C); however, this is not always possible as the surrounding sequence may still be limiting.

A significant risk of RFLP analysis is that the partially digested PCR product when run out on a gel for analysis can create a false result making a sample which is homozygous for a cut site appear to be heterozygous and while controls can be added to a sample set, individual digestion reactions are not easily controlled.

1.3.3 Oligo Ligation Assay

When primers anneal immediately adjacent to each other on the same template strand they can be joined together using a ligase to form a single stranded molecule. By designing one of those primers such that the 3' base is a compliment to one allele of a SNP and applying stringent annealing conditions, ligation will only occur where that 3' base finds a true compliment, this forms the basis of the OLA technique.

The original technique (Landegren et al., 1988) uses a discriminating probe with a 5' biotin moiety and the common probe with a 3' radiolabel. Following the ligation reaction streptavidin is used to isolate the biotin probes and only those molecules where ligation has been successful will be radio labelled. The results of individual reactions are hybridised to a specific location on a membrane and the results of an assay can be determined. Several developments have improved the throughput whilst retaining the core technique.

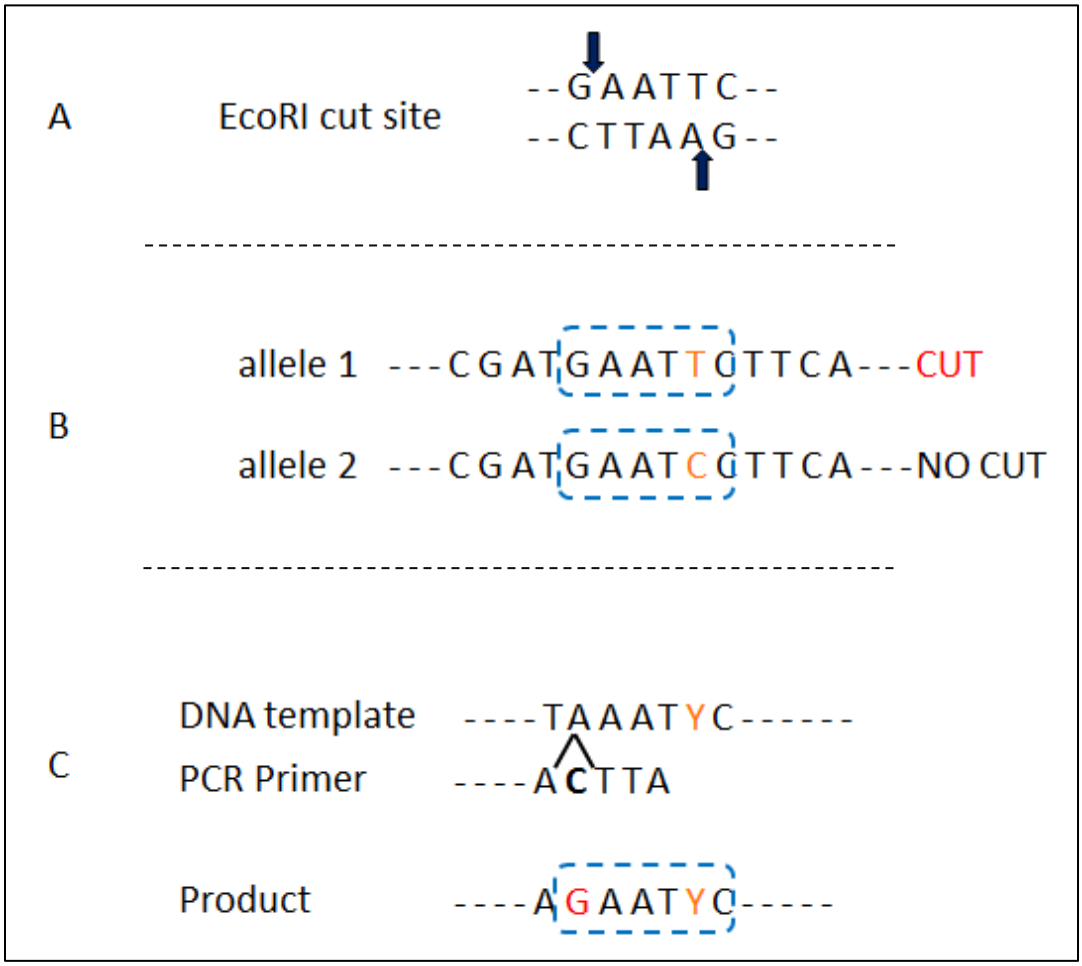


Figure 1.6: RFLP.

(A) The natural cut site for *EcoRI*. (B) Allele 1 will be cut by *Eco*RI but allele 2 does not match the *Eco*RI cut site, digestion will demonstrate the sequence at this loci. (C) The PCR primer is designed with a base that is not complimentary to the template sequence (**C** in the primer). In subsequent rounds of amplification the C allele becomes incorporated into the product in place of the naturally occurring A and a corresponding G allele is created on the opposite strand producing an *Eco*RI cut sit which includes the SNP locus 'Y'.

1.3.4 Amplification Refractory Mutation System (ARMS)

Several acronyms are used to describe this technique including ARMS, PASA (PCR Amplification of Specific Alleles), ASA (Allele Specific Amplification) and ASPE (Allele Specific Primer Extension), the assay uses a three primer system similar to that of the SNPlex technique with 2 allele discriminating primers and a common primer for the opposite strand. The DNA template dictates which discriminating primer anneals during the PCR cycle, where annealing is successful a product can be generated, where the discriminating primer does not anneal to the template no product is produced (Newton et al., 1989).

Separate reactions are set up to test for the presence of a reference allele and alternative allele, where both assays produce a fragment the sample is heterozygous, where only one PCR is generated the sample is homozygous for the respective allele. The assay uses simple lab techniques and equipment and individual reactions can be controlled for by the inclusion of a separate set of PCR primers with a different overall size, this control will identify failed PCRs from those where one allele is not present in the template.

Optimisation is critical for these assays as a low annealing temperature can lead to a primer binding in spite of a 3' mismatch, it is common to introduce a mismatch at the n-2 position of the discriminating primers so that this instability increases the likelihood of a primer not binding where the 3' match is not perfect (Figure 1.7). The internal control can also be problematic as a result of competition for PCR reagents; however, this remains a widely used technique that has contributed significantly to the discovery of disease genes (Johannsson et al., 1996).

An ARMS assay generally requires 2 assays to test a sample though a variation on the technique which tests for both alleles in one reaction has been described and termed Tetra –ARMS-PCR (Piccioli et al., 2006).

1.3.5 FRET-OLA

Fluorescence resonance energy transfer (FRET) has been used to distinguish between ligated and un-ligated probes. In this technique one primer carries a 'donor' molecule the other carries a 'reporter' molecule both of which will emit a specific frequency of light when excited. When the two primers are ligated the donor and reporter molecules are brought into close proximity where energy may transfer between them, excitation of the donor with a specific light frequency will cause it to emit light at a frequency that excites the reporter, this in turn emits light at a new frequency, the presence of this new frequency of light demonstrates successful ligation (Chen et al., 1998).

1.3.6 Padlock Probes

Nilsson et al (1994) described the use of 'padlock' probes where two end of a single DNA molecule align at the SNP loci. Ligation causes circularisation of the molecule and Rolling Circle Amplification (RCA) is used to produce PCR product. SNP assays are carried out in duplicate testing for each allele separately with specific probes, the technique is also known as Ligation-Rolling Circle Amplification (Qi et al., 2001).

1.3.7 Minisequencing / Primer Extension

As the name suggests this is a version of a standard Sanger sequencing reaction that Sokolov (1990) published whilst working with *CFTR*. With the absence of normal deoxynucleotides (dNTPs) which would be present in a standard sequencing reaction, the polymerase can only extend a sequencing primer by a single fluorescent di-deoxynucleotide (ddNTP). By designing the sequencing primer so that its 3' nucleotide is adjacent to a SNP locus the single incorporated fluorescent ddNTP will be the compliment to the SNP allele of the template effectively creating an oligo with a 5' fluorescent label specific to the amplified allele. Subsequent analysis of the products can be performed in several different

ways and the chosen technique is likely to depend upon the technology that is already available within a laboratory.

1.3.8 Taqman

The Taqman assay determines the genotype of a SNP locus during PCR amplification, techniques that do this are known as 'real-time' assays. A set of normal PCR primers are designed which span the region of interest and an allele specific probe is designed which anneals between them. Covalently bound to the 5' end of the allele specific probe is a fluorophore and at the 3' end there is a quencher molecule, the close association of these two molecules when they are bound to the probe prevents the fluorophore from emitting light. As the polymerase travels along the nascent strand its inherent 5´→3´ exonuclease activity means the fluorophore is cleaved from the probe releasing it from the inhibiting effects of the quencher. Light is released where an allele specific probe had hybridised successfully demonstrating the presence of the corresponding allele (Figure 1.8) (Holland et al., 1991).

Assays are now available for over 4.5 million SNPs, the set up process is simple and the technology required for analysis is relatively common in genetics labs as numerous 'real time' PCR machines are suitable making this a very popular format for screening individual SNPs; however, the assays cannot be easily multiplexed so for genome wide studies other techniques are more widely used.

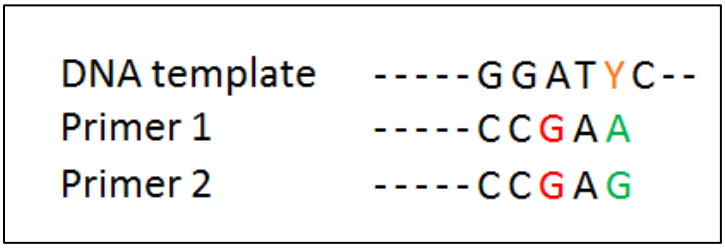


Figure 1.7: Designing primers for an ARMS assay.

The 3' nucleotide in primers 1 and 2 are complimentary to the different alleles of the SNP loci. The n-2 base of both primers (red) is a deliberate mismatch to the template sequence, this introduces instability that reduces the chance that a product will be created where the 3' nucleotide is a mismatch.

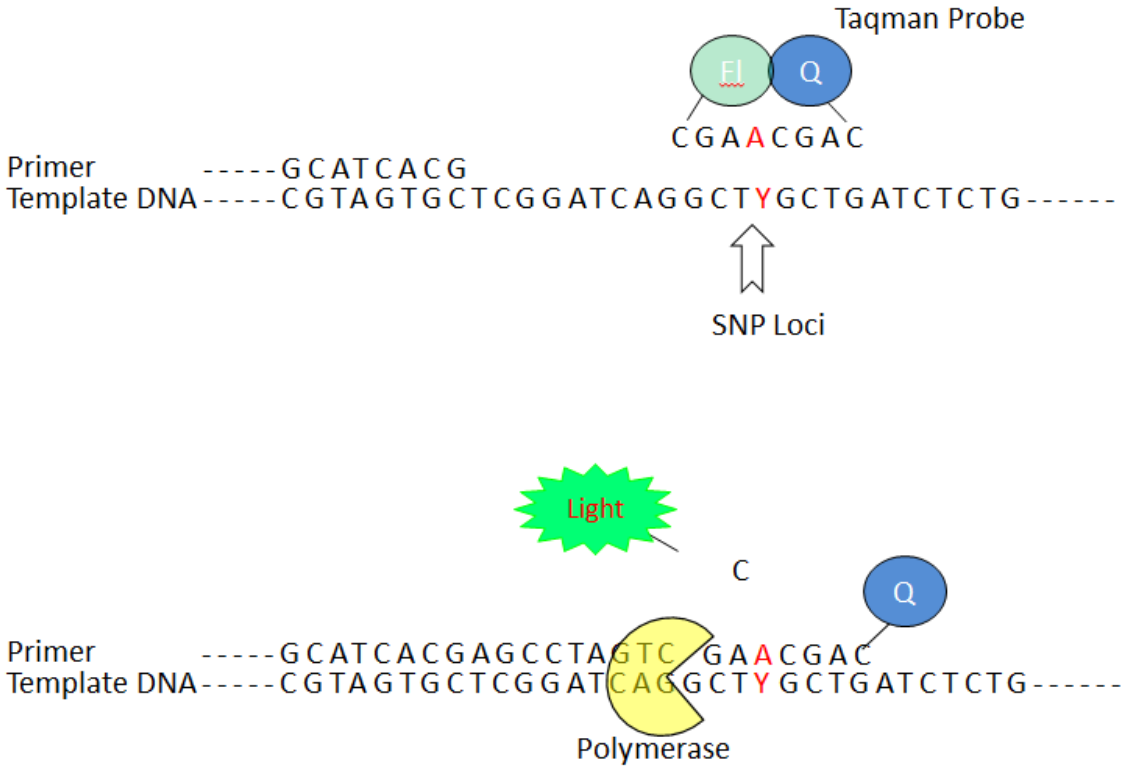


Figure 1.8: Taqman Genotyping Assay.

An allele specific Taqman probe is labelled with a fluorescent molecule at the 5' and a quencher at the 3', the quencher inhibits fluorescence while the molecules are in close proximity. During a PCR the polymerase incorporates dNTP along the template until it reaches an annealed probe at which point the 3' exonuclease activity of polymerase digests the probe DNA separating the quencher and fluorophore and allowing the fluorescence to be detected. Different fluorophores are used for each allele of the SNP allowing the reaction to occur in a single tube.

1.3.9 Molecular Beacons

DNA probes are designed with complimentary sequence at the 3 and 5' ends so that in their native form the ends hybridise creating a stem/loop structure. The sequence within the loop is specific to one allele of a SNP locus and as it is longer than the complimentary stem sequence, when it binds a template the new structure is more stable than the stem-loop. Covalently bound to the probe ends are a fluorescent molecule and a quencher, in the native form these are held in close proximity but where hybridisation occurs the molecules are separated sufficiently by the intervening sequence that the fluorophore can emit light when excited hence demonstrating a sample genotype.

In solution the probes bind the template, if the sequence is fully complimentary the fluorophore is released from the quencher, the alternative allele's probe will remain in its native form and its fluorescence is quenched. As the temperature increases the template/probe hybrid molecules will denature causing a reduction in the fluoresce associated with that allele as the probe returns to its native form, eventually all of the probes will become denatured and fluorescence of both alleles will be detected (Tyagi and Kramer, 1996).

Molecular Beacon assays require the same equipment for analysis as TaqMan assays (a real time PCR machine) and have therefore been 'competing' against an established and well supported TaqMan community hence the technique is not as well used as it might otherwise have been.

1.3.10 Fluorescent Minisequencing

Fluorescently labelled ddNTPs enable the use of standard PAC gel technology to detect the incorporated fluorophore. The size resolution of the polyacrylamide gel enables a certain amount of sample pooling to be used for the purpose of increasing throughput, this can be achieved by designing assay primers with different lengths and using an internal size standard to separate the results during analysis (Kobayashi et al., 1995).

1.3.11 Matrix-Assisted Laser Desorption/Ionization (MALDI)

Also known as Mass-spec, this technology 'fires' DNA molecules along a tube, the Time-of -flight (TOF) is recorded and can be used to detect the difference between primers molecules that have incorporated either the reference or alternative allele during the extension reaction.

Samples or 'analytes' are mixed with an organic matrix molecule which forms a co-crystallised product when spotted onto a metal MALDI plate, the plate is then placed into an ionisation chamber which contains a vacuum. A single spot on the plate is targeted with a laser of a specific wave length (depending upon which matrix was chosen). The matrix absorbs the laser energy whilst protecting the analyte molecule from direct exposure to the laser. Energy is transferred between the matrix and analyte which becomes ionised and evaporated into the gas phase within the vacuum. The ionised analytes charge is used to draw the molecules through the mas spec flight tube to the detection plate at the far end. Mass spec has sufficient resolution to detect the difference between molecules with differing bases at their 3' end hence deducing genotypes (Braun et al., 1997).

Unlike the fluorescent method no labels or internal ladders are required and though samples are analysed sequentially run times are a fraction of a second. The negative charge of the phosphate backbone of a DNA molecule complicates the sample preparation and ultimately leads to a reduced signal when compared to analysis of other peptides with a similar mass. Complicated preparation steps are more difficult to automate which can make MALDI-TOF a less attractive option for genotyping but where the skills and equipment are available the technique is still utilized (Roberts et al., 2011).

1.3.12 Fluorescent Polarisation

Polarised light can be used to excite fluorescent molecules in solution. If the molecules are relatively small (such as an unincorporated fluorescent dNTP) they will be rotating as they are excited and emit scattered light. If the molecules are large

(such as the products of minisequencing reaction) then they will be rotating more slowly and the light which is emitted will remain polarised. This emitted polarised light can be detected making it is possible to distinguish incorporated and free fluorescent dNTP and therein identifying which dNTP is incorporated following a minisequencing reaction (Chen et al., 1999).

1.3.13 dHPLC Genotyping

Hoogendorn *et*. *al*. (2000) described a variation of the process which removed the need for fluorophores. Assay specific mixtures of dNTP and ddNTP are created such that the presence of one allele will result in the immediate termination of extension and the presence of the other allele will result in extension of a small number of bases which can be limited by excluding specific dNTPs from the reaction mixture. The two molecules can be run on a dHPLC machine to resolve the sizes therein detecting the genotype.

This technique is relatively low throughput and complicated by the fact that separate reaction mixes are used for each SNP. In theory any technique that measures the sizes of small oligos precisely could be used if dHPLC equipment was not available.

1.3.14 Dynamic Allele Specific Hybridisation (DASH)

Based on the same principal as the SNP discovery technique HRM (Section 2.3.4) DASH uses the release of an intercalated dye to demonstrate which SNP allele a DNA strand contains. PCR of a region containing a SNP is carried out where one of the primers is biotinilated, this adapter molecule can then be used to fix PCR products to a surface that has been coated with streptavidin such as the bottom of a well in a microtitre plate. The fixed PCR product is denatured to remove the complimentary strand which is washed away, a specific probe can then hybridise to the fixed single strand in the presence of an intercalating dye. This probe must cover

the SNP of interest where it will either form a fully complimentary double strand structure or a mismatch will form creating a heteroduplex, when the sample is slowly heated the heteroduplex fragments will be first to release their dye enabling genotypes to be determined (Howell et al., 1999).

1.3.15 Pyrosequencing

The incorporation of a nucleotide during a Pyrosequencing reaction is followed by an enzyme cascade that ultimately results in a flash of light (Ahmadian et al., 2000, Nordstrom et al., 2000). Detection of light confirms incorporation of the nucleotide that was introduced at that point and the process is repeated to generate a short DNA sequence read. As nucleotides are added in turn it is possible to detect the presence or absence of an allele at a given location and because the quantity of light is proportional to the number of incorporations it is also possible to detect heterozygosity where half the quantity of a homozygous allele is available.

In choosing the sequencing primer location and order in which nucleotides will be added it is possible to have a negative control within each well of a plate of samples where an allele is added to the reactions knowing that it is absent in the template. Likewise, before the SNP locus is interrogated a positive control can be created by testing for known sequence that lies between the primer and SNP. Internal control reactions are unique to this genotyping technology though there is a cost to every nucleotide addition which needs to be considered.

The process of determining sequence by detecting the release of phosphate was originally described by Ronaghi (1999) who identified the problem that homopolymer runs of greater than 4 bases were difficult to accurately convert into sequence. Though the chemistry has been improved in recent years the linear relationship seen between light intensity and homopolymer runs breaks down after 8 bases (Margulies et al., 2005). This creates a problem for a small number of SNPs but like many assays the opposite DNA strand can be used to avoid difficult sequence.

1.3.16 Invader Assay

The Invader® assay is based on the unique cut site of Flap Endonuclease (FEN) which recognises a specific DNA structure rather than a specific DNA sequence and is highly sensitive to sequence mismatches. The assay has an invader probe which binds sequence immediately upstream of the SNP locus and has a 3' non matching base which overlaps but does not bind the SNP locus. Allele specific probes are designed to hybridise immediately downstream of the SNP locus but also extends in the opposite direction creating a 'flap' over the location where the Invader probe has already aligned. The discerning base of the allele specific primer is towards the centre of the sequence, if it matches the template the two probes bind with the template creating a 3 molecule structure that FEN recognises and cuts such releasing the 3' end of the allele specific primer (the flap) which is covalently labelled with a fluorophore. As with the molecular beacon technology a fluorophore/quencher arrangement can be used to detect the separation of this 3' fragment but fluorescent polarisation and MALDI techniques have also been used (Lyamichev et al., 1999).

In its original form each allele has to be tested for separately but developments describe by Hall (2000) utilised a secondary FEN digestion step to allow detection of both alleles in one tube and array technology to enable both alleles and multiple assays to be tested simultaneously.

1.3.17 SNPlex

The SNPlex assay by Applied Biosystems (Tobler et al., 2005) is considerably more complicated than the other OLA based techniques but OLA is still at the core of the discrimination. A single reaction is used to test for both alleles so three probes are used for the allele discrimination stage; one for each of the alleles (the discriminating probes) and a common probe which can be ligated to which ever probe overlaps the SNP locus successfully during the hybridisation. The two discriminating probes have 5' 'Zipcode' sequences that can later be used to identify

which allele was present at a specific locus. The 5' sequence of the common probe and the Zipcode sequences are recognised by a second set of three probes which attach common primer sequences to the ends of the ligated products so that multiple assays can undergo PCR with one set of conditions. The PCR includes a 5' biotin molecule which can be used to fix the amplicons (containing the allele and assay specific Zipcode sequence) onto a solid surface. Fluorescent ZipChute ™ probes which are specific to one allele of a given locus are allowed to hybridise to the immobilised DNA molecules, those that do not bind are washed away and those retained can be separated from their compliments and identified using electrophoresis.

1.3.18 KBiosciences Competitive Allele-Specific PCR (KASPar)

The KASPar assay starts with a three probe system similar to that used in an ARMS assay. The first primer is specific to one allele, the second is specific to the alternative allele and the third is a common PCR primer. During amplification both discriminating primers are available but will only be incorporated if there is a suitable allele at the SNP locus which is complimentary to the 3' nucleotide of the discriminating probe. Where there is a match, the primer anneals, PCR occurs and a fragment is generated. As well as their specific 3' nucleotide, each discriminating probe has a specific 5' sequence, when a round of amplification is complete this sequence is used as an annealing site for a fluorescent primer which is initially bound to a complimentary sequence carrying a quencher molecule. When the probe and compliment separate the probe can be used as an amplification primer and the light signal can be detected as the association with the quencher has been disrupted. Subsequent PCR cycles will strengthen the fluorescent signal (Figure 1.9) (Cuppen, 2007).

As with the Taqman assay the whole process is completed in a single tube; however, as the fluorescent probes are generic and the assays specific probes are

unlabelled the overall cost is very low, what is more KBioscience (the provider for the assay) have a very high assay conversion rate (>90%).

1.3.19 Array Techniques

Microarrays were the first technology to provide a global view of the genome due to the number of assays that can be simultaneously genotyped. The techniques used to discriminate between alleles are simply developments of those described above; however, array technology has enabled the miniaturisation and parallelisation of assays to provide greater genotyping throughput than any of the previously described techniques. Modern arrays are capable of analysing hundreds of thousands of assays simultaneously; however, in the case where a small number of SNPs need to be assayed they tend to be impractical.

An array is a solid surface to which oligo probes can be attached, commonly used materials include glass slides or polystyrene microspheres or 'beads'. DNA sequences are covalently bound to the surface; they may span SNP loci and use allele specific hybridisation or terminate at the base upstream of the loci and use a minisequencing reaction to create a discriminating molecule. The two most commonly used SNP chips are; Illumina's Infinium II technique which uses a combination of assay specific beads and minisequencing to analyse 500,000 assays in parallel, and Affymetrix SNP Chips which use glass slides and specific hybridisation, over 900,000 SNPs.

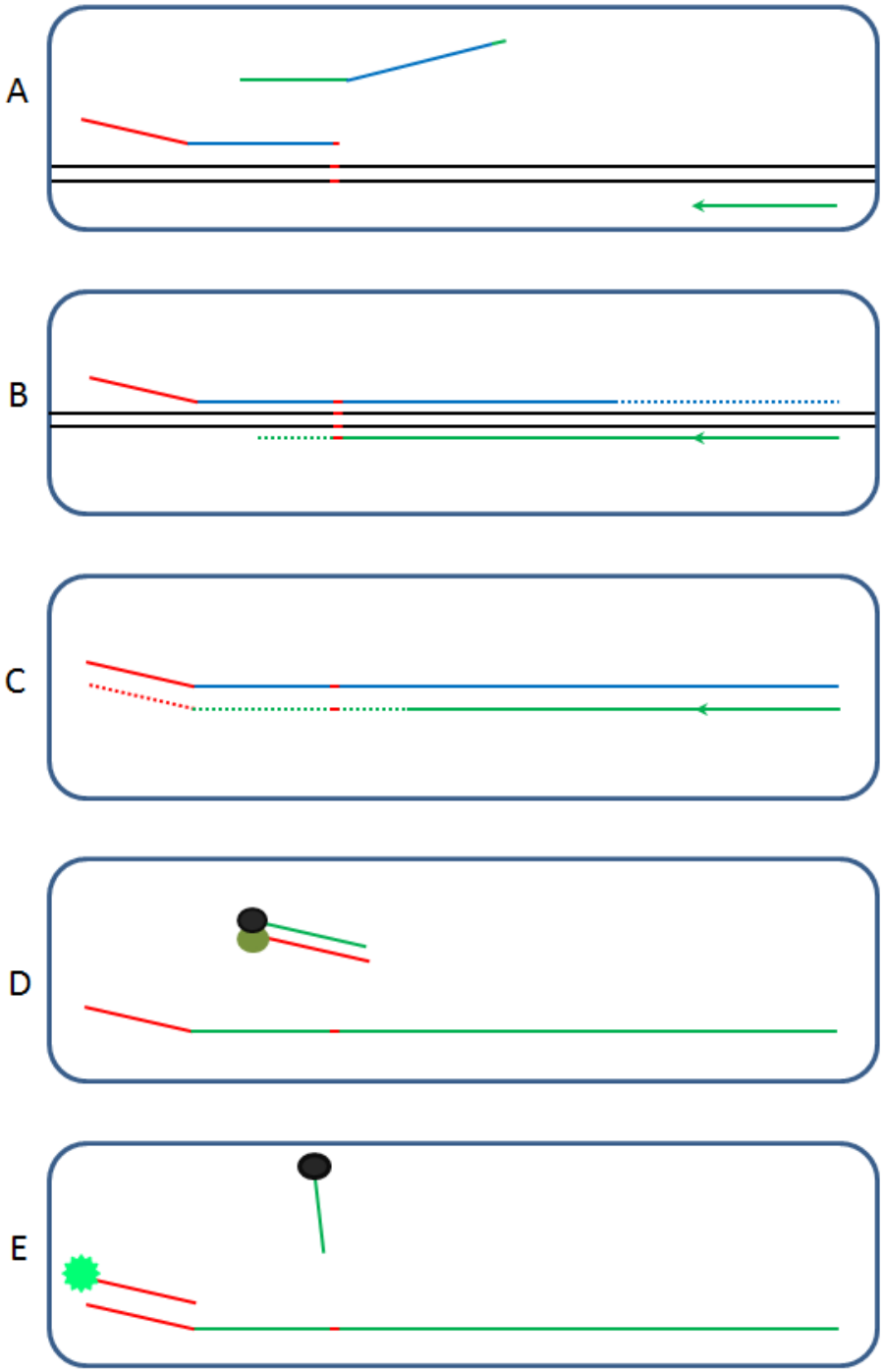


Figure 1.9: KASPar assay.

(A) Two allele specific primers and one common primer are mixed under specific annealing conditions such that the SNP allele determines which specific primer binds (red). (B) The primers are extended. (C) Subsequence rounds of PCR generate products that include the allele specific primers 5' sequence. (D) The allele specific double stranded quencher-fluorophore complex is added to the single stranded PCR template. (E) Where the probe binds to the allele specific 5' sequence the quencher-fluorophore complex is separated and fluorescence can be detected.

1.3.20 Golden Gate Genotyping Assay (Illumina)

As with an ARMS assay the Golden Gate (Fan et al., 2006) reaction starts with three primers; discriminating primers will anneal to the SNP adjacent sequence and the 3' base will be specific to one allele of the SNP locus. In addition to this there is a further section of sequence upstream which does not align to the genomic DNA but will later be used as a generic amplification target. This target sequence identifies which allele lies at the 3' end of the discriminating primer (the wild type or alternative). The third primer anneals immediately adjacent to the SNP locus and has three sections, the first is the assay specific sequence, in the middle is an 'address' sequence which will be used to link the ultimate product to a specific array location (and hence a particular assay) and the 5' sequence is a generic amplification target sequence.

During the reaction the probes anneal around the SNP locus, the template sequence determines which of the discriminating primers is incorporated and the rest of the sequence from both primers remains free from the template. A ligase connects the two primers under stringent annealing conditions to create a single product which can be pooled with hundreds of similar products from other assays. The pooled products are amplified with a set of three generic primers, the discriminating primers are fluorescently labelled and anneal to the 5' end of the discriminating primers from the ligation reaction hence demonstrating the presence of a reference or alternative allele. When amplification is complete the mixed pool is hybridised to an array which contains assay specific probes to bind the 'address' section of the products, now the pooled assay products are separated out to specific locations the fluorescent label can be translated back into sequence information (Figure 1.10).

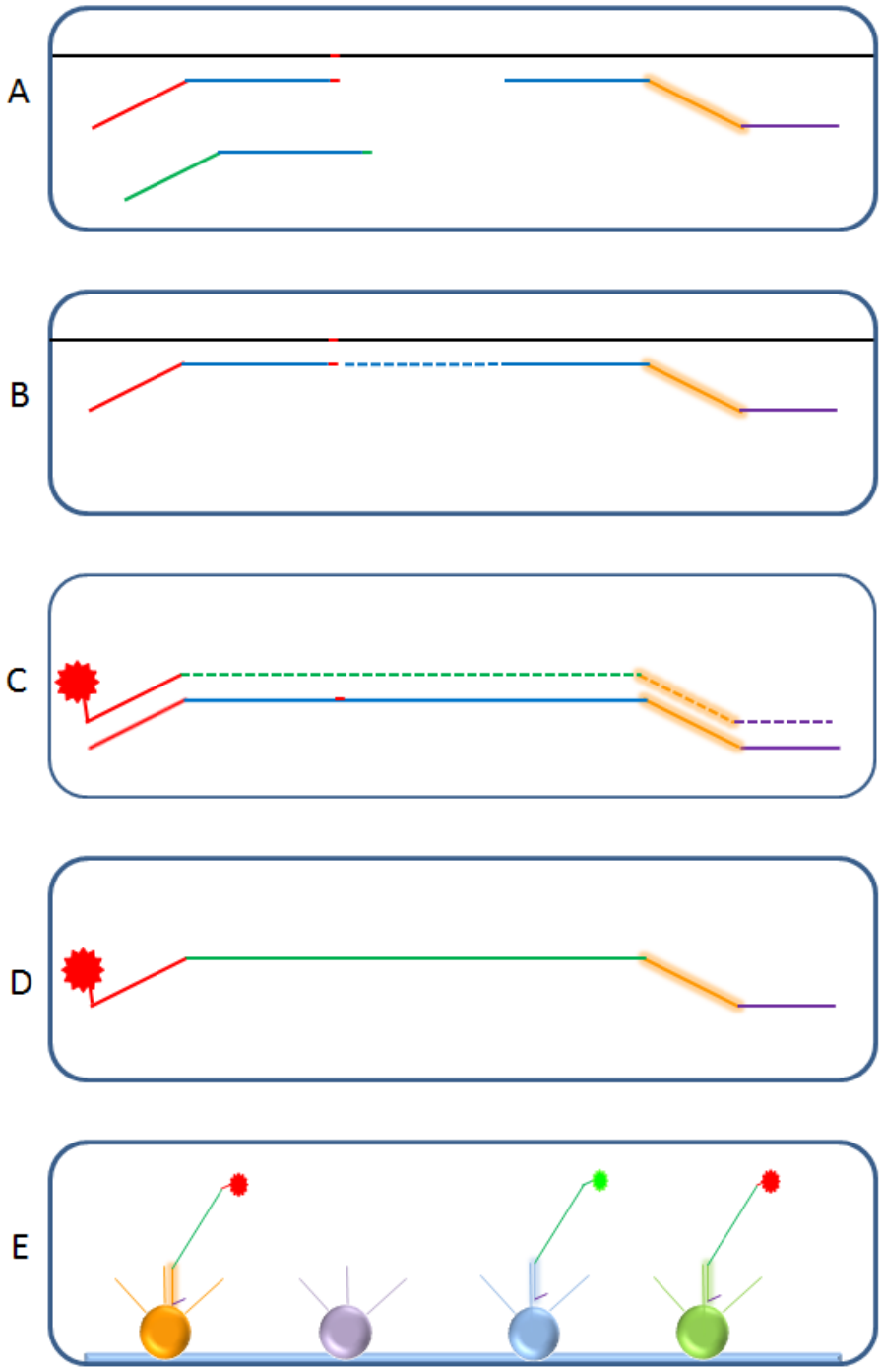


Figure 1.10: Golden Gate.

(A) Two allele specific primers and one common primer are mixed under specific annealing conditions such that the SNP allele determines which specific primer binds. (B) The primers are extended and ligated. (C) Generic primers amplify the single stranded template; this can be done in a single tube for multiple assays. (D) The PCR products include a fluorescent 5' molecule which relates to the template SNP allele and the locus address sequence. (E) The address sequence binds specific compliment sequence at a given location, when the fluorescence is read the allele is determined for a specific locus.

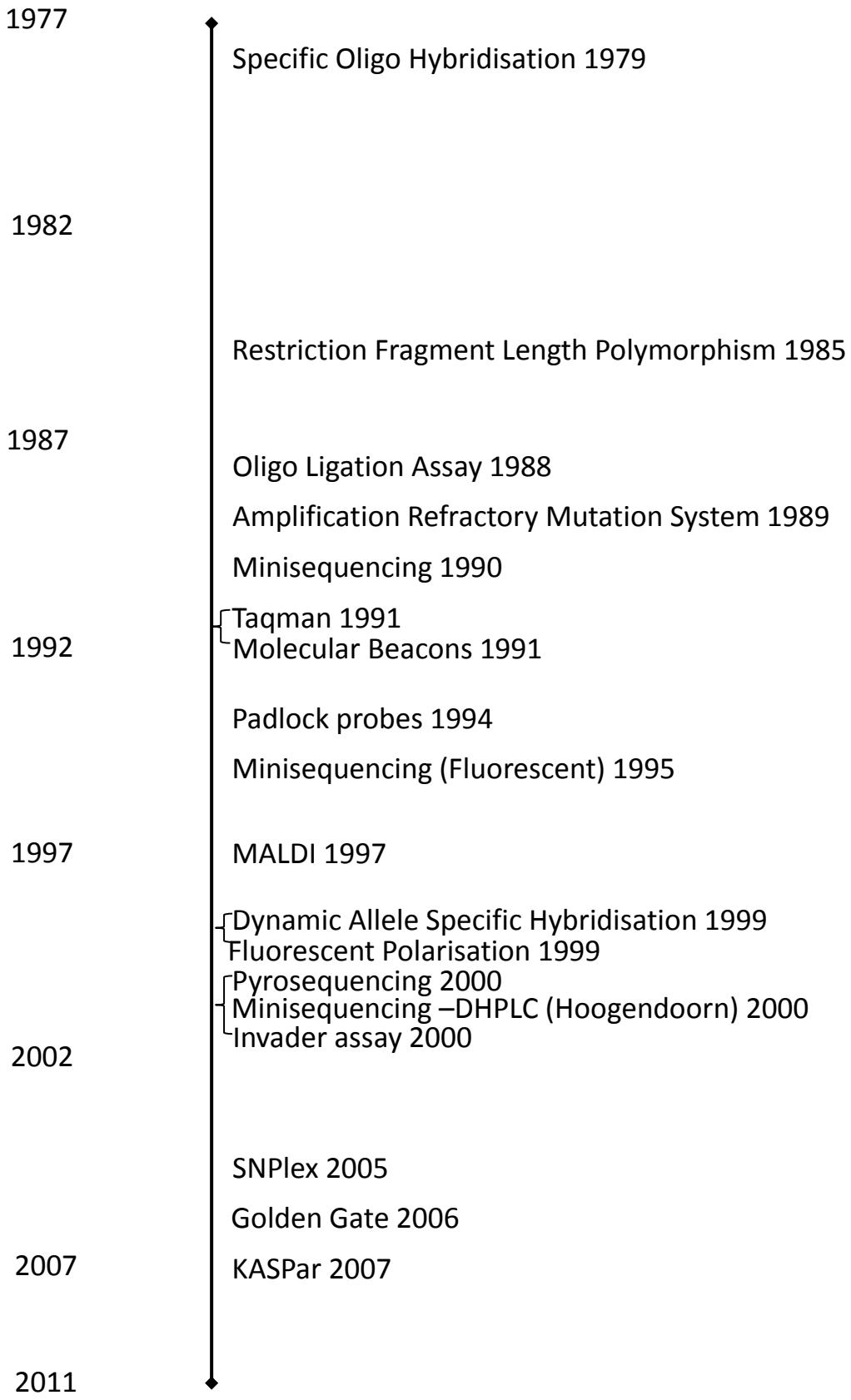


Figure 1.11: Genotype Assays Time Line.

1.4 DNA Sequencing

Central to the field of genetics is the ability to determine the order of nucleotides in a DNA molecule. For the past three decades a single method has been the dominant approach (Sanger Sequencing); however, the field is currently undergoing a revolution with numerous new techniques due to appear over the next few years.

1.4.1 Sanger Sequencing

DNA Sequencing is considered the 'gold standard' SNP discovery technology and though 'Next Generation Sequencing' (NGS) platforms have recently become available, re-sequencing of PCR amplicons is still primarily performed using the Sanger method (Sanger et al., 1977).

Sanger Sequencing utilises the properties of the chain terminating ddNTPs, once a ddNTP has been incorporated into a strand of nucleotides, no further 3' extension of that strand can take place.

During thermal cycling of a sequencing mixture a primer will attach to template DNA (e.g. cleaned PCR) and the polymerase will drive extension using the normal dNTPs, as soon as a ddNTP is incorporated extension of the strand is terminated leaving a single strand that is fluorescently labelled at the 3' end with a fluorophore that corresponds in colour to the type of ddNTP (A, C, G or T). The reaction mixture is set up with enough template, primer and reaction mix so that multiple fragments will be created that terminate at every base of the template and by separating these fragments by size, the sequence may be determined (Figure 1.12).

By using fluorescently labelled ddNTPs Sanger sequencing can be analysed by capillary electrophoresis, each fragment that passes the CCD camera is a single base longer than the previous and that base is the 5' fluorescently labelled ddNTP. The CCD Camera emits the laser light and 'captures' the fluorescence emitted from the

'excited' fluorophore. The data is fed back to the computer where the software converts the observed fluorescence into chromatograms which can be viewed and screened for polymorphisms.

Size separation of these fragments can be achieved by capillary electrophoresis, the fluorophores are excited by a laser and the specific fluorescence given off is captured with a CCD camera that translates the data into a chromatogram (Figure 1.13). As genomic DNA templates contain both a maternal and paternal allele a single 'peak' actually represents the product of both alleles and where heterozygosity occurs two peaks of different colour occur at the same loci with reduced heights (compared to the peaks seen in a wild type sample).

Each 'peak' represents the product from the maternal and paternal chromosomes, where the alleles match one peak is observed, but where the alleles differ two 'half height' peaks are observed at the same location with different colours, these characteristics are indicative of a heterozygous allele.

1.4.2 Maxam and Gilbert

Maxam and Gilbert devised a method of determining DNA sequence by placing the same DNA sample in four separate tubes and performing a different chemical cleavage reaction in each tube. The incomplete cleavage reactions cause the fragments in each tube to all terminate in the same base (or combination thereof) the combination of bases are; G, G and A, C and T or T (Maxam and Gilbert, 1977).

The technique was initially very popular as template preparation was more simplified than the Sanger method (Section 1.4.1) which initially required cloned single stranded templates; however, as the Sanger reaction improved it became the dominant sequencing technology for the next three decades.

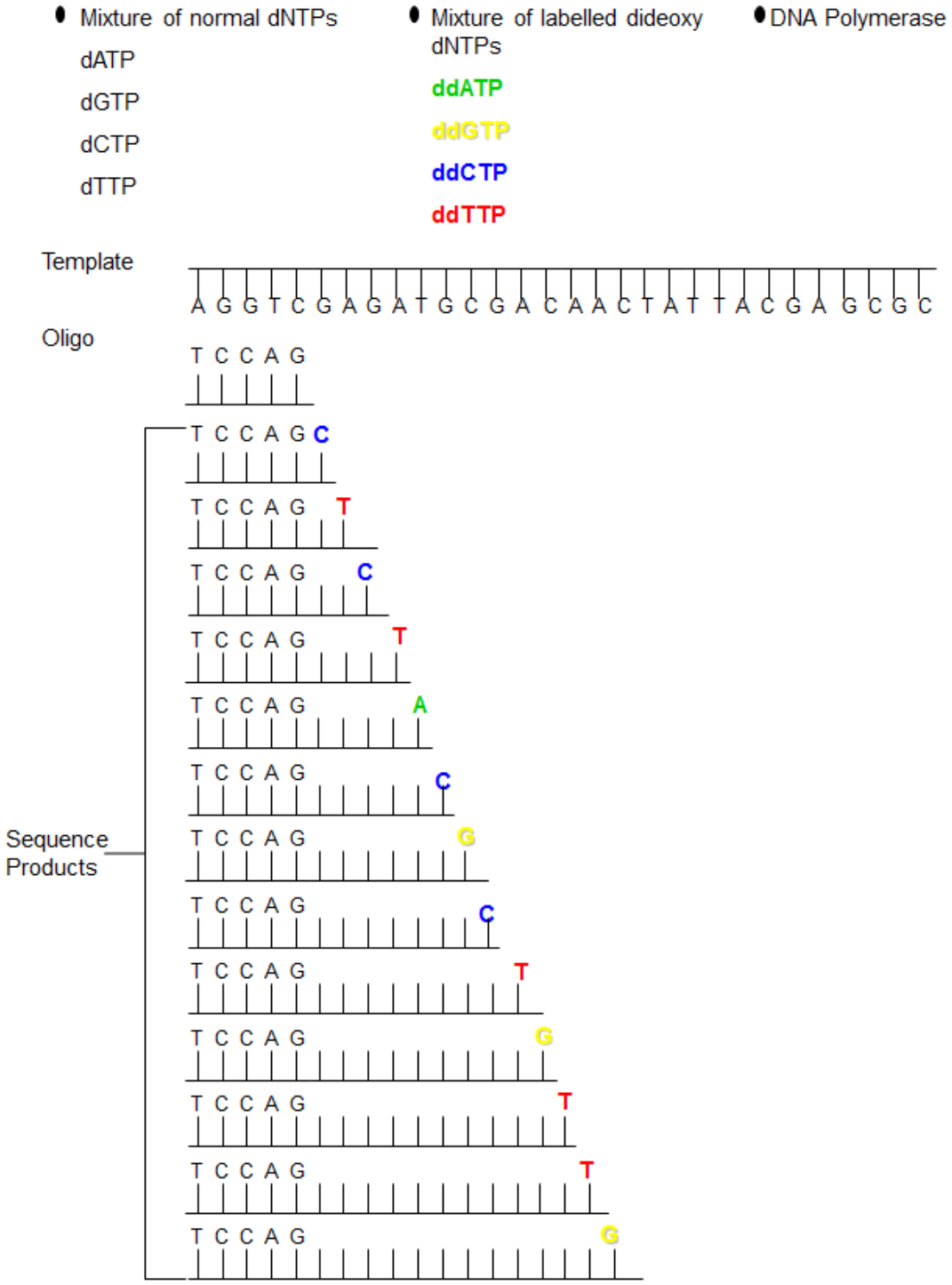


Figure 1.12: Sanger Sequencing.

A DNA template is mixed with dNTP, fluorescent ddNTP, DNA polymerase and a single sequencing primer. During a thermal cycle products will be generated that terminate in fluorescent ddNTPs. Sufficient cycles of amplification are carried out such that all loci are represented by fragments terminating at that position. Fragments are separated by size to determine DNA sequence order.

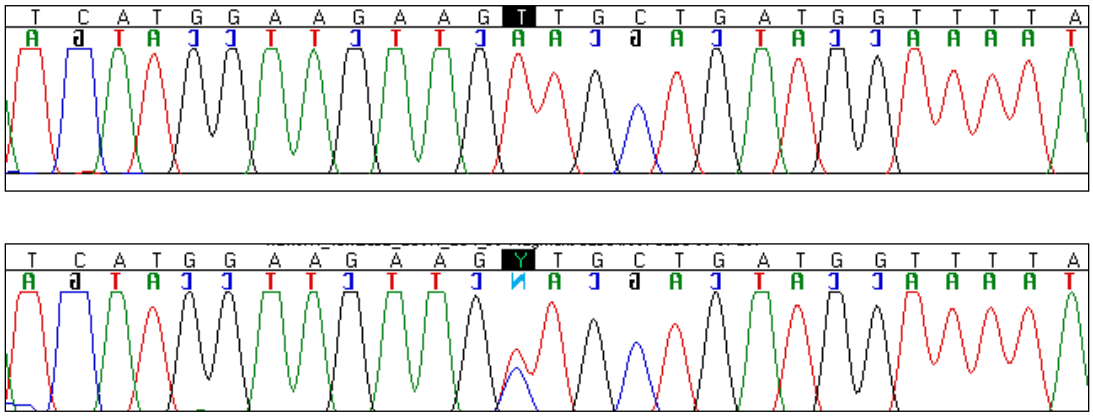


Figure 1.13: Heterozygous SNP locus.

The upper chromatogram shows a normal homozygous ‘T’ allele at the highlighted SNP loci. The lower chromatogram shows heterozygosity at the same locus.

1.4.3 Next Generation Sequencing

The name 'Next Generation Sequencing' was given to the first group of techniques that represent a significant departure from Sanger Sequencing. The techniques are radically different from each other in how they generate sequences but there is a common approach in that fragmented genomic DNA is sequenced in a massively parallel manner before being re-assembled in software for analysis. The first publication of a genome that was sequenced by NGS (Roche -454) was that of James Watson (Wheeler et al., 2008) at a cost of $1.5 Million it was a significant improvement on the Sanger sequenced human genome (Levy et al., 2007) which has been reported as costing over $100 million with some estimates as high as $4 billion. The techniques described below are by no means comprehensive, there are several other techniques in development, but those described represent the most commonly used techniques to date.

1.4.3.1 Roche – 454

All NGS techniques start with shearing a genomic DNA sample selecting fragments of a specific size and tagging the ends with primer sequences which will be used for amplification and sequencing, this collection of fragments is known as a library. With the 454 technique the library is denatured and mixed with primer coated beads in an oil/water emulsion. A specific ratio is used to achieve the maximum number of beads binding a single library fragment whilst minimising the quantity of beads with multiple templates. Titrations of 'bead to template' are undertaken to minimise the creation of polyclonal beads (where 2 or more fragments attach) whilst reducing the number of un-templated beads. Un-templated beads can be removed by an enrichment process but polyclonal beads will carry through to the sequencing phase and ultimately reduce the amount of high quality data that can be produced.

The primer sequence of the bead anneals to one of the two sequence tags that tail each fragment of the library. The second tag on the library template is used

to generate a complimentary strand which will become covalently attached to the bead. The template strand is then washed off to leave a bead with a single stranded covalently attached template or a 'library bead'. In using both tags in this process library fragments that incorporated the same tag at both in the initial tagging library preparation fail to amplify, the process is known as enrichment as it positively selects beads with the products that are suitable for sequencing.

The 454 technique then utilises emulsion PCR (emPCR) to amplify the template strand such that each primer on the bead becomes elongated by the same sequence (clonal amplification). The aqueous phase of the emPCR contains the PCR reagents and when mixed with oil forms droplets or 'micro-reactors'. Those micro-reactors containing a library bead plus the necessary reagents and one template strand will replicate their product to cover the bead in single stranded template DNA. The library beads are used to fill a picotitre plate containing picolitre volume wells created such that a single well may only contain a single library bead, once the library beads are in place the wells are packed with beads which are coated with the enzymes required for the sequencing phase.

The Sequence reaction occurs within the sequence instrument, the core technology is the Pyrosequencing technique (1.3.15). Reagents buffers and nucleotides are distributed across the plate in a sequential fashion by a microfluidics system the picotitre plate is heated to induce the chemical cascade and the quantity of light produced within an individual well is recoded. For each well 1 unit of light translates as the presence of a single template base, where homopolymer runs occur within a template strand there is a linear relationship between the quantity of light and the number of alleles up to a run of 8 after which sequence becomes unreliable (Margulies et al., 2005).

The 454 instrument has the lowest output of the three main NGS instruments (currently up to 750Mb per day) but read lengths of 700bp are considerably larger than the Illumina and SOLiD systems and comparable to Sanger technology. Comparatively long reads make the 454 particularly suitable for

Metagenomics and *deNovo* sequencing where fragment length simplifies the downstream sequence mapping and assembly process.

1.4.3.2 Illumina Sequencing

The Illumina platform was the first available short read sequencer and has a considerable market lead over the Applied Biosystems instrument, the current read length is 100bp and the instrument is capable of generating 50Gb of data per day.

Following the creation of a library the molecules are denatured and allowed to anneal to a 'lawn' of oligos which are covalently attached to an array surface. The array bound oligos then act as primer sites to the adaptors on the annealed strand and are extended by a polymerase to create a covalently bound template, the original strand is washed away. These covalently bound single stranded template molecules then under go 'bridge-amplification', the free end of the molecule is encouraged to anneal to another oligo on the array which again acts as a primer and creates a strand of sequence that is a compliment to the template, both strands are now covalently bound to the surface and the process is repeated cyclically until a 'cluster' of clonal fragments is created.

As the library is of a fixed length there is a natural limit to the cluster size that bridge amplification will tend to produce before clusters start to overlap, this introduces an inherent limit on the maximum theoretical length of a sequence that can be determined with this technique.

Once the clusters have been formed a primer is introduced which anneals to a section of the original tag that was added to the sheared DNA, a single fluorescent nucleotide is incorporated to each strand and the entire array is imaged. Each nucleotide is a reversible terminator, which prevents multiple incorporations and means that the sequence strands can be built up and imaged one base at a time, the process is named 'sequencing by synthesis' (Figure 1.14).

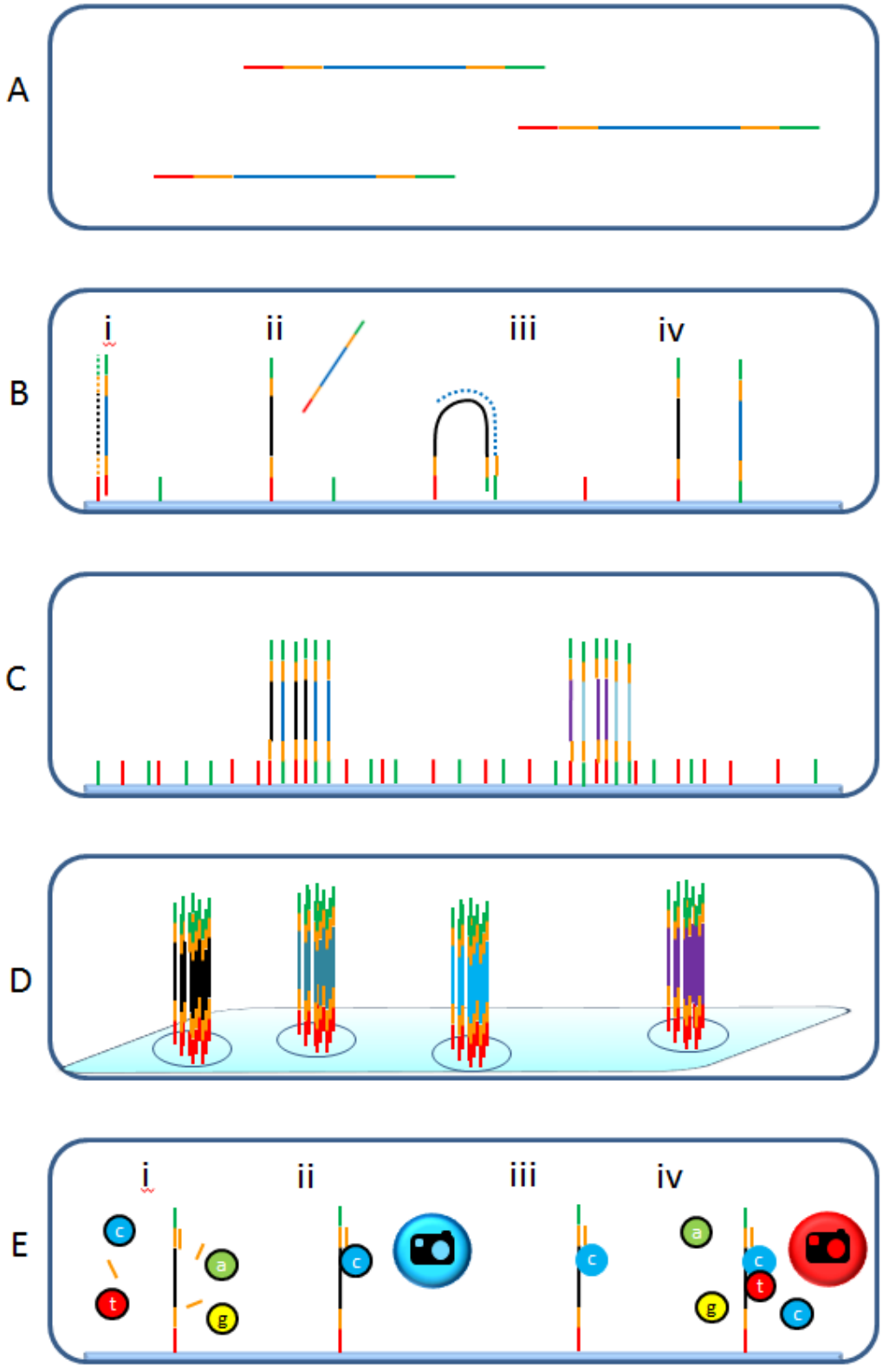


Figure 1.14: Illumina Sequencing.

Legend on following page.

(A) Adapters (red and orange, green and orange) are added to each end of a DNA fragment (blue) to create a library. (B) i. Library fragments bind to an oligo that is bound to the array surface, ii. The oligo is used as a primer for extension and a compliment (black) to the library fragment is generated on the surface, iii. The fragments are encouraged to form 'bridges' with complimentary oligos, iv. Both strands of a fragment are attached to the slide in close proximity. (C) Clusters of fragments are generated. (D) One of the two complimentary strands in each cluster is removed. (E) i, Primers, polymerase and fluorescent ddNTP are added, ii. The primer is extended by a single base and an image is recorded, iii. The incorporated ddNTP is modified so that a second ddNTP may be incorporated, iv. Subsequent cycles build up a sequence of colours at each location on the array surface.

1.4.3.3 Sequencing by Oligonucleotide Ligation and Detection (SOLiD)

The Applied Biosystems instrument produces the shortest read length (75bp) and claims to generate 15Gb per day but the method of sequence generation leads to the lowest error rate and the most reliable SNP calling (Liu et al., 2012). The technique is based on ligation of short oligos rather than extension by incorporation of individual nucleotides.

Libraries are prepared for sequencing by emPCR the process is very similar to that of the 454 technique with the ultimate goal of producing beads with covalently bound, single stranded, clonally amplified products. The templated beads are then immobilised onto a glass slide where the sequence reactions take place.

In the first step the array is treated with a mixture of sixteen eight-mer oligo probes which represent the sixteen possible combinations of sequence in the first two bases of each probe, (AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG and TT). Each probe is labelled with one of four fluorescent moieties so each fluorophore represents four of the probes. The probes align to the template sequence according to the complementarity of their first two bases, the next three bases are degenerate bases and the last three are universal bases but none of these affect binding. When the eight-mer is bound to the template an enzyme ligates the 3' end of the probe to the 5' end of the primer the entire array is scanned and the colour of the fluorophore is recorded. The scanned fluorophore relates to one of four possible sequences of the first two bases of the template. The last three bases of the probe along with the fluorophore are then cleaved and washed away ready for a new ligation reaction. For this second round of ligation bases 6 and 7 of the template are interrogated, followed by 11 and 12, 16 and 17 etc. Once the reaction has reached the end of the possible 75 bases of sequence the strand made up of ligated probes is removed and the process is repeated with a new primer that is one base shorter that the first, this is termed the 'n-1 primer'. The n-1 read determines the sequence of the most 5' base of the tagging primer sequence and the first base of the genomic DNA (bases '0' and 1) subsequent rounds of ligation the identify the

sequence at 5 and 6, 10 and 11, 15 and 16 etc. In total there are 5 rounds of cyclical probe ligation each round is staggered by one base, eventually each loci of the template is interrogated twice leading to very high sequence accuracy. Individual incorporation errors are readily spotted and a true SNP will cause two adjacent colour changes in two separate reads. Translation of the colour codes into bases is possible as the n-1 sequences (the last base of the tagging primer) are known. When the first base is known and the colour is observed the second base of that probe can be determined, in turn this 'second' base relates to the first base of the next probe, the colour is known and the subsequent base is identified so by knowing the first base and viewing the colours from all five rounds of ligation DNA sequence can be determined.

Though determining sequence in 'colour-space' is a strength of the technique it can also be considered a hindrance in terms of analysis as the raw data is not in the 'base-space' format that the majority of analysis tools require limiting the number of applications that can be used.

1.4.3.4 Ion Torrent

The Ion Torrent system is commercially available technique that utilises the properties of semiconductors. The approach to generating sequence is similar to that of the Pyrosequencing technique but instead of detecting light, pH changes are detected. Nucleotide incorporation releases Hydrogen ions which are converted into an electrical signal that can be detected by the semiconductor sensor. As with Pyrosequencing the array is treated with nucleotides in sequence and the number of Hydrogen molecules has a linear relationship with the number of sequential incorporations at a single location on the array making the technique prone to the same homopolymer run issue as Pyrosequencing (a 5-mer is called with greater than 97.5% accuracy) http://www.iontorrent.com/lib/images/PDFs/performance_overview_application_note_041211.pdf. The current run length is 100bp but that is set to increase to 400bp in 2012.

1.4.3.5 Polony Sequencing

The Polonator G.007 is a completely 'open' system for next generation sequencing all aspects of the software and protocols are open-source, and the chemistry is composed of 'off the shelf' items making the system very low cost. Read length is very short (13bp) so the system has never really challenged for a significant market share. The technique is similar to SOLiD in that beads are coated in clonally amplified template during an emPCR stage and then attached to a glass slide. The sequencing chemistry follows the Illumina method of incorporation a single fluorescent molecule, imaging the array, cleaving the fluorescence and repeating he process. Throughput is 1 to 2Gb per day.

1.4.4 Capture Technology

Genome sequencing identifies around 3.5 million SNPs per sample (Wheeler et al., 2008) the majority of which are in non-coding DNA. By sequencing only the coding regions of a genome sequence costs are reduced, sample throughput is increased and the data sets generated are of a more manageable size.

There are currently two main providers of capture technology which extracts fragments of DNA that contain coding sequence from fragmented genomes. The Roche Applied Science "SeqCap EZ Library" and Agilent Technologies – 'Sureselect' each work by creating biotinilated oligo 'baits' which are mixed with the fragmented DNA. The baits are designed to anneal to exomic sequence so when hybridisation occurs it is possible to isolate the fragments of DNA that have become bound and therefore sequence only the exonic sequence of a sample (Figure 1.15) (Ng et al., 2009).

1.4.5 Third Generation Sequencing

The term "Next Generation" has already become somewhat dysfunctional as a new wave of techniques are starting to come to the market, the term 'Third

Generation' is sometimes used but there is some discussion over what constitutes a Third Generation technology, whether it is those that can determine the sequence of single DNA molecules or those which do not use light for detection. Possibly the 'Generation' categories are no longer suitable as there are techniques available which are not universally considered Second or Third Generation (e.g. Life Technologies 'Ion Torrent' system).

1.4.5.1 Heliscope

The Helicos Heliscope was the first 'single molecule' instrument to be placed in research laboratories but problems soon lead to their recall. The system has many similarities to the Illumina technique, attaching a library to a glass array covered with a lawn on oligos and undertaking sequence by synthesis; however, this is a single colour system so one reversible terminator is added a time and the array is scanned to see where single incorporation events have occurred.

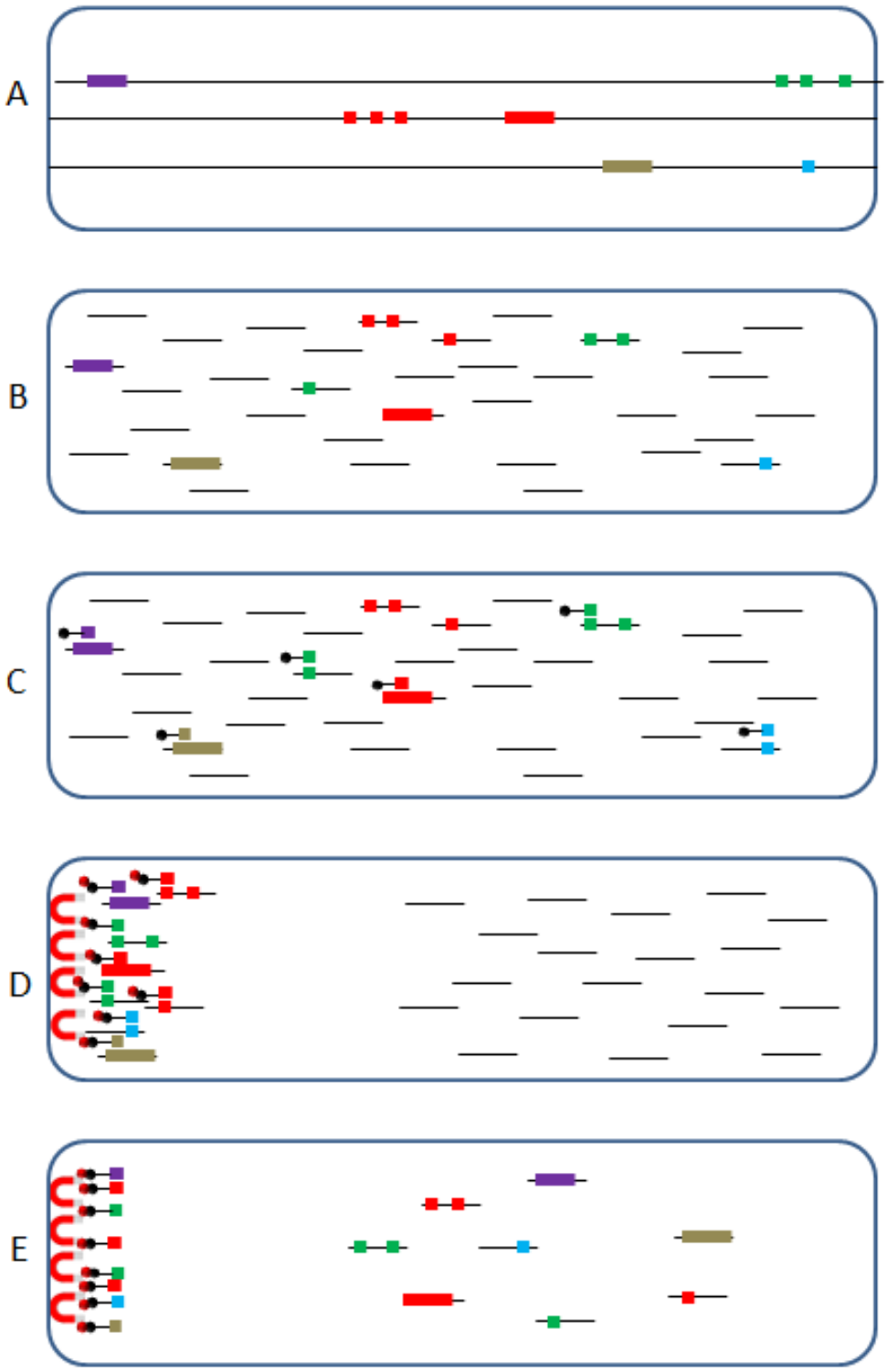


Figure 1.15: Exome Capture.

(A) A representation of genomic DNA with exons shown as colour blocks. (B) Genomic DNA is randomly fragmented. (C) Biotinilated probes that bind exonic sequence are allowed to anneal complimentary sequence. (D) The Biotinilated probes are drawn out of solution by their interaction with magnetic streptavidin coated beads. (E) Fragments of DNA are eluted, now enriched for exonic sequence.

1.4.5.2 Single Molecule Real Time (SMRT)

Pacific Bioscience SMRT technology works by immobilising a polymerase within a 'nano photonic visualisation chamber' on a microarray. While a fluorescent dNTP is being incorporated into a strand it is held in place for several milliseconds by the polymerase enzyme during which time the fluorescence is detected before being cleaved from the nucleotide by the polymerase (as part of the incorporation process) and released into solution. The technical term for a 'chamber' is Zero Mode Waveguide (ZMW) they have a glass surface at the base through which fluorescence is detected and a total volume of 20 zeptolitres. Fluorescent nucleotides that naturally diffuse into and out of the chamber move in and out in microseconds enabling the difference between noise and incorporation to be easily separated. Thousands of ZMWs are analysed simultaneously compared to hundreds of millions clusters (Illumina) or beads (SOLiD) but the rate of sequencing is 10 bases per second and read lengths can reach 1,000bp in length enabling whole human genomes to be read in a matter of minutes.

1.4.5.3 Nanopores

Oxford Nanopore Technologies (ONT) are developing two different methods both based around a nanopore which is created between a lipid bi layer (Howorka et al., 2001). In the original approach the nanopore is coupled with an enzyme which cleaves single stranded molecules. The cleaved nucleotides are drawn through the nanopore in sequence by an electrical current. Within the pore an adapter molecule transiently binds nucleotides as they pass through, when this occurs the binding disrupts the electrical current in a characteristic way enabling identification of that molecule and therefore the sequence. Their second approach may have been in response to a problem that was observed where by the cleaved nucleotides did not always pass through the pores, this second technique has a polymerase attached to the pore, a single stranded DNA molecule passes through the pore and into the polymerase, as that polymerase incorporates nucleotides the single strand is

'ratcheted' through the pore one base at a time controlling movement though the pore and allowing detection based again upon disruption to the electrical current that has been created to pass through the pores.

The method of passing a DNA molecule through a pore and detecting nucleotides at they pass through is also at the heart of several other developing third generation technologies (e.g. Bionanomatrix (Das et al., 2010) NABsys, IBM, Noblegen, Graphene Nanopres and Stratos [all unpublished]).

1.4.5.4 Others

Visigen technology is based on the transfer of energy between a donor fluorophore attached to a polymerase and different coloured fluorophores attached to each nucleotide. Transfer of light between the two molecules produces a characteristic signature which is detected and recorded. The technology promises to sequence over a million bases per second.

Halcyon Molecular are developing a technology where 'ultra-high-throughput' transmission electron microscopy (EM) will be used to literally see the order of bases in the DNA chain when it is stretched over an EM substrate. There is little available detail on the technology but they aim to sequence reads of up to 4Mb in length. ZS Genetics is also developing a method based on EM but their method focuses on determining sequence by observing DNA strands constructed from "heavy-atoms/modified dNTPs".

Complete Genomics have a proprietary technology and provide sequencing as a service. The technique involves concatenating identical fragments and causing those molecules to 'ball up' creating nanoballs of DNA which are tightly packed onto a glass substrate. A ligation based technology (Combinatorial Probe-Anchor Ligation [cPAL™]) determines the sequence.

GnuBio, Lightspeed Genomics, Mobious and QuantuMDx are also developing technology but little information in available on their approaches.

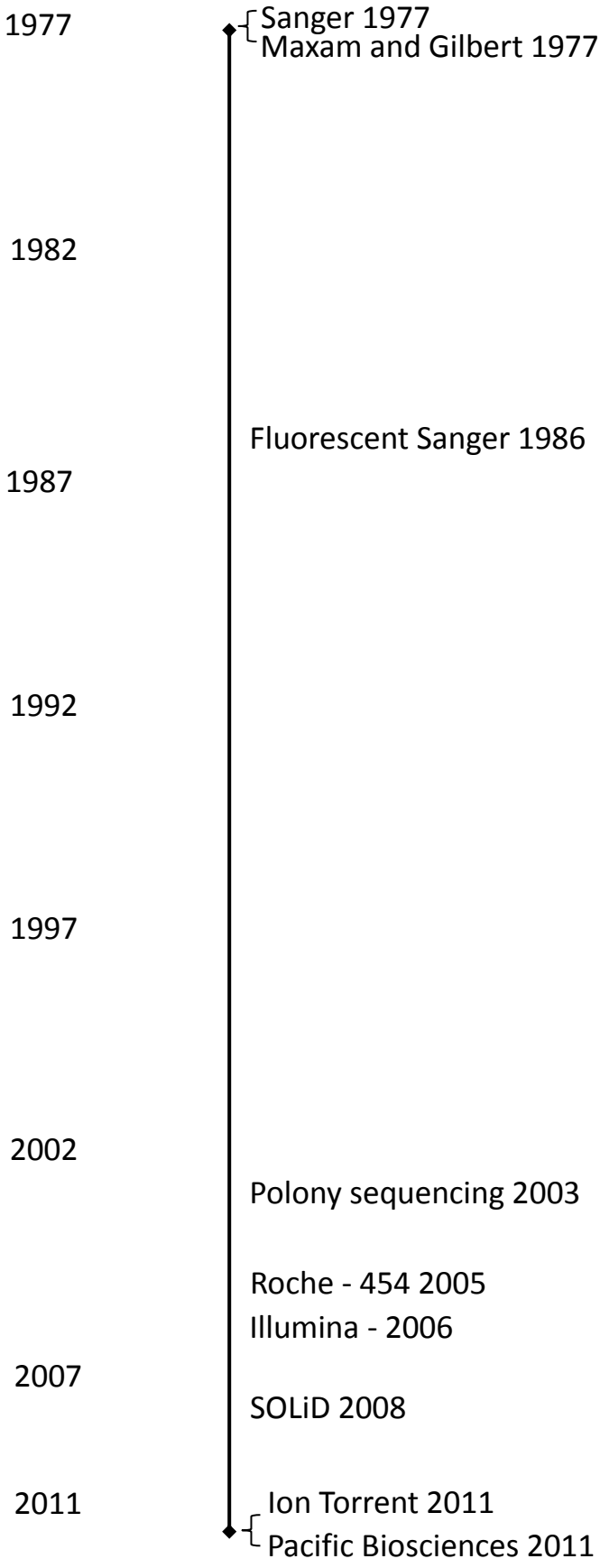


Figure 1.16: Sequencing Technology Time Line.

1.5 Databases

As sequence technology developed from slab gels to capillary arrays even relatively small research groups were able to discover large numbers of SNPs and other genetic variation as thousands of amplicons could be scanned each week. Throughout the 1990's several online resources were created in order for researchers to share that information via the internet. e.g. The Human Gene Mutation Database (HGMD) - (Cooper et al., 1998) GenBank (Benson et al., 1993) and dbSNP (Sherry et al., 2001) followed by HapMap (Couzin, 2002) 1000 genomes (2010) and UK10K (to be released) (Table 1.1).

HGMD grew out of a project to identify mutational mechanisms in human genes (Cooper and Krawczak, 1993) in 1996 it was made publically available and contained information on 11,972 instances of genetic variation that had been shown to underlie human disease. The database covers a broad range of variation types including SNPs, small and large indels, duplications, complex rearrangements, inversions and repeat expansions. Data is acquired by manual publication searches and submission from journals and the utility has grown to contain information on over 113,000 disease loci, though the full data set can now only be accessed via a subscription.

The GenBank database (Benson et al., 1993) is an annotated sequence repository (protein and DNA) set up by the National Centre for Biotechnology Information (NCBI) in the USA though it also has collaborations with the UK EMBL data Library (Stoesser et al., 1998) and the Databank of Japan (Tateno and Gojobori, 1997). Sequencing centres and authors submit data to the utility where it is annotated and made publically available. The project came from the Los Alomos Sequence Database and underwent exponential growth between 1982 and 2009. The database currently 126,551,501,141 bases of information and 62,715,288 sequence records.

dbSNP was created to supplement the GenBank sequence database by providing information exclusively for genetic variation. Despite the title, a broad

range of variation is included (e.g. SNPs, insertions, deletions and microsatellite markers). The database is open to submission from many sources; industrial and academic, individual research groups and international collaborations. It has been designed to deal with multiple submissions of the same variation by recording each submission with a submitted sequence (ss) number and attaching that information to a variation page with a Reference SNP (rs) number. An algorithm within the database uses the Basic Local Alignment Search Tool (BLAST) tool to search the surrounding sequence identifies other NCBI resources that are relevant, therein creating a vast web of interconnected data. The rs number has become accepted by most databases as the definitive identifier for a SNP in many cases databases use direct hyperlinks from their own web sites to the dbSNP rs pages.

The PharmGKB database has accumulated the information generated in research labs regarding the relationship between DNA and drug efficacy (Hewett et al., 2002), the information held draws relationships between; genes, variants, pathways, drugs and diseases. As genomics enters a phase where its relationship with pharmacology is regularly being investigated this facility will continue to provide a valuable insight into the complexities of the relationships between drugs and genotypes.

Table 1.1: On-line Databases.

Database	Content
1000 genome	All genetic variation with a 1% minor allele frequency in a pre-determined population.
dbSNP	A DNA sequence variation repository.
Encode (UCSC)	A comprehensive parts list of functional elements in the human genome.
Ensemble	Genome databases for vertebrates.
GenBank	A DNA sequence repository.
GWAS central (formerly HGVbase)	A centralized compilation of summary level findings from genetic association studies.
HapMap	Compares haplotypes between defined individuals.
HGMD	A list of disease causing mutation.
JSNP	SNPs from the Japanese population.
OMIM	A comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes.
PharmGKB	Knowledge about the impact of human genetic variations on drug response.
SNP500Cancer (NGS related)	Genomes for 102 reference samples.
SRA	Raw sequencing data from NGS platforms.
UK10K	The genomes of 10,000 UK samples with and without disease.

1.6 Colorectal Cancer

One third of colorectal cancers are thought to be caused by inherited factors (Kinzler and Vogelstein, 1996) but less than 6% of cases carry known high-penetrance mutations (Aaltonen et al., 2007) such as those found in the *APC* gene which cause Familial Adenomatous Polyposis Coli (FAP; MIM 175100) (Fearnhead et al., 2001) several mismatch repair (MMR) genes (*MLH1*, *MSH2*, *MSH6* and *PMS2*) which cause hereditary non-polyposis CRC (HNPCC; MIM 114500) (Peltomaki, 2001) and the BER gene MUTYH which causes *MUTYH* associated polyposis (MAP; MIM 608456) (Al-Tassan et al., 2002, Sampson et al., 2003).

Current approaches to understanding the remaining proportion of inherited CRC have focused on Genome Wide Association Studies (GWAS) looking at thousands of cases and controls and identifying low-penetrance SNPs that may confer increased susceptibility.

1.6.1 FAP (MIM 175100)

FAP is an autosomal dominant disorder with a population incidence of 1:8,000. It is caused by mutations in the *APC* (Chromosome 5) gene which are the first stage in the progression to malignancy for FAP patients and are also critical in sporadic CRC occurring in almost 85% of all colorectal tumours (Hisamuddin and Yang, 2004). FAP is characterised by the development of hundreds to thousands of pre-malignant, benign adenomatous polyps in the epithelium of the large colon and the rectum from the third decade of life onwards (Kinzler and Vogelstein, 1996). Numerous extra colonic features may also develop such as congenital hypertrophy of the retinal pigment epithelium (CHRPE) desmoids and upper gastro intestinal (GI) tract tumours (especially following surgery), Gardner's syndrome and Turcot syndrome. Without clinical intervention and given the number of adenomas that occur a few will inevitably progress to become tumours.

1.6.1.1 The APC gene

APC is expressed in a variety of different human tissue but only appears to have a 'gatekeeper' function in the gut, other tissues expressing a mutant allele will not necessarily produce adenomas. *APC* is known to function in numerous pathways the most important of which (in relation to CRC) is to act as an antagonist to Wnt signalling preventing the inappropriate proliferation of cells in the colon by suppressing the activity of the β-catenin oncogene.

Normal β-catenin regulated transcription occurs when the Wnt peptide stimulates the cell transmembrane receptor 'LRP' to recruit Frizzled, Discheveled, Axin and GSK- β3 to form a membrane bound complex that dephosphorylates β-catenin. This prevents APC recognising and degrading the protein as it would phosphorylated β-catenin. Dephosphorylated β-catenin builds up in the cytoplasm and eventually enters the nucleus to drive appropriate transcription of factors such as the oncogene c-*myc* and cyclin D1. APC's normal roll in the Wnt signalling pathway is to degrade β-catenin when the Wnt signalling pathway is not activated and therefore prevent proliferation in the absence of the Wnt peptide.

Mutation of the *APC* gene leads to aberrant regulation of β-catenin and consequently constituently active transcription of β-catenin regulated factors leading do dysplasia of epithelium cells. Damaging mutations tend to occur in the Mutation Cluster Region (MCR) towards the centre of the gene and most often causing truncation of the coded protein via nonsense mutation or small insertions/deletions (Indels) (though Loss if Heterozygosity (LOH), whole exon deletion and splicing errors are also known(Aretz et al., 2004)). Centrally located Non-synonymous changes (such as p.I1307K) have been shown to lead to the disease with non-synonymous SNPs towards the 3' and 5' ends of the gene tending to lead to Attenuated FAP (AFAP) where the adenomas observed in the bowel are less florid and less frequent than when mutations are found towards the centre of the gene.

Germline mutation in *APC* leads to FAP being inherited in an autosomal dominant fashion with the most common mutations at residues 1061 and 1309, (within the MCR). Mutations downstream of residue 1600 are extremely rare.

Predisposition to CRC is conferred by a single inherited germline mutation in *APC*, for the disease to progress a second mutation is required to alter the inherited wild type copy of the gene (Knudson's "2 hit hypothesis" (Knudson, 1971)), with no fully functioning *APC* protein produced by a cell, the suppressor function is removed and normal epithelium become dysplasic, the first step towards malignancy.

Somatic mutation in *APC* is the more common cause of CRC, as sporadic mutations affect the epithelium later in life the number of adenomas resulting from sporadic mutation tend not as high as seen in FAP patients.

1.6.2 Hereditary nonpolyposis colorectal cancer (HNPCC; MIM120435)

HNPCC is a heterogeneous disorder that is characterised by increased risk of Colorectal (and other GI tract), brain, skin, ovarian and endometrial cancers and is diagnosed using the revised 'Amsterdam criteria' (AC II) (Vasen et al., 1999). It is not known whether the number of colonic polyps in these patients is greater than in the general population but the development of a malignant adenoma in an HNPCC patient is considerably faster taking 1 to 2 years instead of 10 or more (Knowles and Selby, 2005). The disease occurs where there is a breakdown in the Mis-Match Repair (MMR) pathway due to a mutation in one of the highly conserved MMR genes.

1.6.2.1 MisMatch Repair (MMR)

MMR (specifically short patch MMR) works by excising and replacing only the incorrect base from a mismatch, this is where the process differs from BER. The two strands of DNA can be referred to as 'parental' and 'daughter' where the daughter strand is the one that is produced during replication from the 'parental' template,

during DNA replication mismatches are introduced by the polymerase, most are dealt with by the 'proof reading' but a few may escape this corrective process. Via an unknown mechanism MMR recognises mismatches and critically excises the incorrect nucleotide(s) exclusively from the daughter strand replacing it with a correct match according to the parental strand therefore maintaining the integrity of the sequence, significantly in HNPCC MMR also undertakes the correction of mismatches that are introduced by mutation events when DNA is not replicating.

In the normal MMR pathway *MSH*2 and *MSH*6 combine to form a heterodimer in the presence of ATP, this molecule is described as a 'sliding clamp' which can travel in either direction along the double strand until it locates a mismatch at which point it combines with a *MLH*1/*PMS*2 heterodimer (again utilising ATP) and stimulates exonuclease activity of the daughter strand, short patch MMR results in the replacement of approximately 10 nucleotides, long patch can replace a few kilobases. There are few physical differences between a daughter and parent strand Stojic ((Stojic et al., 2004)) suggests that the Okazaki fragments are the basis for discrimination; small sections of single stranded DNA correctly paired with their parental template but with gaps in the phosphate backbone.

Because HNPCC mutations affect a repair pathway (as opposed to cell cycle) adenomas are only formed when chance mutation occurs in a suitable location such as a tumour suppressor or oncogene hence the number of adenomas observed (as with *MYH* associated polyposis) is considerably lower than in FAP however the ratio of progression of adenoma to carcinoma in HNPCC patients is very high (1:1) as opposed to FAP patients (30:1) (Jass et al., 2002) probably due to the increased rate of somatic mutations that will occur in a cell with germline MMR mutation.

1.6.2.2 Mutations in HNPCC Genes

Mutations in the MMR Glycosylase genes MLH1 and MSH2 account for 90% of all cases but mutations in *MLH*3, *MSH*6, *PMS*, *PMS*2 and *TFGBR*2 can also cause the disease by preventing the MMR. Unlike *APC* there are no known mutation 'hot

spots' and the position of the mutation in not known to affect the phenotype. Missense, nonsense, small insertions and deletions are all known to cause the disease (Papadopoulos and Lindblom, 1997).

An important feature of the mutations in HNPCC genes is that they lead to Microsatellite instability (MI) in tumour tissue whereby small repetitive sequences (most often di-nucleotide CA repeats) are no longer reliably copied during DNA replication and the tracts can expand or contract. Expansions in oncogenes or TGSs lead to frame shift mutations and eventually adenomas.

1.6.3 MYH Polyposis (MIM 608456)

Multiple Ademoa Polyposis (MAP) (also known as *MUTYH* Associated Polyposis or FAP,2) is caused by mutations in the *MUTYH* gene (Sampson et al., 2005) and in contrast to FAP, AFAP and HNPCC, it is an autosomal recessive disorder. The clinico-pathological features are very similar to FAP but the disease tends to be less severe (Sieber et al., 2003b). Carcinoma is more common that in HNPCC but progression from the adenoma stage is slower.

The *MUTYH* gene was originally identified as playing a significant role in CRC in by looking at three British Siblings who were negative for APC mutation and showed no signs on MI in their tumour tissue. A higher than expected number of G:C>T:A transversions were noticed in the *APC* gene from tumour tissue which suggested a Base Excision Repair (BER) gene may be involved. The human homologues of the *E.Coli* oxidative damage genes (including MYH as the homologue to MutY) were sequenced in the siblings who were found to be compound heterozygous for *MUTYH* mutations (Al-Tassan et al., 2002).

1.6.3.1 Base Excision Repair (BER)

MUTYH is involved in the BER process of the DNA repair pathway. Reactive Oxygen Species (ROS) such as free radicals (oxygen molecules with a single unpaired

electron in an outer shell) are generated by aerobic metabolism, these molecules can damage guanine nucleotides so that they form a base pair 'mismatch' with adenine instead of cytosine. During DNA replication incorporation errors will become permanent mutations and left unchecked the accumulation of mutations in a cell's DNA can lead to aberrant growth and eventually the development of cancer.

Within the BER pathway the function of MUTYH is to remove the normal adenine base that incorporates opposite the damaged guanine before replication perpetuates the error. Glycosylases 'tag' affected guanine nucleotides by breaking the β-N glycosidic bond, separating the base from the deoxyribose backbone, the break is then recognised by an exonuclease which cuts the affected or 'abasic' site at the 5' phosphodiester and leaves a free 3'-OH, this site is recognised by DNA polymerase (POL I) which replaces the nucleotide by virtue of its inherent exonuclease activity. DNA Ligase completes the process by connecting the newly added nucleotide to the nucleotide upstream. Where mutations in MUTYH lead to the accumulation of mutations in the APC gene, FAP like pathology is observed known as MYH associated polyposis.

Naturally other BER glycosylases such as *OGG1*, *MTH1*, *NTH1* are considered candidates to explain at least some of the inherited CRC not explained by the known high penetrance mutations.

1.6.3.2 *MUTYH* Mutations

The first two mutation identified by Al-Tassan (2002), (p.Y176C and p.G393D, originally described as p.Y165C and p.G382D respectively) are conserved residues in the *E.Coli* homologue MutY and their mutation in known to lead to a large reduction in the Adenine removal stage of the BER. Nonsense mutation and small deletions have also been shown to cause the disease as well as individuals who are homozygous for a mutant such as p.Y165C (Jones et al., 2002, Sieber et al., 2003b).

1.7 Aim of the project

1. To assess the specificity of the Transgenomic Navigator software for the identification of genomic variation following dHPLC analysis of DNA.
2. To apply this approach to analysis of the base excision repair genes, *OGG1*, *NEIL1*, *NEIL2*, *NEIL3*, *MTH1* and *NTH1* in patients with a multiple colorectal adenoma phenotype.
3. To assess the sensitivity and specificity of software for mutation calling of automated Sanger sequencing in comparison to manual inspection.
4. To investigate whether rare missense variants in the β-catenin down-regulating domain of *APC* are associated with the phenotype of multiple colorectal adenomas.
5. To investigate SNP representation in build 129 of the dbSNP database
6. To profile 480 reference human cell lines in ECCAC for non-synonymous SNPs in 134 human DNA repair genes and to use *in silico* approaches to investigate their potential functional significance.
7. To investigate non-synonymous SNPs in DNA repair genes as potential determinants of CRC predisposition and safety and efficacy of therapy.
8. To investigate genetic determinants of severe peripheral neuropathy after oxaliplatin-based chemotherapy using exome re-sequencing.

2 Chapter two

Materials and Methods.

2.1 Commercial Materials and Suppliers

Item	Supplier
10X Capillary Electrophoresis Running Buffer	Sigma
10X PCR Buffer	Applied Biosystems
100bp DNA Ladder	NEB
3100 Injection Solution	GRI
5X Sequencing Buffer	Applied Biosystems
Agarose MP	Roche
Ampli Taq Gold	Applied Biosystems
AmpliTaq Gold 10 X reaction Buffer	Applied Biosystems
BigDye Terminators	Applied Biosystems
BigDye Terminator 5X Reaction Buffer	Applied Biosystems
dHPLC Buffers (A, B, C and D)	Transgenomic
dNTP	Amersham (GE HealthCare)
Ethidium Bromide	Sigma (Fluka)
Ethylenediaminetetraacetic acid	Sigma
Exonuclease I	New England Biolabs
Gene Amp 10X Buffer	ABI
Mutation Standards	Transgenomic
Oligonucleotides	MWG

Orange G	Sigma
POP6 3100 Polymer	Applied Biosystems
Restriction Enzymes	New England Biolabs
Shrimp Alkaline Phosphatase	Amersham (GE HealthCare)
Sterile Water	Fresenius Kabi
Sucrose	Sigma
Tris	USB

2.1.1 Kits

Item	Supplier
Nimble Gen Exome capture kit	Roche
Illumina Sequencing kits	Illumina
DNAeasy Blood and Tissue Kit	Qiagen

2.2 Solutions

Gel Loading Buffer - Orange G (trace), 40% Sucrose

TAE buffer - 40mM Tris-acetate, 1mM EDTA, pH to 8.0

2.3 Equipment

Item	Supplier
96 well plate Vacuum Manifold	Millipore
ABI 3100 Genetic Analyser	Applied Biosystems
Agarose Gel Electrophoresis tanks	Abgene
Beckman Biomek FX	Beckman Coulter
Covaris S series	Covaris

Genome Analyser II	Illumina
CFX965^{TM} Real-Time PCR Detection System	Bio Rad
MJ Research DNA Engine Tetrad PTC-225	GRI/MJ
Power Supplies	Bio Rad
Vacuum Pump	Millipore
Wave HT3500	Transgenomic

2.4 Methods

2.4.1 Primer Oligonucleotide Primer Design

All primers were designed manually. The following criteria were applied where possible:

1 – Primers were designed with a 50% G:C content

2 – An annealing temperature of 60^0C was targeted with pairs not differing by more than 3 degrees

2 – The 3' end would have a high GC content

3 – Repetitive and palindromic sequences would be avoided

4 – Between primer pairs complimentary should be avoided particularly at the 3' end.

2.4.2 Polymerase Chain Reaction (PCR)

Thermal cycling conditions consisted of an initial denaturation at 94°C for 10 minutes, followed by 35 cycles of 94°C for 30 seconds, 30 seconds at the annealing temperature and 72°C for 30 seconds, followed by a final elongation stage of 72°C for 10 minutes. Reactions were carried out on MJ Tetrads.

Each 20ul reaction contained; 0.2U AmpliTaq gold (ABI), 0.2mM dNTPs (Amersham), 25pmol of each primer (MWG), 10X reaction Buffer (ABI) and 20ng of DNA.

2.4.3 PCR Purification

PCR products were purified using ExoSap, in which 10 to 15µl of PCR product was incubated with 2U of exonuclease I (New England Biolabs) and 2U of shrimp alkaline phosphatase (GE Healthcare). The reaction mixture was incubated at 37°C for one hour, followed by denaturation at 80°C for 15 minutes.

2.4.4 Sanger Sequencing

PCR products were sequenced using the BigDye™ v3.1 Sequencing kit (Applied Biosystems). Thermal cycling parameters consisted of 24 cycles of 94°C for 10 second, 50°C for 5 seconds and 60°C for 3 minutes and 30 seconds.

Each reaction consisted of 2ul of PCR template 2.5pmol of sequencing primer and 1ul BigDye reaction mix and 5X Sequence Reaction Buffer. Cleaned sequence reactions were run on ABI 3100 or 3730 genetic analysers (Applied Biosystems).

2.4.5 Sequencing Clean Up

Purification of sequencing products was carried out using the Millipore Montage SEQ96 Sequencing Reaction Cleanup kit under the recommended conditions.

2.4.6 Agarose Gel Electrophoresis

2% Agarose gels were run in 1xTAE buffer. Approximately 0.05ug/ml of Ethidium Bromide was used as an intercalating dye to allow visualisation of PCR products under UV exposure. 2 to 10ul of product was mixed with 5ul of Gel loading Buffer.

Electrophoreses was carried out between 80 and 120 volts in a BioRad Gel tank. 100bp ladder (NEB) was added to each gel to determine fragment sizes.

2.4.7 Denaturing High Performance Liquid Chromatography (dHPLC)

dHPLC (Section 1.2.2.3) was carried out using the 3500HT WAVE nucleic acid fragment analysis system (Transgenomic).

To allow the detection of samples harbouring homozygous variants, 5µl of wild type PCR product was added to 20µl of sample product, denatured at 95ºC then cooled to 50ºC at a rate of 0.5ºC per 30 seconds. dHPLC was carried out using the melting temperatures predicted using the Transgenomic Navigator software (Version 1.5.4, Build 23). A 10 - 12% ACN gradient was run over 2 – 2.5 minutes**.**

2.4.8 Restriction Digestion Assay

20ul of PCR product was digested under the recommended conditions (NEB) for each enzyme (usually 1U of enzyme with a 2 hour, 37^0C incubation in an appropriate digestion buffer). Digested products were analysed by size separation on an agarose gel (Section 2.4.6).

2.4.9 Amplification Refractory Mutation System (ARMS)

Allele specific ARMS primers were designed with a mismatch at the n-3 base. PCR reactions were set up as described in Section 2.4.3. Optimum amplification temperature was determined by using a temperature gradient on the MJ tetrad.

Each ARMS assay contained a second set of PCR primers to act as a control. These primers would amplify a PCR fragment with a significant size different to the assay fragments but have the same annealing temperature.

2.5 Suppliers

ABgene, Surrey, UK.

Amersham Bioscience, Amersham, Aylesbury, Buckinghamshire, UK.

Applied Biosystems, Cheshire, UK.

Bio-Rad, Watford, Hertfordshire, UK.

Covaris, Inc. Woburn, Massachusetts, USA.

GRI, Braintree Essex/ MJ Research, Massachusetts, USA.

Illumina, Chesterford Research Park, Little Chesterford, Essex, UK.

Millipore, Hertfordshire, UK.

MWG-Biotech, Buckinghamshire, UK.

Myriad Genetics, Salt Lake City, USA.

New England Biolabs, Ipswich MA, USA.

Roche Biochemicals, East Sussex, UK.

Sigma, Chemical Company Ltd, Poole, Dorset, UK.

Transgenomic, Cheshire, UK.

United States Biochemical (USB), Cambridge Bioscience, UK.

2.6 Websites

1000Genomes	http://www.1000genomes.org/
Align-GVGD	http://agvgd.iarc.fr/
CGAP	http://cgap.nci.nih.gov/
DAVID	http://david.abcc.ncifcrf.gov/home.jsp
dbSNP	http://www.ncbi.nlm.nih.gov/snp
Ensembl	http://www.ensembl.org/index.html
GeneSNPs	http://www.genome.utah.edu/genesnps/

HapMap	http://hapmap.ncbi.nlm.nih.gov/
homologene	ftp://ftp.ncbi.nlm.nih.gov/pub/HomoloGene/current
HWSIM software	http://krunch.med.yale.edu/hwsim/
Koask prediction	http://www.cbs.dtu.dk/services/NetStart
MutationDiscovery.com	http://mutationdiscovery.com/md/MD.com/home_page.jsp
NCBI Blast	http://blast.ncbi.nlm.nih.gov/Blast.cgi
Polyphen	http://genetics.bwh.harvard.edu/pph/
Pupasuite	http://pupasuite.bioinfo.cipf.es/
SNPper	http://snpper.chip.org/
SNPHunter	http://www.hsph.harvard.edu/ppg/software.htm
t-coffee	http://www.igs.cnrs-mrs.fr/Tcoffee/tcoffee_cgi/

3 Chapter three

Assessment of the Transgenomic Navigator™ software to rapidly detect aberrant dHPLC elution profiles

3.1 Introduction

Mutation detection is a key tool in both clinical diagnostic settings and research environments. A number of techniques have been developed including single-stranded conformational polymorphism analysis, denaturing gradient gel electrophoresis, heteroduplex analysis, chemical or enzymatic cleavage, denaturing dHPLC and DNA sequencing [reviewed (Cotton et al., 1998), discussed in Section 1.2]. Despite the availability of numerous technologies, only a few of these methods have been formatted to facilitate automated mutation calling and, to-date, most attention has focused on automated heterozygote detection by DNA sequence analysis [(Hattori et al., 1993, Nickerson et al., 1997, Versluis et al., 1993)]. dHPLC is a highly sensitive technique, that resolves homo- and heteroduplex DNA molecules by using ion-pair reverse phase HPLC under conditions of partial helix denaturation [reviewed in (Xiao and Oefner, 2001), discussed in Section 1.2.2.3]. Several investigators have shown that samples carrying sequence variants often generate characteristic dHPLC elution profiles which may be used to facilitate mutation detection [(O'Donovan et al., 1998, Young et al., 2002)]. Recently, Transgenomic Inc. (www.transgenomic.com) has developed software for automated dHPLC-based mutation calling using customisable pattern recognition capabilities. Here, we evaluate the utility of the Transgenomic Navigator software to facilitate automated detection of aberrant dHPLC elution profiles after analysing amplicons from *MSH6*, *NEIL2*, *NEIL3* and *OGG1* in 172 patients with multiple colorectal adenomas with or without cancer. These genes were selected from on-going projects and were selected as they were known to contain polymorphisms and rare variants. DNA fragments harbouring both a polymorphism and a mutation often generate complex elution profiles that may hinder mutation detection. Therefore, we also assessed the

utility of the software to rapidly detect samples carrying novel sequence variants in addition to common polymorphisms

Data in this chapter was presented in the manuscript “Rapid recognition of aberrant dHPLC elution profiles using the Transgenomic Navigator software” (Colley et al., 2005) of which James Colley was the primary author and as a consequence there are some overlaps in parts of the text.

3.2 Materials and Methods

3.2.1 Samples

DNA was extracted from the venous blood samples of 172 patients with multiple colorectal adenomas with or without cancer.

3.2.2 PCR

We analysed fragments that were known to be polymorhpic spanning exons 2, 3, 4 (fragments 1-4, 6, 8, 11), 5 (fragments 1-2), 6 and 9 of *MSH6* (NM_000179.); exons 1, 2 (fragment 2), 3 and 4 (fragment 1) of *NEIL2* (NM_001135746); exons 1, 2, 5, 8 (fragment 1) and 10 of *NEIL3* (NM_018248); and, exons 5 and 7 of *OGG1* (NM_002542).

Reaction and thermal cycling conditions are described in Section 2.4.2, primers and annealing temperatures are in Appendix A. Fragments were designed to span exons with 50bp excess on either side (to account for the poor sequence often seen at the start of sequence traces), where these fragments would exceeded 600bp multiple overlapping fragments were designed.

3.2.3 Sequencing

Where Sanger sequencing was required PCR products were purified as described in Section 2.4.3 and sequenced as described in Section 2.4.4.

Purification of sequencing products was carried as described in Section 2.4.5. Samples were run on an ABI 3100 genetic analyser (Applied Biosystems) and sequence data viewed using Sequencher v4.6.

3.2.4 Denaturing High Performance Liquid Chromatography (dHPLC) and Analysis with the Navigator Software

dHPLC was carried out as described in Section 2.4.7 at the melting temperatures predicted by Navigator (Version 1.54) software (Appendix C).

3.2.5 Assays for Common Polymorphisms and Automated Sequencing

Common polymorphisms were typed using ARMS (Section 2.4.9) or restriction digestion based assays (Section 2.4.7) (Assay details in Appendix B).

3.2.6 Navigator Analysis

dHPLC elution data was imported into the analysis page of the Navigator software and samples with weak amplification profiles were excluded. Three levels of analysis were performed: Level 1 – elution profiles were automatically normalised and grouping information was collected from the affinity tree. Level 2 – whilst observing the overlaid profiles and 3D plot, a slide tool on the left axis of the 3D scatter graph was manually adjusted to alter the grouping parameters (profiles lying immediately adjacent to a major group were incorporated into that group) grouping information was collected from the affinity tree. Level 3 – the user re-assessed the outlying profiles from Level 2, and, if necessary, manually re-assigned them using the 'discrimination tool' (Figure 3.2).

3.2.7 Author's Contribution

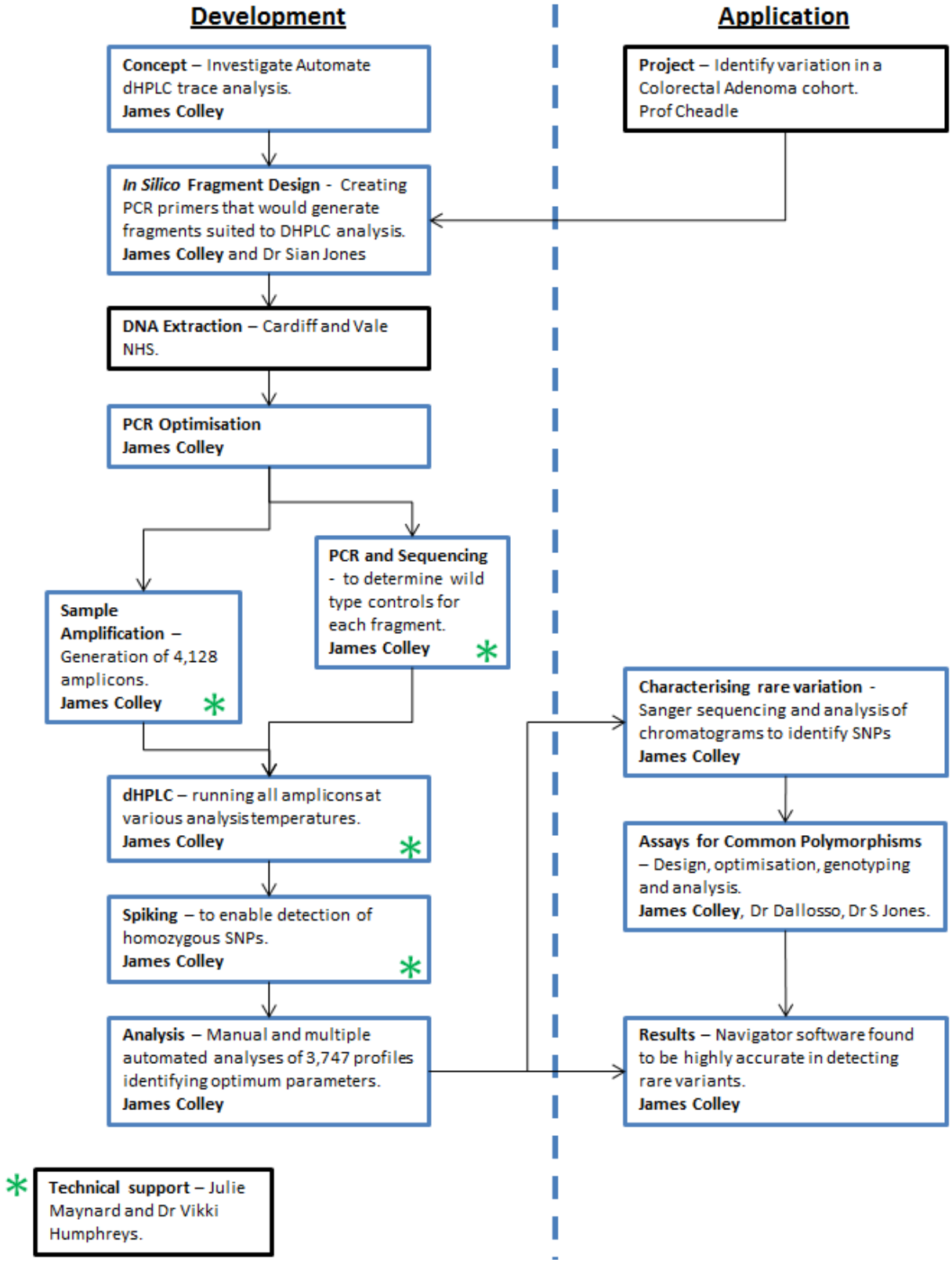


Figure 3.1: Author's contribution to Chapter 3.

Table 3.1: dHPLC Analysis Temperatures.

Amplicon	dHPLC Temp 1(ºC)	dHPLC Temp 2(ºC)
MSH6X2newF	57	60
MSH6X3F	58	59.5
MSH6X4.1	57.5	59
MSH6X4.2	59.5	/
MSH6X4.3	56	58.5
MSH6X4.4	57.5	59
MSH6X4.6	56	58
MSH6X4.8	56.5	58
MSH6X4.11	57	58.5
MSH6X5.1	57	58.5
MSH6X5.2	54.5	58
MSH6X6	57	58.5
MSH6X9	55	57

Amplicon	dHPLC Temp 1(ºC)	dHPLC Temp 2(ºC)
NEIL2X1new	61	62
NEIL2X2.2	62	/
NEIL2X3new	59	60.5
NEIL2X4.1	59	62.5
NEIL3X1	62.5	64.5
NEIL3X2	57	/
NEIL3X5	57	/
NEIL3X8.1	53	55.5
NEIL3X10	55.5	56.5
OGG1X5	63	/
OGG1X7	62	/

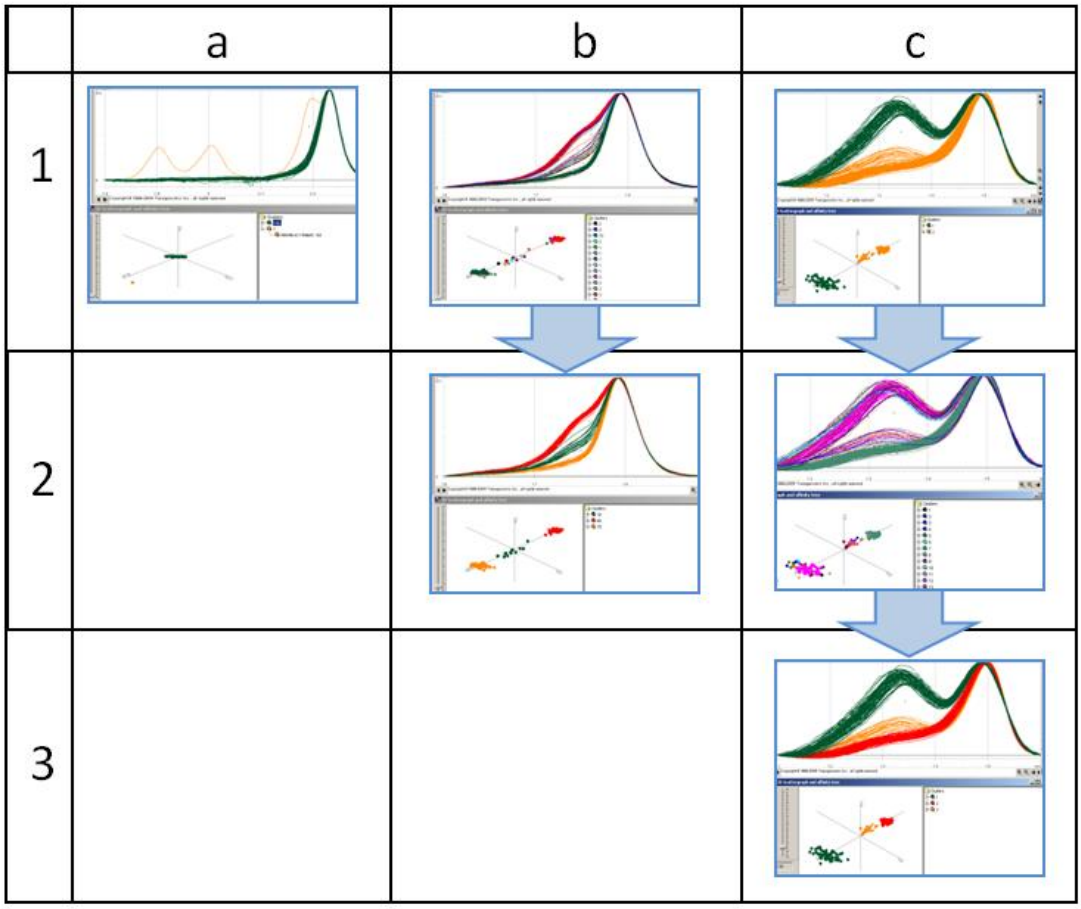


Figure 3.2: Three levels of Analysis.

Columns a, b, and c show the analyses of different fragments, rows 1, 2, and 3 show those fragments at different levels of analysis. **1**) *MSH6* exon 4 fragment 4 analysed at 57.5ºC (sample 76 is heterozygous for p.Y214 (642C>T). The fully automated Level 1 analysis (**1a**) provides sufficient separation between the wild type and variant traces; **(2)** *MSH6* exon 3 analysed at 58ºC (common SNP p.D180 (540T>C). Level 1 analysis (**1b**) has obvious errors, adjusting the slide tool (Level 2) gives correct separation (**2b**); **(3)** *MSH6* exon 3 analysed at 59.5ºC (common SNP p.D180 (540T>C). Levels 1 and 2 (**1c** and **2c** respectively) have obvious errors manually altering the automatic data points Level 3 analysis (**3c**) separates the three trace types.

3.3 Results

3.3.1 Evaluation of the Navigator Software to Detect Rare Variants

We sought germline variants in *MSH6* (exon 4 fragments 2-4, 6, 8, 11; exon 5, fragment 1; exon 6; and exon 9), *NEIL2* (exon 1; and exon 2, fragment 2) and *NEIL3* (exon 2; and exon 10) in 172 patients with multiple colorectal adenoma with or without carcinoma. In total, thirteen different fragments were amplified and all except three were analysed at two dHPLC temperatures (Table 3.1), thereby generating twenty-three data sets (consisting of 3,747 elution profiles) for analysis with the Navigator software (Table Table 3.2). Under Level 1 analysis, 138 products were identified as having profiles outlying the wild type group. Direct sequence analysis showed that 60 (43.5%) of these products harboured novel variants (heterozygous and homozygous) and the remaining 78 were wild type. One of the analysis conditions (for exon 9 of *MSH6* at 57ºC) resulted in poorly resolved elution profiles in which only one of the sixty outlying product profiles was confirmed as carrying a genuine variant. Repeat dHPLC analysis on independent PCR products from this amplicon again yielded poorly resolved elution profiles with a high proportion of outlying samples (under Level 1 analysis) that were shown to be wild type upon sequencing. Excluding this result, 59/78 (75.6%) products with outlying profiles were confirmed as harbouring novel variants under Level 1 analysis. Repeat dHPLC analysis of 16 of the remaining 19 PCR products (3 samples had insufficient DNA for repeat amplification) characterised as having outlying profiles but no sequence variant, showed that all were subsequently classified correctly to the wild type group under Level 1.

The results from Level 2 and 3 analyses were identical to each other, with 60 products identified as having outlying profiles, 59 of which (98.3%) were shown by sequence analysis to carry genuine sequence variants (examples are shown in Figure 3.4 and 3.6). Only a single product (from exon 4 fragment 3 of MSH6 analysed at 56ºC) was incorrectly classified, displaying an outlying trace but proven wild type on sequence analysis. This profile lay close to the wild type profile and was correctly identified as wild type under repeat dHPLC analysis of a fresh PCR product and at the alternative dHPLC conditions using the original product.

All 3,747 dHPLC elution profiles were also analysed by visual inspection, by an independent investigator, to allow for a direct comparison to the results generated using the software. In total, 58 products with outlying profiles were identified, all of which were shown to harbour novel variants upon sequence analysis. In comparison to the Level 3 analyses using the Navigator software, one product with a novel variant (p.H388 [c.1164C>T] in exon 4, fragment 3 of *MSH6* at 56ºC) was missed by visual inspection of the traces because the elution profile was very similar in shape to a sample that was heterozygous for p.Y397F only; this product was identified as having a unique profile by manual analysis at the second temperature used for dHPLC (Figure 3.3).

In contrast, the single product identified as an outlier under Level 3 analysis but shown to be wild type upon sequencing was not misclassified by manual inspection. We assessed the level of false negatives identified using the Navigator software by comparing dHPLC classifications and sequencing results for four rare variants (c.157-18A>G, c.278+50G>A and p.A547S [c.1639G>T] in *NEIL3* and p.G289E [c.866G>A + c.867C>A] in *MSH6*) that were identified during this study. Ten out of 113, 1/115, 1/162 and 1/158 samples were correctly classified as carrying the variants c.157-18A>G, c.278+50G>A, p.A547S (c.1639G>T) and p.G289E (c.866G>A + c.867C>A), respectively – no false negatives were identified at any of the Levels of analysis.

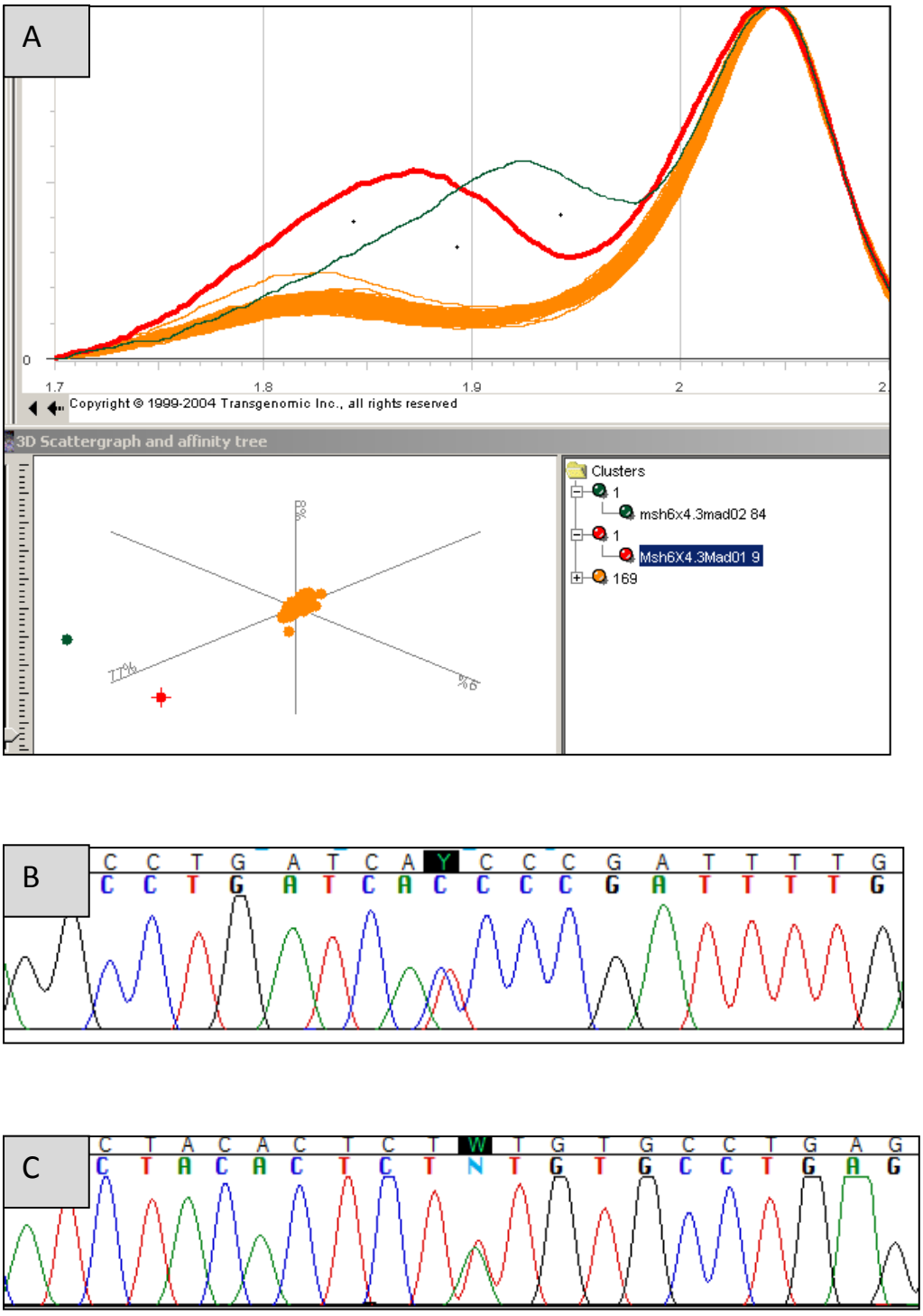


Figure 3.3: MSH6 exon4, fragment 3.

(A) *MSH6* exon 4, fragment 3 Navigator plot. Manual analysis incorrectly identified the red and green trace as being the same variant. Images B and C show the two rare SNPs that cause the dHPLC trace profiles. (B) *MSH6* exon 4, fragment 3, p.H388. (C) *MSH6* exon 4, fragment 3, p.Y397F.

Table 3.2: Evaluation of the Navigator Software to Detect Rare Variants.

Gene	Exon	dHPLC Temp (^{o}C)	No. of products analysed	Level 1	Level 2	Level 3
MSH6	4(Fr.2)	59.5	161	2/2	2/2	2[a]/2
MSH6	4(Fr.3)	56	170	2/3	2/3	2[b]/3
MSH6	4(Fr.3)	58.5	171	2/3	2/2	2[b]/3
MSH6	4(Fr.4)	57.5	169	1/2	1/1	1[c]/1
MSH6	4(Fr.4)	59	169	1/2	1/1	1[c]/1
MSH6	4(Fr.6)	56	172	1/3	1/1	1[d]/1
MSH6	4(Fr.6)	58	171	1/3	1/1	1[d]/1
MSH6	4(Fr.8)	56.5	165	1/2	1/1	1[e]/1
MSH6	4(Fr.8)	58	169	1/1	1/1	1[e]/1
MSH6	4(Fr.11)	57	149	1/1	1/1	1[f]/1
MSH6	4(Fr.11)	58.5	156	1/1	1/1	1[f]/1
MSH6	5(Fr.1)	57	169	1/1	1/1	1[g]/1
MSH6	5(Fr.1)	58.5	169	1/1	1/1	1[g]/1
MSH6	6	57	146	3/3	3/3	3[h]/3
MSH6	6	58.5	149	3/3	3/3	3[h]/3
MSH6	9	55	168	1/1	1/1	1[i]/1
MSH6	9	57	170	1#/60	0/0	0/0
NEIL2	1	61	154	3/3	3/3	3[j]/3
NEIL2	1	62	159	3/4	3/3	3j/3
NEIL2	(2Fr.2)	62	159	1/6	1/1	1[k]/1
NEIL3	2	57	162	13/13	13/13	13[l]/13
NEIL3	10	55.5	162	8/12	8/8	8[m]/8
NEIL3	10	56.5	158	8/8	8/8	8[m]/8
Total			3747	60/138 (43.5%)	59/60 (98.3%)	59/60 (98.3%)

Three levels of analysis (Levels 1-3) were undertaken for all 3,747 products and results are presented as the number of products that contained a novel variant (shown by sequence analysis) out of the total number of products with 'outlier' profiles. Seventy eight outlier products identified under Level 1 analysis and one outlier product under Level 2 and 3 analyses were shown to be wild type (none of the twelve sub-options assessed under Levels 1-3 affected the results of this analysis).

Legend continued on following page.

[#]The variant 4001+12_15del was not detectable by dHPLC at 57ºC (identified by sequence analysis). Novel variants identified were: [a]p.K295R (c.884A>G), p.G289E (c.866G>A + c.867C>A); [b]p.Y397F (c.1190A>T), p.H388 (c.1164C>T); [c]p.S503C (c.1508C>G); [d]p.A630 (c.1890A>C); [e]p.L758 (c.2272C>T); [f]p.R1005X (c.3013C>T); [g]p.T1102 (c.3306T>A); [h]c.3439-16C>T (2 cases), c.3439-57_58del; [i]c.4001+12_15del; [j]p.Q38 (c.114G>A) (3 cases); [k]p.R164T (c.491G>C); [l]c.157-18A>G (12 cases), c.278+50G>A; [m]c.1636-73A>G (7 cases), p.A547S (c.1639G>T). Fr. – fragment.

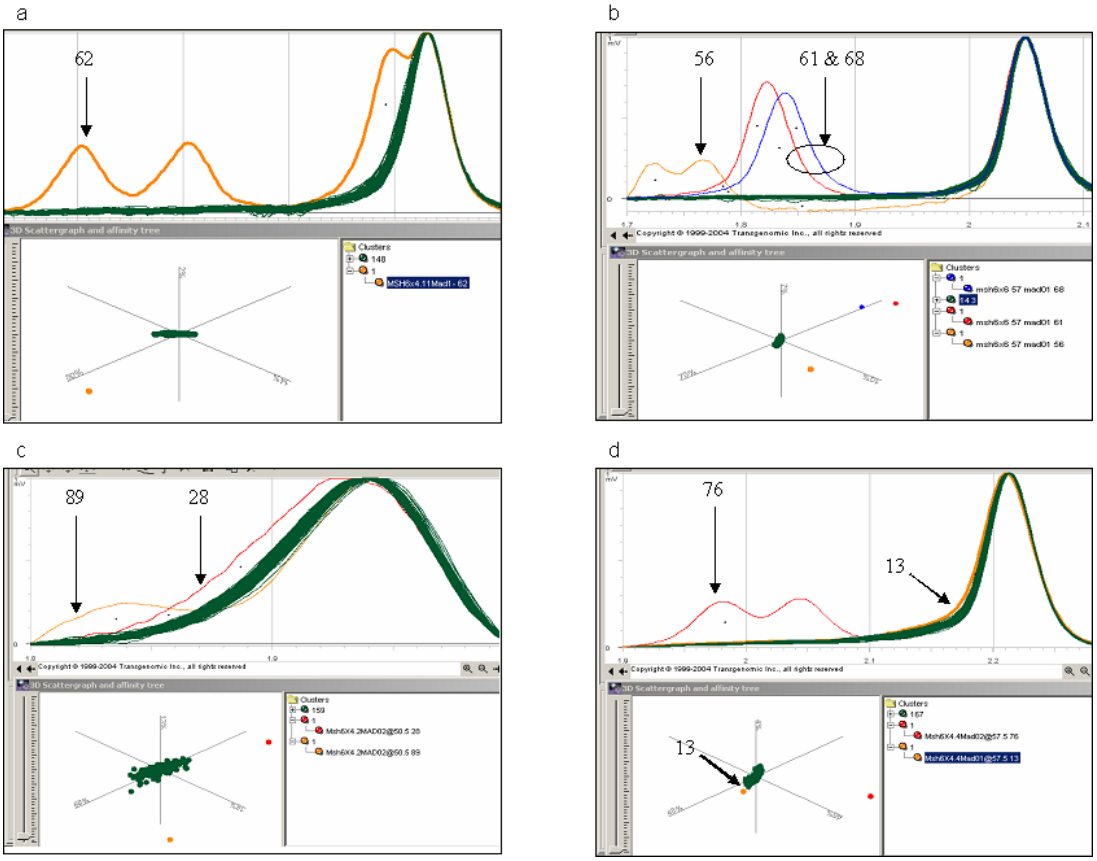


Figure 3.4: Evaluation of the Navigator software to detect rare variants.

Under both Level 2 and Level 3 analysis conditions, 59/60 (98.3%) products with outlying profiles were shown to harbour novel sequence variants. Examples shown are: **(a)** exon 4 fragment 11 of *MSH6* analysed at 57ºC (sample 62 is heterozygous for p.R1005X [c.3013C>T]); **(b)** exon 6 of *MSH6* analysed at 57ºC (sample 56 is heterozygous for c.3439-57_58del, and samples 61 and 68 are heterozygous for c.3439-16C>T); **(c)** exon 4 fragment 2 of *MSH6* analysed at 59.5ºC (sample 28 is heterozygous for p.G289E [c.866G>A + c.867C>A] and sample 89 is heterozygous for p.K295R [c.884A>G]); **(d)** exon 4 fragment 4 of *MSH6* analysed at 57.5ºC (sample 76 is heterozygous for p.S503C [c.1508C>G]) - sample 13 was identified as an outlier under Level 1 analysis conditions but shown to be wild type by sequence analysis; this sample was not considered an outlier under Level 2 analysis conditions. Each figure shows a 3D plot (bottom left) an affinity trees (bottom right) and the overlaid, normalised dHPLC profiles (top). For each image Level 2 and 3 analysis was identical.

3.3.2 Evaluation of the Navigator Software to Detect Common Polymorphisms

To determine whether the Navigator software could be used to reliably detect samples carrying common polymorphisms, the 172 samples were analysed for selected regions of MSH6 (exon 2, exon 3, exon 4 fragment 1 and exon 5 fragment 2), NEIL2 (exon 3 and exon 4 fragment 1), NEIL3 (exon 1, exon 5 and exon 8 fragment 1) and OGG1 (exons 5 and 7) – all of which were known to harbour polymorphisms with minor allele frequencies between 8 and 40% (our unpublished data). All samples were typed for the polymorphisms using ARMS or restriction digest based assays, the accuracy of the genotypes was confirmed by sequence analysis: 952/952 (100%) genotypes from restriction digestion assays, 427/427 (100%) genotypes from ARMS-based assays and 100/100 (100%) genotypes from Genescan size analysis assays were identical to results generated by DNA sequencing, genotypes were compared to dHPLC profile classes determined using the Navigator software. Samples heterozygous or homozygous for the minor allele were considered identical for these analyses (i.e. 'variant'), since all samples underwent dHPLC analysis in the presence of 20% wild type PCR product. (Kaler et al., 2000, Pirulli et al., 2000, Schaeffeler et al., 2001). Eight of the eleven amplified fragments were assayed at two melt temperatures which resulted in nineteen data sets (consisting of 2,784 dHPLC elution profiles). We found discrepancies between the Navigator elution profile classes and the genotypes at each of the levels of analysis and with each of the selected sub-options; however, by selecting the 'rightmost peak' and 'trailing edge' sub-options, we achieved the highest concordance between dHPLC groupings and genotypes. Using these optimal parameters, Level 1 showed 85.1% concordance, Level 2 showed 95.5% concordance, and Level 3 showed 97.2% concordance. If we excluded products with profiles that were obvious outliers to the main groups and that were subsequently shown to also harbour novel variants, then under Level 3 analysis, 1,573/1,580 (99.6%) products were correctly called as wild-type and 1,137/1,145 (99.3%) products were correctly identified as harbouring the polymorphism under assessment (Table 3.3). The 15 incorrect calls were either due to samples homozygous for the rare polymorphic allele being identified as outliers (2 cases), samples carrying a novel variant being classified as carrying a common polymorphism (1 case), or, samples carrying common polymorphisms being poorly

resolved due to non-optimal melting conditions (12 cases); 13 of these products were assayed at a second temperature and all were subsequently called correctly. When considering data from both melt temperatures together, there was 99.9% (1,597/1,599) concordance between Navigator elution profile classification and genotype under Level 3 analysis conditions.

3.3.3 Evaluation of the Navigator Software to Detect Rare Variants in Fragments Harbouring Common Polymorphisms

The presence of two or more sequence variants within a single DNA fragment often results in a complex elution profile. Therefore, we determined the specificity of the software to correctly identify profiles corresponding to the presence of a rare variant in the context of a co-existing frequent polymorphism. Among the 2,784 dHPLC elution profiles analysed for common polymorphisms, we identified 61 products as having distinct profiles under our optimal parameters within a Level 3 analysis. Direct sequencing showed that 59 (96.7%) of these harboured novel variants (Table 3.3) (examples are shown in Figure 3.6). Both remaining outlier products (in exon 5 of *NEIL3* at 57ºC) were shown to correspond to rare homozygotes for the polymorphism present within that exon (c.701+34_40del).

All 2,784 elution profiles were also analysed by visual inspection. In total, 56 products with outlying profiles were identified, all of which were shown to harbour novel variants upon sequence analysis. In comparison to the analyses using the Navigator software, three products with novel variants (all heterozygous for both p.R15 [c.45C>A] and p.R38C [c.112C>T] in exon 1 of *NEIL3*]) were missed by visual inspection because the elution profiles lay very close to the profile for the common polymorphism p.R15 [c.45C>A]. In contrast, the two products identified as outliers using the software, but shown to be homozygous for the rare allele of the polymorphism c.701+34_40del in exon 5 of *NEIL3*, were correctly classified to the polymorphic group upon manual inspection.

3.3.4 Sensitivity

In identifying the optimum conditions for Navigator analysis a single variant (p.R103Q) was discovered in a fragment that was previously thought to be non-polymorphic (*NEIL2* exon 2, fragment 1) and had therefore not been included as part of this study (Figure 3.5).

Table 3.3: SNP Discovery Summary.

Gene	Exon	dHPLC Temp (°C)	No. of products analysed	SNP Assayed	Wild types	Variants	Outliers
MSH6	2	57	147	p.P92 (c.276A>G)	93/95	46/52	2[a]/2
MSH6	2	60	150	p.P92 (c.276A>G)	94/96	54/54	2[a]/2
MSH6	3	58	150	p.D180 (c.540T>C)	69/69	81/81	-
MSH6	3	59.5	152	p.D180 (c.540T>C)	72/72	80/80	-
MSH6	4(Fr.1)	57.5	160	p.Y214 (c.642C>T)	129/129	31/31	-
MSH6	4(Fr.1)	59	156	p.Y214 (c.642C>T)	126/126	30/30	-
MSH6	5(Fr.2)	54.5	157	c.3438 + 14A>T	59/61	96/96	2[b]/2
MSH6	5(Fr.2)	58	156	c.3438 + 14A>T	59/61	95/95	1[c]/1
NEIL2	3	59	95	p.P188 (c.564A>G)	45/47	45/48	5[d]/d
NEIL2	3	60.5	118	p.P188 (c.564A>G)	54/58	55/60	9[e]/9
NEIL2	4(Fr.1)	59	144	c.689 - 13C>T	79/79	62/65	3[f]/3
NEIL2	4(Fr.1)	62.5	144	c.689 - 13C>T	79/79	63/65	2[g]/2
NEIL3	1	62.5	151	p.R15 (c.45C>A)	68/82	62/69	15[h]/15
NEIL3	1	64.5	151	p.R15 (c.45C>A)	74/82	65/69	12[i]/12
NEIL3	5	57	135	c.701 + 34_40del	113/113	20/22[+]	0/2[+]
NEIL3	8(Fr.1)	53	159	c.1040 - 85T>C	85/86	72/73	2[j]/2
NEIL3	8(Fr.1)	55.5	154	c.1040 - 85T>C	82/83	70/71	2[j]/2
OGG1	5	63	151	c.748 - 15C>G	94/95	55/56	2[k]/2
OGG1	7	62	154	p.S326C (c.977C>G)	99/99	55/55	-
Total			2784		1573/1612 (97.6%)	1137/1172 (97.0%)	59/61 (96.7%)
[#]excl. outliers					1573/1580* (99.6%)	1137/1145^ (99.3%)	

Analyses were performed by selecting the 'rightmost peak' and 'trailing edge' sub-options under Level 3; dHPLC profiles were classified as wild types, variants or outliers. The wild type and variant columns show the number of products classified by the Navigator software out of the number of products genotyped to a category. The outlier column shows the number of products that contained a novel variant (shown by sequence analysis) out of the total number of products with outlier profiles ([+]two products were identified as outliers, but genotyping and sequencing showed that they were homozygous for minor allele of the polymorphism 701+34_40del).

Legend continued on next page.

[#]Analyses excluding 59 products with outlying profiles and proven to harbour novel sequence variants: [*]7 incorrect calls were from the wild type group (exon 5 fragment 2 of *MSH6* at 58ºC in 1 case and exon 1 of *NEIL3* at 62.5ºC in 6 cases), and, [^]8 incorrect calls were from the variant group (exon 2 of *MSH6* at 57ºC in 6 cases and exon 5 of *NEIL3* at 57ºC in 2 cases). Novel variants identified were: [a]p.S144I (c.431G>T) (2 cases); [b]c.3438+17G>C, p.T1102 (c.3306T>A); [c]c.3438+17G>C; [d]c.492–50G>C (2 cases), c.492–50G>C and p.P188 (c.564A>G) heterozygote (2 cases), c.492–8C>T and p.P188 (c.564A>G) heterozygote; [e]c.492–50G>C (4 cases), c.492–50G>C and p.P188 (c.564A>G) heterozygote (4 cases), c.492–8C>T and p.P188 (c.564A>G) heterozygote; [f]p.R257L (c.770G>T) and c.689-13C>T homozygote (3 cases); [g]p.R257L (c.770G>T) and c.689-13C>T homozygote (2 cases); [h]c.1–37G>A heterozygote (6 cases), c.1–37G>A and p.R15 (c.45C>A) (4 cases), p.R38C (c.112C>T) and p.R15 (c.45C>A) (3 cases), c.1-37G>A homozygotes (2 cases); [i]c.1-37G>A heterozygote (6 cases), c.1-37G>A and p.R15 (c.45C>A) (4 cases); and c.1-37G>A homozygotes (2 cases); [j]c.1040–181A>G and p.R381 (c.1143G>A), c.1040–107A>G and c.1040–85T>C; [k]p.A288V (c.863C>T) (2 cases).

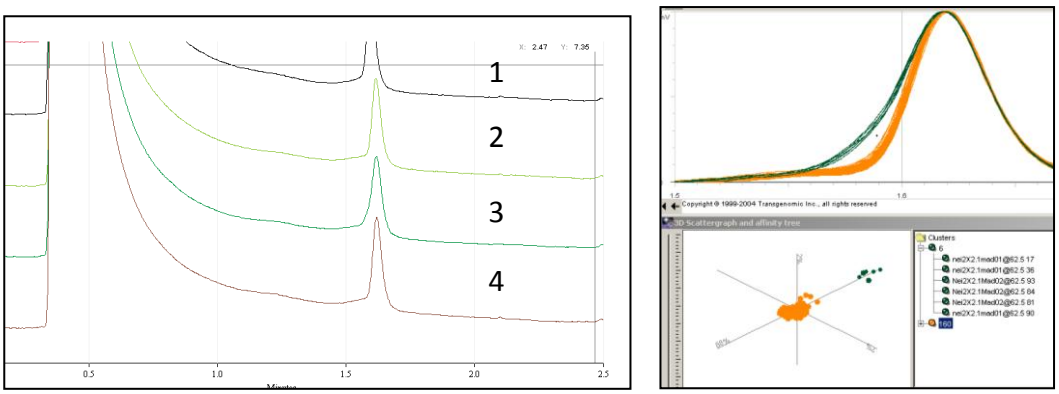


Figure 3.5: Navigator Sensitivity.

The dark green trace (3) in the left image contains the variant *NEIL2* p.R103Q which was detected by the software and missed by the manual analysis when looking at traces individually. The overlaid plots (right) clearly show 6 outlying samples which were subsequently shown to contain the SNP.

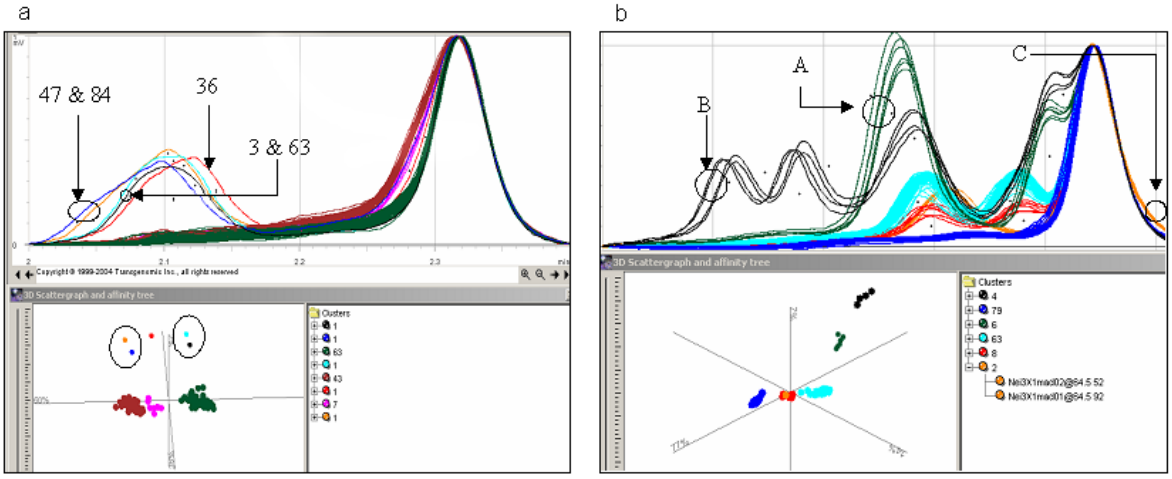


Figure 3.6: Detecting novel variants in fragments harbouring common polymorphism.

Fifty nine out of 61 (96.7%) outliers from the main polymorphism groups harboured novel variants (Level 3 analysis). Examples shown are: **(a)** exon 3 of *NEIL2* at 59ºC - Five outlier products were identified (samples 3 and 63 were heterozygous for c.492-50G>C; samples 47 and 84 were heterozygous for c.492-50 G>C and p.P188 (c.564A>G); and sample 36 was heterozygous for c.492-8C>T and p.P188 [c.564A>G]). **(b)** exon 1 of *NEIL3* at 64.5ºC – Twelve outlier products were identified (6 samples [marked A] were heterozygous for c.1-37G>A, 4 samples [marked B] were heterozygous for c.1-37G>A and p.R15 (c.45C>A) and 2 samples [marked C] were homozygous for c.1-37G>A).

3.4 Discussion

Our analyses of over 6,500 amplicons generated from four different genes have shown that the Transgenomic Navigator software is an excellent tool for the assessment of dHPLC elution profiles, although some user input is necessary. For example, in fragments that do not harbour common polymorphisms, Level 1 analyses revealed that only 43.5% of outlier products were correctly identified as containing new sequence variants, whereas under Level 2 and 3 analyses that require more operator input, 98.3% were correctly called. Similarly, for complex profiles such as those generated from fragments harbouring common polymorphisms, we found that a high level of operator input was necessary to achieve accurate grouping of dHPLC profiles (we achieved 97.2% concordance between dHPLC grouping and polymorphism genotype under Level 3).

An initial concern was that by increasing the stringency for calling outliers (from Level 1, to Levels 2 and 3), we might miss a large number of mutations lying close to the main groups. Our data does not support this concern since of the 138 products classified as having outlying profiles under Level 1 analysis, 60 were shown to harbour novel sequence variants and 78 were wild type upon sequencing. By comparison, 59/60 products classified as outliers under Level 2 or 3 analyses (Table 3.2), were confirmed to contain sequence variants and only one outlier product was wild type upon sequence analysis; the only mutation missed by this more stringent analysis was correctly classified at the alternative dHPLC temperature.

We have assessed the utility of the software in comparison to manual inspection of the dHPLC elution profiles. Under Level 3 analysis with the Navigator software, a total of 118/121 products with outlying profiles were shown to harbour novel sequence variants. In comparison, by visual inspection, 114/114 products with outlying profiles were shown to harbour novel sequence variants. Therefore, four (3.4%) products with novel variants were missed by manual inspection. However, the software also identified three products with outlying profiles that did not contain new sequence variants and these products were correctly classified by manual inspection. In our experience, for fragments with complex elution profiles (such as those harbouring common polymorphisms), the Navigator software allows for a

more rapid analysis of the data when compared to manual inspection, and reduces the possibility of operator error.

Although dHPLC cannot be used as a genotyping technology (since it cannot always discriminate between samples heterozygous and homozygous for a rare allele), we show that the Navigator software can be used to rapidly group samples containing common polymorphisms and allow outliers to be easily recognised. Indeed, we showed that 96.7% of products with profiles outlying the main groups harboured new variants. Furthermore, of those fragments that did not contain a second variant, 99.4% were classified correctly in relation to common polymorphisms. We conclude that the Navigator software provides a powerful tool for automating the identification of novel variants irrespective of the complexity of the dHPLC elution profiles.

While dHPLC has been replaced in our institution by higher throughput HRM (Section 1.2.2.5) technology it is still widely used in for SNP discovery, PubMed details over 20 publications in the first quarter of 2012. Zahary (2012) describes use of the technology in identify germline mutation in Malaysian HNPCC patients in the MMR genes *MLH1* and *MSH2* as part of an expression profile experiment demonstrating the current value of the technique where equipment is available.

4 Chapter four

Using the Transgenomic Navigator™ software to identify novel germline variants in the *OGG1*, *NEIL1*, *NEIL2*, *NEIL3*, *MTH1* and *NTH1* genes and investigating their potential role in susceptibility to colorectal adenomas

4.1 Introduction

Established predisposition genes account for only a small proportion of familial colorectal cancer (CRC) (Kinzler and Vogelstein, 1996). We recently showed that biallelic germline defects in *MUTYH* predispose to multiple colorectal adenoma (CRA) and carcinoma [(Al-Tassan et al., 2002), (Jones et al., 2002), (Sampson et al., 2003), reviewed in (Cheadle and Sampson, 2003)]. This autosomal recessive disorder has been termed *MUTYH* associated polyposis (MAP). *MUTYH* functions as a base excision repair (BER) DNA glycosylase and is responsible for removing adenines mis-incorporated opposite 8 -oxo-7,8-dihydro-2'-deoxyguanosine (8-oxoG), one of the most stable products of oxidative DNA damage (Slupska et al., 1999). The highly conserved BER system plays a major role in the protection against oxidative DNA damage and is particularly important in the colon where high levels of reactive oxygen species are generated by commensal bacteria and dietary carcinogens [(Ames and Gold, 1991) (Bartsch et al., 2002) (Huycke and Gaskins, 2004)]. A number of additional BER enzymes also protect against oxidative DNA damage (Wood et al., 2005). *OGG1* and *NTH1* remove the damaged base from 8-oxoG:C (Roldan-Arjona et al., 1997) and 8-oxoG:G base-pairs (Matsumoto et al., 2001) respectively, and a recently discovered family of enzymes called *NEIL1-3* have overlapping substrate specificities with *OGG1* and *NTH1* (Hazra et al., 2003). In addition, *MTH1* functions as an 8-oxoGTPase that hydrolyses oxidised guanine triphosphates in the nucleotide pool, thereby preventing their incorporation into the nascent strand (Sakumi et al.,

1993). Here, we determined whether germline variants in *OGG1, NEIL1, NEIL2, NEIL3, MTH1* and *NTH1,* may, like *MUTYH*, predispose to multiple CRA and carcinoma.

Data in this chapter was presented in the manuscript "Inherited predisposition to colorectal adenomas caused by multiple rare alleles of MUTYH but not OGG1, NUDT1, NTH1 or NEIL 1, 2 or 3" (Dallosso et al., 2008) of which James Colley was an author who helped write the manuscript and as a consequence there are some overlaps in parts of the text.

4.2 Materials and Methods

4.2.1 Samples

One hundred and thirty four unrelated index cases with multiple CRA with or without carcinoma were recruited from regional UK polyposis registers in Birmingham, Cambridge, Cardiff, Liverpool, Manchester, Southampton and Surrey. Given the pattern of inheritance observed in MAP, patients were selected for a family history consistent with autosomal recessive transmission of the disease. No cases carried biallelic *MUTYH* mutations, nor a truncating mutation in *APC* (associated with familial adenomatous polyposis). Patients were categorised according to the number of CRAs recorded at colonoscopy or colectomy in order to identify a set that closely resembled a MAP like phenotype (Group B). Group A consisted of 55 patients with 3-10 CRAs and presented at a mean age of 48 years, Group B consisted of 53 patients with 11-100 CRAs and presented at a mean age of 48 years, Group C consisted of 26 patients with more than 100 CRAs presenting at a mean age of 44 years. DNA was prepared from venous blood samples using standard methods. As is expected where the cause of disease severity is genetic variation, adenoma frequency was greater in younger patients ($p = 0.005$; linear regression).

396 British Caucasian control blood DNA samples were obtained from the European Collection of Animal Cell Cultures (ECACC) (Dallosso et al., 2008). 47 Dutch patients with 11-100 CRAs and a family history consistent with recessive inheritance

and 262 matched control samples were obtained from Dr F.J Hes in Leiden University.

4.2.2 PCR

We amplified exons 1-8 of *OGG1* and 2 -5 of *MTH1* as previously described (Al-Tassan et al., 2002). The open reading frames (ORFs) of *NEIL1*, *NEIL2, NEIL3* and *NTH1* were amplified as 9, 6, 11 and 5 fragments, respectively. We amplified exons 2 to 4 of *NEIL2* (spanning the variants p.R103Q and p.R257L) as a 6.5kb long distance PCR fragment (Appendix C).

4.2.3 Denaturing High Performance Liquid Chromatography (dHPLC)

dHPLC was carried out as described in Section 2.4.7 at the melting temperatures predicted by Navigator (Version 1.54) software (Appendix C). Samples with aberrant elution profiles were identified using Navigator assisted trace calling (Colley et al., 2005) and were sequenced directly.

4.2.4 Sequencing

PCR products were purified as described in Section 2.4.3 and sequenced as described in Section 2.4.4. Purification of sequencing products was carried as described in Section 2.4.5. All mutations were confirmed by sequencing at least two independent PCR products.

4.2.5 Assays for Sequence Variants

We used ARMS assays (2.4.9) or restriction digests (2.4.8) to genotype the common variants in the patient samples and to determine the frequency of the missense, nonsense and splice site variants in at least 650 British Caucasian control chromosomes (ECACC, Salisbury, UK). Primer sequences and conditions are available in Appendix B.

4.2.6 Statistical Analysis

Variants were assessed for departure from the Hardy-Weinberg equilibrium using the chi-square test or, when the number of genotypes was <5, by using the Monte-Carlo permutation test with 10,000 permutations (HWSIM software). Differences between the proportion of cases and controls with each variant were analysed using the chi-square or Fisher's exact tests. Corrections for multiple testing were performed by a permutation test, randomly reshuffling case-control status (when individual genotypes of cases and controls were available). When individual genotypes for controls were unavailable, correction for multiple testing was performed using spectral decomposition of matrix of pair-wise r^2 measures of linkage disequilibrium (Nyholt, 2004).

4.2.7 Author's Contribution

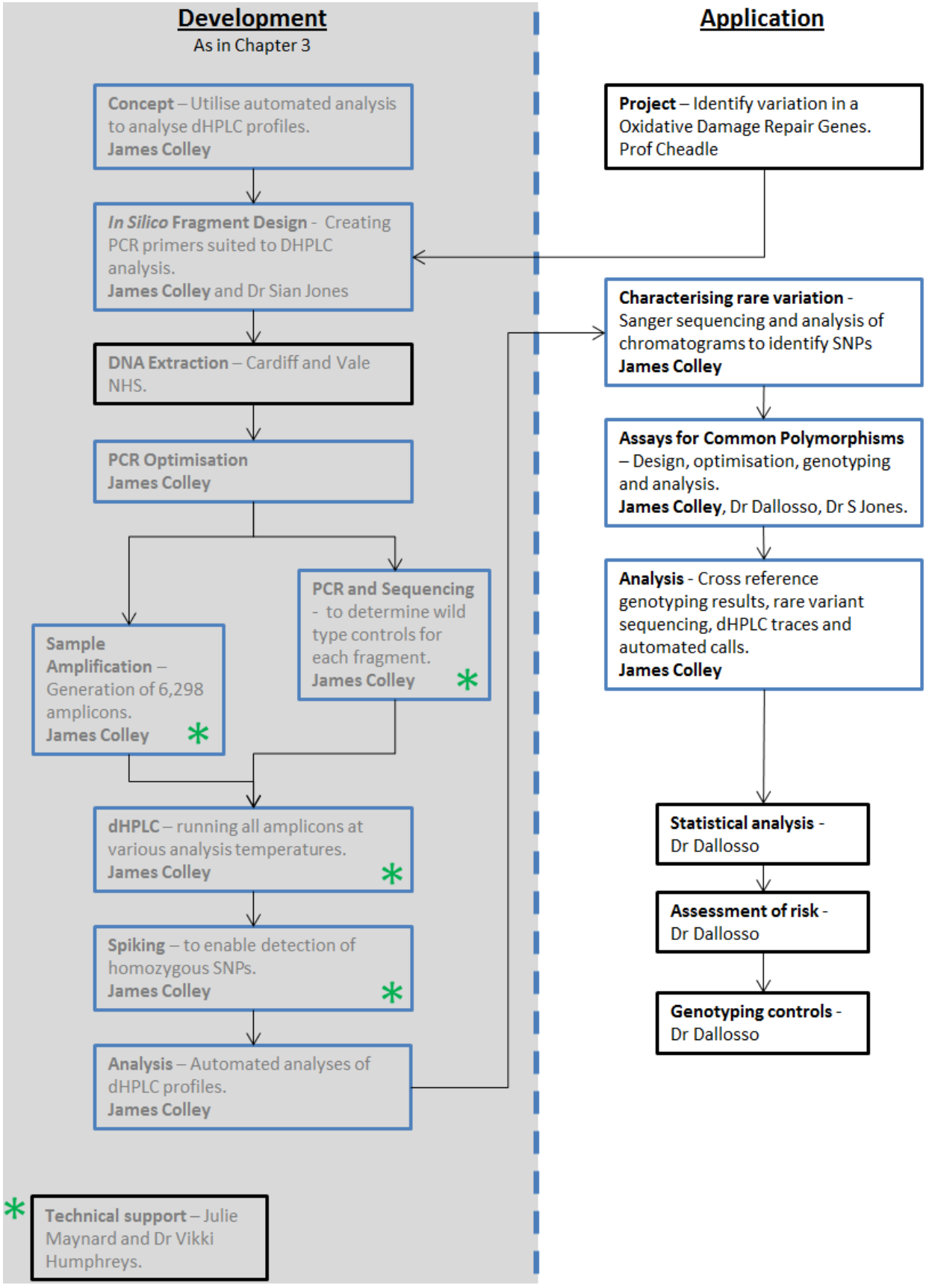


Figure 4.1: Authors Contribution to Chapter 4.

4.3 Results

We screened the entire ORFs of *OGG1*, *NEIL1*, *NEIL2*, *NEIL3*, *MTH1* and *NTH1* for germline variants in 134 patients with multiple CRA, by using a combination of dHPLC and direct DNA sequencing. In total, we identified 65 variants, 41 of which were silent or intronic changes (Table 4.1) and 24 were coding sequence alterations (22 missense changes, a nonsense change and a splice site alteration) (Table 4.2). Of the 22 missense changes, 5 were identified in *OGG1* (p.R46Q, p.A85S, p.A288V, p.I321T and p.S326C), 2 in *NEIL1* (p.P208S and p.R339Q), 3 in *NEIL2* (p.R103Q, p.R164T and p.R257L), 10 in *NEIL3* (p.R38C, p.P117R, p.D132V, p.Q172H, p.H286R, p.R315Q, p.L443P, p.H471Q, p.R520G and p.A547S) and 2 in *MTH1* (p.V106M and p.K155I). In total, 11 of these missense variants were novel at the time. All patients with missense variants with MAFs <15% were sequenced for the entire ORF of the corresponding gene to search for biallelic mutations (as found in MAP).

We also designed assays for each of the missense variants and screened for their presence in a panel of 396 British Caucasian control blood DNA samples. Four patients carried both p.R103Q and p.R257L in *NEIL2*; however, these variants were also identified together in control samples and cloning and sequencing of an *NEIL2* genomic region encompassing exons 2 and 4 revealed that both variants were present on the same allele. Three patients carried the missense changes p.Q172H and p.R38Cin *NEIL3*; however, these variants were detected together in three control samples. Five patients carried the relatively common variant p.P117R (minor allele frequency [MAF]~12%) in *NEIL3* in combination with rare missense changes (with both p.R38C and p.Q172H in 1 case, with p.R38C in 1 case, with p.R315Q in 1 case, and with p.R520G in 2 cases); however, four control samples also harboured p.P117R in combination with these rare variants (with p.R38C in 1 sample, with p.R315Q in 2 samples, and with p.R520G in 1 sample) and p.P117R was also identified in a homozygous state in five (out of 384) control samples. No other patients carried biallelic missense variants (apart from those harbouring the common variants p.S326C in *OGG1* and p.H286R, p.L443P and p.H471Q in *NEIL3*).

We determined whether any missense variants were over-represented in cases (analysed by polyp number and as a whole) versus controls. All except four of the missense variants (p.I321T in *OGG1*, p.P208S in *NEIL1*, p.R164T in *NEIL2* and p.D132V in *NEIL3*) were detected in control samples. No variants showed evidence for departure from the Hardy-Weinberg equilibrium amongst the cases. A single variant, p.P117R in NEIL3, was over-represented in patients with 11-100 CRAs (*P*=0.01) (Table 4.2). This apparent over-representation was re-assessed in an independent cohort. An analysis of 47 Dutch patients with 11-100 CRAs (and with a family history consistent with recessive inheritance) and 262 matched control samples failed to demonstrate a significant association, either alone or in combination with the original data. Indeed, after correcting the original data for multiple testing, p.P117R was no longer found to be significantly over-represented (*P*=0.13 by the permutation test and 0.21 by the spectral decomposition of LD matrix).

A single sporadic case harboured a novel nonsense mutation (p.Q90X) in *NTH1*. Complete sequencing of the *NTH1* ORF failed to identify a second mutation in this case and extended family members were not available to determine the segregation of this variant. Furthermore, no tumour samples were available to assay for somatic inactivation of the wild type *NTH1* allele. p.Q90X was screened for and found in a heterozygous state in two (out of 359) control samples. We identified five cases that harboured the previously reported splice site variant c.434+2 T>C in *NEIL1* (dbSNP reference rs5745908). Sequencing of the *NEIL1* ORF in these cases failed to reveal any second mutations and fresh tissue samples were not available to determine the effect of this variant on RNA splicing. Eight out of 360 control samples were also found to be heterozygous for this variant.

Table 4.1: Intronic and silent variants identified.

Gene / location	Nucleotide change	Amino acid	Minor allele freq.	Reference
OGG1				
IVS2	c.389-111 T>G	-	<1%	-
IVS3	c.5659 G>C	-	<1%	-
IVS3	c.566-4 G>A	-	1%	(Al-Tassan et al., 2002)
IVS4	c.748-15 C>G	-	17%	rs2072668
3' UTR	c.1038 + 279 A>G	-	14%	-
NEIL1				
5' UTR	c.1-50 C>G	-	<1%	-
IVS1	c.434+34 G>A	-	<1%	-
Exon 8	c.1071 G>A	p.K357	<1%	-
IVS8	c.1103-17 G>T	-	<1%	-
IVS8	c.1103-45A>T	-	<1%	-
NEIL2				
Exon	c.1+114 A>G	p.Q38	<1%	-
IVS1	c.138+25 T>C	-	ND	-
IVS1	c.138+35 C>T	-	ND	-
IVS2	c.492-8C>T	-	<1%	rs8191641
IVS2	c.492-50 G>C	-	3%	rs8191640
Exon 3	c.564 A>G	p.P188	21%	rs8191642
IVS3	c.689-13 C>T	-	22%	rs8191663
3' UTR	c.999+21 C>T	-	22%	rs1534862
3' UTR	c.999+34 delC	-	22%	rs8191667
NEIL3				
Exon 1	c.45 C>A	p.R15	24%	rs10013040
5'UTR	c.1-37 G>A	-	4%	-
IVS1	c.157-18 A>G	-	5%	rs3792606
IVS2	c.278+50 G>A	-	<1%	-
IVS5	c.627+13 A>T	-	12%	-
IVS5	c.627+41 A>G	-	42%	rs2271102
IVS6	c.701+34 del 7bp	-	7%	-
Exon 6	c.756 T>C		18%	rs17676249
IVS7	c.869+9 T>C	-	19%	-
IVS7	c.1040-85 T>D	-	49%	-
IVS7	c.1040-107 A>G	-	<1%	-
IVS7	c.1080-181 A>G	-	<1%	-
Exon 8	c.1143 G>T	p.R381	<1%	rs2048074

Exon 8	c.1272 C>T	p.V424	29%	rs10007075
IVS9	c.1636-73 A>G	-	2%	rs6820069
MTH1				
IVS3	c.221+7 A>T	-		-
Exon 4	c.366 C>T	p.D122		-
Exon 5	c.426 C>T	p.D142		rs1799832
3' UTR	c.540+32 C>T	-		-
NTH1				
IVS2	c.378+67 C>T	-		rs2233519
Exon 4	c.600 C>T	p.Y200		-
IVS5	c.939+30 C>T	-		-

Table 4.2: Coding region SNP statistics.

Gene / exon	AA change	Nucleotide change	Type of Evolut[1] change	Allele Freq in cases			Freq in controls[2]	P-value[3]	Reference / dbSNP[4]
				A (110)	**B (106)**	**C (52)**			
OGG1									
1	p.R46Q	c.137 G>A	SC > C	1	0	1	3/1044	0.16	(Kohno et al., 1998)
2	p.A85S	c.253 G>T	C > NC	0	1	0	7/688	0.75	(Chevillard et al., 1998)
5	p.A288V	c.863 C>T	C > NC	2	0	0	3/716	0.33	rs3219012 (1%)
7	p.S326C	c.977 T>C	SC > NC	23	21	7	79/362	0.17	(Shinmura et al., 1998) rs1052133 (28%)
7b	p. I321T	c.962 T>C Transcript 2d	NC >n/a	0	0	1	0/720	0.07	
NEIL1									
1	N/A	c.434 +2 T>C	n/a>n/a	1	2	5	8/720	0.13	rs5745908 (1%)
4	p.P208S	c.622 C>G	SC > SC	1	0	0	0/624	0.15	
8	p.R339Q	c.1016 G>A	NC > NC	1	0	0	1/704	0.24	
NEIL2									
2	p.R103Q	c.308 G>A	SC > NC	2	2	0	7/649	0.31	rs8191613 (7%)
2	p.R164T	c.491 G>C	SC > SC	1	0	0	0/710	0.13	
4	p.R257L	c.770 G>T	NC > NC	2	2	0	7/553	0.43	
NEIL3									
1	p.R38C	c.112 C>T	SC > NC	3	1	2	10/706	0.16	
3	p.P117R	c.349 C>G	NC > SC	8	19	3	75/724	0.01	rs7689099 (7%)
						6/94*	60/524*	0.09*	
3	p.D132V	c.395 AC>TG	NC > SC	1	0	0	0/580	0.13	
4	p.Q172H	c.516 G>C	SC > SC	2	0	1	4/462	0.2	rs17064658 (4%)
6	p.H286R	c.857 A>G	C > NC	22	17	7	113/724	0.07	
7	p.R315Q	c.944 G>A	SC > SC	0	1	0	2/586	0.38	
8	p.L443P	c.1328 T>C	NC > NC	23	28	13	185/726	0.03	rs13112358 (40%)
8	p.H471Q	c.1413 C>A	SC > SC	20	23	10	154/718	0.67	rs13112390
9	p.R520G	c.1558 A>G	NC > SC	0	4	1	18/622	0.29	rs1876268 (5%)
10	P.A547S	c.1639 G>T	SC > SC	1	0	0	5/722	0.5	

MTH1									
4	p.V106M	c.316 G>A	C > SC	1	1	0	6/718	0.62	(Yoshimura et al., 2003) rs4866 (1%)
5	p.K155I	c.464 A>T	SC > NC	0	1	0	2/702	0.34	
NTH1									
2	p.Q90X	c.268 C>T	n/a >NC	0	1	0	2/718	0.34	

Allele frequencies of coding region sequence variants in *OGG1, NEIL1, NEIL2, NEIL3, MTH1* and *NTH1* in patients with multiple CRAs and controls.

[1]Alignment of human (NP_002533; NP_078884; NP_659480; NP_060718; NP_002443), mouse (NP_035087; PAB28790; XP356749; NP_666320; NP_32663), *Drosophila* (NP_572499, OGG1 only), *Arabidopsis* (GI25349383; NP_564608; NP_849798; NP_565965), *S. cerevisiae* (NP_013651; S49801, *OGG1* and *NEIL2* only) and *E. coli* (NP_418092- formamidopyrimidine DNA glycosylase; NP_415242- endonuclease VIII; BAB33526 - 7,8-dihydro-8-oxoguanine-triphosphatase) *OGG1*, *NEIL1*, *NEIL2*, *NEIL3* and *MTH1* homologues was carried out using ClustalW (v.1.82). Residues were determined to be conserved (C), semi-conserved (SC) or non-conserved (NC) through evolution based on the retention of identical or similar amino acids in prokaryote, vertebrates or neither, respectively. The allele frequencies of each variant in patients with 3-10 CRAs (Group A - 55 patients), 11-100 CRAs (Group B - 53 patients) and more than 100 CRAs (Group C - 26 patients), is shown (over 95% of samples were successfully genotyped).

[2]Variants were typed in British Caucasian control samples (supplied by ECACC);

[3]*P*-values were determined for each cohort of patients vs. controls.

[4]dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) reference numbers and published minor allele frequencies are shown.

*The *NEIL3* missense variant p.P117R was assayed in an independent cohort of Dutch patients with 11-100 CRAs and matched controls - no significant association was identified. AA, amino acid; n/a, not applicable.

4.4 Discussion

We have sought pathogenic mutations in six genes that function in concert with *MUTYH* to protect against oxidative DNA damage. Although it is difficult to ascertain the likely pathogenic effect of novel variants without performing functional assays, we have used three complimentary genetic approaches to help determine pathogenicity: (i) since MAP is an autosomal recessive disease and since five out of the six genes analysed function as BER glycosylases (like MUTYH), we sought biallelic mutations in MAP-like cases, (ii) we looked for the presence of both single and combinations of variants in normal control samples, and, (iii) we sought over-representation of variants in cases versus control samples. No patients harboured clearly pathogenic biallelic sequence variants, nor was there evidence that any single variant was disease causing. Although we initially found that p.P117R in *NEIL3* was over-represented in a sub-group of cases, this finding was not replicated in an independent cohort and became insignificant upon correcting for multiple testing and therefore probably represents a chance observation. Furthermore, despite finding CRA patients carrying the nonsense mutation p.Q90X in *NTH1* and the splice site variant c.434+2 T> C in *NEIL1*, no biallelic mutations were identified in these cases and these variants were also identified in control samples suggesting that they are not associated with a CRA phenotype. Interestingly, p.Q90X lies in an N-terminal domain of NTH1 that is not present in the *E. coli* homolog endonuclease III (Aspinwall et al., 1997) and is closely followed by an in frame methionine at codon 102 that fulfils the criteria of a Kozak translation initiation signal (Pedersen and Nielsen, 1997). Therefore, it is possible that partially truncated, but functional, NTH1 is produced from this allele. Indeed, Liu and Roy (Liu and Roy, 2002) have shown that truncation of this N-terminal domain of NTH1 actually stimulates its activity.

Four rare missense variants were not identified in control samples although we found no supporting evidence for their pathogenicity: p.I321T (only present in a rare alternatively spliced transcript of *OGG1*) and p.D132V in *NEIL3* are non-conservative amino acid substitutions, and, p.P208S in *NEIL1* and p.R164T in *NEIL2* are semi-conservative amino acid substitutions.

This study demonstrates that mutations of likely functional significance in *OGG1, NEIL1, NEIL2, NEIL3, MTH1* and *NTH1* are not common in patients with CRA and these findings support those of a smaller study (Sieber et al., 2003b), that failed to identify pathogenic mutations in *OGG1* and *MTH1* in patients with a similar phenotype. Not only do the substrate ranges of the BER glycosylases overlap, but the mismatch repair complex MSH2/MSH6, the CSB protein and BRCA1 and BRCA2 have all been implicated in the repair of 8-oxoG (reviewed (Cheadle and Sampson, 2003)). Further studies are therefore warranted to determine whether it is a specific, and as yet undetermined, facet of *MUTYH* function that underlies its apparently unique (amongst BER enzymes) association with CRA.

Since the completion of this work in 2008 the *OGG1* variant p.R46Q has been identified as a pathogenic mutation (Morak et al., 2011) affecting the last nucleotide of exon 1 and altering splicing *in vitro*. Morak did not find p.R46Q in 70 control samples however in our set of 1,044 it occurred with a 0.14% MAF leading us to assume it was benign.

5 Chapter five

Investigating sequence analysis software for high throughput variant detection and their application to study the 'rare variant hypothesis' of multifactorial inherited predisposition

5.1 Introduction

With the arrival of multi-capillary platforms and robust sequencing chemistry, the cost of a sequencing reaction is coming to level where single direction sequencing reads are an option for SNP discovery projects. However, the data production capacity of these systems has risen to a level where manual checking of chromatograms is often impractical. We aimed to evaluate four sequence analysis programs for automated identification of SNPs from ABI sequence data files. We used Sequencher (Gene Codes version 4.5), InSNP (Christian-Albrechts-University Kiel), Mutation Surveyor (SoftGenetics, version 3.0), and the Staden package (MRC LMB Cambridge, version 1.6) to search for inherited sequence variants in the *APEX1*, *MBD4, MPG*, *NEIL3, NUDT5, PCNA, POL λ* and *TDG* genes (a collection that was relevant to several on-going projects at the time) in 67 patients with multiple colorectal adenomas. In total, 64 PCR amplicons were designed to span the open reading frames of these 8 genes, and over 4,000 chromatograms were generated by ABI 3100 and ABI 3730 Genetic Analysers. 77 SNPs were identified by manual analysis of the chromatograms and this was considered the 'gold-standard' to which the software packages were compared.

We also compared the performance of an optimised version of the Staden package (through Agencourt, Beckman Coulter) against manual assessment by Sequencher in another large scale variant detection project. We analysed a 2,555bp region of the adenomatous polyposis coli (*APC*) gene (encompassed by 6 PCR amplicons) in a large cohort of 969 healthy controls (a total of >5,800 .ab1 files).

It has been proposed that multiple rare variants in numerous genes collectively account for a substantial proportion of multifactorial inherited predisposition to a variety of diseases including CRAs. We have studied this hypothesis by comparing the frequency of rare inherited non-synonymous variants identified by sequence analyses of the β-catenin down-regulating domain of *APC* in patients who did not carry conventional pathogenic mutations in *APC* or *MUTYH* ('non-FAP non-MAP patients') as compared to healthy controls using *In silico* analyses to predict function.

Data in this chapter was presented in the manuscript "Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas" (Azzopardi et al., 2008) of which James Colley was an author who helped write the manuscript and as a consequence there are some overlaps in parts of the text.

5.1.1 Sequencing as a tool for SNP-discovery

Access to multi-capillary sequencers is now common place for many molecular biology groups. The platforms and chemistry are robust and the cost of a single sequencing reaction is relatively low. Sequencing is therefore likely to replace the more traditional techniques such as single-stranded conformational polymorphism analysis (1.2.1.2), denaturing gradient gel electrophoresis (1.2.1.1), chemical or enzymatic cleavage (1.2.2.1, 1.2.2.2) and dHPLC (1.2.2.3); all of these techniques identify samples that contain novel SNPs but DNA sequencing is required for subsequent characterisation.

Where sequencing is chosen as the SNP Discovery method, the throughput demands of research and clinical diagnostic laboratories have meant that large amounts of data are generated very quickly making sequence analysis a significant bottleneck in the discovery process. A large proportion of time is spent manually checking each chromatogram by eye, a problem which can be reduced by relying on automated analysis to identify SNPs. Numerous software packages are available for the analysis of the raw data files generated by ABI sequencers, including the

commercially available Mutation Surveyor software (SoftGenetics) and the freely available Staden package (Medical Research Council Laboratory of Molecular Biology, Cambridge, UK). Both of these programmes can be set up to provide SNP reports based on automated calling but also provide a list of 'potential SNPs' for user review, and by focusing on these two sets of identified SNPs (and not the rest of the chromatogram), the amount of user time can be greatly reduced. In addition, Sequencher V4.6 (GeneCodes) has a SNP calling feature that calls a variant base which differs from the reference trace or consensus sequence and the stringency of this feature can be altered by the 'Call Secondary Peak' function. Here, we evaluated the utility of these three programs plus the freely available InSNP software (Manaster et al., 2005) (Christian-Albrechts-University Keil, Germany) to identify inherited variants in the *APEX1*, *MBD4, MPG*, *NEIL3, NUDT5, PCNA, POL λ* and *TDG* genes.

5.1.2 The 'rare variant hypothesis' of multifactorial inherited predisposition

In support of the 'rare variant hypothesis' of multifactorial inherited predisposition to common diseases (Fearnhead et al., 2005), it has recently been shown that rare non-synonymous variants in the genes encoding apolipoprotein A1, the adenosine triphosphate binding cassette transporter A1 and lecithin cholesterol acyltransferase, are over-represented in individuals with low plasma levels of high-density lipoprotein cholesterol, a major risk factor for coronary atherosclerosis (Cohen et al., 2004). It is likely that rare non-synonymous variants in numerous genes also predispose to colorectal tumours since epidemiological studies predict that inherited factors play a role in 15-30% of colorectal cancers (CRCs), but only a minority of these can be accounted for by established CRC predisposition alleles (Fearnhead et al., 2005, Kinzler and Vogelstein, 1996). Germline truncating mutations in the *APC* gene cause familial adenomatous polyposis (FAP [MIM 175100]), an autosomal dominant disorder characterised by hundreds or thousands of colorectal adenomas (CRA), some of which progress to cancer (Fearnhead et al., 2001) and inherited mutations in the human MutY homologue (*MUTYH*) gene cause *MUTYH*-associated polyposis (MAP [MIM 608456]), an autosomal recessive disorder

with a multiple CRA and CRC phenotype (Al-Tassan et al., 2002, Sampson et al., 2003). Whether rare inherited non-synonymous variants in *APC* might act as low penetrance alleles remains highly speculative; however, the variant p.I1307K has been shown to create a hypermutable tract that predisposes to somatic mutations (Laken et al., 1997) and p.E1317Q has been shown to be over-represented in the germline of patients with multiple CRAs (Lamlum et al., 2000) and has also been identified as a somatic change in sporadic CRC (Frayling et al., 1998). *APC* encodes a large protein which plays a role in signal transduction in the wnt-signalling pathway and it has been proposed that a specific level of β-catenin signalling, mediated by selection for *APC* genotypes that retain some 20 amino acid β-catenin down-regulating repeats, is required for colorectal tumour formation (Albuquerque et al., 2002).

To address the potential role of rare non-synonymous variants of *APC* in inherited predisposition to CRAs, we sequenced and compared the frequency of these variants in the β-catenin down-regulating domain in 691 unrelated North American patients with CRAs and 969 matched healthy controls. We also assessed the performance of an optimised Staden package in these genomic analyses.

5.2 Materials and Methods

5.2.1 Samples

To assess the four software packages DNA was extracted from the venous blood samples of 67 patients with multiple CRAs with or without cancer recruited from regional UK polyposis registers in Birmingham, Cambridge, Cardiff, Liverpool, Manchester, Southampton and Surrey

5.2.2 PCR

We analysed fragments spanning the coding region of *APEX1* (Accession number NC_003076.4), *MBD4* (NC_000003.10), *MPG* (NC_000016.8), *NEIL3* (NC_000004.10), *NUDT5* (NC_000010.9, *PCNA* (NC_000020.9)*, POL λ* (NC_000010.9)

and *TDG* (NC_000012.10). For *NUDT5*, the fragment covering exon 4 was not included due to poor amplification.

Reaction and thermal cycling conditions are described in Section 2.4.2, primers and annealing temperatures for *APEX1*, *MBD4, MPG*, *NEIL3, NUDT5, PCNA, POL λ* and *TDG* genes are in Appendix D.

5.2.3 Sequencing

PCR products were purified as described in Section 2.4.3 and sequenced as described in Section 2.4.4. Purification of sequencing products was carried as described in Section 2.4.5. Mutations were described according to the established nomenclature system (Antonarakis, 1998).

5.2.4 Base Calling

All four programs used the base calls from the ABI software, so any call errors were carried through to the software analysis. Sequencher's 'call secondary peak' function over rides the base calls and provides its own call. Mutation Surveyor has a similar function. Staden creates its own confidence calls for variant identification - if a variant has been missed by the ABI calling it will retain the wrong base call in the 'Contig view' but it can still be highlighted by the software as a high or low confidence variant. Where the ABI base caller has introduced a false positive call, Staden does not necessarily highlight them for checking as it makes its own assessment. InSNP retains the ABI base calls throughout. To identify SNPs not called by the ABI software, it provides a SNP plot from which SNPs can be picked out by the user.

Table 5.1: PCR Primers.

Primers for the APC β-catenin down regulation domain.

Gene/primer name	Primer Sequence (5' to 3')	Product Size (bp)	Annealing Temp. (ºC)
APC 01 For	CAGTGAGAATACGTCCACACC	513	60
APC 01 Rev	CTAAACATGAGTGGGGTCTCC		
APC 02 For	TCAGCTGAAGATCCTGTGAGC	574	60
APC 02 Rev	GGCTGGATGAACAAGAAAATCC		
APC 03 For	CTCCTCAAACAGCTCAAACC	564	60
APC 03 Rev	CATGCTTTTGGGGTTGCAACC		
APC 04 For	TCTGCCATGCCAACAAAGTC	490	62
APC 04 Rev	TCACACGGAAAGGCTTGTGAC		
APC 05 For	TACAGATGAGGCTCAAGGAG	598	60
APC 05 Rev	GGTAACTTTAGCCTCTGATTCC		
APC 06 For	CATCATTACACGCCTATTGAAGG	667	60
APC 06 Rev	TGGGACTATGTTTTTCATTATCACC		

5.2.5 Sequencher

Manual - For a single fragment, all the .ab1 files were imported from the 'File' menu into a Sequencher file containing the consensus sequence of the gene. Using the 'Contig' menu, traces were aligned manually and poor sequence and trailing 'N's at the 3' and 5' ends were trimmed. All chromatograms were scanned by eye and variants recorded.

Automated - The .ab1 trace files were imported into a Sequencher file, trimmed according to the default parameters and assembled automatically. The 'compare to function' was used to compare the traces to a pre-determined reference trace using six increments from 90 to 15% (calls a SNP when the secondary peak height is 90-15% of that of the primary peak) and the resulting variance table was analysed.

5.2.6 InSNP

The PCR fragment consensus sequence was imported together with the PCR oligonucleotide sequences. The .ab1 trace files were imported according to their direction (the software requires forward and reverse reads to be imported separately). Traces were 'aligned' to the reference sequence and 'sorted' by quality. InSNP provides 4 relevant options for analysis; the 'Find Indels' and 'Trim' functions were carried out, 'Filtered' to remove the low quality sequence and 'find SNPs' to generate a list of variations.

5.2.7 Staden

Automated - .ab1 files were imported into preGap4 which recalls bases and assigns quality values, detects point mutations and assembles multiple reads into a Contig, the output of preGap4 is a database that is suitable to use as input into Gap4. Gap4 is the Contig editor of the Staden suite, within Gap4 a reference trace was selected for each analysis and a "Mutation Report" containing all information on potential variants was exported to a file.

Manual - In Gap4 file, for each analysis all highlighted loci were reviewed and edited before the new Mutation Report was exported. This review included the user making an informed decision regarding low quality SNPs, this helped to identify indels that were incorrectly named or over looked in the mutation report.

5.2.8 Mutation Surveyor

The GenBank reference sequence was imported into the software together with the .ab1 trace files. The 'R>S' button was used to compare all files to each other and the 'import' function was selected to start the analysis. All variants (high and low confidence) were reviewed before the SNP report was exported.

5.2.9 Analysis of the APC β-catenin down-regulating domain

5.2.9.1 Sequencing and Staden-based analysis of healthy controls

We sought sequence variants in 969 unrelated North American healthy controls that were matched to the unrelated North American non-FAP non-MAP patients (Table 5.2) for age (mean 41.6yrs for controls and 47.8yrs for patients), sex (~50% males/50% females in both groups) and self-reported ethnic backgrounds (data not shown). All control samples were anonymised. Peripheral blood DNA samples from healthy controls were sequenced over a ~2.5kb region of *APC* spanning the β-catenin down-regulating domain by Agencourt Biosciences using 6 overlapping PCR fragments (Table 5.1).

Agencourt analysis was performed using an optimised version of the Staden package (details of the optimisation would not be disclosed) and compared to our manual analyses using Sequencher.

5.2.9.2 Mutation detection of cases with CRAs

Myriad Genetics undertook comprehensive mutation analysis of the *APC* and *MUTYH* genes in 691 unrelated North American patients that were referred by their

physicians for genetic testing because of a clinical diagnoses of either FAP or "multiple" colorectal polyps. Peripheral blood DNA samples from all patients were sequenced for the entire ORF and splice sites of *APC* and exons 7 and 13 of *MUTYH* (which harbour the two common *MUTYH* mutations, (Sampson et al., 2005)) and screened for deletions at the *APC* locus by MLPA and by Southern blot analysis (Eliason et al., 2005). Samples with a single *MUTYH* mutation were then sequenced for the ORF and splice sites of *MUTYH* to identify biallelic mutations. Comparisons of numbers of patients versus controls harbouring variants were performed using either the Chi-squared test or Fisher's Exact test.

5.2.9.3 *In silico* analyses

Predictions using Polyphen, were based on *Homo Sapiens APC* and predictions using Align-GVGD, were based on a multiple sequence alignment (created using using t-coffee) of APC orthologues from *Homo Sapiens*, *Echinops Telfairi*, *Pan Troglodytes*, *Macaca Mulatta*, *Oryctolagus Cuniculus*, *Bos Taurus*, *Xenopus Tropicalis*, *Monodelphis Domestica*, *Rattus Norvegicus*, *Mus Musculus*, *Dasypus Novemcinctus* and *Loxodonta Africans*.

Table 5.2: Ethnicity.

Self-reported ethnic backgrounds of the patients and healthy controls from North America.

	Non_FAP Non-MAP patients	Healthy controls
African	5%	5%
Ashkenazi	4%	4%
Asian	2%	2%
Caucasian	61%	61%
Latin American	4%	4%
Native American	2%	2%
Near Eastern	1%	1%
None Specified	15%	15%
Other	5%	5%

5.2.10 Author's Contribution

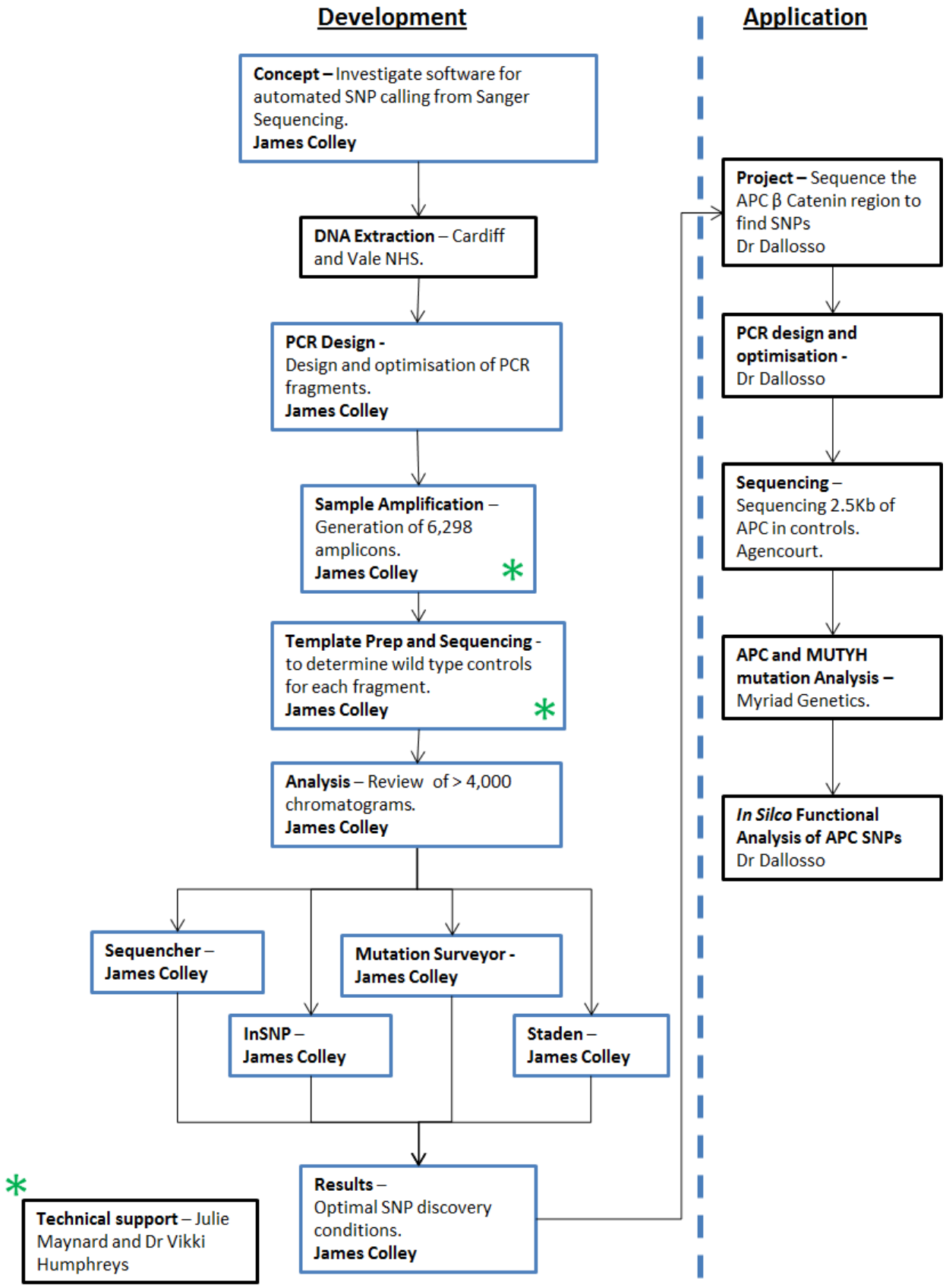


Figure 5.1: Author's Contribution to Chapter 5.

5.3 Results

5.3.1 Evaluating software for automated variant detection

We sought germline variants in 67 patients with multiple CRAs with or without carcinoma by sequencing the ORFs of *APEX1*, *MBD4, MPG*, *NEIL3, NUDT5, PCNA, POL λ* and *TDG*. In total, 64 fragments were amplified and sequenced, which generated 4,046 out of a potential 4,288 chromatograms. 242 (5.6%) chromatograms were unavailable due to PCR failures or insufficient DNA. Each chromatogram was analysed using four different variant detection programs (Sequencher, InSNP, Staden and Mutation Surveyor), each with a variety of sensitivity settings to generated 14 data sets for analysis. In total therefore, 56,644 analyses were carried out.

A manual analysis of the sequencing traces observed using Sequencher (used solely as a chromatogram viewer), was considered the 'gold-standard' and a reference for the subsequent analyses of the four software packages. In total, 77 SNPs were identified, 41 of which had MAFs <5% and 36 had MAFs >5%, thereby generating 878 samples with genotype calls heterozygous or homozygous for the rare allele.

5.3.2 Sequencher

We analysed the sensitivity of Sequencher for variant detection using the automated function together with six increments (from 90 to 15%) of the 'Call Secondary Peak' function. Using the automated function, 58/77 (75.3%) SNPs (Table 5.3, Figure 5.2) and 626/878 (71.3%) genotypes (Table 5.4, Figure 5.3) were called correctly, but with 3,944 false positives (Table 5.5, Figure 5.4). 19 SNPs were not identified accounting for 73 of the missing genotypes. By reducing the 'Call Secondary Peak' function from 90% to 15%, we increased the detection rate of SNPs from 54/77 (70.1%) to 67/77 (87.0%) and the correct genotype calls from 464/878 (52.8%) to 748/878 (85.2%), but at the expense of significantly increasing the number of false positives (from 4,450 to 44,903).

5.3.3 InSNP

Under the automated analyses, we identified 26/77 (33.8%) SNPs. Upon manual review of the 'Sequence Frequency' chart, a further 25 SNPs were identified, thus allowing the correct identification of 51/77 (66.2%) SNPs and 594/878 (67.7%) genotypes, together with 4,249 false positives. In total, 26 SNPs were missed accounting for 134 missing genotypes.

5.3.4 Staden

The automated output correctly identified 70/77 (90.9%) SNPs and 752/878 (85.6%) genotypes. User review identified a further 6 SNPs, generating a total of 76/77 (98.7%) correctly called SNPs and 849/878 (96.7%) genotypes with 818 false positives. One SNP was not identified accounting for 28 of the missing genotypes.

5.3.5 Mutation Surveyor

73/77 (94.8%) SNPs were correctly identified as 'high confidence' (blue) with 817/878 (93.1%) correctly called genotypes and 1,130 false positives. Three additional SNPs were identified as low confidence calls (red) together with 55/878 (6.3%) correctly called genotypes and 1,838 false positives. By combining these data, we identified 76/77 (98.7%) SNPs and correctly called 872/878 (99.3%) genotypes, together with 2,968 false positives.

5.3.6 Time Analyses

The average time taken to complete the manual analysis of 67 samples for a single amplicon was 15.4 minutes (range of 6-32 minutes depending on the number of variants that had to be recorded, Figure 5.5). On average, 3.3 minutes was taken for Sequencher under the automated setting which included data import, auto trim and auto assembly plus generation, review and exporting of the SNP Report. In 11 fragments, common polymorphisms were missed by the software but were detected by the manual review of the variation plot - in these cases the time taken

significantly increased as the names of the samples showing variation were recorded manually. InSNP provided a very fast analysis with data entry, analysis and output of a mutation report taking on average 1.3 minutes to perform. However, the reliability of SNP calling was acutely affected by the sequence quality - anything but high quality data caused significant analysis problems. The first stage of the Staden analysis procedure (Pregap4 and Gap4) took 2 minutes to perform. The average time taken for fragment analysis of the mutation report was 6.7 minutes including a manual review of all highlighted SNPs (high and low confidence calls); this package was not capable of fully automated SNP detection. Data entry, analysis and output of a mutation report in Mutation Surveyor took on average 10.8 minutes to perform.

5.3.7 An analysis of an optimised version of Staden for the identification of non-synonymous variants in the β-catenin down-regulating domain of APC

Using a re-sequencing approach, we sought rare non-synonymous variants in the ~2.5kb β-catenin down-regulating domain of APC in 969 unrelated North American healthy controls (Section 5.2.9). Sequencing was carried out by Agencourt using 6 PCR fragments and 5,814 .ab1 files were analysed by two methods. Agencourt used an optimised version of the Staden package and we used a manual, Sequencher-based, analysis.

In total, eighteen different rare (MAFs <2%) non-synonymous variants were identified in the β-catenin down-regulating domain in 37 healthy controls using both the Staden and manual Sequencher analyses (100% concordance). Thirty-six controls carried single heterozygous variants and one carried three heterozygous variants (p.T1633K, p.T1655A and N1761T) (Table 5.8, Figure 5.6). Five out of eighteen of these non-synonymous variants (p.I1307K, p.E1317Q, p.M1413V, p.K1454E and p.R1676G) were also identified in the patient cohort (details below), whereas the remaining thirteen variants were unique to the control group.

In total, five common polymorphisms (MAFs >5%) were identified in the β-catenin down-regulating domain using both the Staden and manual Sequencher

analyses. The genotype concordance rate was 100% (2,958/2,958 genotypes were identified correctly by both analyses).

5.3.8 Identification of non-synonymous APC variants in cases with CRAs

Myriad Genetics identified truncating mutations of *APC* in 178 patients and biallelic mutations of *MUTYH* in 33, confirming FAP and MAP, respectively. These patients with conventional CRA-predisposition alleles were excluded from the subsequent analyses. Amongst the 480 non-FAP non-MAP patients, 15.4% had ≤10 CRAs, 33.5% had 11-99 CRAs, 9.2% had ≥100 CRAs and 41.9% had 'multiple CRA, number unknown', as recorded at colonoscopy or colectomy.

In total, 81 of the 480 (16.9%) patients without FAP or MAP carried rare non-synonymous variants in the *APC* ORF. When non-FAP non-MAP patients were classified according to the number of CRAs, the group with 11-99 CRAs had a higher frequency of rare non-synonymous variants (18.6% of patients) as compared to the groups with ≤10 CRAs (13.5% of patients) or ≥100 CRAs (13.6% of patients) (Table 5.6).

5.3.9 Comparison of rare non-synonymous variants in the β-catenin down-regulating domain in non-FAP non-MAP cases versus controls

We found that significantly more non-FAP non-MAP patients carried rare non-synonymous variants in the β-catenin down-regulating domain as compared to the healthy controls (32/480 vs. 37/969, χ^2=5.74, *P*=0.0166) and this over-representation was highest in the non-FAP non-MAP patients with 11-99 CRAs (13/161 vs. 37/969, χ^2=5.914, *P*=0.015) (Table 5.6). In terms of individual variants, we did not observe an over-representation of p.I1307K in the non-FAP non-MAP patients versus controls (6/480 vs. 9/969, respectively). However, we did find that significantly more non-FAP non-MAP patients carried p.E1317Q as compared to healthy controls (13/480 vs. 11/969, χ^2=4.88, *P*=0.0272), although even when this variant was excluded from the analyses, significantly more non-FAP non-MAP

patients with 11-99 CRAs carried other rare non-synonymous variants in the β-catenin down-regulating domain as compared to controls (9/161 vs. 26/969, χ^2=3.887, *P*=0.0487). These data suggest that a proportion of non-synonymous variants in the β-catenin down-regulating domain of *APC* are likely to alter β-catenin signalling to promote tumourigenesis.

In contrast, the frequencies of rare synonymous variants in this region and common polymorphisms were almost identical between the non-FAP non-MAP patients and controls (Table 5.7). Both groups of patients also had similar self-reported ethnic backgrounds so it was highly unlikely that these findings could be attributed to population stratification.

5.3.10 In silico predictions of likely pathogenicity of APC SNPs

Non-synonymous variants could interfere with APC's roles in β-catenin degradation, mediation of intercellular adhesion, stabilization of the cytoskeleton, chromosome stability or regulation of the cell cycle and apoptosis (Al-Tassan et al., 2002, Fodde et al., 2001). We used the programs PolyPhen and Align-Grantham Variation/Grantham Deviation (Align-GVGD) (Mathe et al., 2006, Ng and Henikoff, 2006) to help predict the functional consequences of all 18 different non-synonymous variants that we identified in *APC*. *In silico* analyses using PolyPhen predicted that 9 non-synonymous variants were likely to be damaging whereas Align-GVGD predicted that 6 variants, 4 of which were also called 'damaging' by PolyPhen, were likely to alter function (Table 5.8). Together, these analyses predicted that 11/18 (61%) non-synonymous variants were likely to alter function.

5.4 Discussion

5.4.1 Comparison of Sequencher, Staden, InSNP and Mutation Surveyor for (semi-) automated variant detection

Here, we showed that Staden and Mutation Surveyor performed extremely well and correctly identified 76/77 (98.7%) SNPs and 96.7% and 99.3% of the genotypes, respectively. Both packages failed to identify a single SNP (c.444T>G,

p.D148E *in APEX1*); however this variant was in a region of poor signal intensity, but was accurately called by manual analysis (Figure 5.7). This illustrates the necessity of having high quality sequence data for accurate (semi-) automated mutation detection. Both of these packages also failed to fully genotype a SNP corresponding to c.117-32 delCAACA in *POLλ*. This variant occurs at the 5' end of the exon 2 PCR product. Staden assessed some of these fragment to be poor sequence (as opposed to 'heterozygous indels') and removed a region of good sequence (approximately 20-25 bases, with only 2 to 5 of these bases having a confidence value over 40) from the analysis. Mutation Surveyor detected this SNP but only in its heterozygous state and homozygotes were missed - the sequences 3' to the deletion were aligned correctly but the 5' sequences were misaligned and treated as poor sequence, and hence the gap was not identified. The relatively low specificity of Mutation Surveyor meant that 2,968 false positives were generated and needed to be checked manually causing the time taken for analysis to be the second longest. Staden, on the other hand, generated only 818 false positives.

5.4.2 Assessment of an optimised Staden package for variant detection in a large cohort of healthy controls

We have assessed an optimised Staden package for the analysis of >5,800 ab1 chromatograms generated from Sanger sequencing of 6 PCR amplicons (spanning 2,555bp) in 969 healthy controls. Staden and manual Sequencher analyses correctly identified 18 nonsynonymous variants, 11 synonymous variants and 5 common polymorphisms within this region and correctly genotyped 3,428/3,428 (100%) variant samples. We conclude that this optimised Staden package offers extremely high performance in a large scale variant detection project.

5.4.3 Role of non-synonymous variants in colorectal tumourigenesis

The role of non-synonymous mutation in *APC* is not clear with various publications supporting (Frayling et al., 1998, Gryfe et al., 1999, Hahnloser et al., 2003, Laken et al., 1997, Lamlum et al., 2000) and opposing (Gismondi et al., 2002, Popat et al., 2000, Sieber et al., 2003a) the claims the p.E1317q an p.I1317Q

predispose to CRA. In our population we found that significantly more non-FAP non-MAP patients with CRAs carry p.E1317Q compared to controls, suggesting a role for this variant in predisposing to CRA, we did not find an over-representation of p.I1307K. Notably our study found that significantly more non-FAP non-MAP patients carried a variety of rare non-synonymous variants, even when p.E1317Q was excluded from the analysis indicating that numerous rare variants in *APC* act as low penetrance disease alleles.

Recent studies (Fearnhead et al., 2004) have reported that rare variants in the Wnt signalling genes *AXIN1* and *CTNNB1* and the mismatch repair genes *hMLH1* and *hMSH2* along with p.E1317Q in *APC*, are over-represented in patients with multiple CRAs. Here, we present genetic and *in silico* data that show rare, inherited non-synonymous variants in *APC* also play a significant role in inherited predisposition to CRA. In support of this, further work from our laboratory (by Dr Azzopardi and Dr Dallosso) has shown that seven out of sixteen non-synonymous variants (43.8%) showed a significantly reduced ability to suppress β-catenin-regulated transcription *in vitro*, and both p.I1307K and p.E1317Q were found to be functionally compromised (Azzopardi et al., 2008). Further characterisation of these and other low penetrance alleles should therefore help contribute to our understanding of CRA and CRC-predisposition.

Table 5.3: SNP identification.

Number of SNPs correctly identified by each software package (Figure 5.2).

	Sequencher								InSNP		Staden		Surveyor	
GENE	Manual	Auto	90%	75%	60%	45%	30%	15%	Auto	Manual	Default	Edited	Blue	Red
APE1	4	2	1	3	3	3	3	3	1	2	4	4	4	4
MBD4	7	6	3	4	6	6	6	6	0	5	7	7	6	7
MPG	7	2	3	3	4	4	5	5	0	2	6	7	6	6
NUDT5	7	5	4	5	5	5	5	5	1	2	7	7	6	7
PCNA	5	5	5	5	5	5	5	5	3	4	5	5	5	5
POLL	11	10	9	9	10	10	10	10	0	6	10	10	11	11
TDG	15	8	9	10	11	13	13	13	5	9	12	15	15	15
NEIL3	21	20	20	20	20	20	20	20	16	21	19	21	20	21
Total	77	58	54	59	64	66	67	67	26	51	70	76	73	76

'Sequencher Manual' gives the number of variant genotypes correctly identified using a manual review of the traces (the 'gold-standard'). 'Sequencher Auto' gives the number of correct variant genotypes identified by the "Compare to, Reference" function. '90%' refers to those identified by adjusting the "Call Secondary Peak" function to 90% (likewise 75%, 60%, 45%, 30% and 15%). InSNP 'Auto' gives the numbers of correct variant genotypes identified in the InSNP report without any manual intervention. 'InSNP Manual' gives the number of correct variant genotypes identified by manual inspection of the base comparison images. 'Staden Default' gives the number of correct variant genotypes identified in the mutation report, 'Edited' are those added following a manual review of the putative SNPs. Surveyor Blue are the high confidence SNPs and Red are low confidence SNPs.

Table 5.4: Genotype identification.

Number of genotypes correctly identified by each software package (Figure 5.3).

			Sequencher						InSNP		Staden		Surveyor	
GENE	Manual	Auto	90%	75%	60%	45%	30%	15%	Auto	Manual	Default	Edited	Blue	Red
APE1	9	5	3	5	8	8	8	8	3	4	9	9	9	9
MBD4	35	20	6	11	20	20	20	20	0	19	32	35	24	35
MPG	16	4	7	7	8	8	10	10	0	3	11	16	11	16
NUDT5	64	61	29	47	60	61	61	60	41	57	64	64	62	64
PCNA	70	18	32	47	47	51	61	61	11	24	56	70	58	70
POLL	204	135	105	126	133	135	144	153	0	102	174	176	191	198
TDG	149	106	128	135	144	147	147	147	93	112	109	149	149	149
NEIL3	331	277	154	223	252	288	291	289	249	273	297	330	313	331
Total	878	626	464	601	672	718	742	748	397	594	752	849	817	872

'Sequencher Manual' gives the number of variant genotypes correctly identified using a manual review of the traces (the 'gold-standard'). 'Sequencher Auto' gives the number of correct variant genotypes identified by the "Compare to, Reference" function. '90%' refers to those identified by adjusting the "Call Secondary Peak" function to 90% (likewise 75%, 60%, 45%, 30% and 15%). InSNP 'Auto' gives the numbers of correct variant genotypes identified in the InSNP report without any manual intervention. 'InSNP Manual' gives the number of correct variant genotypes identified by manual inspection of the base comparison images. 'Staden Default' gives the number of correct variant genotypes identified in the mutation report, 'Edited' are those added following a manual review of the putative SNPs. Surveyor Blue are the high confidence SNPs and Red are low confidence SNPs.

Table 5.5: False Positives identified.

False Positives identified by each software package (Figure 5.4).

		Sequencher							InSNP	Staden	Surveyor	
GENE	Manual	Auto	90%	75%	60%	45%	30%	15%	Auto		Blue	Red
APE1	0	372	560	609	782	1,104	2,572	12,101	326	39	40	95
MBD4	0	332	140	147	159	229	849	8,035	1,092	15	84	234
MPG	0	285	482	533	621	741	963	1,957	401	50	87	212
NUDT5	0	65	91	80	78	93	150	2,747	300	59	63	118
PCNA	0	538	670	693	729	768	928	2,256	69	43	93	148
POLL	0	153	151	124	138	177	476	4,818	441	71	302	832
TDG	0	1,258	1,344	1,400	1,466	1,759	2,472	7,508	1,129	26	108	166
NEIL3	0	941	1,012	1,093	1,274	1,562	2,400	5,481	491	515	353	1,163
Total	0	3,944	4,450	4,679	5,247	6,433	10,810	44,903	4,249	818	1,130	2,968

'Sequencher Manual' gives the number of variant genotypes correctly identified using a manual review of the traces (the 'gold-standard'). 'Sequencher Auto' gives the number of correct variant genotypes identified by the "Compare to, Reference" function. '90%' refers to those identified by adjusting the "Call Secondary Peak" function to 90% (likewise 75%, 60%, 45%, 30% and 15%). InSNP 'Auto' gives the numbers of correct variant genotypes identified in the InSNP report without any manual intervention. 'InSNP Manual' gives the number of correct variant genotypes identified by manual inspection of the base comparison images. 'Staden Default' gives the number of correct variant genotypes identified in the mutation report, 'Edited' are those added following a manual review of the putative SNPs. Surveyor Blue are the high confidence SNPs and Red are low confidence SNPs.

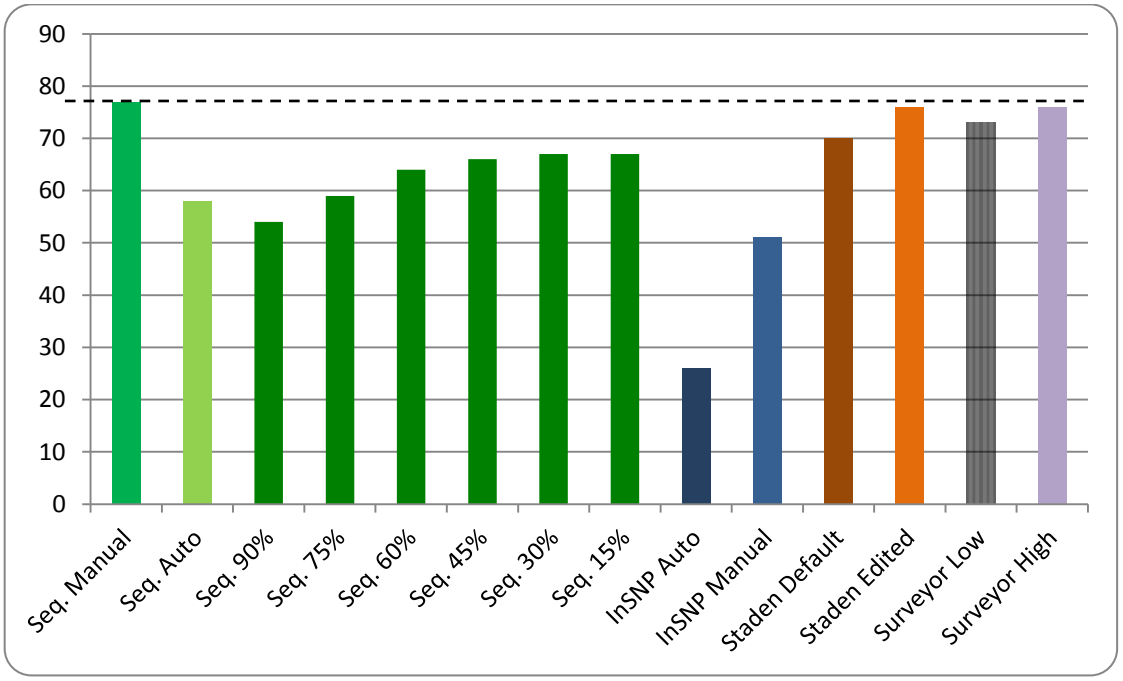


Figure 5.2: SNP Count.

Number of SNPs identified by the software packages under various sensitivity settings (Table 5.3). Shown are manual, automatic and 'Call Secondary Peak' increments for Sequencher (Seq.); automatic, manual and combined for InSNP; and high and low confidence calls plus a combined result for both.

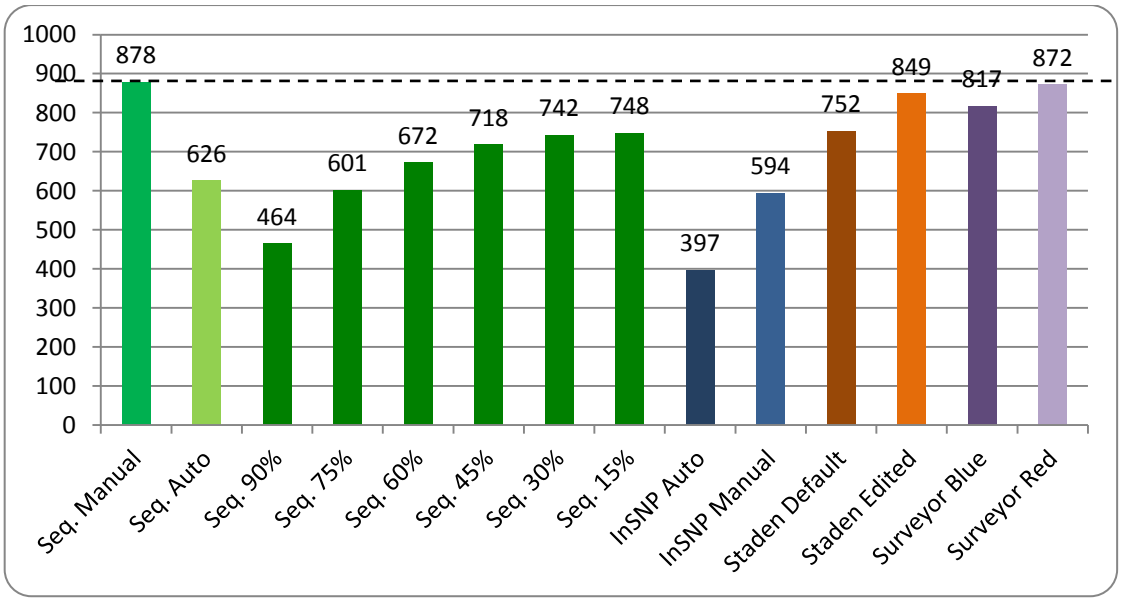


Figure 5.3: Genotypes.

Number of correct genotypes identified using the different programmes (Table 5.4).

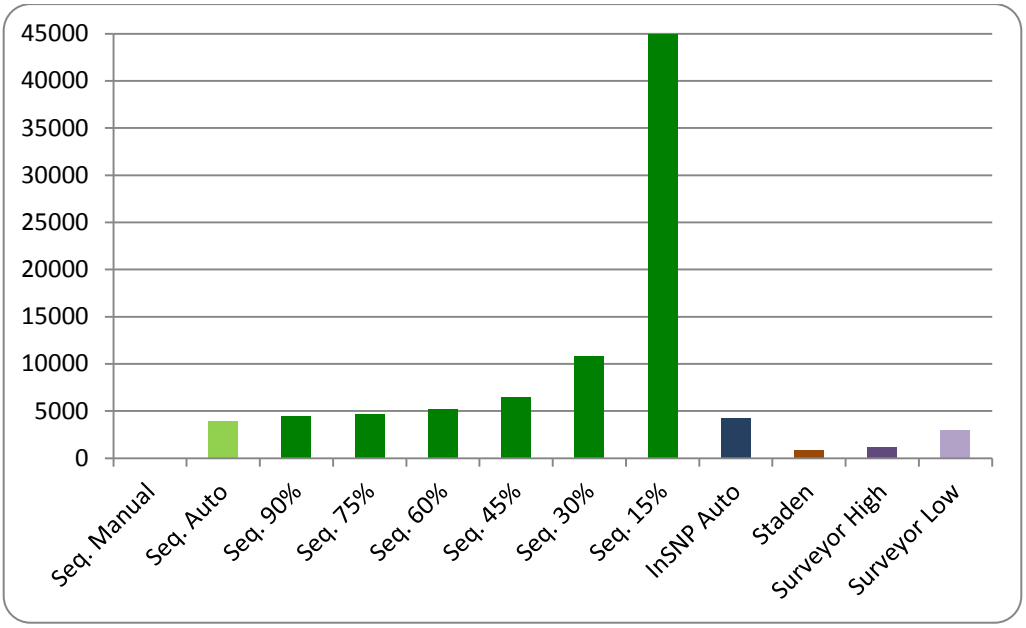


Figure 5.4: False Positives identified.

Number of false positives identified using the different programmes (Table 5.5)

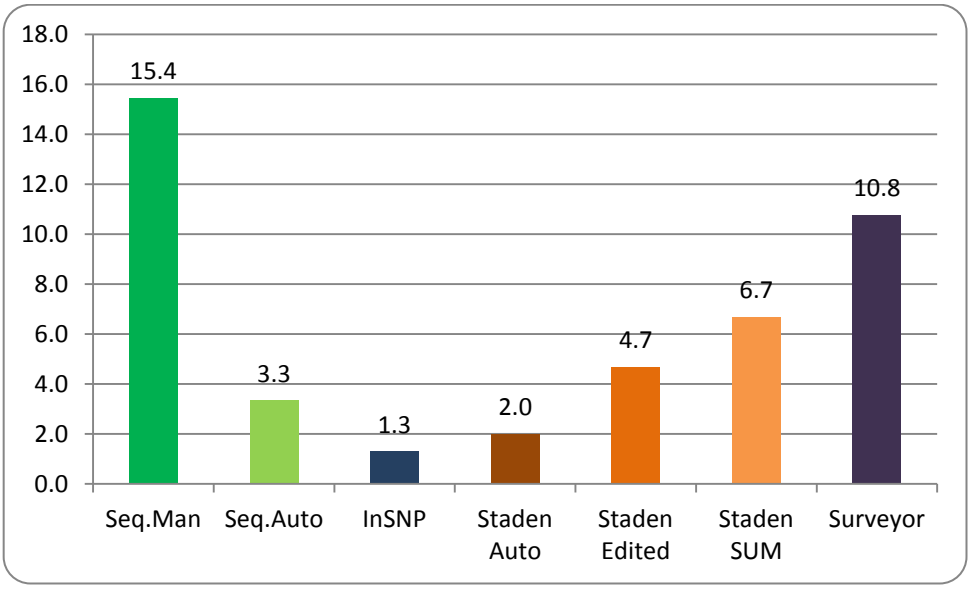


Figure 5.5: Analysis Times.

Average time taken (in minutes) to analyse a single exon for 67 samples for each programme.

Table 5.6: APC SNPs in Non-FAP non-MAP patients.

Inherited non-synonymous variants spanning the *APC* ORF in 480 non-FAP non-MAP North American patients with CRAs (classified according to number of adenomas) and in the β-catenin down-regulating domain in 969 North American healthy controls.

Category	Non-synonymous variants	Total (and frequency)
SNPs in bold lie within the β-catenin down-regulating domain		
non-FAP non-MAP patients (spanning the entire *APC* ORF)		
≤10 CRAs	p.P981R, **p.I1307K (2), p.E1317Q (2), p.T1445A,** p.G2502S, p.R2505Q, p.N2593S, p.S2621C	10/74 (13.5%)
11-99 CRAs	p.R216Q, p.P981R, p.V1125A, p.L1129S, p.T1160K, **p.I1307K (3), p.E1317Q (4), p.V1352A, p.M1413V, p.C1578G, p.I1579V, p.D1714N, p.G1921S**, p.P2158R, p.H2232D, p.A2274V, p.G2502S (5), p.R2505Q, p.I2573V, p.S2621C, p.A2795T	30/161 (18.6%)
≥100 CRAs	p.K150R, p.S643P, p.R653K, p.G2502S (3)	6/44 (13.6%)
Multiple CRAs, number unknown	p.K150R, p.E538V, p.P870S, p.C947S, p.P870S+p.M949I, p.L1129S, p.T1160K, **p.I1307K, p.E1317Q (7), p.A1446T, p.K1454E, p.P1467S, p.A1474T, p.I1572T, p.P1934L**, p.R2066G, p.I2329V, p.G2502S (10), p.I2541V, p.I2756V	35/201 (17.4%)
	Total	81/480 (16.9%)
		32/480 (6.7%)* in the β-catenin down-regulating domain
Healthy controls (β-catenin down-regulating domain)		
	p.A1247T, p.I1307K (9), p.E1317Q (11), p.K1363I, p.M1413V (3), p.K1454E (2), p.P1458S, p.T1493M, p.P1584S, R1589G, p.R1589C, p.T1633K+p.T1655A+p.N1761T, p.R1676G, p.S1730F, p.Q1916K, p.I1975F	37/969 (3.8%)* *χ^2=5.74, *P*=0.0166

In total, 54 different rare non-synonymous variants in APC were identified in this study. Each non-synonymous variant was identified in a single patient/control within each group unless otherwise indicated in parentheses.

Table 5.7: Frequency of rare synonymous variants in patients and healthy controls.

Eleven different rare synonymous variants were found spanning the β-catenin down-regulating domain in 55 (5.7%) healthy controls.

	non-FAP non-MAP patients	**Healthy controls**
β-catenin down-regulating domain	6.0%	5.7%

Table 5.8: *In silico* predictions.
In silico predictions of the likely functional consequences of the non-synonymous variants identified in *APC*.

Non-synonymous variant	**PolyPhen prediction**	**Align-GVGD prediction**
p.A1247T	benign	neutral 1
p.I1307K	**possibly damaging**	neutral 2
p.E1317Q	benign	neutral 1
p.K1363I	**probably damaging**	**deleterious 1**
p.M1413V	**possibly damaging**	**deleterious 2**
p.K1454E	benign	**deleterious 2**
p.P1458S	benign	neutral 2
p.T1493M	**possibly damaging**	**deleterious 1**
p.P1584S	**probably damaging**	**deleterious 1**
p.R1589G	**probably damaging**	neutral 2
p.R1589C	benign	neutral 2
p.T1633K	**possibly damaging**	neutral 1
p.T1655A	benign	**deleterious 2**
p.R1676G	benign	neutral 1
p.S1730F	**possibly damaging**	neutral 1
p.N1761T	**possibly damaging**	unclassified
p.Q1916K	benign	neutral 2
p.I1975F	benign	unclassified

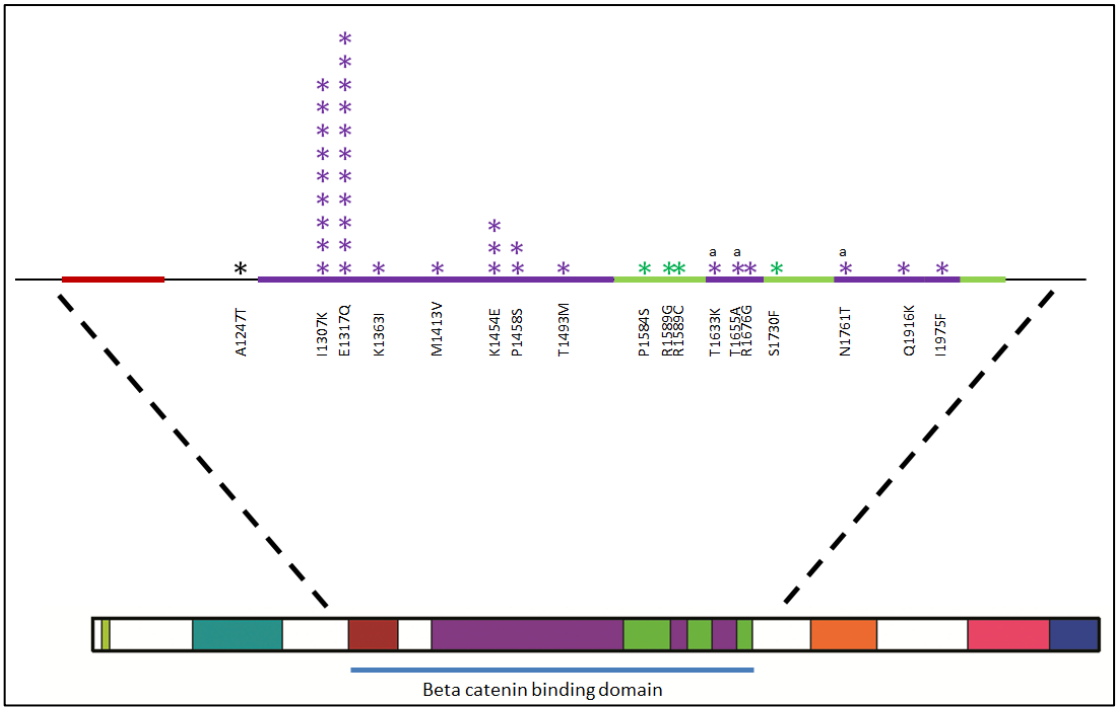


Figure 5.6: APC β-catenin down-regulating domain SNPs.

Distribution of inherited *APC* non-synonymous variants in 480 non-FAP non-MAP patients with CRAs and in 969 healthy controls (spanning the β-catenin down-regulating domain). 36 patients carried a single variant and 1 patient[a] carried three variants. Functional domains are coloured: ■ oligomerisation domain (codons 6-57), ■ armadillo region (codons 453-767), ■ 15 amino acid repeat β-catenin binding domain (codons 1020-1169), ■ 20 amino acid repeat β-catenin down-regulating domain (codons 1262-2033), ■ SAMP repeats/axin binding domain (codons 1562-2056), ■ basic domain (codons 2200-2400), ■ EB1 binding domain (codons 2559-2771), and, ■ HDLG binding domain (codons 2771-2843) (domains not drawn to scale). Non-synonymous variants are plotted as stars, coloured according to the domains in which they lie (black stars denote non-synonymous variants that do not lie within known functional domains).

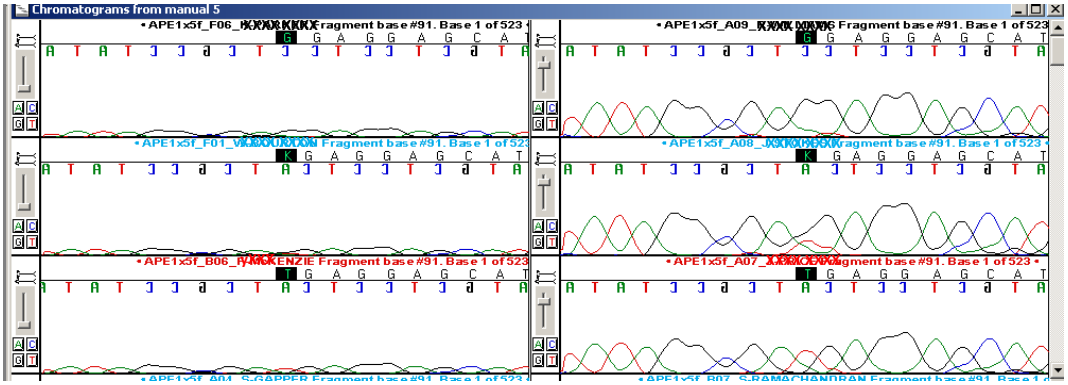


Figure 5.7: APEX1 p.D148E.

The variant c.444T>G, p.D148E in *APEX1* occured in a region of poor quality sequence. The left hand side shows the defualt view, the right hand side has been enhanced. Only manual analysis was able to identfy this variant.

6 Chapter six

Identification and characterisation of common nonsynonymous variants in every DNA repair gene in the human genome

6.1 Introduction

The ever increasing pool of on-line SNP data means that 'in house' laboratory analyses may be unnecessary when identifying common variations. Though numerous databases exist (1000Genomes, HapMap, Ensembl, GeneSNPs, CGAP, MutationDiscovery.com), the majority refer to the dbSNP 'refSNP cluster ID number' (rs) number as the definitive identifier. dbSNP is a central repository for genetic variation covering an array of species which is only limited by the data that is submitted (Sherry et al., 2001). The database was established in 1998 and all laboratories are encouraged to contribute their own data. Despite the name, the database is not exclusive to SNPs as it also contains entries for variations such as indels and microsatellite markers, but the majority of records do refer to SNPs as they are the most common source of genetic variation. SNPs are considered to be validated within dbSNP if they fulfil any one of five criteria: (i) it has been submitted more than once by independent investigators, (ii) it has been submitted with genotype data showing the minor allele in at least two chromosomes, (iii) it has been confirmed as validated by the submitter, (iv) all alleles have been observed in at least two chromosomes each, or, (v) it has been genotyped by HapMap.

6.1.1 Utility of dbSNP (build 129) as a resource for common ORF variation

In 2003, Carlson *et al.(2003)* showed that build 104 of dbSNP was not sufficient to create a panel of SNPs for whole genome association studies despite containing information on 2.7 million human SNPs. However, there has since been an explosion of data deposited into this database. Build 127 (March 2007) of dbSNP

contained nearly 12 million human SNPs, 5.7 million (47.5%) of which had been validated and build 129 contained 18 million human SNPs, 7 million (41.7%) of which had been validated (Figure 6.1). Between builds 130 and 131 the number of human SNPs more than doubled for validated and un-validated entries largely due to the submission of SNPs from a number of samples which have undergone whole genome sequencing (i.e. the Watson and Venter genomes). The current build of dbSNP at the time of writing (build 132, April 2011) contains information on nearly 20 million validated human SNPs. Here, we evaluated the data held in dbSNP for common variants in the Caucasian population within the open reading frames of ten DNA repair genes.

6.1.2 Automated searching of dbSNP

Manually mining SNPs from dbSNP is a laborious process with each SNP page requiring individual scrutiny to identify frequency information for specific populations. Several programmes have been designed to automate this process (PolyMAPr, SNPselector, SNPper, SNPHunter, and Pupasuite); however, all of these have significant pitfalls. PolyMAPr (Freimuth et al., 2005) was unavailable at the time of analysis, SNPselector (Xu et al., 2005) used only dbSNP v126, SNPper (Riva and Kohane, 2002) and SNP Hunter (Wang et al., 2005) were limited to the analysis of a single gene at a time and Pupasuite (Conde et al., 2005) returned only a limited amount of frequency data in 'single gene mode' and none for the 'gene lists' mode. We therefore designed our own software to extract selected information from the database in an automated fashion.

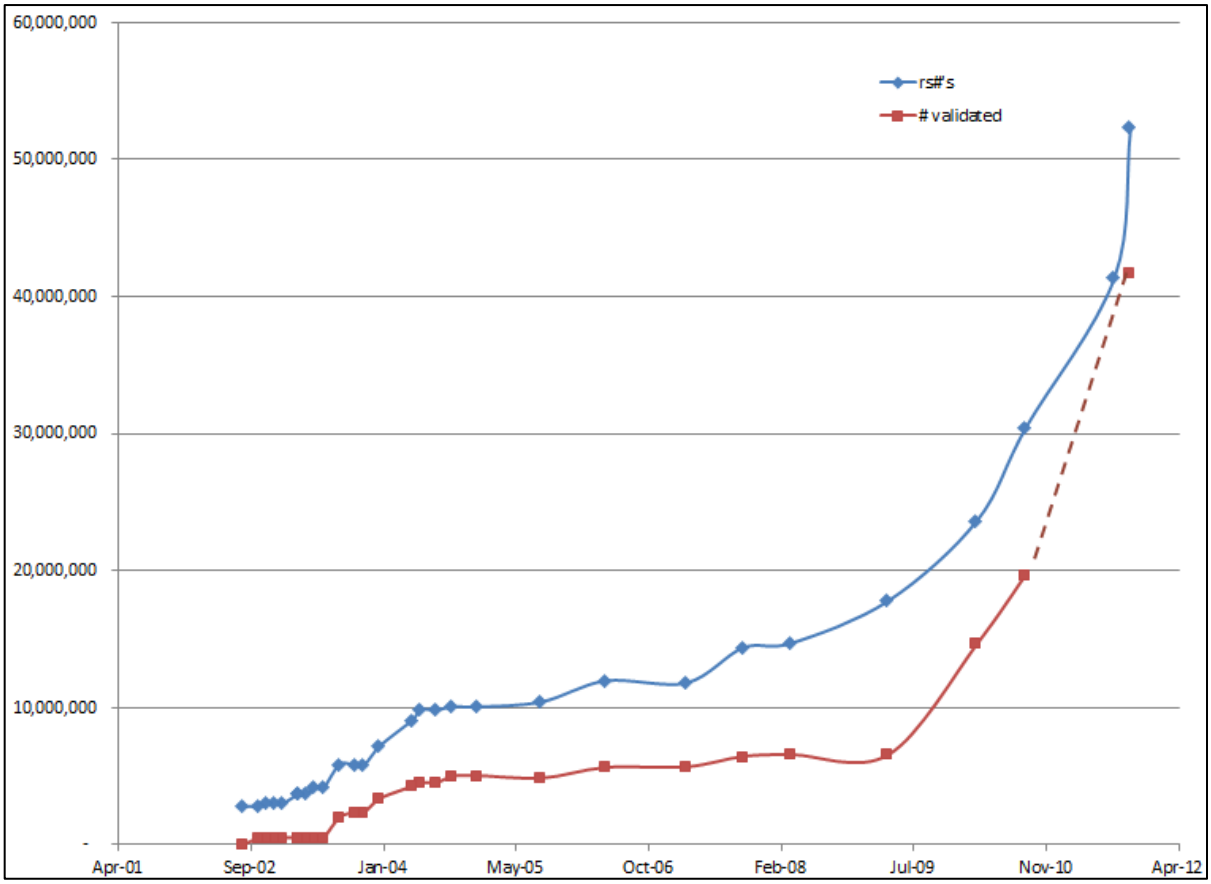


Figure 6.1: Human SNPs in dbSNP from build 106 to the present build (135).

Since July 2009 there has been a sudden and profound increase in both the number of validated and un-validated SNPs in the dbSNP database.

The process that determines the validated SNPs in build 134 was affected by a software bug (communicated by e-mail) so the data point is not plotted here, data points for build 133 to 135 have been joined with a dotted line.

6.1.3 Creating a functional resource

Damage to DNA is caused during replication, by environmental and dietary mutagens and by endogenous products of intercellular processes. Lodish *et. al.* (2000) estimated that up to a million lesions could occur per cell per day. As a consequence, the human genome is constantly undergoing repair and cells have developed several repair mechanisms to deal with the different forms of damage that may occur. Although it is well documented that rare high penetrance mutations in many of the genes from different DNA repair pathways predispose to cancers (e.g. BRCA1, BRCA2, MUTYH, PMS2, XPC.) the role of common nonsynonymous variants in these genes and their functional consequences are unclear. This is primarily due to a lack of sensitivity of the currently available DNA repair assays and a lack of appropriately profiled cell lines on which the assays can then be performed. Here, we aimed to profile 480 normal human lymphoblastoid cell lines for every common nonsynonymous SNP in every DNA repair gene in the human genome and to provide this resource free of charge to researchers to facilitate their functional analyses. We have also carried out *in silico* assessments of the likely functional consequences of these SNPs.

6.2 Materials and Methods

6.2.1 PCR

Reaction and thermal cycling conditions are described in Section 2.4.2, primers and annealing temperatures are in Appendix E.

6.2.2 Initial assessment of dbSNP

SNPs were initially sought in dbSNP with frequency information in the Caucasian population in a total of ten genes - *ERCC5* (Accession number NC_000013.9)*, XPA* (Accession number NC_000009.10)*, XPC* (Accession number NC_000003.10)*, RPA1* (Accession number NC_000017.9)*, RPA2* (Accession number NC_000001.9)*, GTF2H1* (Accession number NC_000011.8)*, GTF2H2* (Accession

number NC_000005.8)*, GTF2H3* (Accession number NC_000012.10)*, GTF2H4* and *GTF2H5* (Accession number NC_000006.10). We selected the 'Gene Model' and also BLAST aligned exons to dbSNP. In addition, we sequenced PCR fragments spanning the coding region of these ten genes in 23 Caucasian DNA samples from healthy controls, purchased from ECACC (Porton Down, Salisbury). SNPs were identified using the Mutation Surveyor package as described in Chapter 5 and described according to the established nomenclature system (Antonarakis, 1998).

6.2.3 Development of novel software to search dbSNP

We used the method of 'stripping' in which the source code for each html SNP page was filtered and processed to provide only the data that we specified. In this way any search returns the most recent data and a user can chose to rapidly repeat the process for subsequent builds if necessary. GeneID was found to be the only reliable method of searching dbSNP for two reasons; firstly, it limits the results instantly to human genes and secondly, searching by Gene names will return coding SNPs from a gene that shares the gene region but are actually intronic for the searched gene. Our PYTHON script was built to adhere to the dbSNP rules concerning automated query submission namely that individual queries were sent 2 seconds apart. Gene names were converted from their 'official symbol' to their Entrez GeneID using the US National Institutes of Health (NIH) Database for Annotation, Visualization and Integrated Discovery (DAVID) tool (release 2008). Our programme (Appendix F) was coded by Dr Jon Giddy in the School of Computer Science and Informatics, Cardiff University.

6.2.4 Cell lines and SNP profiling

We purchased DNA from 480 EBV-transformed lymphoblastoid cell lines from unrelated healthy individuals from ECACC. Fetch2.py was used to search dbSNP (build 129) for nonsynonymous SNPs from all 151 genes known to be involved in DNA repair (Wood et al., 2001) and that had a MAFs>4% in the Caucasian population (N.B. detailed discussion of the MAF selected is given in Section 7.2.3). SNPs were genotyped by the Illumina FAST-track service using Goldengate assays, by GeneService using Taqman assays or by KBiosciences using KASPar technology.

6.2.5 In silico analyses

In silico analyses were carried out using PolyPhen, AlignGVGD and SIFT, as described in Section 5.3.10. The three software packages attempt to predict the likely magnitude of a SNP on a protein sequence by examining basic biophysical properties and conservation of amino acid residues.

The NCBI database Homologene was used to source the protein sequences for each gene and their homologues. In order to provide a degree of normalisation, only mammalian sequences were used.

6.2.6 Author's Contribution

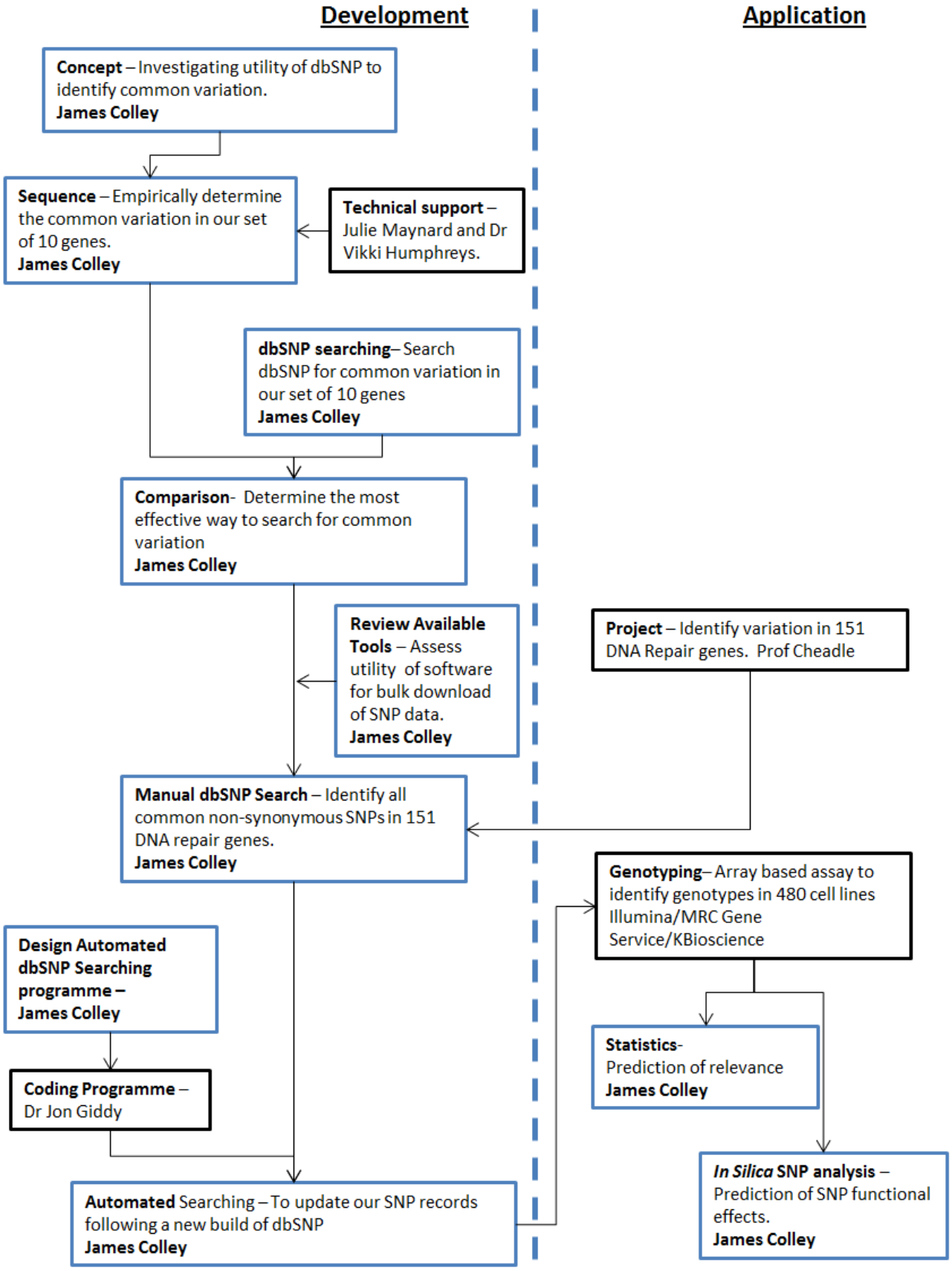


Figure 6.2: Author's Contribution of Chapter 6.

6.3 Results

6.3.1 Assessment of dbSNP

We evaluated the data held in dbSNP (build 129) for common variants in the Caucasian population within the ORFs of ten DNA repair genes (*ERCC5, XPA, XPC, RPA1, RPA2, GTF2H1, GTF2H2, GTF2H3, GTF2H4* and *GTF2H5)*. We compared the variants found in dbSNP to those found by sequencing the same genes in 23 DNA samples from healthy European controls. Using 23 control samples, we had >90% power to detect variants with MAFs >4%.

Mining of dbSNP revealed no coding region SNPs (with frequency information in the Caucasian population) in four of the genes analysed (*RPA2, GTF2H1, GTF2H3, GTF2H5*). For the remaining six genes *(ERCC5, GTF2H2, GTF2H4, RPA1, XPA* and *XPC*) we identified a total of 23 coding SNPs of which 8 were synonymous variants and 15 were nonsynonymous variants (Table 6.1). 19 variants were identified by reviewing the cSNP lists in the Gene Model section of the respective dbSNP page and the remaining 4 were found after BLAST analysis of the individual exons. Of the 23 SNPs, 17 had MAFs >4% in the Caucasian population.

In total, 105/115 (91.3%) fragments were successfully amplified and Sanger sequenced in a single direction with 92.65% coverage for the ERCC5 ORF, 95.70% for XPA, 94.26% for XPC, 95.87% for RPA1, 96.05% for RPA2, 91.35% for GTF2H1, 95.87% for GTF2H2, 94.95% for GTF2H3, 95.13% for GTF2H4 and 97.85% for GTF2H5. All common SNPs found by sequencing (in ≥2/23 samples; MAFs ~4%) were already present in dbSNP (Table 6.2) which suggests that this database is a comprehensive resource of common variation. As expected a small number of rare (MAFs<2%) SNPs (p.P19L in ERCC5, p.V147M and p.V147M in XPC) were identified by sequencing that we not present within dbSNP.

6.3.2 Development of a new programme for extracting SNP information

Nonsynonymous SNPs with a MAF of >4% in the Caucasian population were initially manually mined from dbSNP build 126 for 151 DNA repair genes (Wood et

al., 2001). When the build 127 was released, repeating the manual search was not practical, so an automated method was developed. Fetch2.py was developed and used on build 129, and returned results on 38 nonsynonymous SNPs across 85 DNA repair genes that had frequency information for Caucasian samples. 195 SNPs had MAFs >5%, 26 SNPs had MAFs of 4%, 16 SNPs had MAFs of 3%, 52 SNPs had MAFs of 2%, 37 SNPs had MAFs of 1% and 54 SNPs had MAFs <1%. No variants were identified in 47 genes analysed. (4% and 5% SNPs combined in Appendix G).

All SNPs identified by the manual review were also identified by our software. In addition, ten SNPs were identified by the software (in build 129) that were not present when manually reviewed (in build 127). These were rs45439799 MAF 0.010, rs2048074 MAF 0.210, rs5744857 MAF 0.436, rs5744944 MAF 0.436, rs5744990 MAF 0.158, rs5745022 MAF 0.411, rs4987206 MAF 0.001, rs4987207 MAF 0.009, rs4987208 MAF 0.010 and rs4251691 MAF 0.460. Four of these SNPs were identified as synonymous upon manual inspection. Annotation error of the *POLE* gene within dbSNP had caused Fetch2.py to extract the data in error.

Six of these new SNPs (including all 2 with high MAFs >5%) were verified as genuine by genotyping of 480 healthy control samples 4 SNPs were assayed by the Illumina assay, 2 by Taqman (Table 6.3).

Table 6.1: Result of searching 10 genes in dbSNP.

Gene	name	rs#	MAF
ERCC5	p.H46H	rs1047768	0.39
ERCC5	p.M254V	rs1047769	0.05
ERCC5	p.E399K	rs4150315	0.01
ERCC5	p.C529S	rs2227869	0.05
ERCC5	p.D320D	rs4150314	0.01
ERCC5	p.N879S	rs4150342	0.08
ERCC5	p.D1104H	rs17655	0.25
GTF2H2	p.I151M	rs2576895	0.4
GTF2H2	p.K133E	rs162961	0.36
GTF2H2	p.I199T	rs162959	0.08
GTF2H2	p.V236L	rs162956	0.26
GTF2H4	p.T98T	rs1419693	0.12
GTF2H4	p.R337Q	rs3218820	0.03
GTF2H4	p.P385P	rs1132408	0.11
RPA1	p.S352S	rs2230930	0.1
RPA1	p.S535S	rs2230931	0.16
XPA	p.R228R	rs16923331	0.02
XPC	p.L16V	rs1870134	0.01
XPC	p.L48F	rs3731062	0.03
XPC	p.R492H	rs2227999	0.1
XPC	p.A499V	rs2228000	0.3
XPC	p.R687R	rs3731151	0.25
XPC	p.K939Q	rs2228001	0.37

Table 6.2: Discovered SNPs.

Shading indicates the 10 common variants (MAFs ~4%) also identified by Sanger sequencing, all of which were already present in dbSNP and thereby validating this database as a comprehensive resource of common SNPs.

Gene	exon	c.Name	p.Name	rs. no	Freq. in dbSNP	Freq. by sequencing
ERCC5	1	c.56 C>T	p.P19L	not found	-	1/46
ERCC5	7	c.760 A>G	p.M254V	rs1047769	0.05	1/44
ERCC5	8	c.1195 G>A	p.E399K	rs4150315	0.01	0/46
ERCC5	8	c.1586 G>C	p.C529S	rs2227869	0.05	1/19
ERCC5	8	c.960 C>T	p.D320D	rs4150314	0.01	1/46
ERCC5	12	c.2636 A>G	p.N879S	rs4150342	0.08	0/46
ERCC5	15	c.3310 G>C	p.D1104H	rs17655	0.25	9/46
GTF2H2	4	c.135_136delAG	frameshift	not found	-	11/46
GTF2H2	8	c.453 A>G	p.I151M	rs2576895	0.4	23/46
GTF2H2	8	c.397A>G	p.K133E	rs162961	0.36	23/46
GTF2H2	10	c.596 T>C	p.I199T	rs162959	0.08	4/46
GTF2H2	11	c.706G>C	p.V236L	rs162956	0.26	2/46
GTF2H4	4	c.294 A>C	p.T98T	rs1419693	0.12	6/46
GTF2H4	11	c.1010 G>A	p.R337Q	rs3218820	0.03	0/46
GTF2H4	13	c.1155 C>A	p.P385P	rs1132408	0.11	1/42
RPA1	15	c.1605T>C	p.S535S	rs2230931	0.16	1/46
XPA	6	c.684A>G	p.R228R	rs16923331	0.02	0/46
XPC	1	c.46C>G	p.L16V	rs1870134	0.01	0/46
XPC	2	c.142 C>T	p.L48F	rs3731062	0.03	1/44
XPC	4	c.439 G>A	p.V147M	not found	-	1/46
XPC	9	c.1475G>A	p.R492H	rs2227999	0.10	4/46
XPC	9	c.1496C>T	p.A499V	rs2228000	0.30	11/46
XPC	11	c.2061 A>G	p.R687R	rs3731151	0.25	9/44
XPC	16	c.2815A>C	p.K939Q	rs2228001	0.37	23/46

Table 6.3: Genotype Assays.

SNP Name	MAF	Assay
rs45439799	0.01	Illumina
rs2048074	0.21	Taqman
rs5744857	0.436	Not assayed*
rs5744944	0.436	Not assayed*
rs5744990	0.158	Not assayed*
rs5745022	0.411	Not assayed*
rs4987206	0.001	Illumina
rs4987207	0.009	Illumina
rs4987208	0.01	Taqman
rs4251691	0.46	Illumina

* Synonymous

6.3.3 Generating a resource of cell lines profiled for every common nonsynonymous SNP in every DNA repair gene in the human genome

221 SNPs with (according to dbSNP) MAFs >4% were submitted for '*in silico* locus conversion' on the Goldengate platform (Illumina). 46 SNPs failed locus conversion either because of their proximity to other known SNPs (within 60bp of the SNP loci) or because Illumina's assay prediction software gave them a low chance of success. 175 SNPs were subsequently genotyped in DNA samples from 480 EBV-transformed lymphoblastoid cell lines from unrelated healthy individuals by Illumina. 33 of the 'locus conversion failed' SNPs were subsequently genotyped via ABI TaqMan assays and 6 were genotyped with KBioscience's KASPAR assays. Only 1 SNP could not be genotyped by any assay due to the presence of several common SNPs in the immediate adjacent sequence.

In total, 135 of the SNPs (Appendix H were found to have MAFs ≥4% in the cell lines described herein and in our population of 2,200 European control DNAs (Section 7.2.1) that were genotyped at the same time. 80 SNPs had a lower than the expected MAFs 26 of which were monomorphic.

6.3.4 In silico analyses predicting the functional effect of SNPs

134/135 SNPs underwent in silico analysis with the 3 most commonly used programmes, NEIL3 p.R381-/R was not suitable for analysis as the recessive allele creates either a synonymous change or a deletion of the loci, both options are not analysed by the software. Polyphen predicted that 38/134 (28.4%) of the nonsynonymous variants with MAFs>4% (Appendix I) were damaging or possibly damaging, with the remainder being classified as benign. AlignGVGD predicted that 22/134 (16.4%) of the nonsynonymous variants were likely to have a functional effect (C45, C55 or C65) 23 (17.2%) were possibly functional (C35, C25 C15) and 89 (66.4%) were classified as neutral variations (C0). SIFT predicted 39/134 (29.1%) of the variants as damaging and 95 (70.9%) were 'tolerated'.

Overall, 72/134 (53.7%) of the variants were predicted by one or more of the algorithms to be alter protein function. 19/134 (14.2%) of the variants were predicted by two algorithms to be damaging and 15 (11.2%) SNPs were predicted by all three algorithms to be damaging. 62/134 (46.3%) SNPs returned results of 'no functional effect' with all algorithms.

6.4 Discussion

6.4.1 Assessment of dbSNP

We showed that from a pilot study of ten genes, dbSNP contained a comprehensive amount of data for common sequence variation. Furthermore, by exploiting our novel software, we are able to rapid search this resource and download lists of relevant SNPs. However, it is clear that there are still limitations to this database:

(i) False Positive Rate

A number of groups have investigated potential false positive data within dbSNP and found that a significant proportion of entries were unreliable. Mitchell (2004) reviewed four of the publications (Carlson et al., 2003, Cutler et al., 2001, Gabriel et al., 2002, Reich et al., 2003) and estimated that 15-17% of SNPs were false positives caused by the submission of sequencing errors or paralogous sequences (Dvornyk et al., 2004, Musumeci et al., 2010).

(ii) False Negative Rate

Three SNPs were not found in dbSNP that were present in our population (Table 6.2). Two of these had low MAF (both 2%) but the third has a MAF of 24% (*GTF3H2* c.135_136 del AG) - this variant does exist in dbSNP but the frequency information for the Caucasian population was absent.

(iii) Full model vs. abbreviate lists

Five SNPs were listed on the *XPC* full 'Gene Model' list and labelled as exonic but did not appear on the abbreviated cSNP list; this is attributed to a known annotation error that dbSNP is trying to resolve.

(iv) Validation

In our study, we only assessed those SNPs within dbSNP that had frequency information in the Caucasian population. However, if we considered all SNPs (regardless of whether frequency data was present) we found 117 coding SNPs (build 129) 85 of which were nonsynonymous or frameshift variants. Only 40 of these were considered validated by dbSNP and this includes five which were also validated by the criteria in HapMap; however, further analyses of HapMap revealed these five SNPs are, in fact, monomorphic in all populations tested (rs4150295, rs3731140, rs3731126, rs3731063 and rs4150665). HapMart (an application for performing bulk downloads from the HapMap database (Smith, 2008) shows that out of 3.9 million Caucasian SNPs, 1.2 million are apparently 'monomorphic' including 28,279 cSNPs, 53% of which are nonsynonymous. Therefore, it is clear than many SNPs in dbSNP are very rare/private and will have limited utility in association studies.

6.4.2 Functional assessment of common nonsynonymous SNPs in DNA repair genes

Xi et al (2004) found that the scores observed from SIFT and Polyphen were highly associated with concordance between 298/478 (62.3%) of their variants. In using their approach we found 87/134 (64.9%) of variants to be concordant, using the same approach we found concordance of 64.2% between SIFT and Align GVGD and 58.2% between Align

GVGD and Polyphen. In combining all three programmes 67/134 (50.0%) SNPs are concordant.

In comparing the number of SNPs called Benign, Class C0 and Tolerated (in Polyphen, Align GVGD and SIFT respectively) we identified 62 SNPs. In comparing the number of SNPs called C65, Probably Damaging and Damaging we identified 5 SNPs, over all the concordance of the three programmes in this respect is 50.0% leaving 72 ambiguous results. By allowing some broader definitions (Possible/C0/Tolerated, Possible/C15-65/Damaging) concordance raises to 61.2% but still leaves 52 ambiguous calls.

It is important to note that when Align-GVGD analyses were undertaken with all homologous sequences from the Homologene database, all SNPs initially listed as C25 or greater (derived from an alignment of only mammalian species) were reduced to C0. Therefore, sequence variation introduced by non-mammalian sequences makes it very difficult for any residue change to be considered significant. So, although species diversity is required to identify conserved residues, use of all homologous sequences may limit the usefulness of these programmes.

As a consequence of the poor performance of *in silico* analyses and the lack of sensitive *in vitro* assays, the functional consequences of common nonsynonymous variants in DNA repair genes remains unclear. Most studies carried out to-date are either based on 'artificial' over-expression systems, knockdown human cell lines or genetically manipulated yeast or *E-coli* cells. We aimed to provide a resource of human cells lines which would provide a better and more natural environment for studying these nonsynonymous variants and, hopefully, would avoid the need to carry out transfections with artificial constructs. We profiled 480 normal human cell lines for every common nonsynonymous SNP in every DNA repair gene in the human genome. This resource will be made available free of charge to researchers to facilitate their analyses. Investigators can now select cell lines that only harbour their variant of interest and avoid those cells that harbour additional common variants within the same gene or pathway. By investigating 'purer' populations of 'variant' cells, researchers may be able to determine which variants have genuine functional effects.

7 Chapter seven

Investigating the role of common nonsynonymous variants in DNA repair genes in CRC-susceptibility and response to treatment

7.1 Introduction

7.1.1 Inherited predisposition to CRC and role of DNA repair

Inherited factors are thought to play a significant role in up to one third of CRCs (Section 1.6) but only a minority of these can be accounted for by established CRC predisposition genes (Kinzler and Vogelstein, 1996). <6% of cases carry high-penetrance germline mutations (Aaltonen et al., 2007) such as those found in the *APC* gene, mismatch repair (MMR) genes (*MLH1*, *MSH2*, *MSH6* and *PMS2*) and the BER gene MUTYH. It has been proposed that a substantial proportion of the remaining inherited predisposition is caused by either rare or common low penetrance variants (reviewed (Bodmer and Bonilla, 2008)). In support of the 'rare variant hypothesis', we have recently shown that individually rare, but collectively common, inherited nonsynonymous variants in *APC* (the gatekeeper of colonic proliferation) play a significant role in multifactorial inherited predisposition to CRAs, which can develop into CRCs if left untreated (Chapter 5 (Azzopardi et al., 2008)). Other investigators have also reported that rare variants in the Wnt signalling genes *AXIN1* and *CTNNB1* and the MMR genes *MLH1* and *MSH2* contribute to colorectal tumourigenesis (Fearnhead et al., 2004).

In the 'common-disease common-variant' model, the risk associated with the individual variants is small; however, they make a significant contribution to the overall disease burden by virtue of their high frequencies in the population. Moreover, by acting in concert with each other, they have the potential to significantly affect an individual's risk of developing CRC. The availability of comprehensive sets of tagSNPs that capture most of the common sequence variation has allowed genome-wide association studies (GWAS) for disease associations to be efficiently conducted. Several studies have conducted multistage GWA for CRC and, to-date, identified fourteen susceptibility loci mapping to 8q24

(Tomlinson et al., 2007), 18q21 (Broderick et al., 2007), 15q13 (Jaeger et al., 2008), 11q23 (Tenesa et al., 2008), 10p14 and 8q23.3 (Tomlinson et al., 2008), 14q22, 16q22, 19q13 and 20p12 (Houlston et al., 2008), 1q41, 3q26.2, 12q13.13 and 20q13.33 (Houlston et al., 2010).

We hypothesized that rare inherited variants in the MMR and BER pathways may predispose to CRAs and/or CRC, we also hypothesized that common nonsynonymous variants in these and other DNA repair pathways may also act as susceptibility alleles (Chapter 6). Here, we studied patients with aCRC from the UK-national, multi-centre, randomised controlled trials (RCTs) COIN and COIN-B. The COIN trial is a comparison between "COntinuous chemotherapy plus cetuximab, or INtermittent chemotherapy with standard continuous palliative combination chemotherapy with oxaliplatin and a fluoropyrimidine in first line treatment of metastatic colorectal cancer". All patients had either previous or current histologically confirmed primary adenocarcinomas of colon or rectum, together with clinical or radiological evidence of advanced and/or metastatic disease, or had histologically/cytologically confirmed metastatic adenocarcinomas, together with clinical and/or radiological evidence of a colorectal primary tumour. (See Section 7.2.2 and Table 7.1 for more details on these patients).

Table 7.1: Clinicopathological data for patients in COIN and COIN-B.

		COIN (%)	COIN-B (%)
n =		2070	113
Age at diagnosis	Mean (S.D.)	62.6 (9.7)	61.9 (10.5)
	<20	1 (0.1)	0 (0.0)
	20-49	207 (10.0)	12 (10.6)
	50-59	476 (23.0)	25 (22.1)
	60-69	852 (41.2)	50 (44.3)
	70-79	521 (25.2)	24 (21.2)
	80-89	13 (0.6)	2 (1.8)
Sex	Female	696 (33.6)	48 (42.5)
	Male	1374 (66.4)	65 (57.5)
WHO-PS	0	969 (46.8)	58 (51.3)
	1	948 (45.8)	46 (40.7)
	2	153 (7.4)	9 (8.0)
Primary Site	Colon	1116 (53.9)	69 (61.1)
	Rectum	653 (31.6)	32 (28.3)
	Rectosigmoid junction	297 (14.4)	12 (10.6)
	Other	3 (0.1)	0 (0.0)
	Missing	1 (0.1)	0 (0.0
Number of metastatic sites	0	14 (0.7)	1 (0.9)
	1	736 (35.6)	43 (38.1)
	2	813 (39.3)	50 (44.3)
	≥3	507 (24.5)	19 (16.8)
Metastatic sites	Liver only	458 (22.1)	24 (21.2)
	Liver + others	1,094 (52.9)	56 (49.6)
	No Liver	518 (25.0)	33 (29.2)

7.1.2 Pharmacogenetics of CRC and role of DNA repair

The treatment of metastatic CRC (mCRC) is rapidly improving. Average survival has increased from around 6 months with best supportive care (BSC) alone, through 10-12 months with single agent Fluorouracil (5FU) regimens (Maughan et al., 2002) and up to 16-20 months in randomised trials including irinotecan and/or oxaliplatin as well as 5FU (Douillard, 2000, Saltz et al., 2000). International results have shown increased response rates (RR) (31-56%) median progression-free survival (PFS) (6.5-9.0 months) and median overall survival (OS) (14.5-21.4 months) with combination chemotherapy in first line treatment. In addition, cetuximab, a chimeric monoclonal antibody targeted against epidermal growth factor receptor (EGFR) has been shown to improve OS (6.1 months in the cetuximab group compared to 4.6 months in the best supportive care group) and PFS in patients with CRC in whom other treatments have failed (Jonker et al., 2007). Despite these advances, it is clear that although some people respond well to chemotherapy and monoclonal antibody treatment, others do not (~50% fail to respond to first line treatment).

There are also considerable side effects associated with chemotherapy. For 5FU, these include gastrointestinal (diarrhoea in 50% of patients, nausea 43%, vomiting 28%, stomatitis 25%, and abdominal pain 12%), cutaneous (hand-foot syndrome 53%, dermatitis 10%), general (fatigue 24%), neurologic (headache 5%, paraesthesia 5%, dizziness 5%), cardiovascular (lower limb oedema 4%), haematologic (grade 3/4 neutropenia 4% and thrombocytopenia 2%) and biochemical (grade 3/4 bilirubin elevation 17%) problems. Oxaliplatin is administered together with 5FU (or capecitabine) and, in general, the adverse events with this combination are more frequent and more severe than with 5FU treatment alone.

Inherited genetic variations can affect a patient's response to chemotherapy and pharmacogenetics aims to use knowledge of these variations to 'tailor' therapy for improved response and reduced toxicity. To-date, most research has focused on single polymorphisms; however, a more comprehensive approach to predict treatment response is to consider genetic variation in entire biological and pharmacological pathways. Damage to cellular DNA is believed to determine the antiproliferative properties of platinum drugs. Oxaliplatin targets DNA forming platinum-DNA adducts which produce 1,2-GG intrastrand cross-links and the removal of this damage is mediated by the nucleotide excision repair

(NER) and homologous recombination repair pathways (Reardon et al., 1999). Fluoropyrimidines (FPs) have three possible mechanisms of action that are exerted by different metabolites. Two of these mechanisms of FP-mediated cytotoxicity act at the level of DNA (disruption of dNTP pools and the direct incorporation of FPs into DNA) and the most important consequences are the mutagenic effects of base analogues/mispairs in DNA and the fragmentation of DNA created in the cell's attempts to repair these lesions. The BER, MMR and double strand break (DSB) repair pathways have all been suggested to modify FP response (Meyers et al., 2003).

Given that numerous repair pathways have been implicated in removing the damage caused by these chemotherapeutic agents, we assayed germline DNA from patients in the COIN and COIN-B clinical trials for every nonsynonymous SNP with a MAF of >4%, in every repair gene characterised in the human genome (Chapter 6). In this way, we intend to generate 'DNA repair SNP profiles' which can then be related to response to and side effects from, the different chemotherapy regimens being used.

7.2 Materials and Methods

7.2.1 Samples

We analysed 2,186 samples from unrelated patients with aCRC from COIN (2,073 samples) and COIN-B (113 samples). All patients gave fully informed consent for their samples to be used for bowel cancer research (approved by REC [04/MRE06/60]). COIN patients were randomised 1:1:1 to receive continuous oxaliplatin and fluoropyrimidine chemotherapy (Arm A), continuous chemotherapy plus cetuximab (Arm B), or intermittent chemotherapy (Arm C). COIN-B patients were randomised 1:1 to receive intermittent chemotherapy and cetuximab (Arm D) or intermittent chemotherapy and continuous cetuximab (Arm E). In all patients, treatment was identical for the first 12-weeks apart from the choice of fluoropyrimidine together with the randomisation of ± cetuximab.

We also analysed 2,176 blood DNA samples from healthy controls from the UK Blood Services collection of Common Controls (UKBS collection) (Wellcome Trust Case Control Consortium 2007, Wellcome Trust Case Control Consortium and Australo-Anglo-American

Spondylitis Consortium 2007). This collection was funded by the Wellcome Trust grant 076113/C/04/Z, by the Juvenile Diabetes Research Foundation grant WT061858, and by the National Institute of Health Research of England. These samples were selected from a total of 3,092 samples within the UKBS collection that best matched the patients with aCRC in terms of place of residence within the UK. The UKBS samples had appropriate ethical approval (REC 05/Q0106/74).

7.2.2 Genotyping

We attempted to assay every nonsynonymous SNP with a MAF >4% in the Caucasian population in every DNA repair gene in the human genome (Wood et al., 2001). We identified relevant SNPs by searching dbSNP (Chapter 6). In total, we catalogued 221 DNA repair SNPs with MAFs >4% of which 46 failed *in silico* locus conversion on the Illumina BeadArray™ platform. The 4% figure was chosen to increase our chance of finding SNPs with a 5% or greater MAF in our population. Using Illumina's Fast-Track Genotyping Services (San Diego, CA) (Section 1.3.20) we assayed 175 nonsynonymous SNPs in 80 DNA repair genes (Section 6.3.3) of which 39 were from genes within the BER, MMR and their associated response pathways and 136 were from other genes known to function in DNA repair.

7.2.3 Statistical analyses (susceptibility alleles)

Genotypes for each variant were tested for deviation from the Hardy Weinberg Equilibrium (HWE) using a chi-squared test. Single marker association analyses were performed using the PLINK version 1.07 software (Purcell et al., 2007). To correct for multiple testing, we carried out Bonferroni correction.

7.2.4 Statistical analyses (pharmacogenetics)

The primary efficacy endpoint was; 12-week response, defined as complete response (CR) or partial response (PR) at 12-weeks, versus stable disease (SD) or progressive disease (PD). The primary endpoints for toxicity were: (i) a dose reduction or delay in chemotherapy in the first 12-weeks of treatment due to any toxicity except peripheral neuropathy (PN) and

(ii) grade ≥2 PN or dose reduction or delay due to PN versus grade 0 or 1 PN and no oxaliplatin dose modification in the first 12-weeks. Since neurotoxicity is a cumulative event, patients who received <12 weeks of treatment for reasons other than neurotoxicity were excluded.

Pharmacogenetic analyses were carried out using a co-dominant model with two degrees of freedom, tested using the likelihood-ratio chi-squared statistic. Analyses were adjusted for cetuximab use and type of fluoropyrimidine.

7.2.5 Author's Contribution

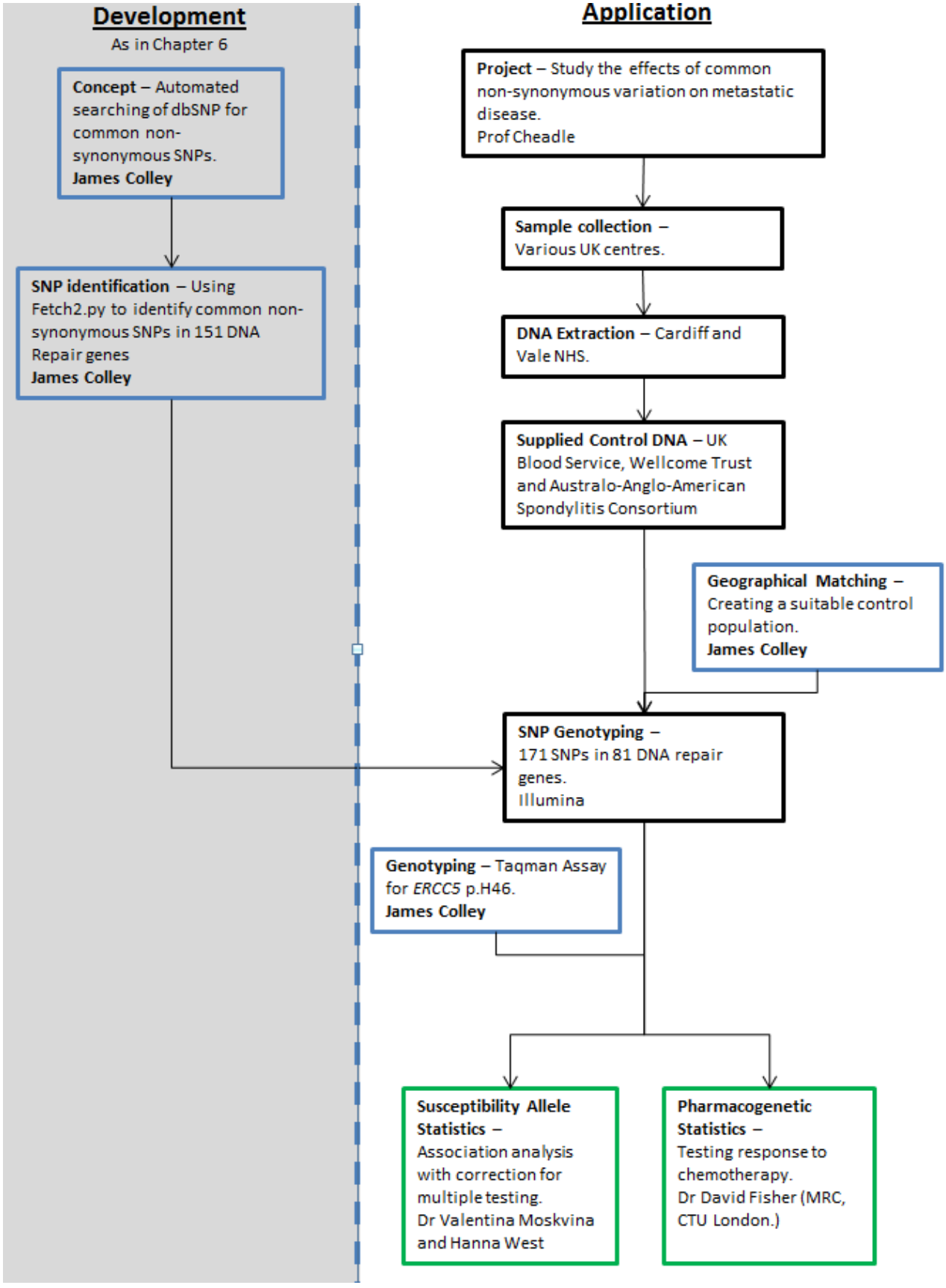


Figure 7.1: Author's contribution to Chapter 7.

7.3 Results

7.3.1 Identifying novel low penetrance susceptibility alleles

Our cohort (Section 7.2.2) has recently been used to validate four novel CRC-susceptibility loci (Houlston et al., 2010). Using ~2,200 cases and 2,200 controls and with a MAF>4%, we had over 95% power to detect alleles with odds ratio of 1.3 or greater. To further demonstrate the utility of this cohort to identify CRC-susceptibility alleles, we assayed a SNP from each of the ten other known loci identified from GWAS. aCRC cases identified as being of non-Caucasian origin (n=38) were excluded from the analyses and 3 samples failed genotyping. We independently validated five SNPs/loci using these samples (rs4939827 at 18q21, $P=1.14x10^{-4}$; rs16892766 at 8q23, $P=1.83x10^{-4}$; rs4779584 at 15q13, $P=2.86x10^{-4}$; rs10795668 at 10p14, $P=3.34x10^{-3}$ and rs6983267 at 8q24, $P=5.65x10^{-3}$).

In an attempt to identify novel CRC-susceptibility alleles, we assayed 175 nonsynonymous SNPs, representing 80 DNA repair genes, in our case-control series. Genotyping concordance rates for duplicate samples was 100% (9,216/9,216 genotypes were concordant) GenTrain scores ranged from 0.466 to 0.966 (Figures 7.2 and 7.3) and the overall genotyping success rate was 99.999% (724,695/725,760 genotypes were called successfully). We found that nine SNPs, representing seven genes, were significantly over-represented at the 4% level (all nine SNPs conformed to the Hardy Weinberg equilibrium) (Table 7.2). Only rs1805327, corresponding to p.E281G in RAD1 (X2=13.60, P=2.4x10-4) remained significant after rigorous correction for multiple testing (P=0.024). Genotype-specific ORs (OR=0.73, 95% CI 0.62 - 0.87) were most compatible with an over dominant model of disease susceptibility. The minor allele was associated with a decreased risk of CRC (OR_{ACvsAA}= 0.697, CI 0.58 - 0.83; OR_{ACvsCC} = 0.649, CI 0.26 – 1.62).

p.E281 in RAD1 is conserved in; *Pan Troglodytes*, *Canis Lupus Familiaries*, *Mus Musculus, Ratus norvegicus*, *Gallus gallus* and *Danio rerio* (Figure 7.4) and *in silico* analyses suggest that p.E281G has a likely functional effect with a PolyPhen score of 1.586 (possibly damaging) an align-GVGD score of class C65 (GD 97.85) (likely to interfere with function) and a SIFT score of 0.03 (affect protein function).

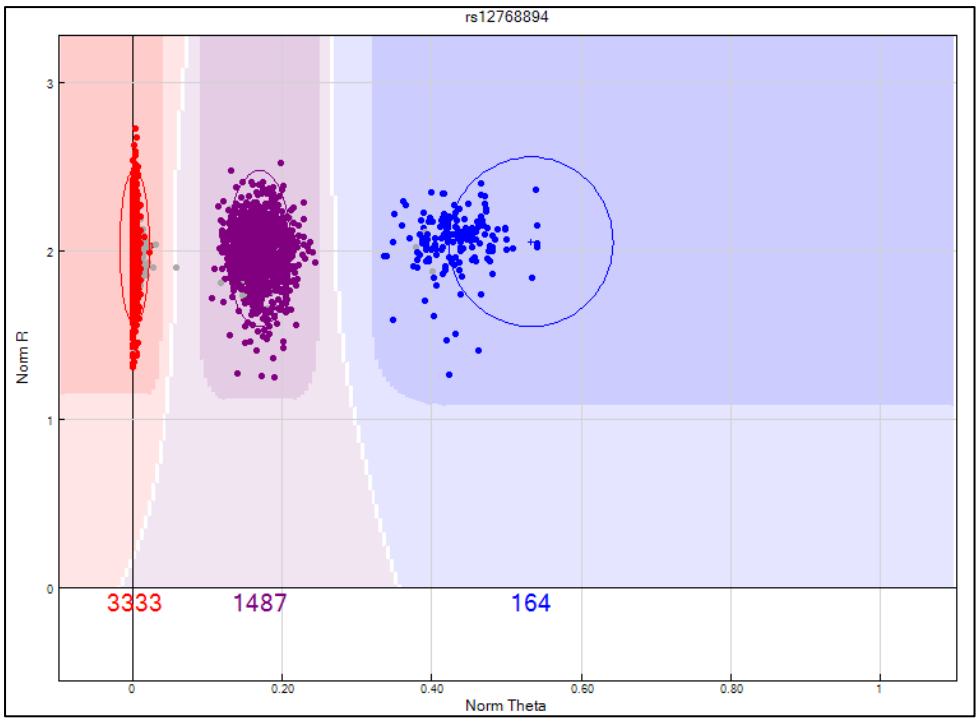


Figure 7.2: rs12768894, GenTrain 0.466.

The lowest GenTrain score from the Illumina genotyping of 175 SNPs shows clear separation between the three genotype groups.

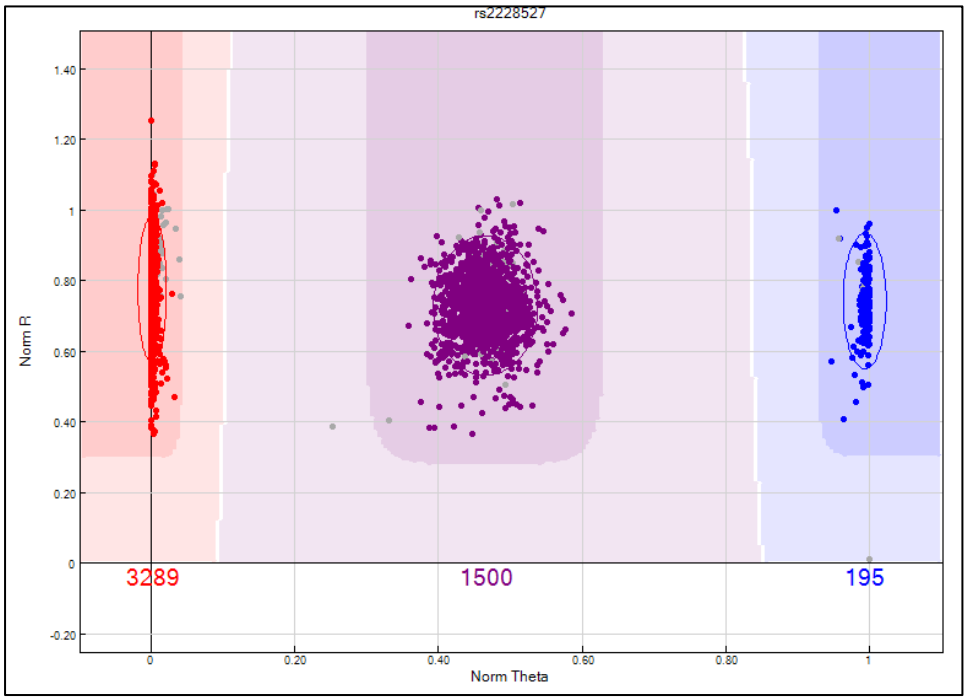


Figure 7.3: rs2228527, GenTrain 0.966.

The highest GenTrain score from the Illumina genotyping of 175 SNPs.

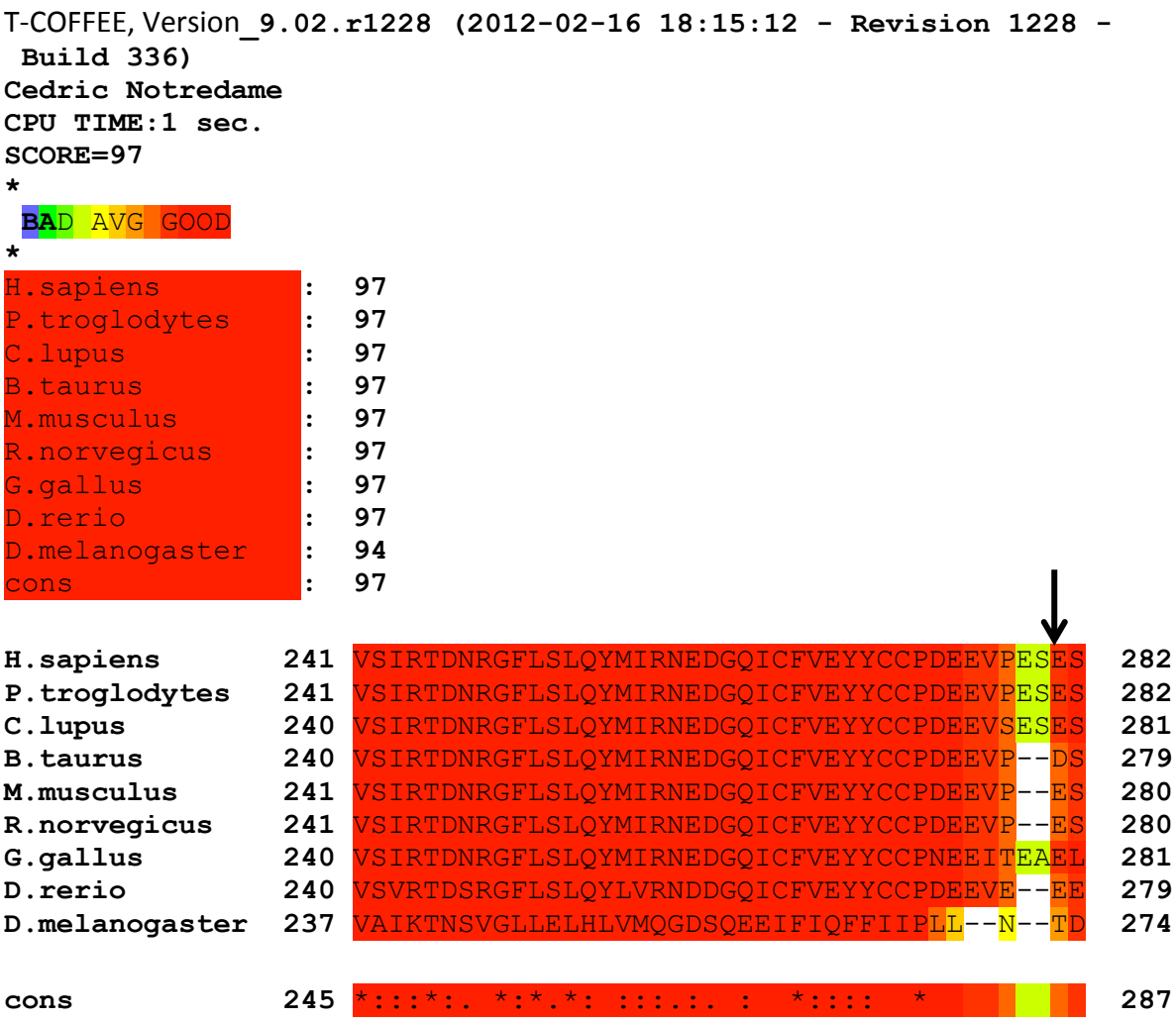


Figure 7.4: p.E281G region of RAD1.

3' section of the RAD1 CLUSTALW analysis shows 'GOOD' conservation over most of the RAD1 sequence. p.E281G (marked with an arrow) has some homology with the included species but it is not completely conserved.

7.3.2 Identifying alleles that affect response to, and side effects from, chemotherapy

We analysed all patients treated with chemotherapy, regardless of whether they also received cetuximab (all arms of COIN and COIN-B) for the 175 nonsynonymous variants in DNA repair genes. All three arms of COIN and the two arms of COIN-B had similar efficacy and toxicity outcomes at 12-weeks (Maughan et al., 2011) therefore these groups were combined to increase power.

The primary outcomes assessed were 12-week response, any toxicity (excluding PN) and PN. Overall, ~45% of patients responded after 12-weeks of treatment. Based on 2186 patients, we had >85% power to detect an OR of 1.31, corresponding to a 7% difference in response, for a variant with a MAF >20% (and an OR of 1.56, corresponding to an 11% difference in response, for a variant with a MAF >5%). The incidence of dose reduction or delay in chemotherapy in the first 12-weeks of treatment due to any toxicity was 35%; we therefore had >85% power to detect an OR of 1.32, corresponding to a 7% difference in toxicity, for a variant with a MAF >20%. The incidence of grade ≥2 PN or dose reduction or delay due to PN in the first 12 weeks was 13%; therefore we had >85% power to detect an OR of 1.43, corresponding to a 5% difference in PN, for a variant with a MAF >20%.

We found that 5 coding region variants in the DNA repair pathways were associated with 12-week response, 4 were associated with any toxicity and 5 were associated with PN at a 5% significance level, respectively (Table 7.3 for significant SNPs and Appendix J for all data). However, none remained significant after correction for multiple testing.

Interestingly, the most significant variants for 12-week response (p.N279S in *EXO1* and p.R399Q in *XRCC1*) were in genes that function in base excision repair. The SNPs are reported as “Probably Damaging” by Polyphen; however, AlignGVGD assigns categories C0 to both SNPs where a broad spectrum of species are aligned. Aligning only available vertebrate species re-assigns p.N279S as C45, p.R339Q remains a C0.

7.4 Discussion

DNA damage triggers signal transduction pathways involving sensor, transducer and effector proteins. Rad9, Rad1 and Hus1 form the '9-1-1' DNA damage sensor complex which is recruited to the site of damage by RAD17-RFC-2-5. The subsequent recruitment of specialised DNA polymerases and glycosylases by 9-1-1 is responsible for the repair of the lesions. 9-1-1 activates the transducers ataxia telangiectasia mutated (*ATM*) and ATM- and Rad3-related (*ATR*) which in turn activate the checkpoint effectors Chk2 (*ATM*) or Chk1 (*ATR*) to initiate cell cycle arrest or apoptosis (Zhou et al. 2000). Disruption of the 9-1-1 complex as a result of functional mutations in all three proteins has been shown to influence the cells ability to undergo cell cycle arrest as a result of DNA damage.

Recent studies have shown that the proteins encoded by many of the known high penetrance CRC predisposition genes interact with, and are stimulated by, the 9-1-1 complex. MUTYH interacts with Hus1 and Rad1, and co-localises to nuclear foci with Rad9 following oxidative damage (Shi et al., 2006). Each subunit of the 9-1-1 complex interacts with MSH2, MSH3 and MSH6 and can stimulate the binding activity of MSH2/hMSH6 with G/T-containing DNA (Bai et al., 2010). Rad9 also interacts with MLH1 and disruption of the interaction by a single-point mutation in Rad9 leads to significantly reduced MMR activity (He et al., 2008). RAD1 consists of 282 amino acids and p.E281G results in the substitution of an acidic residue to a neutral amino acid. Given that residues in the C terminal domain of RAD1 bind to the N terminal domain of RAD9 in the formation of the ring complex, it is possible that p.E281G affects ring formation (Dore et al., 2009).

Hannah West (a Ph.D. student in our laboratory) has carried out further analyses of *RAD1*. She has carried out direct sequencing of the *RAD1* open reading frame and promoter region in 15 aCRC cases carrying the rs1805327 AG genotype in an attempt to identify any other potentially functional variants that may be in linkage disequilibrium with rs1805327. She did not find any other likely pathogenic variants, suggesting that p.E281G itself is directly responsible for the observed association. She has also tested whether p.E281G might affect the stability of RAD1 RNA and/or protein. Although *RAD1* was expressed in most tumours, she did not observe any differences in the RNA expression levels in CRCs from patients with the p.E281 as compared to the p.G281 genotypes. Furthermore, she did not find any differences in the RAD1 protein levels as determined by immunohistochemical

analysis of CRCs from patients with a range of p.E281G genotypes. Further studies are therefore warranted to understand the functional effects of this variant.

7.4.1 Pharmacogenetics

Several previous studies have suggested that DNA repair variants effect response to therapy for aCRC. A study of 91 patients looked at a synonymous SNP in *ERCC1* (p.N118) which was postulated to impair mRNA translation. The relative response to 5-FU/oxaliplatin was significantly higher in the T/T genotype group compared with those with C/T and C/C genotypes: 61.9%, 42.3%, and 21.4%, respectively (P=0.018) (Viguier et al., 2005). A study of 71 patients treated with 5-FU/oxaliplatin looked at the effect of p.K751Q in *ERCC2* - 24% of patients with the K/K genotype responded versus 10% of those with either K/Q or Q/Q genotypes (P=0.015). A study of 106 patients who received 5-FU/oxaliplatin showed that four alleles, including *ERCC2* p.751Q, were associated with shorter OS (relative risk: 3.33, P=0.037) (Stoehlmacher et al., 2004). In the Federation Francophone de Cancerologie Digestive 2000-05 Trial (349 patients with DNA) *ERCC2* p.K751Q was associated with an increased risk of OxFU induced G3+ hematologic toxicity (P=0.01) (Boige et al., 2010). In a study of 42 patients, who received 5-FU/oxaliplatin, two SNPs were shown to have a predictive value: a C to T change in *ERCC5* leading to a synonymous SNP p.H46 and a T to C change in the 5'UTR of *XPA*. Patients with the *ERCC5* C/C genotype had a PFS of 9.8 compared with 7.8 month for those with T/C or C/C genotypes (P=0.009). Also, patients with a combination of *ERCC5* C/C and *XPA* C/C or T/C had a better PFS in comparison to the rest of the group: 13.7 versus 7.8 months (P=0.0001) (Monzo et al., 2007). A study of 61 patients who received 5-FU/oxaliplatin has looked at the predictive value of p.R399Q in *XRCC1* - 66% of non-responders had at least one Q allele compared to only 27% of the responder group (P=0.038) (Stoehlmacher et al., 2001).

In our study, we did find modest effects for p.N279S in *EXO1*, p.R399Q in *XRCC1,* p.N372H in *BRCA2,* p.F63L in *EME* and p.T9I in LIG4 on response to therapy (Table 7.3); however, none of these findings were close to being statistically significant after rigorous correction for multiple testing. Our data therefore suggest that some of the aforementioned studies are likely to have identified false-positive associations due to the use of small under-

powered sample sets however the field of pharmacogenetics remains a high research priority.

The successful identification of a relationship between *K-ras* mutation and cetuximab (Karapetis et al., 2008) has the potential to bring a level of stratification to CRC treatment and other similar relationships are sure to be identified in the coming years. CRUK has also launched a stratified medicine project with the aim of establishing a national genetic testing service as they expect treatments to become increasingly targeted to groups of patients with the ultimate goal of personalised medicine.

Table 7.2: Nonsynonymous SNPs in DNA repair genes over-represented at the 4% level.

			CASE				Control				Allele freq. in cases		Allele freq. in controls					
SNP	*Gene*	*Amino acid*	*AA*	*AB*	*BB*	fail	*AA*	*AB*	*BB*	fail	A	B	A	B	*OR*	X^2	*P value*	*Corrected P value*
rs1805327*	RAD1	p.E281G	1889	245	10	1	1825	340	9	1	4023	265	3990	358	0.7342	13.6	0.00023	0.038
rs3087374	POLG	p.Q1236H	13	286	1846		15	350	1810		312	3978	380	3970	1.2204	6.27	0.01225	1
rs3087403	REV1	p.V138M	152	843	1149	1	172	918	1085		1147	3141	1262	3088	1.1191	5.5	0.01906	1
rs2228527‡	ERCC6	p.R1213G	1385	668	92		1470	624	81		3438	852	3564	786	1.1237	4.51	0.03369	1
rs799917†	BRCA1	p.L871P	243	975	927		237	917	1021		1461	2829	1391	2959	0.9103	4.22	0.03992	1
rs12360068	MMS19L	p.A579V		164	1981		2	201	1971	1	164	4126	205	4143	1.2449	4.2	0.04039	1
rs1800282	FANCA	p.V6D	10	353	1781	1	25	384	1766		373	3915	434	3916	1.1632	4.17	0.04124	1
rs2228529‡	ERCC6	p.Q1413R	1386	666	92	1	1468	626	81		3438	850	3562	788	1.1176	4.1	0.04292	1
rs16942†	BRCA1	p.K1183R	947	968	228	2	1040	913	222		2862	1424	2993	1357	1.0934	4.07	0.04362	1

AA = Wildtype, AB= Heterozygote = BB = homozygous recessive

Note: rs1805327 in RAD1 remained significant after correction for multiple testing (P=0.025). The BRCA1 variants p.L871P†, p.K1183R† and p.S1613G† and the ERCC6 (CSB) variants p.Q1413R‡ and p.R1213G‡ were in complete linkage disequilibrium.

Table 7.3: Variants with P<0.05 for the primary outcomes in patients treated with chemotherapy ± cetuximab.

Outcome	rs no.	Gene	Variant	n	AA	AB	BB	X^2 (df)	P-value	Corrected P-value
12-week response	rs4149909	EXO1	N279S	1847	1741	106	0	8.67 (1)	0.0032	0.538
	rs25487	XRCC1	R399Q	1847	246	847	754	9.56 (2)	0.0084	1.000
	rs144848	BRCA2	N372H	1848	958	732	158	7.21 (2)	0.027	1.000
	rs17714854	EME1	F63L	1847	1800	47	0	4.74 (1)	0.029	1.000
	rs1805388	LIG4	T9I	1846	47	515	1284	6.32 (2)	0.042	1.000
Any Toxicity (except peripheral neuropathy)	rs4986850	BRCA1	‡D397N	2167	19	282	1866	7.30 (2)	0.026	1.000
	rs5745459	MSH4	Y589C	2167	2124	43	0	4.06(1)	0.044	1.000
	rs12022378	DCLRE1B	H61Y	2167	60	620	1487	6.18 (2)	0.046	1.000
	rs1799966	BRCA1	‡S430G	2167	962	976	229	6.13 (2)	0.047	1.000
peripheral neuropathy	rs3750898	DCLRE1A	H317D	2019	132	722	1165	8.55 (2)	0.014	1.000
	rs1800058	ATM	L72F	2017	0	70	1947	5.50 (1)	0.019	1.000
	rs3093921	PARP2	D222G	2019	1945	74	0	4.86 (1)	0.027	1.000
	rs13181	ERCC2	K751Q	2018	829	905	284	6.66 (2)	0.036	1.000
	rs9352	CHAF1A	V923A	2019	594	984	441	6.06 (2)	0.048	1.000

‡The BRCA1 variants D397N and S430G variants were in LD (r^2=0.16, D'=1) so likely to be associated with the same signal.

8 Chapter Eight

Using exome re-sequencing to identify the cause of severe peripheral neuropathy after oxaliplatin-based chemotherapy

8.1 Introduction

Oxaliplatin, a third-generation platinum drug, is the standard of treatment in combination with 5-fluorouracil/leucovorin for locally advanced and metastatic cancer of the colon or rectum. It improves survival in the adjuvant setting amongst stage III patients compared with 5-fluorouracil/leucovorin treatment, as well as in the first-line treatment of metastatic disease as compared with 5-fluorouracil/irinotecan therapy (Goldberg et al., 2004). In addition, oxaliplatin has shown promising activity against rectal, pancreatic and gastric malignancies (Becouarn et al., 2001). Oxaliplatin and other platinum agents exert their effects by forming inter-strand and intra-strand cross-links on DNA (Brabec and Kasparkova, 2005) which stall the cell cycle, inhibit DNA synthesis (Johnson et al., 1980) and triggers apoptosis (Faivre et al., 2003).

Peripheral neuropathy is a well-recognised dose-limiting toxicity of oxaliplatin (Hartmann and Lipp, 2003). Oxaliplatin induces two clinically distinct forms of peripheral neuropathy (Ocean and Vahdat, 2004). The acute syndrome, consisting of distal or perioral paresthesias and pharyngolaryngeal dysethesias, appears soon after the administration of oxaliplatin and is usually transient and reversible within hours or days. The chronic form is a pure sensory, axonal neuronopathy, closely resembling cisplatin-induced peripheral neuropathy (Pasetto et al., 2006). High cumulative doses of oxaliplatin are associated with occurrence of chronic peripheral nerve damage (Quasthoff and Hartung, 2002) and chronic sensory neuropathy has been observed in ~50% of patients who received oxaliplatin with infusional 5-fluorouracil/leucovorin. Most importantly, it is neurotoxicity, rather than tumour progression, which is often the cause of treatment discontinuation (McWhinney et al., 2009).

Since neurotoxicity is not correlated with response to treatment (McWhinney et al., 2009) it is likely that this is an avoidable side effect. However, the mechanism underlying oxaliplatin induced peripheral neuropathy has not been defined. Here, we sought to delineate the underlying mechanism by exome re-sequencing patients with severe peripheral neuropathy after treatment with oxaliplatin-based chemotherapy who were negative for known peripheral neuropathy mutations.

8.2 Materials and Methods

8.2.1 Patients

We analysed blood DNA samples from unrelated patients with advanced CRC from the UK national trial COIN (Maughan et al., 2011) (Sections 7.1.1 and 7.2.2). In all patients, treatment was identical for the first 12 weeks apart from the choice of fluoropyrimidine together with the randomisation of ± cetuximab.

We obtained the maximum grade of peripheral neuropathy observed in each patient after 12 weeks of oxaliplatin-based chemotherapy. Patients with Grade 3/4 peripheral neuropathy or that had had oxaliplatin-dose reduction due to severe peripheral neuropathy were classified as suffering from peripheral neuropathy associated with oxaliplatin (PNAO).

8.2.2 Molecular analyses and exome re-sequencing

PMP22 gene dosage analysis was carried out at the Bristol Genetics Laboratory, Southmead Hospital, North Bristol NHS Trust. Library fragments containing exomic DNA were collected using the Roche Nimblegen SeqCap EZ Exome (version 2) Library solution-based method. Massively parallel sequencing was performed on the Illumina Genome Analyser at the University of North Carolina. FASTQ files were processed through a sequence analysis pipeline using BWA (version 0.5.1) (Li and Durbin, 2009) for sequence alignment and modules from the Broad Institute's Genome Analysis Toolkit (GATK) (version 2.0) (McKenna et al., 2010) to recalibrate quality scores, refine alignments around potential indels, eliminate duplicate reads, call indel and SNP genotypes, generate QC metrics, and

apply quality filters to the genotype calls. SNP calls were annotated using the analysis package ANNOVAR (subversion 322) (Wang et al., 2010).

8.2.3 PCR and Sanger sequencing

PCR reaction and thermal cycling conditions are described in Section 2.4.2, primers and annealing temperatures are in Table 8.3. Sanger sequencing was carried out as described in Section 2.4.3, 2.4.4 and 2.4.5. Samples were run on an ABI 3100 genetic analyser (Applied Biosystems) and sequence data viewed using Sequencher v4.6.

8.2.4 Author's Contribution

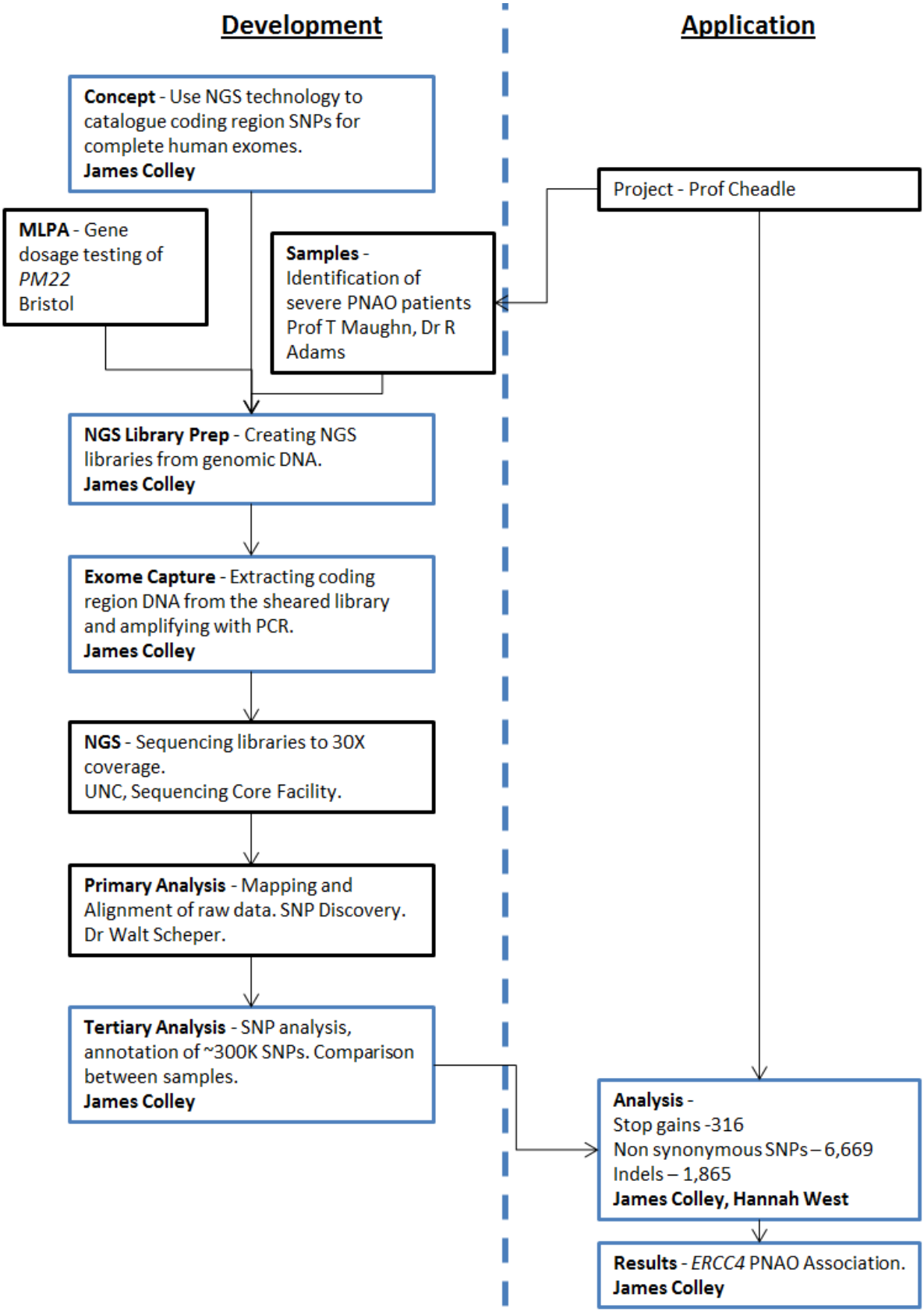


Figure 8.1: Author's Contribution to Chapter 8.

8.3 Results

8.3.1 Patient selection

Patients were selected from a total of 2,445 patients with advanced CRC undergoing treatment with oxaliplatin and a fluoropyrimidine (and sometimes cetuximab) as part of the COIN trial (Section 7.1.1). Severe peripheral neuropathy was observed in 23% of patients receiving oxaliplatin and fluorouracil-based therapy and 16% of those receiving oxaliplatin and capecitabine-based therapy, over the entire treatment period (Maughan et al., 2011). We selected ten patients that suffered from severe PNAO; these were considered the most severe sufferers in terms of peripheral neuropathy so as to enrich for potential inherited genetic defects.

8.3.2 Exome re-sequencing 10 patients with PNAO

We used the Roche Nimblegen SeqCap EZ Exome capture (version 2) kit together with the Illumina Genome Analyser to carry out exome re-sequencing of 10 patients with PNAO to help elucidate any underlying genetic defect. Coverage of the whole exome for each patient ranged from 23 to 39 fold. We identified on average 68.6 (range 59-81) stop-gains and 187.5 (range 158-209) indels predicted to result in frameshift mutations per patient exome. We reasoned that those rare variants not present in dbSNP v.129 ('novel') or with MAFs <1% were the most likely to cause PNAO and warranted further investigation. We identified on average 9.7 (range 3-15) and 27.7 (range 14-56) novel stop-gains and truncating indels, respectively, per patient and 17.3 (range 14-21) and 60 (range 55-69) stop-gains and truncating indels with MAFs <1%, respectively (Table 8.1).

Table 8.1: Oxaliplatin transport, metabolism and associated repair pathway SNPs.

Numbers of stop-gains and truncating indels identified in the whole exome or the oxaliplatin transport, metabolism and associated repair pathways in ten patients with PNAO.

		1	2	3	4	5	6	7	8	9	10
Stop-gains											
Whole exome	Total	67	85	86	76	83	86	81	79	73	96
	Novel	3	10	8	10	12	12	12	8	7	15
	(MAF <1%)	(15)	(19)	(21)	(14)	(20)	(18)	(14)	(18)	(16)	(18)
In Ox transport / metabolism / repair pathway	Total	1	1	1	2	1	1	1	2	1	2
	Novel	0	0	0	0	0	0	0	1	0	0
	(MAF <1%)	(1)	(1)	(1)	(2)	(1)	(1)	(1)	(1)	(1)	(2)
Truncating indels											
Whole exome	Total	73	111	81	87	83	100	90	78	94	81
	Novel	16	56	21	20	14	41	27	18	39	25
	(MAF <1%)	(57)	(55)	(60)	(66)	(69)	(59)	(63)	(60)	(55)	(56)
In Ox transport / metabolism / repair pathway	Total	2	2	1	1	1	0	3	0	0	0
	Novel	0	0	0	0	0	0	0	0	0	0
	(MAF <1%)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)

We considered that variants absent from dbSNP v.129 (termed novel) or with MAFs <1% (in parenthesis) were the most likely to affect function and warranted further investigation.

8.3.3 Excluding known inherited neuropathies

The hereditary neuropathies are a clinically and genetically heterogeneous group of disorders with overall prevalence of 1 in 2,500. Although there are over 30 known genes associated with inherited peripheral neuropathies, approximately 75% of patients with Charcot Marie Tooth disease and 85% of patients with hereditary neuropathy with liability to pressure palsies have a duplication or deletion of *PMP22* at 17p11.2, respectively. The Bristol Genetics Laboratory therefore carried out multiplex ligation-dependent probe amplification (MLPA) analysis of *PMP22* in all ten patients with PNAO and excluded any abnormalities in *PMP22* gene dosage.

We also examined the exome re-sequencing data for PMP22 and other genes associated with rare inherited neuropathies such as *MPZ, LITAF, EGR2, NEFL, GJB1, PRPS1, DNM2, YARS, MFN2, RAB7, GARS, HSPB1, HSPB8, GDAP1, LMNA, MED25, MTMR2, SBF2/MTMR13, SH3TC2, NDRG1, PRX, FGD4, FIG4, BSCL2, DCTN1* and *IGHMBP2*. On average, we had >10x coverage for 79.6% of each of the respective ORFs in each patient (Table 8.2).

We failed to find any stop-gain mutations or truncating insertions or deletions (indels) in these genes in our 10 patients with PNAO. Although we did find a several nonsynonymous variants these were also found in dbSNP at a similar frequencies and were therefore considered to be benign polymorphisms. Therefore, we excluded all known genes associated with inherited neuropathies as the likely cause of PNAO.

Table 8.2: 10X coverage of selected genes.

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7	Patient 8	Patient 9	Patient 10
BSCL2	86.2	75.0	86.2	86.2	86.2	80.6	79.8	86.2	81.8	83.2
DCTN1	97.9	78.8	94.9	97.9	99.8	88.9	98.3	97.0	92.8	99.7
DNM2	86.8	79.9	86.6	90.7	92.7	88.1	91.7	86.9	93.0	91.8
EGR2	89.3	85.5	99.0	97.9	96.8	92.2	99.7	93.7	87.6	100.0
FGD4	99.2	97.9	99.1	100.0	99.7	97.3	100.0	100.0	100.0	99.6
FIG4	99.9	96.3	93.4	100.0	100.0	95.0	99.9	99.9	98.0	100.0
GARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GDAP1	100.0	100.0	99.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0
GJB1	97.4	90.7	96.2	93.3	100.0	91.2	95.0	89.8	94.1	100.0
HSPB1	10.4	4.7	15.9	29.9	29.3	0.0	20.6	26.9	17.6	25.2
HSPB8	67.5	66.8	80.0	71.1	100.0	90.4	75.5	68.5	82.7	81.4
IGHMBP2	95.8	87.0	96.3	96.2	96.5	92.2	97.9	95.4	95.8	96.6
LITAF	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
LMNA	92.6	73.5	99.9	86.6	95.4	92.9	99.9	83.9	88.6	99.5
MED25	62.8	47.1	67.5	71.3	73.9	60.2	69.8	64.2	60.7	72.1
MFN2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
MPZ	98.7	86.2	96.8	100.0	94.2	89.0	100.0	98.3	86.1	100.0
MTMR2	95.9	95.3	95.8	95.9	96.5	95.9	99.0	95.9	95.9	97.9
NDRG1	100.0	96.3	100.0	100.0	96.2	97.1	100.0	99.1	97.2	100.0
NEFL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PRPS1	100.0	100.0	90.5	100.0	100.0	91.1	94.9	100.0	100.0	100.0
PRX	5.2	4.9	5.1	5.1	6.7	5.1	5.1	5.1	6.4	5.1
SBF2	99.0	98.1	97.4	98.9	99.0	98.5	98.7	99.0	98.8	98.9
SH3TC2	98.1	88.6	96.8	99.5	100.0	98.1	99.0	98.2	92.7	99.8
YARS	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

8.3.4 Identification of truncating mutations in the oxaliplatin transport, metabolism and associated DNA repair pathways

We hypothesized that rare germline truncating mutations in genes involved in oxaliplatin transport, metabolism or the repair of its associated damage might be responsible for the phenotype of PNAO. We identified genes potentially involved in the pharmacokinetics and mechanism of action of platinum compounds via literature reviews. In total we identified eighty-one genes that may play a role including four genes involved in drug influx (*SLC22A1, SLC22A2, SLC31A1* and *SLC47A1*), eight genes involved in detoxification (*CCS*, *SOD1, MT1A, MT2A, NQO1, GSTT1, GSTP1* and *GSTM1*), two genes involved in oxalate metabolism (*AGXT* and *GRHPR*), three genes involved in sequestration (*ATOX1, ATP7A* and *ATP7B*), twenty-five genes involved in apoptosis due to DNA damage (*SUPT16H, SSRP1*, *HMGB1, p53, COX17, AURKA, RB1, ABL1, p73, MAP2K3, MAP2K6, MAP3K1, MAPK8, MAPK14, RPS6KA5, EHMT2, DUSP1,2,4,8,10* and *16*, *CASP3*, *CASP8* and *CASP9*), thirty-three genes involved in DNA repair and the associated response pathways (*POLH, POLM, REV3L, POLB, ATM, ATR, CHK1, CHK2, CDC25C, CDC2, BRCA1, BRCA2, MLH1, MSH2, FANCA, FANCC, FANCD2, FANCE, FANCF, FANCG, MSH6, PMS2, MPO, ERCC1-6, XPA, XRCC1, XRCC3* and *MGMT*), and six genes involved in drug efflux (ABCC1-5 and ABCG2).

We sought stop-gain and truncating indels in these 81 genes in all ten patients with PNAO. We identified p.Q73X in *MAP2K3* in multiple patients which was also shown to be common in dbSNP**;** therefore it was considered likely to be a common benign polymorphism. c.1275delTGGTAAGT *SLC22A1* was predicted to be present in multiple patients (rs113569197, no frequency data) but Sanger sequencing of independent PCR products failed to confirm the presence of this deletion suggesting that it was an exome sequencing or bioinformatic artefact.

We also identified a stop-gain in *BRCA2* (p.K3326X) in two patient samples which was verified by Sanger sequencing independent PCR products. Although this is a well-studied rare variant (rs11571833, MAF 0.8%) shown not to play a role in breast cancer susceptibility (Mazoyer et al., 1996) it has been shown to be over represented in patients with inherited pancreatic cancer (Martin et al., 2005) we considered whether it could play a role in PNAO. We therefore assayed for p.K3326X in all other available cases from COIN. Overall, we found

similar proportions of cases with (2/63, 3.17% of patients) and without (37/1753, 2.1% of patients) PNAO harbouring this variant, thereby failing to support a causal role (P=0.395).

We identified a single patient (Patient 8) who carried the novel stop gain p.S613X in exon 9 of *ERCC4* (*XPF*) which was verified by Sanger sequencing of an independent PCR product. *ERCC4* plays a role in nucleotide excision repair (NER) the main DNA repair pathway involved in the removal of bulky and DNA-distorting adducts, such as those formed by oxaliplatin (Scheeff et al., 1999). Patients carrying biallelic *ERCC4* mutations are known to suffer from the UV sensitivity disorder xeroderma pigmentosum group F (XPF) characterised by an elevated risk of cancer, in particular skin and oral cancers (Lehmann et al., 2011, Matsumura et al., 1998). Interestingly, XP patients with mutations in other complementation groups (*XPA, XPC* and *XPG*) have been reported to suffer from peripheral neuropathy prior to treatment (Anttinen et al., 2008, Kanda et al., 1990, Robbins et al., 2002, Thrush et al., 1974). We therefore looked for a mutation in the second *ERCC4* allele in Patient 8 by direct sequence analysis of their entire ORF and flanking intronic sequences. We failed to find any other coding region variants.

8.3.5 Phenotype of Patient 8

Retrospective review of notes and clinical trials data indicate that this patient was a 79 year old lady at the time of oxaliplatin therapy and was diagnosed with metastatic CRC after an Ultrasound scan of her liver in March 2006. Her presenting symptoms to the primary care physician were of right upper quadrant pain and a 2 month history of intermittent diarrhoea. Her carcinoembryonic antigen (CEA) on presentation was 130. She went on to have a CT which demonstrated multiple metastases throughout both lobes of the liver. There was also a large caecal mass consistent with caecal tumour. Tru-cut biopsy of liver provided histological diagnosis of adenocarcinoma from a synchronous colonic primary cancer.

Past medical history included tubular adenoma excised 2001, peri-orbital rosacea diagnosed in 2003, and a seborrhoeic wart excised. Allergy skin tests at this time had suggested nickel sensitivity, she had moderate macular degeneration consistent with her age group and mild osteoarthritis. She was allergic to lidocaine. Of particular note in this

case, there was no past medical history of skin cancers, no immunodeficiency disorders or related diseases, no ataxia, memory loss or muscle weakness. It is evident that this lady did not have xeroderma pigmentosum.

8.3.6 Analysis of ERCC4 in other cases with PNAO

We investigated whether other cases with PNAO were caused by variants in *ERCC4* by Sanger sequencing the ORF and flanking intronic sequences in the nine remaining patients (Patients 1-7, 9 & 10) in the original cohort (to search for variants missed via exome re-sequencing) together with 54 other cases with PNAO from COIN. We did not identify any additional cases with truncating mutations. In total, we identified five nonsynonymous variants: p.P379S was found in 3 patients (MAF = 2.34%) and was previously documented in dbSNP (rs1799802, MAF=1.8%); p.R415Q in 9 patients (MAF = 7.03%) and in dbSNP (rs1800067, MAF=4.9%) p.R576T in 1 (MAF = 0.78%) patient and in dbSNP (rs1800068, MAF=0.00%) p.E875G in 4 patients (MAF = 3.13%) and in dbSNP (rs1800124, MAF=1.3%) and p.H466Q in a single patient and not in dbSNP, this SNP is part of on-going an investigation. Apart from one case that carried both p.R576Tand p.E875G, all other cases carried a single *ERCC4* nonsynonymous variant in a heterozygous state. p.P379S, p.R576T and p.E875G were all predicted to interfere with function by *in silico* analyses using Align-GVGD, whereas the variant p.R415Q was predicted as less likely to interfere with function. We also identified 3 synonymous variants (p.A11 [rs3136042], p.S835 [rs1799801] and p.T885 [rs16963255]) and three variants in the 5' un-translated region (UTR) (c.1-30A>T [rs1799797], c.1-356C>A [rs6498486] and c.1-69G>C [Novel]) all of which were unlikely to affect function.

Since *ERCC4* and *ERCC1* function together to form a 5' incision complex in NER (Sijbers et al., 1996) we also sought likely causal variants in *ERCC1* via direct sequence analysis of the ORF and intronic boundaries in all 64 patients with PNAO. We did not identify any cases with *ERCC1* truncating mutations. We found three synonymous variants (p.T75 [rs3212947], p.N118 [rs11615], p.P128 [rs139827427]) and five variants in the promoter region/5'UTR (c.1-230C>A [rs41559012], c.1-96T>G [rs2298881], c.1-303C>T [rs41540513], c.1-495C>A [rs3212931], c.1-790T>C [rs3212930]). None of the variants in the promoter

region/5'UTR lay within potential transcription factor binding sites, so none were considered likely to affect function.

8.3.7 Identification of truncating mutations in other genes potentially involved in neuropathy

We also considered whether rare germline truncating mutations in genes involved in nerve function might also be responsible for PNAO. Therefore, we filtered our data against the HapMap CEU population data and manually reviewed dbSNP where necessary to identify every gene predicted to carry a novel or rare (MAF<1%) stop-gain or truncating indel (n=59 genes) from the whole exome analyses. These genes were assessed in the literature for a potential role in neuronal function. 51 genes were identified as potentially relevant; however, the variants p.Y54X in *ANXA7*, c.188_191del in *NEFM*, c.2711_2714deland c.188_191del in *NRP2*, c.1416_1417insG in *APPL1*, c.1944_1945insG in *SEMA4C* and c.1684_1685insGin *PPP1R13L* were not confirmed upon sequencing an independent PCR product from the relevant patient, so were excluded. A stop-gain variant in *STOML3* (p.R155X) identified in a single patient and absent from dbSNP, was confirmed in an independent PCR product. *STOML3* encodes stomatin like protein, a mechanosensory channel, deletion of which leads to loss of mechanoreceptor function and loss of mechanosensitive currents in isolated neurons from mice (Wetzel et al., 2007). We therefore sequenced the entire ORF, flanking intronic sequences and promoter region of *STOML3* in 54 other patients with PNAO. No additional coding region variants were found and only a single intronic variant (rs9548577) was identified in one patient and was considered unlikely to affect function.

Table 8.3: PCR Primers.

Gene/ region	Forward primer (5'-3')	Reverse primer (5'-3')	Product Size (bp)
ERCC1			
Promoter A	GCTGTCGTTGGTCACTGCT	AGACTGCAGAGGGATCGAG	463
Promoter B	CCTGCTCTATGCTCTACTCTCC	AGAGCTCCATAGCGTCAGGT	482
Exon 1	TGCGGGATGAGAACGTAGAC	CCCCATCCTATCCTCTTCGT	237
Exon 2	AAAGGGGAGAGGAACTCACA	GGAGAACAAAGTGGCTGGAA	405
Exon 3	GTGCAAGAAGAGGTGGAGGA	TCCAGAACACTGGGACATGA	263
Exon 4	ATTCCAGTGAGAGGGAAAAGG	CTGCATTTCCTCTTGGAAGG	265
Exon 5	CCACCACGCCTGGCTAAT	ACAGGAAGGAGAAGGGAAGG	241
Exon 6	GGCAATTCTTATGACTGACCA	TGGAACTGAAGCTCAACCAC	255
Exon 7	CAGGCAGTCTGGGGACAC	CAGGGAGATGGAAGGAAATG	260
Exon 8	CCCTGGGGAATATCTGAGG	AGGCTGGTCTCCAACTTCTG	
Exon 9	TAAAGAACCAAAACCCCACTC	CAGAATCCCTCCCCAGAGAC	238
Alt. Exon 3	AAGTGATCCTCCTGCCTCAG	CTGGCTACAGGCCAGCTCTT	169
ERCC4			
Promoter A	AGCCTGGGCAACATATCAAC	TTCATGGAGTCACCTGTAGGG	332
Promoter B	AGGGGATGTGGAAACTCAAA	TAGCCGAGGAGAGCTGAGA	238
Exon 1A	CTCTCGGACTCGGCTCTCT	GTGCAGCTGGAGAAAGTGG	
Exon 1B	CCGCTGCTGGAGTACGAG	TGTCATCGCGTAGTGTCAGG	433
Exon 2	TCAGAGAAAGACAGCACATTATTT	TGGAGAAAAATAAAATGGAAATTG	357
Exon 3	CTCTGTTCTGTGCGTGGCTA	CCATCAAATTGCTCTCGACTT	547
Exon 4	TTTGTTGTTTTGCTTTTCGTG	GCTATGTTTTTAAGTGACCTCCA	425
Exon 5	GATACACAGGAAATAATCCTTTTGA	CACACCTGATTCCCCCTAAA	354
Exon 6	CGGTGTGGTTGGTAGGAAGA	TTTCACATGGCCAAAGAAGAC	348
Exon 7	TGATGCTCGTGTTATCTGTTG	AAATAGAGACAGGGTTTCACCA	327
Exon 8A	ATGTCTTCCCTTCGGGTGA	AGCCCGTTCTTTGTTTTGG	314
Exon 8B	GAGCGGAGGCCTTCTTATTG	AGTGAGGGGTTCTTTCAGGA	377
Exon 8C	AAGGAGATGTCGAGGAAGGA	AAGCAGCATCGTAACGGATA	401
Exon 9	GCGCTCTAGGTTGCTGATTT	CTTCCTTGCCCTATCCTTCC	287
Exon 10	TCCTTGTTTTTGTTTTTGTTTTTC	CCAACCCCCATTTTTAAGAG	361
Exon 11A	CCATCCATCAGAGTTAACAACA	CCTCGGGAAGTGAGAGAGAA	403
Exon 11B	TGGAGCGCAAGAGTATCAGT	ATCAAGGAGCGGCAGTTTTT	430
Exon 11C	CTGAAACAAAGCAAGCCACA	TCTGGTCCACCGTACAATCA	442
STOML3			
Exon 1A	TTTCAAAAGCTCACTCACTGC	TGTGAAGAACAGGCAGCAAC	419
Exon 1B	CTTCCCTCACCAGGGTAACT	TGCTACAACTCCTGCTTTGC	293
Exon 2	TTCTATGCAGCCACATCAGG	CCAGACGGAATACAACAGCA	342
Exon 3	CATTACCTTCCCCATCTCCA	AATAGGCACCACCAGGAAAA	281
Exon 4	CATGTATCGCCCCATGTAAA	GCGGGTACTCAGCTCATCTT	302
Exon 5	GCCAGGACAGGTTTTAGGTG	GTGGGGGATGCTTTGAACT	441
Exon 6	TCACTCCAAATGCTGTAAATGC	AACCCCTTTCTCATGCAAAT	381
Exon 7	CTGGGGAGAGGGGTATCAA	GTGTTGGAATTCTCACCGTTT	410

8.4 Discussion

To our knowledge, our data provide the first definite insight into the mechanism underlying oxaliplatin induced peripheral neuropathy in humans. Interestingly, several other studies help support our findings: (i) Patients with XP and with mutations in other complementation groups suffer from peripheral neuropathy prior to treatment (Anttinen et al., 2008, Kanda et al., 1990, Robbins et al., 2002, Thrush et al., 1974). Importantly, the patient with PNAO and the truncating mutation in *ERCC4* described herein, carried a single mutant allele and was not reported as suffering from XP. These data may suggest that haploinsufficiency for a mutant *ERCC4* allele is sufficient to induce peripheral neuropathy upon exposure to oxaliplatin. (ii) An Ercc1-/Δ murine model, which has reduced expression of the ERCC1-ERCC4 complex, develops accelerated spontaneous peripheral neurodegeneration with significant alterations of the sciatic nerves upon morphologic and ultrastructural analysis (Goss et al. 2011). (iii) In $Xpa^{-/-}$ and $Xpc^{-/-}$ mice, chronic exposure to cisplatin resulted in an accelerated accumulation of unrepaired intrastrand cross-links in neuronal cells. Furthermore, the augmented adduct levels in dorsal root ganglion cells of these mice coincided with an earlier onset of peripheral neuropathy-like functional disturbance of their sensory nervous system (Dzagnidze et al., 2007). (iv) Multiple myeloma patients carrying the variants rs1799800 and p.S835 (in linkage disequilibrium) in *ERCC4*, were at a 2.74 and 2.48-fold greater risk, respectively, of developing late onset peripheral neuropathy after treatment with bortezimab (Broyl at el. 2010).

Biallelic mutations in *ERCC4* or *ERCC1* have been shown to cause XP, XPF-ERCC1 (XFE) progeroid syndrome or cerebro-oculo-facio-skeletal syndrome, characterised by increased risk of cancer, accelerated aging and severe developmental abnormalities, respectively (Gregg et al., 2011). Interestingly, the variant p.P379S, which we identified in 1 patient with PNAO in a heterozygous state, has previously been identified as a pathogenic mutation causing XFE progeroid syndrome (when it co-occurs with another mutation) (Gregg et al., 2011). It has recently been shown that mutations in *ERCC4* associated with XFE progeroid syndrome cause mislocalisation of XPF-ERCC1 into the cytoplasm probably due to protein mis-folding (Ahmad et al., 2010) analogous to the ΔF508 mutation in *CFTR*. Further studies are therefore warranted to determine whether rare nonsynonymous variants in *ERCC4*,

together with truncating mutations, contribute to PNAO. These studies are on-going in our laboratory.

9 Chapter nine

General Discussion

At the commencement of this study 'Large Scale' or 'High Throughput' genomic technology could generally be thought of as any instrumentation that was capable of utilising 96 well plates for automated or parallel processing. In reality the capacity of any instrument can only be described as 'Large Scale' relative to the other technology available at the time. A 96 capillary automated sequencer was once considered to be a significant improvement over slab gels hence providing 'high-throughput'; however, when comparing that change to the difference between the capillary and NGS instruments the increase in throughput from gel to capillary seems marginal. Over the period of study much of the technology that was assessed has been effectively replaced by the massively parallel processing NGS techniques. SNP discovery on an exome or genome wide scale is becoming relatively common amongst research groups and NGS is poised to revolutionise clinical genetics as it has research by identifying disease genes for both Mendelian and multifactorial disease (Table 9.1).

9.1 dHPLC

In 2005 dHPLC was a commonly used technology in SNP discovery projects. Automated sampling of 96 well plates provided a high throughput method of screening for the presence of heteroduplexes (and recessive alleles where sample spiking was used). By screening samples in this way a researcher could avoid costly sequencing of homozygous fragments that would were uninformative. Furthermore being able to cross reference dHPLC and Sanger sequence chromatograms aided the identification of SNPs where chromatograms alone might not have been conclusive.

As the number of samples and fragments that were required for analysis grew, manually testing dHPLC traces became a bottle neck and automated software was introduced to solve the problem. Here we have demonstrated that the Transgenomic Navigator software is effective in identifying rare variants on their own or from fragments that also contain common variation. While many research groups have moved onto new technology recent publications demonstrate that the technique is still well used for SNP discovery (Zahary et al., 2012).

Optimum Navigator analysis conditions were applied to a Glycosylase project where MAP like 134 samples were screened for the ORF of 5 glycosylase genes to investigate the relationship between rare variation and common disease. None of the SNPs discovered where found to be in association with CRC however, recent publications have found *OGG1* p.R46Q to be pathogenic (Anderson and Daggett, 2009, Morak et al., 2011) . Within our study this variant was excluded from further analysis due to its presence in our control population however, redundancy in the repair pathway may allow low penetrance variation such as this to exist in healthy individuals.

9.2 Sequence Analysis Software

As the cost of capillary sequencing dropped (Figure 9.1: The falling cost of Sanger Sequencing reactions) and the availability of 96 capillary machines increased, researchers have been able to afford large scale projects using sequencing as the method for SNP discovery. Though SNP detection by Sanger sequencing alone may not be as effective as using a combined dHPLC/Sequencing approach the improved throughput was considered to compensate for the loss in sensitivity. Here the bottleneck quickly became the analysis of chromatograms and as with dHPLC (Section 9.1) automated calling software was introduced to streamline the SNP discovery process. Here we have assessed various chromatogram analysis programmes and while excellent results were obtained from the Staden and Mutation Surveyor programmes Sequencher remained the primary sequence analysis software in our institute due to the wealth of additional sequence analysis features beyond chromatogram analysis.

In analysing a large quantity of data we were able to show a significant over representation of non-synonymous SNPs were found in the β catenin down regulating domain of *APC* therefore supporting hypothesis that the unexplained proportion of hereditary CRC is caused by numerous rare variants in multiple genes each conferring a small increase in risk.

9.3 dbSNP

The number of validated SNPs in dbSNP continues to rise rapidly making the resource more valuable with each build. Extracting data can be laborious when researchers wish to obtain information that fulfils a strict set of criteria. Several programmes had been designed that accessed dbSNPs database directly but subsequent changes to that database left each programme in need of editing. Here we developed a piece of software to 'strip' html pages (rather than access the database directly) with the thought that this would avoid the pitfalls of database updates by dealing only with the user interface. Though the programme would have survived the previous updates changes in build 133 have prevented the script form running, changes have yet to be written into the code as we have not since required the function.

In genotyping all common non-synonymous in every human DNA repair gene for 480 normal human lymphoblastoid cell lines we have generated a resource of cell lines with DNA repair profiles. Our intention was to select cell lines with specific SNP profiles and test the variation in repair capacity within a set of control cell lines however, we were not able to reproducibly show the variation between cells response to oxidative damage, possibly due to the sensitivity of available assays.

9.4 Pharmacogenetics

It is though that an individual's genetic variation plays a significant role in determining the success of any given therapeutic agent and we now have the technology to rapidly identify that variation and select or tailor treatments based on this knowledge.

Several recent studies with relatively small cohorts of samples have attempted to show association between genotypes and drug response (Section 7.4.1) however these associations were not reproduced in our set of over 2,000 samples. It is likely that the effects of SNPs on drug response will have a similar pattern to the effects of SNPs on disease, with a broad range of penetrance. Though a proportion of SNPs will be detectable from small studies with tightly controlled 'phenotype' a large number will be part of complex multi-gene systems and identification will require large studies to identify those variants with small ORs.

Table 9.1: Examples of NGS enabled discovery.

	Miller syndrome	**Kabuki syndrome**	**Autism**	**Autosomal recessive intellectual disability**
MIM #	MIM263750	MIM147920	MIM209850	none
Type	Monogenic	Monogenic	Complex	Complex
Method	Exome Sequencing	Exome Sequencing	Exome Sequencing	Targeted Sequencing
Gene	*DHODH*	*MLL2*	FOXP1, GRIN2B, SCN1A, LAMC3	50 genes identified
Reference	(Ng et al., 2010b)	(Ng et al., 2010a)	(O'Roak et al., 2011)	(Najmabadi et al., 2011)

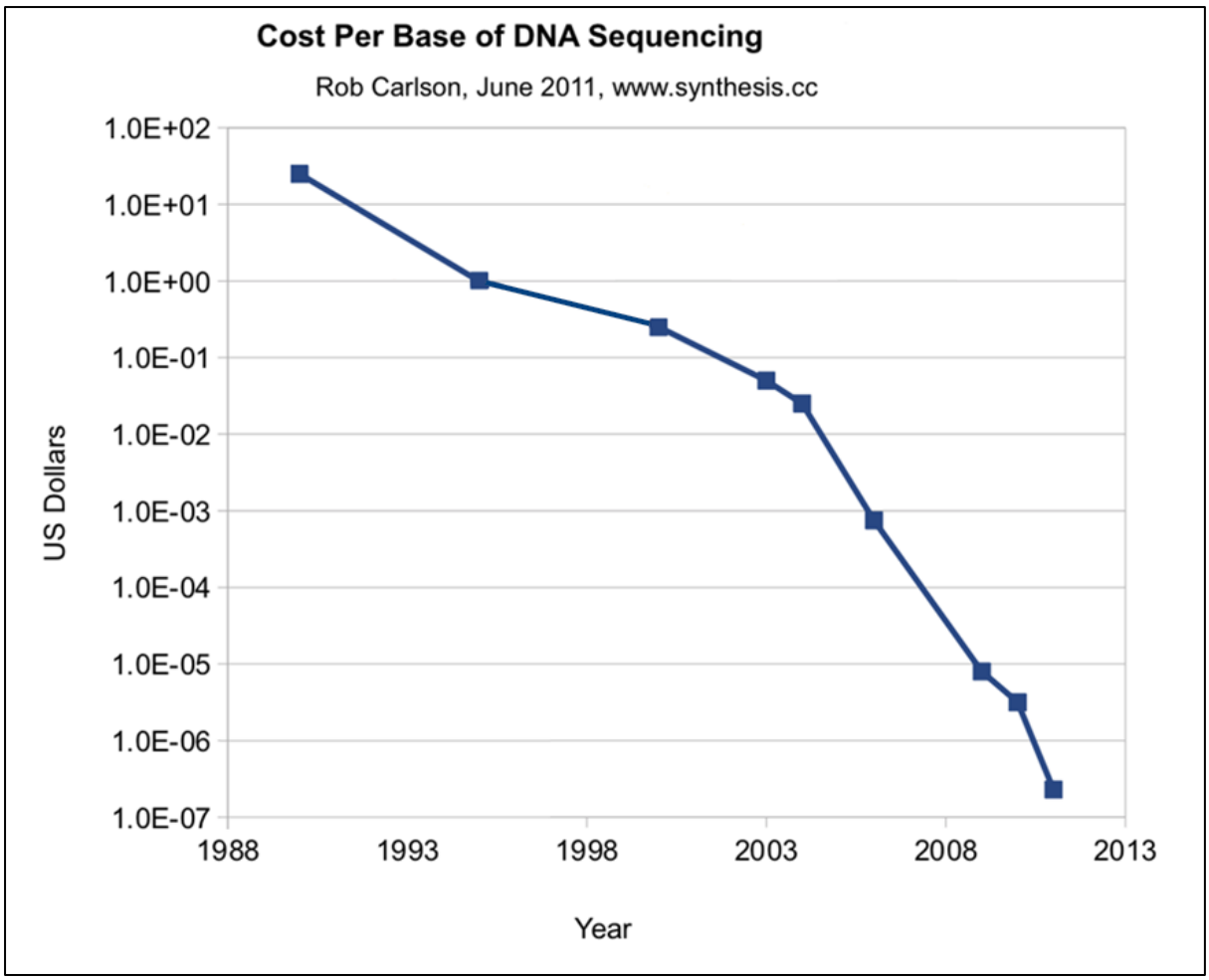


Figure 9.1: The falling cost of Sanger Sequencing reactions

9.5 Next Generation Sequencing

9.5.1 NGS in research

In essence NGS technology can provide a view of the entire genome, multiple exomes, deep coverage targeted regions and more recently the epigenome. NGS is being applied to advance research into disease mechanisms, cancer biomarkers, diagnostics, drug development and any other discipline with an interest in genetics.

By determining the sequence of an entire human exome it is now possible to apply an increasingly 'hypothesis-neutral' approach to mutation discovery making NGS an ideal tool for rare disease research. Before NGS was available, searching for a causative mutation in a monogenic disorder would require either a strong hypothesis suggesting a causative gene or vast resources such that numerous genes could be screened. Biochemical information, Association studies and animal models may lead to a list of candidate genes for testing but with an exome sequence the hypothesis is not necessary. In a recent presentation at Hinxton (Genomic Disorders 2011 - The Genomics of Rare Diseases. 23–26 March 2011) Han Brunner presented results from the exome sequencing of 200 patients representing 30 different rare diseases. For 15/30 diseases, new disease genes were discovered and a further 5/30 diseases had causative mutation identified in known disease genes demonstrating an unprecedented ability to discover a large proportion of rare disease genes.

In one of the first 'rare monogenic disease – NGS' publications (Ng et al., 2010b) mutations causing disease were discovered by sequencing the exomes of just four individuals, 2 siblings and 2 unrelated cases. Common variations (dbSNP129 and 8 HapMap exomes) were excluded and the study focussed on nonsynonymous, splice-site and coding Indel variation. Shared variation between the siblings was compared to the unrelated individuals leading to the identifications of a single candidate gene *DHODH*. By limiting SNP searching to exomes this approach cannot be considered entirely hypothesis neutral; however, it represents a considerable move in that direction compared to the candidate gene approach.

SNP discovery and genotyping can be expanded into non-exonic DNA by sequencing entire genomes hence moving closer to a truly hypothesis neutral approach. Currently there

are financial and practical limitations to genome sequencing on a large scale; however, when these obstacles are overcome NGS may enable researchers to return to analysing multi-case families as was seen in early positional cloning methods with linkage analysis. When using a limited set of markers this approach was surpassed by Association studies using large populations as a mechanism of finding disease gene loci. With a complete picture of variation supplied by genome sequencing it may become a realistic option in some cases to return to a Linkage analysis approach to identify co-segregating mutation in families affected by complex disease.

By focusing on small sections of the genome (a few Megabases or less) to create libraries NGS can provide deep coverage of a targeted sequence. This approach has the potential to identify low level variation (e.g. mosaicism) or rare transcription products depending upon the starting material. “Targeted Deep Sequencing” of mRNA has recently exposed levels of complexity to the human transcriptome which were not previously detectable. Mercer *et. al.* (2011) have described the sequencing of 0.77Mb of DNA including; exons, introns, known non-coding RNA and intergenic regions extracted from a primary foot fibroblast cell line. NGS enabled the region to be covered with an average 4,607-fold coverage, in doing so they describe 204 previously unknown isoforms of 55 protein coding loci. This demonstrates the utility of the technology to expose a new level of detail to our knowledge of transcription. The numerous known GWAS loci that as yet have no relationship to causative mutation (NHGRI data http://www.genome.gov/26525384) may benefit from this understanding of low level products. Variation in non-coding RNA (such as micro RNA) and rare transcripts may come to explain a functional aspect to some of these ‘orphan’ GWAS loci.

Transcriptome sequencing demonstrates that NGS is not limited to analysis of DNA. RNA sequencing on this scale can give a more comprehensive overview of expression than microarrays. By studying the expression profiles of cells treated with drugs and contrasting them with profiles from cells in a disease state it may be possible to find new uses for existing drugs. This approach is known as 'drug repositioning’, compounds that have failed their original purpose are screened for alternative therapeutic effects (a classic example being Pfizer’s Viagra, a failed Angina treatment). Given the failure rate and cost of creating new drugs, approaches that will demonstrate a potential use for a ‘failed’ compound are an

attractive option especially where trials have already been completed showing the drug to be suitable for humans.

9.5.2 NGS in clinics

Recent reports from Life Technologies and Oxford Nanopore suggest that the current cost of NGS is likely to continue to fall dramatically over at least the next 12 months and, ultimately a point will be reached where it is more cost effective for clinical laboratories to undertake NGS rather than individual genetic tests.

It has been demonstrated that through NGS it is possible to obtain a clinical diagnosis by identifying variation. In a patient with a "suspected diagnosis" of Bartter syndrome (MIM 607364) Choi *et*. *al*. sequenced the exome and identified a missense mutation in the highly conserved residue (p.D652N) of *SLC26A3* a gene known to cause congenital chloride diarrhoea (Choi et al., 2009). A further five patients were found to have deleterious mutations in the same gene.

In the case of Nicholas Volker (Worthey et al., 2011) NGS identified a mutation causing a previously undefined type of Inflammatory Bowel Disease (IBD). The patient presented with symptoms similar to Crohn disease at the age of 15 months. Before he had turned four he had undergone over one hundred surgeries relating to gastrointestinal inflammation including the removal of his colon. The patient's exome was sequenced for coding region mutation. After filtering nonsynonymous variation and testing with Polyphen p.C203Y was discovered in the X-linked apoptosis gene *XIAP*, a gene not previously associated with the observed pathology. The diagnosis based on NGS data led to a bone marrow transplant that has led to an effectively normal life. These cases demonstrate that NGS is already a solution for diagnosis of rare disease in situations where traditional diagnosis techniques fail to uncover a cause.

9.5.3 NGS Problems

The sequencing of a human genome has dropped from a reported $4B to under $4,000 in ten years. Whilst this is empowering researchers to investigate DNA on a new

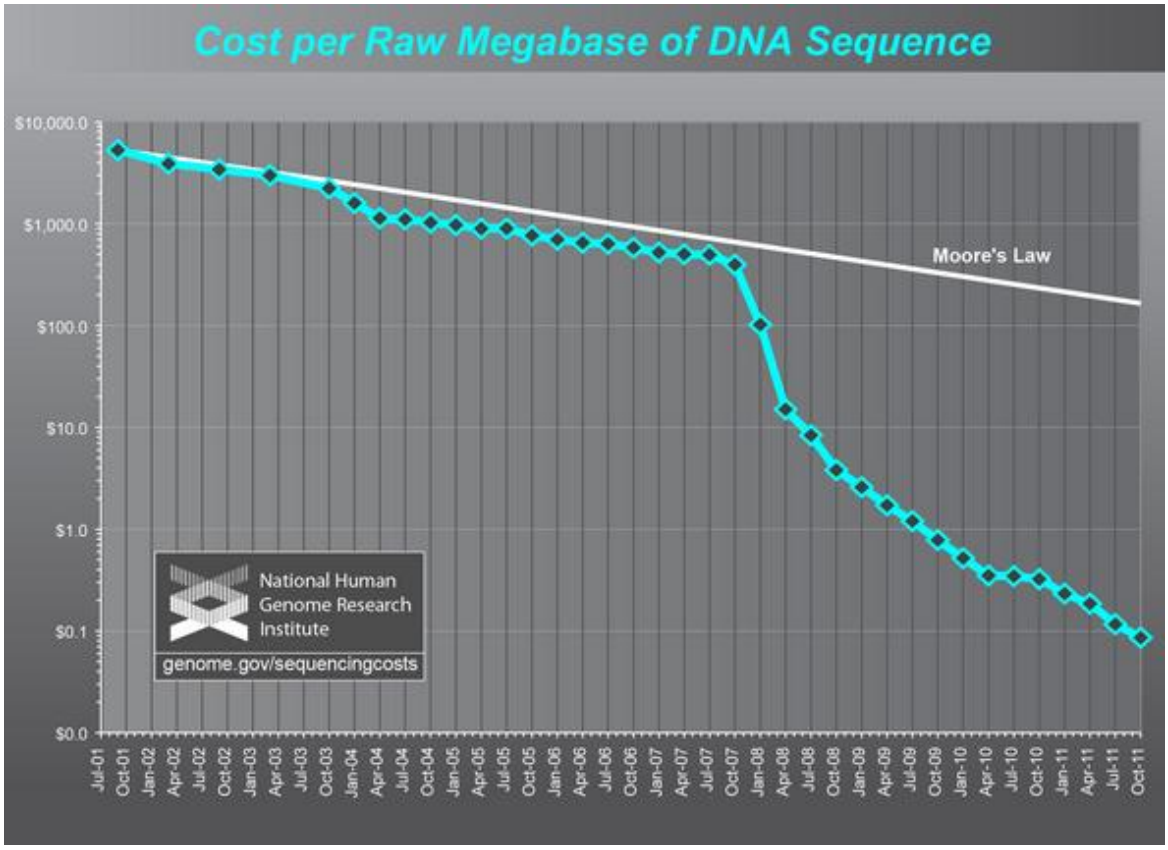
scale it has also brought problems. Technology providers are heavily promoting the concept that these instruments will soon be fast and cheap enough for routine clinical practice but analysis is not trivial and while the cost of generating the data may soon seem insignificant the cost and practicality of analysis and storage are not improving at the same rate. Moore's Law (Moore, 1965) is used in computing to predict the improvement over time of functions such as processor capacity and memory speed. It has proven to be highly accurate in predicting a doubling in performance annually. Production of NGS data far outpaces Moore's Law as demonstrated in Figure 9.1.

9.5.3.1 Data Analysis

Data analysis is an area that contains a mass of variables. Currently a single data set when analysed with two separate pipelines will show inconsistency between the reported SNPs. Between pipelines decisions are made on elements such as mapping algorithms, quality calling, accounting for phasing, aligning non-unique sequence. The combination of these variables will culminate in differences in mapping and hence differences in SNP calling from the same raw data. Broadly speaking there are NGS analysis packages that tend to be the most commonly used (BWA aligner, GATK, SAM Tools) but there is not yet a consensus on settings within these programmes even when the objectives are identical.

9.5.3.2 In vitro variation

In a laboratory experiment with the same; protocol, reagents, instruments and library you would not necessarily expect to see the same result following NGS analysis. Two aliquots from the same library will not be identical at a molecular level. Even if it were possible to provide identical library aliquots (molecule for molecule) the numerous processes of library preparation are likely to result in differing sequence success at each loci. At the level of an individual project (within a single analysis pipeline) thresholds are set that determine whether a SNP is called or not. If a locus has 30X coverage and 15 of those sequences identify a minor allele then heterozygote calling is straight forward; however, if that sample has 10X coverage with 4, 3 or 2 minor alleles then calling relates to the judgement of the user.



(http://www.genome.gov/sequencingcosts)

Figure 9.2: Falling cost of DNA Sequencing.

The cost of producing raw sequence (blue) fell steeply in 2008 when centres started to move away from Sanger sequencing technology. The decrease in cost continues to outpace Moore's law which predicts improvements in computer technology.

9.5.3.3 Potential for misuse

Fears of misuse of information are well founded and far reaching. An inability to obtain health insurance, a mortgage or even employment could potentially occur if an individual is known to carry strong risk factors for disease. In reality genetic predisposition does not necessitate disease and it is likely that all individuals will carry several significant risk factors.

9.5.3.4 External Perception

The NGS revolution has not yet begun to slow its pace and while Research is reaping the rewards the gap in public understanding is likely to grow at an equally exponential rate unless there is a significant change in the teaching practice from schools to professional medical courses (Ware et al., 2012).

Understanding our own risk of disease and acting accordingly are not always closely aligned. The technology is now available to provide individuals with detailed information on their known risk factors for various diseases; however, providing appropriate advice does not necessarily result in appropriate behavioural/lifestyle changes (Bloss et al., 2011). Indeed it can be readily demonstrated by the fact that in the UK alone there are an estimated 10 million cigarette smokers (http://www.ash.org.uk) despite the well-publicised risks.

9.5.4 The Future of NGS

Current NGS projects tend to focus on exomic sequences ignoring the rest of the genome for reasons of cost and perhaps perceived irrelevance of non-coding variation.

While the cost of sequencing an exome is lower than the cost of a genome the difference is not as great as might be expected. The exome capture kits represent a significant proportion of the total cost as well as adding a significant time to the library preparation procedure (3-4 days). By extending the library preparation procedure with

fragment capture we introduce opportunities for errors and necessitate a qPCR validation step to prove enrichment has occurred.

Exomes provide the most readily interpretable SNPs. Our knowledge of non-coding variation is minute in comparison to the quantity of sequence it refers to and it is not likely to expand until genome sequencing becomes the standard. With the developments that are promised for faster, cheaper sequence generation whole genome NGS is likely to become a core approach in clinical diagnostics, simplify library preparation and analysis and enabling our knowledge of mutation to expand into further into non-coding DNA.

The Baylor College of Medicine has already announced its aims to routinely genome sequence all new patients at the Texas Children's Cancer Centre (Genome web August 03, 2011 http://www.bcm.edu/news/packages/pediatricgenomecenter/) and eventually to sequence all new paediatric cancer cases in the state of Texas, approximately 1,000 per year. This demonstrates the perceived importance this technology has to health care and will create a vast resource in childhood cancer research.

In the short term the most significant impact of NGS is that will underlie an improved understanding of heterogeneous disease and drug response/efficacy. In the long term there is likely to come a tipping point where all current genetic testing for mutation is replaced by genome sequencing. This would simplify the work load of NHS diagnostic facilities and provide economies of scale whilst producing patient information that would be relevant for all subsequent medical considerations and treatments.

In 2009-10 Welsh Assembly Government (WAG) budget for health care was £5.5B, or £1,781 per person (http://wales.gov.uk/topics/statistics/headlines/health2012/120328/?lang=en) when pharmacogenomics has developed to the point where Personalised Medicine is a reality, the proportion of this cost that is spent on individual genetic tests, treatment of Adverse Drug Reaction (ADR) and Cancer therapy will have to be balanced against the cost of nationwide sequencing programme. Should that information become available to the research community it would represent an invaluable resource.

9.5.5 Third Generation

Third generation technology (Section 1.4.5) is set to provide the platforms with throughput and costs that will enable genome sequencing to surpass exome sequencing as the preferred approach. Oxford Nanopore announced two systems at the AGBT 2012 conference that will require “minimal sample preparation“ and provide GBs of reads that are 10s of Gb in length for a total predicted cost of around US$1,000. However, the 4% error rate (presented at the conference) is a significant drawback in comparison to current Next Generation technology which operates with error rates that are a fraction of a per cent.

While this third generation will undoubtedly provide improved de novo genome assembly, structural variant analysis, and Metagenomics a considerable improvement will be required before it is suitable for SNP related work.

Publications resulting from this work

AZZOPARDI, D., DALLOSSO, A. R., ELIASON, K., HENDRICKSON, B. C., JONES, N., RAWSTORNE, E., COLLEY, J., MOSKVINA, V., FRYE, C., SAMPSON, J. R., WENSTRUP, R., SCHOLL, T. & CHEADLE, J. P. 2008. Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas. *Cancer Res,* 68**,** 358-63.

COLLEY, J., JONES, S., DALLOSSO, A. R., MAYNARD, J. H., HUMPHREYS, V., DOLWANI, S., SAMPSON, J. R. & CHEADLE, J. P. 2005. Rapid recognition of aberrant dHPLC elution profiles using the Transgenomic Navigator software. *Hum Mutat,* 26**,** 165.

DALLOSSO, A. R., DOLWANI, S., JONES, N., JONES, S., COLLEY, J., MAYNARD, J., IDZIASZCZYK, S., HUMPHREYS, V., ARNOLD, J., DONALDSON, A., ECCLES, D., ELLIS, A., EVANS, D. G., FRAYLING, I. M., HES, F. J., HOULSTON, R. S., MAHER, E. R., NIELSEN, M., PARRY, S., TYLER, E., MOSKVINA, V., CHEADLE, J. P. & SAMPSON, J. R. 2008. Inherited predisposition to colorectal adenomas caused by multiple rare alleles of MUTYH but not OGG1, NUDT1, NTH1 or NEIL 1, 2 or 3. Gut, 57, 1252-5.

HOULSTON, R. S., CHEADLE, J., DOBBINS, S. E., TENESA, A., JONES, A. M., HOWARTH, K., SPAIN, S. L., BRODERICK, P., DOMINGO, E., FARRINGTON, S., PRENDERGAST, J. G., PITTMAN, A. M., THEODORATOU, E., SMITH, C. G., OLVER, B., WALTHER, A., BARNETSON, R. A., CHURCHMAN, M., JAEGER, E. E., PENEGAR, S., BARCLAY, E., MARTIN, L., GORMAN, M., MAGER, R., JOHNSTONE, E., MIDGLEY, R., NIITTYMAKI, I., TUUPANEN, S., COLLEY, J., IDZIASZCZYK, S., THOMAS, H. J., LUCASSEN, A. M., EVANS, D. G., MAHER, E. R., MAUGHAN, T., DIMAS, A., DERMITZAKIS, E., CAZIER, J. B., AALTONEN, L. A., PHAROAH, P., KERR, D. J., CARVAJAL-CARMONA, L. G., CAMPBELL, H., DUNLOP, M. G. & TOMLINSON, I. P. 2010. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. Nat Genet, 42, 973-7.

References

2010. A map of human genome variation from population-scale sequencing. *Nature,* 467**,** 1061-73.

AALTONEN, L., JOHNS, L., JARVINEN, H., MECKLIN, J. P. & HOULSTON, R. 2007. Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. *Clin Cancer Res,* 13**,** 356-61.

AHMAD, A., ENZLIN, J. H., BHAGWAT, N. R., WIJGERS, N., RAAMS, A., APPLEDOORN, E., THEIL, A. F., JH, J. H., VERMEULEN, W., NG, J. J., SCHARER, O. D. & NIEDERNHOFER, L. J. 2010. Mislocalization of XPF-ERCC1 nuclease contributes to reduced DNA repair in XP-F patients. *PLoS Genet,* 6**,** e1000871.

AHMADIAN, A., GHARIZADEH, B., GUSTAFSSON, A. C., STERKY, F., NYREN, P., UHLEN, M. & LUNDEBERG, J. 2000. Single-nucleotide polymorphism analysis by pyrosequencing. *Anal Biochem,* 280**,** 103-10.

AL-TASSAN, N., CHMIEL, N. H., MAYNARD, J., FLEMING, N., LIVINGSTON, A. L., WILLIAMS, G. T., HODGES, A. K., DAVIES, D. R., DAVID, S. S., SAMPSON, J. R. & CHEADLE, J. P. 2002. Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors. *Nat Genet,* 30**,** 227-32.

ALBUQUERQUE, C., BREUKEL, C., VAN DER LUIJT, R., FIDALGO, P., LAGE, P., SLORS, F. J., LEITAO, C. N., FODDE, R. & SMITS, R. 2002. The 'just-right' signaling model: APC somatic mutations are selected based on a specific level of activation of the beta-catenin signaling cascade. *Hum Mol Genet,* 11**,** 1549-60.

AMES, B. N. & GOLD, L. S. 1991. Endogenous mutagens and the causes of aging and cancer. *Mutat Res,* 250**,** 3-16.

ANDERSON, P. C. & DAGGETT, V. 2009. The R46Q, R131Q and R154H polymorphs of human DNA glycosylase/beta-lyase hOgg1 severely distort the active site and DNA recognition site but do not cause unfolding. *J Am Chem Soc,* 131**,** 9506-15.

ANTONARAKIS, S. E. 1998. Recommendations for a nomenclature system for human gene mutations. Nomenclature Working Group. *Hum Mutat,* 11**,** 1-3.

ANTTINEN, A., KOULU, L., NIKOSKELAINEN, E., PORTIN, R., KURKI, T., ERKINJUNTTI, M., JASPERS, N. G., RAAMS, A., GREEN, M. H., LEHMANN, A. R., WING, J. F., ARLETT, C. F. & MARTTILA, R. J. 2008. Neurological symptoms and natural course of xeroderma pigmentosum. *Brain,* 131**,** 1979-89.

ARETZ, S., UHLHAAS, S., SUN, Y., PAGENSTECHER, C., MANGOLD, E., CASPARI, R., MOSLEIN, G., SCHULMANN, K., PROPPING, P. & FRIEDL, W. 2004. Familial adenomatous polyposis: aberrant splicing due to missense or silent mutations in the APC gene. *Hum Mutat,* 24**,** 370-80.

ASPINWALL, R., ROTHWELL, D. G., ROLDAN-ARJONA, T., ANSELMINO, C., WARD, C. J., CHEADLE, J. P., SAMPSON, J. R., LINDAHL, T., HARRIS, P. C. & HICKSON, I. D. 1997. Cloning and characterization of a functional human homolog of Escherichia coli endonuclease III. *Proc Natl Acad Sci U S A,* 94**,** 109-14.

AZZOPARDI, D., DALLOSSO, A. R., ELIASON, K., HENDRICKSON, B. C., JONES, N., RAWSTORNE, E., COLLEY, J., MOSKVINA, V., FRYE, C., SAMPSON, J. R., WENSTRUP, R., SCHOLL, T. & CHEADLE, J. P. 2008. Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas. *Cancer Res,* 68**,** 358-63.

BAI, H., MADABUSHI, A., GUAN, X. & LU, A. L. 2010. Interaction between human mismatch repair recognition proteins and checkpoint sensor Rad9-Rad1-Hus1. *DNA Repair (Amst),* 9**,** 478-87.

BAKKER, E., HOFKER, M. H., GOOR, N., MANDEL, J. L., WROGEMANN, K., DAVIES, K. E., KUNKEL, L. M., WILLARD, H. F., FENTON, W. A., SANDKUYL, L. & ET AL. 1985. Prenatal diagnosis and carrier detection of Duchenne muscular dystrophy with closely linked RFLPs. *Lancet,* 1**,** 655-8.

BARTSCH, H., NAIR, J. & OWEN, R. W. 2002. Exocyclic DNA adducts as oxidative stress markers in colon carcinogenesis: potential role of lipid peroxidation, dietary fat and antioxidants. *Biol Chem,* 383**,** 915-21.

BECOUARN, Y., AGOSTINI, C., TRUFFLANDIER, N. & BOULANGER, V. 2001. Oxaliplatin: available data in non-colorectal gastrointestinal malignancies. *Crit Rev Oncol Hematol,* 40**,** 265-72.

BENSON, D., LIPMAN, D. J. & OSTELL, J. 1993. GenBank. *Nucleic Acids Res,* 21**,** 2963-5.

BLOSS, C. S., SCHORK, N. J. & TOPOL, E. J. 2011. Effect of direct-to-consumer genomewide profiling to assess disease risk. *N Engl J Med,* 364**,** 524-34.

BODMER, W. & BONILLA, C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet,* 40**,** 695-701.

BOIGE, V., MENDIBOURE, J., PIGNON, J. P., LORIOT, M. A., CASTAING, M., BARROIS, M., MALKA, D., TREGOUET, D. A., BOUCHE, O., LE CORRE, D., MIRAN, I., MULOT, C., DUCREUX, M., BEAUNE, P. & LAURENT-PUIG, P. 2010. Pharmacogenetic assessment of toxicity and outcome in patients with metastatic colorectal cancer treated with LV5FU2, FOLFOX, and FOLFIRI: FFCD 2000-05. *J Clin Oncol,* 28**,** 2556-64.

BRABEC, V. & KASPARKOVA, J. 2005. Modifications of DNA by platinum complexes. Relation to resistance of tumors to platinum antitumor drugs. *Drug Resist Updat,* 8**,** 131-46.

BRAUN, A., LITTLE, D. P. & KOSTER, H. 1997. Detecting CFTR gene mutations by using primer oligo base extension and mass spectrometry. *Clin Chem,* 43**,** 1151-8.

BRODERICK, P., CARVAJAL-CARMONA, L., PITTMAN, A. M., WEBB, E., HOWARTH, K., ROWAN, A., LUBBE, S., SPAIN, S., SULLIVAN, K., FIELDING, S., JAEGER, E., VIJAYAKRISHNAN, J., KEMP, Z., GORMAN, M., CHANDLER, I., PAPAEMMANUIL, E., PENEGAR, S., WOOD, W., SELLICK, G., QURESHI, M., TEIXEIRA, A., DOMINGO, E., BARCLAY, E., MARTIN, L., SIEBER, O., KERR, D., GRAY, R., PETO, J., CAZIER, J. B., TOMLINSON, I. & HOULSTON, R. S. 2007. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet,* 39**,** 1315-7.

CARLSON, C. S., EBERLE, M. A., RIEDER, M. J., SMITH, J. D., KRUGLYAK, L. & NICKERSON, D. A. 2003. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat Genet,* 33**,** 518-21.

CHEADLE, J. P. & SAMPSON, J. R. 2003. Exposing the MYtH about base excision repair and human inherited disease. *Hum Mol Genet,* 12 Spec No 2**,** R159-65.

CHEN, X., LEVINE, L. & KWOK, P. Y. 1999. Fluorescence polarization in homogeneous nucleic acid analysis. *Genome Res,* 9**,** 492-8.

CHEN, X., LIVAK, K. J. & KWOK, P. Y. 1998. A homogeneous, ligase-mediated DNA diagnostic test. *Genome Res,* 8**,** 549-56.

CHEVILLARD, S., RADICELLA, J. P., LEVALOIS, C., LEBEAU, J., POUPON, M. F., OUDARD, S., DUTRILLAUX, B. & BOITEUX, S. 1998. Mutations in OGG1, a gene involved in the repair of oxidative DNA damage, are found in human lung and kidney tumours. *Oncogene,* 16**,** 3083-6.

CHOI, M., SCHOLL, U. I., JI, W., LIU, T., TIKHONOVA, I. R., ZUMBO, P., NAYIR, A., BAKKALOGLU, A., OZEN, S., SANJAD, S., NELSON-WILLIAMS, C., FARHI, A., MANE, S. & LIFTON, R. P. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A,* 106**,** 19096-101.

COHEN, J. C., KISS, R. S., PERTSEMLIDIS, A., MARCEL, Y. L., MCPHERSON, R. & HOBBS, H. H. 2004. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science,* 305**,** 869-72.

COLLEY, J., JONES, S., DALLOSSO, A. R., MAYNARD, J. H., HUMPHREYS, V., DOLWANI, S., SAMPSON, J. R. & CHEADLE, J. P. 2005. Rapid recognition of aberrant dHPLC elution profiles using the Transgenomic Navigator software. *Hum Mutat,* 26**,** 165.

CONDE, L., VAQUERIZAS, J. M., FERRER-COSTA, C., DE LA CRUZ, X., OROZCO, M. & DOPAZO, J. 2005. PupasView: a visual tool for selecting suitable SNPs, with putative pathological effect in genes, for genotyping purposes. *Nucleic Acids Res,* 33**,** W501-5.

COOPER, D. N., BALL, E. V. & KRAWCZAK, M. 1998. The human gene mutation database. *Nucleic Acids Res,* 26**,** 285-7.

COOPER, D. N. & KRAWCZAK, M. 1993. *Human Gene Mutation,* Oxford, BIOS Scientific Publishers.

COTTON, R. G., RODRIGUES, N. R. & CAMPBELL, R. D. 1988. Reactivity of cytosine and thymine in single-base-pair mismatches with hydroxylamine and osmium tetroxide and its application to the study of mutations. *Proc Natl Acad Sci U S A,* 85**,** 4397-401.

COTTON, R. G. H., EDKINS, E. & FORREST, S. 1998. *Mutation Detection. A Practical Approach,* New York, Oxford University Press, Inc.

COUZIN, J. 2002. Human genome. HapMap launched with pledges of $100 million.

CUPPEN, E. 2007. Genotyping by Allele-Specific Amplification (KASPar). *CSH Protoc,* 2007**,** pdb prot4841.

CUTLER, D. J., ZWICK, M. E., CARRASQUILLO, M. M., YOHN, C. T., TOBIN, K. P., KASHUK, C., MATHEWS, D. J., SHAH, N. A., EICHLER, E. E., WARRINGTON, J. A. & CHAKRAVARTI, A. 2001. High-throughput variation detection and genotyping using microarrays. *Genome Res,* 11**,** 1913-25.

DA COSTA AGUIAR, V. R., DE FREITAS CORDEIRO-SILVA, M., DE CARVALHO, A. A. & LOURO, I. D. 2011. Comparison of DGGE and immunohistochemistry in the detection of TP53 variants in a Brazilian sample of sporadic breast tumors. *Mol Biol Rep,* 38**,** 3351-4.

DALLOSSO, A. R., DOLWANI, S., JONES, N., JONES, S., COLLEY, J., MAYNARD, J., IDZIASZCZYK, S., HUMPHREYS, V., ARNOLD, J., DONALDSON, A., ECCLES, D., ELLIS, A., EVANS, D. G., FRAYLING, I. M., HES, F. J., HOULSTON, R. S., MAHER, E. R., NIELSEN, M., PARRY, S., TYLER, E., MOSKVINA, V., CHEADLE, J. P. & SAMPSON, J. R. 2008. Inherited predisposition to colorectal adenomas caused by multiple rare alleles of MUTYH but not OGG1, NUDT1, NTH1 or NEIL 1, 2 or 3. *Gut,* 57**,** 1252-5.

DAS, S. K., AUSTIN, M. D., AKANA, M. C., DESHPANDE, P., CAO, H. & XIAO, M. 2010. Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Res,* 38**,** e177.

DORE, A. S., KILKENNY, M. L., RZECHORZEK, N. J. & PEARL, L. H. 2009. Crystal structure of the rad9-rad1-hus1 DNA damage checkpoint complex--implications for clamp loading and regulation. *Mol Cell,* 34**,** 735-45.

DOUILLARD, J. Y. 2000. Irinotecan and high-dose fluorouracil/leucovorin for metastatic colorectal cancer. *Oncology (Williston Park),* 14**,** 51-5.

DVORNYK, V., LONG, J. R., XIONG, D. H., LIU, P. Y., ZHAO, L. J., SHEN, H., ZHANG, Y. Y., LIU, Y. J., ROCHA-SANCHEZ, S., XIAO, P., RECKER, R. R. & DENG, H. W. 2004. Current limitations of SNP data from the public domain for studies of complex disorders: a test for ten candidate genes for obesity and osteoporosis. *BMC Genet,* 5**,** 4.

DZAGNIDZE, A., KATSARAVA, Z., MAKHALOVA, J., LIEDERT, B., YOON, M. S., KAUBE, H., LIMMROTH, V. & THOMALE, J. 2007. Repair capacity for platinum-DNA adducts determines the severity of cisplatin-induced peripheral neuropathy. *J Neurosci,* 27**,** 9451-7.

ELIASON, K., HENDRICKSON, B. C., JUDKINS, T., NORTON, M., LECLAIR, B., LYON, E., WARD, B., NOLL, W. & SCHOLL, T. 2005. The potential for increased clinical sensitivity in genetic testing for polyposis colorectal cancer through the analysis of MYH mutations in North American patients. *J Med Genet,* 42**,** 95-6.

FAIVRE, S., CHAN, D., SALINAS, R., WOYNAROWSKA, B. & WOYNAROWSKI, J. M. 2003. DNA strand breaks and apoptosis induced by oxaliplatin in cancer cells. *Biochem Pharmacol,* 66**,** 225-37.

FAN, J. B., CHEE, M. S. & GUNDERSON, K. L. 2006. Highly parallel genomic assays. *Nat Rev Genet,* 7**,** 632-44.

FEARNHEAD, N. S., BRITTON, M. P. & BODMER, W. F. 2001. The ABC of APC. *Hum Mol Genet,* 10**,** 721-33.

FEARNHEAD, N. S., WILDING, J. L., WINNEY, B., TONKS, S., BARTLETT, S., BICKNELL, D. C., TOMLINSON, I. P., MORTENSEN, N. J. & BODMER, W. F. 2004. Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc Natl Acad Sci U S A,* 101**,** 15992-7.

FEARNHEAD, N. S., WINNEY, B. & BODMER, W. F. 2005. Rare variant hypothesis for multifactorial inheritance: susceptibility to colorectal adenomas as a model. *Cell Cycle,* 4**,** 521-5.

FISCHER, S. G. & LERMAN, L. S. 1979. Length-independent separation of DNA restriction fragments in two-dimensional gel electrophoresis. *Cell,* 16**,** 191-200.

FODDE, R., SMITS, R. & CLEVERS, H. 2001. APC, signal transduction and genetic instability in colorectal cancer. *Nat Rev Cancer,* 1**,** 55-67.

FRAYLING, I. M., BECK, N. E., ILYAS, M., DOVE-EDWIN, I., GOODMAN, P., PACK, K., BELL, J. A., WILLIAMS, C. B., HODGSON, S. V., THOMAS, H. J., TALBOT, I. C., BODMER, W. F. & TOMLINSON, I. P. 1998. The APC variants I1307K and E1317Q are associated with

colorectal tumors, but not always with a family history. *Proc Natl Acad Sci U S A,* 95**,** 10722-7.

FREIMUTH, R. R., STORMO, G. D. & MCLEOD, H. L. 2005. PolyMAPr: programs for polymorphism database mining, annotation, and functional analysis. *Hum Mutat,* 25**,** 110-7.

GABRIEL, S. B., SCHAFFNER, S. F., NGUYEN, H., MOORE, J. M., ROY, J., BLUMENSTIEL, B., HIGGINS, J., DEFELICE, M., LOCHNER, A., FAGGART, M., LIU-CORDERO, S. N., ROTIMI, C., ADEYEMO, A., COOPER, R., WARD, R., LANDER, E. S., DALY, M. J. & ALTSHULER, D. 2002. The structure of haplotype blocks in the human genome. *Science,* 296**,** 2225-9.

GISMONDI, V., BONELLI, L., SCIALLERO, S., MARGIOCCO, P., VIEL, A., RADICE, P., MONDINI, P., SALA, P., MONTERA, M. P., MARENI, C., QUAIA, M., FORNASARIG, M., GENTILE, M., PIETRO, G., ROSSINI, P., ARRIGONI, A., MEUCCI, G. M., BRUZZI, P. & VARESCO, L. 2002. Prevalence of the E1317Q variant of the APC gene in Italian patients with colorectal adenomas. *Genet Test,* 6**,** 313-7.

GOLDBERG, R. M., SARGENT, D. J., MORTON, R. F., FUCHS, C. S., RAMANATHAN, R. K., WILLIAMSON, S. K., FINDLAY, B. P., PITOT, H. C. & ALBERTS, S. R. 2004. A randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer. *J Clin Oncol,* 22**,** 23-30.

GREGG, S. Q., ROBINSON, A. R. & NIEDERNHOFER, L. J. 2011. Physiological consequences of defects in ERCC1-XPF DNA repair endonuclease. *DNA Repair (Amst),* 10**,** 781-91.

GROSS, E., ARNOLD, N., GOETTE, J., SCHWARZ-BOEGER, U. & KIECHLE, M. 1999. A comparison of BRCA1 mutation analysis by direct sequencing, SSCP and DHPLC. *Hum Genet,* 105**,** 72-8.

GRYFE, R., DI NICOLA, N., LAL, G., GALLINGER, S. & REDSTON, M. 1999. Inherited colorectal polyposis and cancer risk of the APC I1307K polymorphism. *Am J Hum Genet,* 64**,** 378-84.

GUNDRY, C. N., VANDERSTEEN, J. G., REED, G. H., PRYOR, R. J., CHEN, J. & WITTWER, C. T. 2003. Amplicon melting analysis with labeled primers: a closed-tube method for differentiating homozygotes and heterozygotes. *Clin Chem,* 49**,** 396-406.

HAHNLOSER, D., PETERSEN, G. M., RABE, K., SNOW, K., LINDOR, N. M., BOARDMAN, L., KOCH, B., DOESCHER, D., WANG, L., STEENBLOCK, K. & THIBODEAU, S. N. 2003. The APC E1317Q variant in adenomatous polyps and colorectal cancers. *Cancer Epidemiol Biomarkers Prev,* 12**,** 1023-8.

HALL, J. G., EIS, P. S., LAW, S. M., REYNALDO, L. P., PRUDENT, J. R., MARSHALL, D. J., ALLAWI, H. T., MAST, A. L., DAHLBERG, J. E., KWIATKOWSKI, R. W., DE ARRUDA, M., NERI, B. P.

& LYAMICHEV, V. I. 2000. Sensitive detection of DNA polymorphisms by the serial invasive signal amplification reaction. *Proc Natl Acad Sci U S A,* 97**,** 8272-7.

HARTMANN, J. T. & LIPP, H. P. 2003. Toxicity of platinum compounds. *Expert Opin Pharmacother,* 4**,** 889-901.

HATTORI, M., SHIBATA, A., YOSHIOKA, K. & SAKAKI, Y. 1993. Orphan peak analysis: a novel method for detection of point mutations using an automated fluorescence DNA sequencer. *Genomics,* 15**,** 415-7.

HAZRA, T. K., IZUMI, T., KOW, Y. W. & MITRA, S. 2003. The discovery of a new family of mammalian enzymes for repair of oxidatively damaged DNA, and its physiological implications. *Carcinogenesis,* 24**,** 155-7.

HE, W., ZHAO, Y., ZHANG, C., AN, L., HU, Z., LIU, Y., HAN, L., BI, L., XIE, Z., XUE, P., YANG, F. & HANG, H. 2008. Rad9 plays an important role in DNA mismatch repair through physical interaction with MLH1. *Nucleic Acids Res,* 36**,** 6406-17.

HEWETT, M., OLIVER, D. E., RUBIN, D. L., EASTON, K. L., STUART, J. M., ALTMAN, R. B. & KLEIN, T. E. 2002. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res,* 30**,** 163-5.

HISAMUDDIN, I. M. & YANG, V. W. 2004. Genetics of colorectal cancer. *MedGenMed,* 6**,** 13.

HOLLAND, P. M., ABRAMSON, R. D., WATSON, R. & GELFAND, D. H. 1991. Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of Thermus aquaticus DNA polymerase. *Proc Natl Acad Sci U S A,* 88**,** 7276-80.

HOOGENDOORN, B., NORTON, N., KIROV, G., WILLIAMS, N., HAMSHERE, M. L., SPURLOCK, G., AUSTIN, J., STEPHENS, M. K., BUCKLAND, P. R., OWEN, M. J. & O'DONOVAN, M. C. 2000. Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. *Hum Genet,* 107**,** 488-93.

HOULSTON, R. S., CHEADLE, J., DOBBINS, S. E., TENESA, A., JONES, A. M., HOWARTH, K., SPAIN, S. L., BRODERICK, P., DOMINGO, E., FARRINGTON, S., PRENDERGAST, J. G., PITTMAN, A. M., THEODORATOU, E., SMITH, C. G., OLVER, B., WALTHER, A., BARNETSON, R. A., CHURCHMAN, M., JAEGER, E. E., PENEGAR, S., BARCLAY, E., MARTIN, L., GORMAN, M., MAGER, R., JOHNSTONE, E., MIDGLEY, R., NIITTYMAKI, I., TUUPANEN, S., COLLEY, J., IDZIASZCZYK, S., THOMAS, H. J., LUCASSEN, A. M., EVANS, D. G., MAHER, E. R., MAUGHAN, T., DIMAS, A., DERMITZAKIS, E., CAZIER, J. B., AALTONEN, L. A., PHAROAH, P., KERR, D. J., CARVAJAL-CARMONA, L. G., CAMPBELL, H., DUNLOP, M. G. & TOMLINSON, I. P. 2010. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat Genet,* 42**,** 973-7.

HOULSTON, R. S., WEBB, E., BRODERICK, P., PITTMAN, A. M., DI BERNARDO, M. C., LUBBE, S., CHANDLER, I., VIJAYAKRISHNAN, J., SULLIVAN, K., PENEGAR, S., CARVAJAL-CARMONA, L., HOWARTH, K., JAEGER, E., SPAIN, S. L., WALTHER, A., BARCLAY, E., MARTIN, L., GORMAN, M., DOMINGO, E., TEIXEIRA, A. S., KERR, D., CAZIER, J. B., NIITTYMAKI, I., TUUPANEN, S., KARHU, A., AALTONEN, L. A., TOMLINSON, I. P., FARRINGTON, S. M., TENESA, A., PRENDERGAST, J. G., BARNETSON, R. A., CETNARSKYJ, R., PORTEOUS, M. E., PHAROAH, P. D., KOESSLER, T., HAMPE, J., BUCH, S., SCHAFMAYER, C., TEPEL, J., SCHREIBER, S., VOLZKE, H., CHANG-CLAUDE, J., HOFFMEISTER, M., BRENNER, H., ZANKE, B. W., MONTPETIT, A., HUDSON, T. J., GALLINGER, S., CAMPBELL, H. & DUNLOP, M. G. 2008. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet,* 40**,** 1426-35.

HOWELL, W. M., JOBS, M., GYLLENSTEN, U. & BROOKES, A. J. 1999. Dynamic allele-specific hybridization. A new method for scoring single nucleotide polymorphisms. *Nat Biotechnol,* 17**,** 87-8.

HOWORKA, S., CHELEY, S. & BAYLEY, H. 2001. Sequence-specific detection of individual DNA strands using engineered nanopores. *Nat Biotechnol,* 19**,** 636-9.

HUYCKE, M. M. & GASKINS, H. R. 2004. Commensal bacteria, redox stress, and colorectal cancer: mechanisms and models. *Exp Biol Med (Maywood),* 229**,** 586-97.

IWAHANA, H., FUJIMURA, M., TAKAHASHI, Y., IWABUCHI, T., YOSHIMOTO, K. & ITAKURA, M. 1996. Multiple fluorescence-based PCR-SSCP analysis using internal fluorescent labeling of PCR products. *Biotechniques,* 21**,** 510-4, 516-9.

JAEGER, E., WEBB, E., HOWARTH, K., CARVAJAL-CARMONA, L., ROWAN, A., BRODERICK, P., WALTHER, A., SPAIN, S., PITTMAN, A., KEMP, Z., SULLIVAN, K., HEINIMANN, K., LUBBE, S., DOMINGO, E., BARCLAY, E., MARTIN, L., GORMAN, M., CHANDLER, I., VIJAYAKRISHNAN, J., WOOD, W., PAPAEMMANUIL, E., PENEGAR, S., QURESHI, M., FARRINGTON, S., TENESA, A., CAZIER, J. B., KERR, D., GRAY, R., PETO, J., DUNLOP, M., CAMPBELL, H., THOMAS, H., HOULSTON, R. & TOMLINSON, I. 2008. Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat Genet,* 40**,** 26-8.

JASS, J. R., YOUNG, J. & LEGGETT, B. A. 2002. Evolution of colorectal cancer: change of pace and change of direction. *J Gastroenterol Hepatol,* 17**,** 17-26.

JOHANNSSON, O., OSTERMEYER, E. A., HAKANSSON, S., FRIEDMAN, L. S., JOHANSSON, U., SELLBERG, G., BRONDUM-NIELSEN, K., SELE, V., OLSSON, H., KING, M. C. & BORG, A. 1996. Founding BRCA1 mutations in hereditary breast and ovarian cancer in southern Sweden. *Am J Hum Genet,* 58**,** 441-50.

JOHNSON, N. P., HOESCHELE, J. D., RAHN, R. O., O'NEILL, J. P. & HSIE, A. W. 1980. Mutagenicity, cytotoxicity, and DNA binding of platinum(II)-chloroammines in Chinese hamster ovary cells. *Cancer Res,* 40**,** 1463-8.

JONES, S., EMMERSON, P., MAYNARD, J., BEST, J. M., JORDAN, S., WILLIAMS, G. T., SAMPSON, J. R. & CHEADLE, J. P. 2002. Biallelic germline mutations in MYH predispose to multiple colorectal adenoma and somatic G:C-->T:A mutations. *Hum Mol Genet,* 11**,** 2961-7.

JONKER, D. J., O'CALLAGHAN, C. J., KARAPETIS, C. S., ZALCBERG, J. R., TU, D., AU, H. J., BERRY, S. R., KRAHN, M., PRICE, T., SIMES, R. J., TEBBUTT, N. C., VAN HAZEL, G., WIERZBICKI, R., LANGER, C. & MOORE, M. J. 2007. Cetuximab for the treatment of colorectal cancer. *N Engl J Med,* 357**,** 2040-8.

KALER, S. G., DEVANEY, J. M., PETTIT, E. L., KIRSHMAN, R. & MARINO, M. A. 2000. Novel method for molecular detection of the two common hereditary hemochromatosis mutations. *Genet Test,* 4**,** 125-9.

KANDA, T., ODA, M., YONEZAWA, M., TAMAGAWA, K., ISA, F., HANAKAGO, R. & TSUKAGOSHI, H. 1990. Peripheral neuropathy in xeroderma pigmentosum. *Brain,* 113 (Pt 4)**,** 1025-44.

KARAPETIS, C. S., KHAMBATA-FORD, S., JONKER, D. J., O'CALLAGHAN, C. J., TU, D., TEBBUTT, N. C., SIMES, R. J., CHALCHAL, H., SHAPIRO, J. D., ROBITAILLE, S., PRICE, T. J., SHEPHERD, L., AU, H.-J., LANGER, C., MOORE, M. J. & ZALCBERG, J. R. 2008. K-ras Mutations and Benefit from Cetuximab in Advanced Colorectal Cancer. *New England Journal of Medicine,* 359**,** 1757-1765.

KINZLER, K. W. & VOGELSTEIN, B. 1996. Lessons from hereditary colorectal cancer. *Cell,* 87**,** 159-70.

KNOWLES & SELBY 2005. *Introduction and Molecular Biology of Cancer*.

KNUDSON, A. G., JR. 1971. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A,* 68**,** 820-3.

KOBAYASHI, M., RAPPAPORT, E., BLASBAND, A., SEMERARO, A., SARTORE, M., SURREY, S. & FORTINA, P. 1995. Fluorescence-based DNA minisequence analysis for detection of known single-base changes in genomic DNA. *Mol Cell Probes,* 9**,** 175-82.

KOHNO, T., SHINMURA, K., TOSAKA, M., TANI, M., KIM, S. R., SUGIMURA, H., NOHMI, T., KASAI, H. & YOKOTA, J. 1998. Genetic polymorphisms and alternative splicing of the hOGG1 gene, that is involved in the repair of 8-hydroxyguanine in damaged DNA. *Oncogene,* 16**,** 3219-25.

LAKEN, S. J., PETERSEN, G. M., GRUBER, S. B., ODDOUX, C., OSTRER, H., GIARDIELLO, F. M., HAMILTON, S. R., HAMPEL, H., MARKOWITZ, A., KLIMSTRA, D., JHANWAR, S., WINAWER, S., OFFIT, K., LUCE, M. C., KINZLER, K. W. & VOGELSTEIN, B. 1997. Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. *Nat Genet,* 17**,** 79-83.

LAMLUM, H., AL TASSAN, N., JAEGER, E., FRAYLING, I., SIEBER, O., REZA, F. B., ECKERT, M., ROWAN, A., BARCLAY, E., ATKIN, W., WILLIAMS, C., GILBERT, J., CHEADLE, J., BELL, J., HOULSTON, R., BODMER, W., SAMPSON, J. & TOMLINSON, I. 2000. Germline APC variants in patients with multiple colorectal adenomas, with evidence for the particular importance of E1317Q. *Hum Mol Genet,* 9**,** 2215-21.

LANDEGREN, U., KAISER, R., SANDERS, J. & HOOD, L. 1988. A ligase-mediated gene detection technique. *Science,* 241**,** 1077-80.

LEHMANN, A. R., MCGIBBON, D. & STEFANINI, M. 2011. Xeroderma pigmentosum. *Orphanet J Rare Dis,* 6**,** 70.

LEVY, S., SUTTON, G., NG, P. C., FEUK, L., HALPERN, A. L., WALENZ, B. P., AXELROD, N., HUANG, J., KIRKNESS, E. F., DENISOV, G., LIN, Y., MACDONALD, J. R., PANG, A. W. C., SHAGO, M., STOCKWELL, T. B., TSIAMOURI, A., BAFNA, V., BANSAL, V., KRAVITZ, S. A., BUSAM, D. A., BEESON, K. Y., MCINTOSH, T. C., REMINGTON, K. A., ABRIL, J. F., GILL, J., BORMAN, J., ROGERS, Y.-H., FRAZIER, M. E., SCHERER, S. W., STRAUSBERG, R. L. & VENTER, J. C. 2007. The Diploid Genome Sequence of an Individual Human. *PLoS Biol,* 5**,** e254.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

LIN, S. Y., SU, Y. N., HUNG, C. C., TSAY, W., CHIOU, S. S., CHANG, C. T., HO, H. N. & LEE, C. N. 2008. Mutation spectrum of 122 hemophilia A families from Taiwanese population by LD-PCR, DHPLC, multiplex PCR and evaluating the clinical application of HRM. *BMC Med Genet,* 9**,** 53.

LIU, L., LI, Y., LI, S., HU, N., HE, Y., PONG, R., LIN, D., LU, L. & LAW, M. 2012. Comparison of next-generation sequencing systems. *J Biomed Biotechnol,* 2012**,** 251364.

LIU, X. & ROY, R. 2002. Truncation of amino-terminal tail stimulates activity of human endonuclease III (hNTH1). *J Mol Biol,* 321**,** 265-76.

LODISH, H., BERK, A., ZIPURSKY, S. L., MATSUDAIRA, P., BLATIMORE, D. & DARNELL, J. 2000. *Molecular Cell Biology,* New York, W. H. Freeman.

LYAMICHEV, V., MAST, A. L., HALL, J. G., PRUDENT, J. R., KAISER, M. W., TAKOVA, T., KWIATKOWSKI, R. W., SANDER, T. J., DE ARRUDA, M., ARCO, D. A., NERI, B. P. & BROW, M. A. 1999. Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nat Biotechnol,* 17**,** 292-6.

MAKINO, R., YAZYU, H., KISHIMOTO, Y., SEKIYA, T. & HAYASHI, K. 1992. F-SSCP: fluorescence-based polymerase chain reaction-single-strand conformation polymorphism (PCR-SSCP) analysis. *PCR Methods Appl,* 2**,** 10-3.

MANASTER, C., ZHENG, W., TEUBER, M., WACHTER, S., DORING, F., SCHREIBER, S. & HAMPE, J. 2005. InSNP: a tool for automated detection and visualization of SNPs and InDels. *Hum Mutat,* 26**,** 11-9.

MARGULIES, M., EGHOLM, M., ALTMAN, W. E., ATTIYA, S., BADER, J. S., BEMBEN, L. A., BERKA, J., BRAVERMAN, M. S., CHEN, Y. J., CHEN, Z., DEWELL, S. B., DU, L., FIERRO, J. M., GOMES, X. V., GODWIN, B. C., HE, W., HELGESEN, S., HO, C. H., IRZYK, G. P., JANDO, S. C., ALENQUER, M. L., JARVIE, T. P., JIRAGE, K. B., KIM, J. B., KNIGHT, J. R., LANZA, J. R., LEAMON, J. H., LEFKOWITZ, S. M., LEI, M., LI, J., LOHMAN, K. L., LU, H., MAKHIJANI, V. B., MCDADE, K. E., MCKENNA, M. P., MYERS, E. W., NICKERSON, E., NOBILE, J. R., PLANT, R., PUC, B. P., RONAN, M. T., ROTH, G. T., SARKIS, G. J., SIMONS, J. F., SIMPSON, J. W., SRINIVASAN, M., TARTARO, K. R., TOMASZ, A., VOGT, K. A., VOLKMER, G. A., WANG, S. H., WANG, Y., WEINER, M. P., YU, P., BEGLEY, R. F. & ROTHBERG, J. M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437**,** 376-80.

MARTIN, S. T., MATSUBAYASHI, H., ROGERS, C. D., PHILIPS, J., COUCH, F. J., BRUNE, K., YEO, C. J., KERN, S. E., HRUBAN, R. H. & GOGGINS, M. 2005. Increased prevalence of the BRCA2 polymorphic stop codon K3326X among individuals with familial pancreatic cancer. *Oncogene,* 24**,** 3652-6.

MATHE, E., OLIVIER, M., KATO, S., ISHIOKA, C., HAINAUT, P. & TAVTIGIAN, S. V. 2006. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res,* 34**,** 1317-25.

MATSUMOTO, Y., ZHANG, Q. M., TAKAO, M., YASUI, A. & YONEI, S. 2001. Escherichia coli Nth and human hNTH1 DNA glycosylases are involved in removal of 8-oxoguanine from 8-oxoguanine/guanine mispairs in DNA. *Nucleic Acids Res,* 29**,** 1975-81.

MATSUMURA, Y., NISHIGORI, C., YAGI, T., IMAMURA, S. & TAKEBE, H. 1998. Characterization of molecular defects in xeroderma pigmentosum group F in relation to its clinically mild symptoms. *Hum Mol Genet,* 7**,** 969-74.

MAUGHAN, T. S., ADAMS, R. A., SMITH, C. G., MEADE, A. M., SEYMOUR, M. T., WILSON, R. H., IDZIASZCZYK, S., HARRIS, R., FISHER, D., KENNY, S. L., KAY, E., MITCHELL, J. K., MADI, A., JASANI, B., JAMES, M. D., BRIDGEWATER, J., KENNEDY, M. J., CLAES, B., LAMBRECHTS, D., KAPLAN, R. & CHEADLE, J. P. 2011. Addition of cetuximab to oxaliplatin-based first-line combination chemotherapy for treatment of advanced colorectal cancer: results of the randomised phase 3 MRC COIN trial. *Lancet,* 377**,** 2103-14.

MAUGHAN, T. S., JAMES, R. D., KERR, D. J., LEDERMANN, J. A., MCARDLE, C., SEYMOUR, M. T., COHEN, D., HOPWOOD, P., JOHNSTON, C. & STEPHENS, R. J. 2002. Comparison of survival, palliation, and quality of life with three chemotherapy regimens in metastatic colorectal cancer: a multicentre randomised trial. *Lancet,* 359**,** 1555-63.

MAXAM, A. M. & GILBERT, W. 1977. A new method for sequencing DNA. *Proc Natl Acad Sci U S A,* 74**,** 560-4.

MAZOYER, S., DUNNING, A. M., SEROVA, O., DEARDEN, J., PUGET, N., HEALEY, C. S., GAYTHER, S. A., MANGION, J., STRATTON, M. R., LYNCH, H. T., GOLDGAR, D. E., PONDER, B. A. & LENOIR, G. M. 1996. A polymorphic stop codon in BRCA2. *Nat Genet,* 14**,** 253-4.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res,* 20**,** 1297-303.

MCWHINNEY, S. R., GOLDBERG, R. M. & MCLEOD, H. L. 2009. Platinum neurotoxicity pharmacogenetics. *Mol Cancer Ther,* 8**,** 10-6.

MERCER, T. R. 2011. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.*

MEYERS, M., HWANG, A., WAGNER, M. W., BRUENING, A. J., VEIGL, M. L., SEDWICK, W. D. & BOOTHMAN, D. A. 2003. A role for DNA mismatch repair in sensing and responding to fluoropyrimidine damage. *Oncogene,* 22**,** 7376-88.

MITCHELL, A. A., ZWICK, M. E., CHAKRAVARTI, A. & CUTLER, D. J. 2004. Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. *Bioinformatics,* 20**,** 1022-32.

MONZO, M., MORENO, I., NAVARRO, A., IBEAS, R., ARTELLS, R., GEL, B., MARTINEZ, F., MORENO, J., HERNANDEZ, R. & NAVARRO-VIGO, M. 2007. Single nucleotide polymorphisms in nucleotide excision repair genes XPA, XPD, XPG and ERCC1 in advanced colorectal cancer patients treated with first-line oxaliplatin/fluoropyrimidine. *Oncology,* 72**,** 364-70.

MOORE, G. E. 1965. Cramming more components onto integrated circuits. *Electronics Magazine,* 38**,** 4.

MORAK, M., MASSDORF, T., SYKORA, H., KERSCHER, M. & HOLINSKI-FEDER, E. 2011. First evidence for digenic inheritance in hereditary colorectal cancer by mutations in the base excision repair genes. *European Journal of Cancer,* 47**,** 1046-1055.

MUSUMECI, L., ARTHUR, J. W., CHEUNG, F. S., HOQUE, A., LIPPMAN, S. & REICHARDT, J. K. 2010. Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. *Hum Mutat,* 31**,** 67-73.

NAJMABADI, H., HU, H., GARSHASBI, M., ZEMOJTEL, T., ABEDINI, S. S., CHEN, W., HOSSEINI, M., BEHJATI, F., HAAS, S., JAMALI, P., ZECHA, A., MOHSENI, M., PUTTMANN, L., VAHID, L. N., JENSEN, C., MOHEB, L. A., BIENEK, M., LARTI, F., MUELLER, I.,

WEISSMANN, R., DARVISH, H., WROGEMANN, K., HADAVI, V., LIPKOWITZ, B., ESMAEELI-NIEH, S., WIECZOREK, D., KARIMINEJAD, R., FIROUZABADI, S. G., COHEN, M., FATTAHI, Z., ROST, I., MOJAHEDI, F., HERTZBERG, C., DEHGHAN, A., RAJAB, A., BANAVANDI, M. J. S., HOFFER, J., FALAH, M., MUSANTE, L., KALSCHEUER, V., ULLMANN, R., KUSS, A. W., TZSCHACH, A., KAHRIZI, K. & ROPERS, H. H. 2011. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature,* 478**,** 57-63.

NELSON, N. C., HAMMOND, P. W., MATSUDA, E., GOUD, A. A. & BECKER, M. M. 1996. Detection of all single-base mismatches in solution by chemiluminescence. *Nucleic Acids Res,* 24**,** 4998-5003.

NEWTON, C. R., GRAHAM, A., HEPTINSTALL, L. E., POWELL, S. J., SUMMERS, C., KALSHEKER, N., SMITH, J. C. & MARKHAM, A. F. 1989. Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic Acids Res,* 17**,** 2503-16.

NG, P. C. & HENIKOFF, S. 2006. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet,* 7**,** 61-80.

NG, S. B., BIGHAM, A. W., BUCKINGHAM, K. J., HANNIBAL, M. C., MCMILLIN, M. J., GILDERSLEEVE, H. I., BECK, A. E., TABOR, H. K., COOPER, G. M., MEFFORD, H. C., LEE, C., TURNER, E. H., SMITH, J. D., RIEDER, M. J., YOSHIURA, K., MATSUMOTO, N., OHTA, T., NIIKAWA, N., NICKERSON, D. A., BAMSHAD, M. J. & SHENDURE, J. 2010a. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet,* 42**,** 790-3.

NG, S. B., BUCKINGHAM, K. J., LEE, C., BIGHAM, A. W., TABOR, H. K., DENT, K. M., HUFF, C. D., SHANNON, P. T., JABS, E. W., NICKERSON, D. A., SHENDURE, J. & BAMSHAD, M. J. 2010b. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet,* 42**,** 30-5.

NG, S. B., TURNER, E. H., ROBERTSON, P. D., FLYGARE, S. D., BIGHAM, A. W., LEE, C., SHAFFER, T., WONG, M., BHATTACHARJEE, A., EICHLER, E. E., BAMSHAD, M., NICKERSON, D. A. & SHENDURE, J. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature,* 461**,** 272-6.

NICKERSON, D. A., TOBE, V. O. & TAYLOR, S. L. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res,* 25**,** 2745-51.

NILSSON, M., MALMGREN, H., SAMIOTAKI, M., KWIATKOWSKI, M., CHOWDHARY, B. P. & LANDEGREN, U. 1994. Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science,* 265**,** 2085-8.

NORDSTROM, T., RONAGHI, M., FORSBERG, L., DE FAIRE, U., MORGENSTERN, R. & NYREN, P. 2000. Direct analysis of single-nucleotide polymorphism on double-stranded DNA by pyrosequencing. *Biotechnol Appl Biochem,* 31 (Pt 2)**,** 107-12.

NYHOLT, D. R. 2004. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet,* 74**,** 765-9.

O'DONOVAN, M. C., OEFNER, P. J., ROBERTS, S. C., AUSTIN, J., HOOGENDOORN, B., GUY, C., SPEIGHT, G., UPADHYAYA, M., SOMMER, S. S. & MCGUFFIN, P. 1998. Blind analysis of denaturing high-performance liquid chromatography as a tool for mutation detection. *Genomics,* 52**,** 44-9.

O'ROAK, B. J., DERIZIOTIS, P., LEE, C., VIVES, L., SCHWARTZ, J. J., GIRIRAJAN, S., KARAKOC, E., MACKENZIE, A. P., NG, S. B., BAKER, C., RIEDER, M. J., NICKERSON, D. A., BERNIER, R., FISHER, S. E., SHENDURE, J. & EICHLER, E. E. 2011. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet,* 43**,** 585-9.

OCEAN, A. J. & VAHDAT, L. T. 2004. Chemotherapy-induced peripheral neuropathy: pathogenesis and emerging therapies. *Support Care Cancer,* 12**,** 619-25.

OEFNER, P. J. & UNDERHILL, P. A. (eds.) 1998. *DNA mutation detection using denaturing high performance liquid chromatography (DHPLC),* New York: Wiley.

ORITA, M., IWAHANA, H., KANAZAWA, H., HAYASHI, K. & SEKIYA, T. 1989. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proc Natl Acad Sci U S A,* 86**,** 2766-70.

PAPADOPOULOS, N. & LINDBLOM, A. 1997. Molecular basis of HNPCC: mutations of MMR genes. *Hum Mutat,* 10**,** 89-99.

PASETTO, L. M., D'ANDREA, M. R., ROSSI, E. & MONFARDINI, S. 2006. Oxaliplatin-related neurotoxicity: how and why? *Crit Rev Oncol Hematol,* 59**,** 159-68.

PEDERSEN, A. G. & NIELSEN, H. 1997. Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. *Proc Int Conf Intell Syst Mol Biol,* 5**,** 226-33.

PELTOMAKI, P. 2001. Deficient DNA mismatch repair: a common etiologic factor for colon cancer. *Hum Mol Genet,* 10**,** 735-40.

PICCIOLI, P., SERRA, M., GISMONDI, V., PEDEMONTE, S., LOIACONO, F., LASTRAIOLI, S., BERTARIO, L., DE ANGIOLETTI, M., VARESCO, L. & NOTARO, R. 2006. Multiplex tetra-primer amplification refractory mutation system PCR to detect 6 common germline mutations of the MUTYH gene associated with polyposis and colorectal cancer. *Clin Chem,* 52**,** 739-43.

PIRULLI, D., GIORDANO, M., PUZZER, D., CROVELLA, S., RIGATO, I., TIRIBELLI, C., MOMIGLIANO-RICHIARDI, P. & AMOROSO, A. 2000. Rapid method for detection of

extra (TA) in the promoter of the bilirubin-UDP-glucuronosyl transferase 1 gene associated with Gilbert syndrome. *Clin Chem,* 46**,** 129-31.

POPAT, S., STONE, J., COLEMAN, G., MARSHALL, G., PETO, J., FRAYLING, I. & HOULSTON, R. 2000. Prevalence of the APC E1317Q variant in colorectal cancer patients. *Cancer Lett,* 149**,** 203-6.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J. & SHAM, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* 81**,** 559-75.

QI, X., BAKHT, S., DEVOS, K. M., GALE, M. D. & OSBOURN, A. 2001. L-RCA (ligation-rolling circle amplification): a general method for genotyping of single nucleotide polymorphisms (SNPs). *Nucleic Acids Res,* 29**,** E116.

QUASTHOFF, S. & HARTUNG, H. P. 2002. Chemotherapy-induced peripheral neuropathy. *J Neurol,* 249**,** 9-17.

REARDON, J. T., VAISMAN, A., CHANEY, S. G. & SANCAR, A. 1999. Efficient nucleotide excision repair of cisplatin, oxaliplatin, and Bis-aceto-ammine-dichloro-cyclohexylamine-platinum(IV) (JM216) platinum intrastrand DNA diadducts. *Cancer Res,* 59**,** 3968-71.

REICH, D. E., GABRIEL, S. B. & ALTSHULER, D. 2003. Quality and completeness of SNP databases. *Nat Genet,* 33**,** 457-8.

RIESNER, D., STEGER, G., ZIMMAT, R., OWENS, R. A., WAGENHOFER, M., HILLEN, W., VOLLBACH, S. & HENCO, K. 1989. Temperature-gradient gel electrophoresis of nucleic acids: analysis of conformational transitions, sequence variations, and protein-nucleic acid interactions. *Electrophoresis,* 10**,** 377-89.

RIVA, A. & KOHANE, I. S. 2002. SNPper: retrieval and analysis of human SNPs. *Bioinformatics,* 18**,** 1681-5.

ROBBINS, J. H., KRAEMER, K. H., MERCHANT, S. N. & BRUMBACK, R. A. 2002. Adult-onset xeroderma pigmentosum neurological disease--observations in an autopsy case. *Clin Neuropathol,* 21**,** 18-23.

ROBERTS, E., DEEBLE, V. J., WOODS, C. G. & TAYLOR, G. R. 1997. Potassium permanganate and tetraethylammonium chloride are a safe and effective substitute for osmium tetroxide in solid-phase fluorescent chemical cleavage of mismatch. *Nucleic Acids Res,* 25**,** 3377-8.

ROBERTS, M. R., SHIELDS, P. G., AMBROSONE, C. B., NIE, J., MARIAN, C., KRISHNAN, S. S., GOERLITZ, D. S., MODALI, R., SEDDON, M., LEHMAN, T., AMEND, K. L., TREVISAN, M., EDGE, S. B. & FREUDENHEIM, J. L. 2011. Single-nucleotide polymorphisms in DNA

repair genes and association with breast cancer risk in the web study. *Carcinogenesis,* 32**,** 1223-30.

ROLDAN-ARJONA, T., WEI, Y. F., CARTER, K. C., KLUNGLAND, A., ANSELMINO, C., WANG, R. P., AUGUSTUS, M. & LINDAHL, T. 1997. Molecular cloning and functional expression of a human cDNA encoding the antimutator enzyme 8-hydroxyguanine-DNA glycosylase. *Proc Natl Acad Sci U S A,* 94**,** 8016-20.

RONAGHI, M., NYGREN, M., LUNDEBERG, J. & NYREN, P. 1999. Analyses of secondary structures in DNA by pyrosequencing. *Anal Biochem,* 267**,** 65-71.

RUDOLPH, J. G., WHITE, S., SOKOLSKY, C., BOZAK, D., MAZZANTI, C., LIPSKY, R. H. & GOLDMAN, D. 2002. Determination of melting temperature for variant detection using dHPLC: a comparison between an empirical approach and DNA melting prediction software. *Genet Test,* 6**,** 169-76.

SAKUMI, K., FURUICHI, M., TSUZUKI, T., KAKUMA, T., KAWABATA, S., MAKI, H. & SEKIGUCHI, M. 1993. Cloning and expression of cDNA for a human enzyme that hydrolyzes 8-oxo-dGTP, a mutagenic substrate for DNA synthesis. *J Biol Chem,* 268**,** 23524-30.

SALTZ, L. B., COX, J. V., BLANKE, C., ROSEN, L. S., FEHRENBACHER, L., MOORE, M. J., MAROUN, J. A., ACKLAND, S. P., LOCKER, P. K., PIROTTA, N., ELFRING, G. L. & MILLER, L. L. 2000. Irinotecan plus Fluorouracil and Leucovorin for Metastatic Colorectal Cancer. *New England Journal of Medicine,* 343**,** 905-914.

SAMPSON, J. R., DOLWANI, S., JONES, S., ECCLES, D., ELLIS, A., EVANS, D. G., FRAYLING, I., JORDAN, S., MAHER, E. R., MAK, T., MAYNARD, J., PIGATTO, F., SHAW, J. & CHEADLE, J. P. 2003. Autosomal recessive colorectal adenomatous polyposis due to inherited mutations of MYH. *Lancet,* 362**,** 39-41.

SAMPSON, J. R., JONES, S., DOLWANI, S. & CHEADLE, J. P. 2005. MutYH (MYH) and colorectal cancer. *Biochem Soc Trans,* 33**,** 679-83.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A,* 74**,** 5463-7.

SCHAEFFELER, E., LANG, T., ZANGER, U. M., EICHELBAUM, M. & SCHWAB, M. 2001. High-throughput genotyping of thiopurine S-methyltransferase by denaturing HPLC. *Clin Chem,* 47**,** 548-55.

SCHEEFF, E. D., BRIGGS, J. M. & HOWELL, S. B. 1999. Molecular modeling of the intrastrand guanine-guanine DNA adducts produced by cisplatin and oxaliplatin. *Mol Pharmacol,* 56**,** 633-43.

SCHMALZING, D., BELENKY, A., NOVOTNY, M. A., KOUTNY, L., SALAS-SOLANO, O., EL-DIFRAWY, S., ADOURIAN, A., MATSUDAIRA, P. & EHRLICH, D. 2000. Microchip electrophoresis: a method for high-speed SNP detection. *Nucleic Acids Res,* 28**,** E43.

SCHWAAB, R., OLDENBURG, J., LALLOZ, M. R., SCHWAAB, U., PEMBERTON, S., HANFLAND, P., BRACKMANN, H. H., TUDDENHAM, E. G. & MICHAELIDES, K. 1997. Factor VIII gene mutations found by a comparative study of SSCP, DGGE and CMC and their analysis on a molecular model of factor VIII protein. *Hum Genet,* 101**,** 323-32.

SHEFFIELD, V. C., BECK, J. S., KWITEK, A. E., SANDSTROM, D. W. & STONE, E. M. 1993. The sensitivity of single-strand conformation polymorphism analysis for the detection of single base substitutions. *Genomics,* 16**,** 325-32.

SHERRY, S. T., WARD, M. H., KHOLODOV, M., BAKER, J., PHAN, L., SMIGIELSKI, E. M. & SIROTKIN, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res,* 29**,** 308-11.

SHI, G., CHANG, D. Y., CHENG, C. C., GUAN, X., VENCLOVAS, C. & LU, A. L. 2006. Physical and functional interactions between MutY glycosylase homologue (MYH) and checkpoint proteins Rad9-Rad1-Hus1. *Biochem J,* 400**,** 53-62.

SHINMURA, K., KOHNO, T., KASAI, H., KODA, K., SUGIMURA, H. & YOKOTA, J. 1998. Infrequent mutations of the hOGG1 gene, that is involved in the excision of 8-hydroxyguanine in damaged DNA, in human gastric cancer. *Jpn J Cancer Res,* 89**,** 825-8.

SIEBER, O., LIPTON, L., HEINIMANN, K. & TOMLINSON, I. 2003a. Colorectal tumourigenesis in carriers of the APC I1307K variant: lone gunman or conspiracy? *J Pathol,* 199**,** 137-9.

SIEBER, O. M., LIPTON, L., CRABTREE, M., HEINIMANN, K., FIDALGO, P., PHILLIPS, R. K., BISGAARD, M. L., ORNTOFT, T. F., AALTONEN, L. A., HODGSON, S. V., THOMAS, H. J. & TOMLINSON, I. P. 2003b. Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH. *N Engl J Med,* 348**,** 791-9.

SIJBERS, A. M., VAN DER SPEK, P. J., ODIJK, H., VAN DEN BERG, J., VAN DUIN, M., WESTERVELD, A., JASPERS, N. G., BOOTSMA, D. & HOEIJMAKERS, J. H. 1996. Mutational analysis of the human nucleotide excision repair gene ERCC1. *Nucleic Acids Res,* 24**,** 3370-80.

SLUPSKA, M. M., LUTHER, W. M., CHIANG, J. H., YANG, H. & MILLER, J. H. 1999. Functional expression of hMYH, a human homolog of the Escherichia coli MutY protein. *J Bacteriol,* 181**,** 6210-3.

SMITH, A. V. 2008. Retrieving HapMap Data Using HapMart. *Cold Spring Harb Protoc,* 2008**,** pdb.prot5026-.

SOKOLOV, B. P. 1990. Primer extension technique for the detection of single nucleotide in genomic DNA. *Nucleic Acids Res,* 18**,** 3671.

STOEHLMACHER, J., GHADERI, V., IOBAL, S., GROSHEN, S., TSAO-WEI, D., PARK, D. & LENZ, H. J. 2001. A polymorphism of the XRCC1 gene predicts for response to platinum based treatment in advanced colorectal cancer. *Anticancer Res,* 21**,** 3075-9.

STOEHLMACHER, J., PARK, D. J., ZHANG, W., YANG, D., GROSHEN, S., ZAHEDY, S. & LENZ, H. J. 2004. A multivariate analysis of genomic polymorphisms: prediction of clinical outcome to 5-FU/oxaliplatin combination chemotherapy in refractory colorectal cancer. *Br J Cancer,* 91**,** 344-54.

STOESSER, G., MOSELEY, M. A., SLEEP, J., MCGOWRAN, M., GARCIA-PASTOR, M. & STERK, P. 1998. The EMBL nucleotide sequence database. *Nucleic Acids Res,* 26**,** 8-15.

STOJIC, L., BRUN, R. & JIRICNY, J. 2004. Mismatch repair and DNA damage signalling. *DNA Repair (Amst),* 3**,** 1091-101.

TATENO, Y. & GOJOBORI, T. 1997. DNA Data Bank of Japan in the age of information biology. *Nucleic Acids Res,* 25**,** 14-7.

TAYLOR, G. R. & DAY, N. M. 2005. *Guide to Mutation Detection*, Wiley.

TENESA, A., FARRINGTON, S. M., PRENDERGAST, J. G., PORTEOUS, M. E., WALKER, M., HAQ, N., BARNETSON, R. A., THEODORATOU, E., CETNARSKYJ, R., CARTWRIGHT, N., SEMPLE, C., CLARK, A. J., REID, F. J., SMITH, L. A., KAVOUSSANAKIS, K., KOESSLER, T., PHAROAH, P. D., BUCH, S., SCHAFMAYER, C., TEPEL, J., SCHREIBER, S., VOLZKE, H., SCHMIDT, C. O., HAMPE, J., CHANG-CLAUDE, J., HOFFMEISTER, M., BRENNER, H., WILKENING, S., CANZIAN, F., CAPELLA, G., MORENO, V., DEARY, I. J., STARR, J. M., TOMLINSON, I. P., KEMP, Z., HOWARTH, K., CARVAJAL-CARMONA, L., WEBB, E., BRODERICK, P., VIJAYAKRISHNAN, J., HOULSTON, R. S., RENNERT, G., BALLINGER, D., ROZEK, L., GRUBER, S. B., MATSUDA, K., KIDOKORO, T., NAKAMURA, Y., ZANKE, B. W., GREENWOOD, C. M., RANGREJ, J., KUSTRA, R., MONTPETIT, A., HUDSON, T. J., GALLINGER, S., CAMPBELL, H. & DUNLOP, M. G. 2008. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet,* 40**,** 631-7.

THRUSH, D. C., HOLTI, G., BRADLEY, W. G., CAMPBELL, M. J. & WALTON, J. N. 1974. Neurological manifestations of xeroderma pigmentosum in two siblings. *J Neurol Sci,* 22**,** 91-104.

TOBLER, A. R., SHORT, S., ANDERSEN, M. R., PANER, T. M., BRIGGS, J. C., LAMBERT, S. M., WU, P. P., WANG, Y., SPOONDE, A. Y., KOEHLER, R. T., PEYRET, N., CHEN, C., BROOMER, A. J., RIDZON, D. A., ZHOU, H., HOO, B. S., HAYASHIBARA, K. C., LEONG, L. N., MA, C. N., ROSENBLUM, B. B., DAY, J. P., ZIEGLE, J. S., DE LA VEGA, F. M., RHODES, M. D., HENNESSY, K. M. & WENZ, H. M. 2005. The SNPlex genotyping system: a flexible and scalable platform for SNP genotyping. *J Biomol Tech,* 16**,** 398-406.

TOMLINSON, I., WEBB, E., CARVAJAL-CARMONA, L., BRODERICK, P., KEMP, Z., SPAIN, S., PENEGAR, S., CHANDLER, I., GORMAN, M., WOOD, W., BARCLAY, E., LUBBE, S.,

MARTIN, L., SELLICK, G., JAEGER, E., HUBNER, R., WILD, R., ROWAN, A., FIELDING, S., HOWARTH, K., SILVER, A., ATKIN, W., MUIR, K., LOGAN, R., KERR, D., JOHNSTONE, E., SIEBER, O., GRAY, R., THOMAS, H., PETO, J., CAZIER, J. B. & HOULSTON, R. 2007. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet,* 39**,** 984-8.

TOMLINSON, I. P., WEBB, E., CARVAJAL-CARMONA, L., BRODERICK, P., HOWARTH, K., PITTMAN, A. M., SPAIN, S., LUBBE, S., WALTHER, A., SULLIVAN, K., JAEGER, E., FIELDING, S., ROWAN, A., VIJAYAKRISHNAN, J., DOMINGO, E., CHANDLER, I., KEMP, Z., QURESHI, M., FARRINGTON, S. M., TENESA, A., PRENDERGAST, J. G., BARNETSON, R. A., PENEGAR, S., BARCLAY, E., WOOD, W., MARTIN, L., GORMAN, M., THOMAS, H., PETO, J., BISHOP, D. T., GRAY, R., MAHER, E. R., LUCASSEN, A., KERR, D., EVANS, D. G., SCHAFMAYER, C., BUCH, S., VOLZKE, H., HAMPE, J., SCHREIBER, S., JOHN, U., KOESSLER, T., PHAROAH, P., VAN WEZEL, T., MORREAU, H., WIJNEN, J. T., HOPPER, J. L., SOUTHEY, M. C., GILES, G. G., SEVERI, G., CASTELLVI-BEL, S., RUIZ-PONTE, C., CARRACEDO, A., CASTELLS, A., FORSTI, A., HEMMINKI, K., VODICKA, P., NACCARATI, A., LIPTON, L., HO, J. W., CHENG, K. K., SHAM, P. C., LUK, J., AGUNDEZ, J. A., LADERO, J. M., DE LA HOYA, M., CALDES, T., NIITTYMAKI, I., TUUPANEN, S., KARHU, A., AALTONEN, L., CAZIER, J. B., CAMPBELL, H., DUNLOP, M. G. & HOULSTON, R. S. 2008. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet,* 40**,** 623-30.

TYAGI, S. & KRAMER, F. R. 1996. Molecular beacons: probes that fluoresce upon hybridization. *Nat Biotechnol,* 14**,** 303-8.

VASEN, H. F., WATSON, P., MECKLIN, J. P. & LYNCH, H. T. 1999. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. *Gastroenterology,* 116**,** 1453-6.

VERSLUIS, L. F., ROZEMULLER, E., TONKS, S., MARSH, S. G., BOUWENS, A. G., BODMER, J. G. & TILANUS, M. G. 1993. High-resolution HLA-DPB typing based upon computerized analysis of data obtained by fluorescent sequencing of the amplified polymorphic exon 2. *Hum Immunol,* 38**,** 277-83.

VIGUIER, J., BOIGE, V., MIQUEL, C., POCARD, M., GIRAUDEAU, B., SABOURIN, J. C., DUCREUX, M., SARASIN, A. & PRAZ, F. 2005. ERCC1 codon 118 polymorphism is a predictive factor for the tumor response to oxaliplatin/5-fluorouracil combination chemotherapy in patients with advanced colorectal cancer. *Clin Cancer Res,* 11**,** 6212-7.

WALLACE, R. B., SHAFFER, J., MURPHY, R. F., BONNER, J., HIROSE, T. & ITAKURA, K. 1979. Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res,* 6**,** 3543-57.

WANG, K., LI, M. & HAKONARSON, H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res,* 38**,** e164.

WANG, L., LIU, S., NIU, T. & XU, X. 2005. SNPHunter: a bioinformatic software for single nucleotide polymorphism data acquisition and management. *BMC Bioinformatics,* 6**,** 60.

WARE, J. S., ROBERTS, A. M. & COOK, S. A. 2012. Next generation sequencing for clinical diagnostics and personalised medicine: implications for the next generation cardiologist. *Heart,* 98**,** 276-81.

WETZEL, C., HU, J., RIETHMACHER, D., BENCKENDORFF, A., HARDER, L., EILERS, A., MOSHOURAB, R., KOZLENKOV, A., LABUZ, D., CASPANI, O., ERDMANN, B., MACHELSKA, H., HEPPENSTALL, P. A. & LEWIN, G. R. 2007. A stomatin-domain protein essential for touch sensation in the mouse. *Nature,* 445**,** 206-9.

WHEELER, D. A., SRINIVASAN, M., EGHOLM, M., SHEN, Y., CHEN, L., MCGUIRE, A., HE, W., CHEN, Y. J., MAKHIJANI, V., ROTH, G. T., GOMES, X., TARTARO, K., NIAZI, F., TURCOTTE, C. L., IRZYK, G. P., LUPSKI, J. R., CHINAULT, C., SONG, X. Z., LIU, Y., YUAN, Y., NAZARETH, L., QIN, X., MUZNY, D. M., MARGULIES, M., WEINSTOCK, G. M., GIBBS, R. A. & ROTHBERG, J. M. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature,* 452**,** 872-6.

WHITTOCK, N. V. 2005. Chemical Cleavage of Mismatch Theory and Clinical Applications *MEDICAL BIOMETHODS HANDBOOK***,** 183-194.

WOOD, R. D., MITCHELL, M. & LINDAHL, T. 2005. Human DNA repair genes, 2005. *Mutat Res,* 577**,** 275-83.

WOOD, R. D., MITCHELL, M., SGOUROS, J. & LINDAHL, T. 2001. Human DNA repair genes. *Science,* 291**,** 1284-9.

WORTHEY, E. A., MAYER, A. N., SYVERSON, G. D., HELBLING, D., BONACCI, B. B., DECKER, B., SERPE, J. M., DASU, T., TSCHANNEN, M. R., VEITH, R. L., BASEHORE, M. J., BROECKEL, U., TOMITA-MITCHELL, A., ARCA, M. J., CASPER, J. T., MARGOLIS, D. A., BICK, D. P., HESSNER, M. J., ROUTES, J. M., VERBSKY, J. W., JACOB, H. J. & DIMMOCK, D. P. 2011. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med,* 13**,** 255-62.

XI, T., JONES, I. M. & MOHRENWEISER, H. W. 2004. Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. *Genomics,* 83**,** 970-9.

XIAO, W. & OEFNER, P. J. 2001. Denaturing high-performance liquid chromatography: A review. *Hum Mutat,* 17**,** 439-74.

XU, H., GREGORY, S. G., HAUSER, E. R., STENGER, J. E., PERICAK-VANCE, M. A., VANCE, J. M., ZUCHNER, S. & HAUSER, M. A. 2005. SNPselector: a web tool for selecting SNPs for genetic association studies. *Bioinformatics,* 21**,** 4181-6.

YOSHIMURA, K., HANAOKA, T., OHNAMI, S., KOHNO, T., LIU, Y., YOSHIDA, T., SAKAMOTO, H. & TSUGANE, S. 2003. Allele frequencies of single nucleotide polymorphisms (SNPs) in 40 candidate genes for gene-environment studies on cancer: data from population-based Japanese random samples. *J Hum Genet,* 48**,** 654-8.

YOUIL, R., KEMPER, B. W. & COTTON, R. G. 1995. Screening for mutations by enzyme mismatch cleavage with T4 endonuclease VII. *Proc Natl Acad Sci U S A,* 92**,** 87-91.

YOUNG, J., BARKER, M., FRASER, L., WALSH, M. D., SPRING, K., BIDEN, K. G., HOPPER, J. L., LEGGETT, B. A. & JASS, J. R. 2002. Mutation searching in colorectal cancer studies: experience with a denaturing high-pressure liquid chromatography system for exon-by-exon scanning of tumour suppressor genes. *Pathology,* 34**,** 529-33.

ZAHARY, M. N., KAUR, G., ABU HASSAN, M. R., SINGH, H., NAIK, V. R. & ANKATHIL, R. 2012. Germline mutation analysis of MLH1 and MSH2 in Malaysian Lynch syndrome patients. *World J Gastroenterol,* 18**,** 814-20.

Appendix

Appendix A: PCR primers for dHPLC.

Gene/primer name	Primer Sequence (5' to 3')	Product Size (bp)	Annealing Temp. (ºC)
MSH6	**(NG_007111.1)**		
MSH6X2newF	GGTATGTATTTCCTTTTGGCAA	243	60
MSH6X2newR	CATGGCAGTAGTGACTCTT		
MSH6X3F	ACCCGGCCCTTATTGTTTAT	260	60
MSH6X3R	CTTCCCCCATCACCCTAA		
MSH6X4.1F	AAATACTCTTTCCTTGCCTGGC	283	60
MSH6X4.1R	CTTTGACAGGGCTGTTCAGG		
MSH6X4.2F	GCAGTGGAGTGGGGGATA G	293	54
MSH6X4.2R	TTCCTCCTTAAGCCATTCTAAA		
MSH6X4.3F	GTCGCCCTACTGTTTGGTATC	277	60
MSH6X4.3R	ATGAATTCCAGCCCCAGTTC		
MSH6X4.4F	GCTGTACCACATGGATGCTCT	289	60
MSH6X4.4R	TCACCTTCCAGCACACTGTAA		
MSH6X4.6F	CGCCATTGTTCGAGATTTAG	286	60
MSH6X4.6R	TGTCAACCCAATGGAATCAG		
MSH6X4.8F	CTGATTCCATTGGGTTGACA	268	60
MSH6X4.8R	ACTGCATCTAGCACCATTCG		
MSH6X4.11F	CCAAAGCAGGCTTTGACTC	288	52
MSH6X4.11R	TCCGTTCTTCAGCATTTATGA		
MSH6X5.1F	CCAAACGATGAAGCCTCACT	283	60
MSH6X5.1R	CCCCCATATTTGGTCCAGTA		
MSH6X5.2F	GCCATCCTTGCATTACGAA	237	60
MSH6X5.2R	TGTTTGGAAAATGATCACCTAAG		
MSH6X6F	TGCCTAGCTCTTACGTAAGGG	248	60
MSH6X6R	CTGAATGAGAACTTAAGTGGGAA		
MSH6X9F	CTTGCTAGCACATGTATCGCT	294	60
MSH6X9R	TTTCTTTGAAACTTAAGGTCAGTT		
NEIL2	**(NM_001135746.1)**		
NEIL2X1newF	TGCTTGGCACCTGTTAAAG	238	56
NEIL2X1newR	CCTTCAGAGGAGTATTAC		
NEIL2X2.2F	GGCGAGGATGATTCTGAGTA	210	55
NEIL2X2.2R	GGCCCTATGACTATGAGAGC		
NEIL2X3newF	ATGTGGGGATGTGTGTATGT	402	60
NEIL2X3newR	AACAGCACTGGCACAACTAG		
NEIL2X4.1F	AGCTGATTTCTGCCTTGTTAA	275	60
NEIL2X4.1R	GGCACTGTTCTTTCTGGTAGA		

NEIL3	**(NM_018248.2)**		
NEIL3X1F	GCACAGCGGTATTCTCAC	315	60
NEIL3X1R	GCCAGAGCAAAGAGATGAC		
NEIL3X2F	ACCTGGGCGTGAAGTAATA	348	55
NEIL3X2R	GGAGCTCTCTTAAAAACACAGT		
NEIL3X5F	CAAGGCGTATTATTTCTTACAG	258	60
NEIL3X5R	CTCCATGCTCTTTCCACTA		
NEIL3X8.1F	CATGTCTGGGTTTCTGTAAGT	428	55
NEIL3X8.1R _new	CTAGTATCTGGTTTGCTTTGT		
NEIL3X10F	TGATAACAGCAGTGAAGACTATA	351	60
NEIL3X10R	GAGGAGGACCAAACATTATAC		
OGG1	**(NM_002542.5)**		
OGG1X5F	CCGGCTTTGGGGCTATA	466	60
OGG1X5R	GTTTCTACCATCCCAGCCCA		
OGG1X7F	ACCTCCCAACACTGTCACTA	489	60
OGG1X7R2	TGAACCGGGAGTTTCTCTGC		

Appendix B: SNP genotyping Assays.

Gene/variant	Fragment	Primer name	Primer sequence (5' to 3')	Product Size (bp)	Anneal temp (°C)	Assay
OGG1 **(NM_002542.5)**						
p.R46Q	Exon 1	M1F	CTTTGGGCGTCGACGAG	338	60	*Hph*I digest
(c.137 G>A)		OGG1_Ex1R1	CCCGTGCTTGTTTCCTCTT			G allele digested (130 + 208bp)
p.A85S	Exon 2	OGG1_A85S_FG	ACCGAGGAGACAAGAGCCCGG	150	56	ARMS[1]
(c.253 G>T)		OGG1_A85S_FT	ACCGAGGAGACAAGAGCCGGT			
		OGG1_A85S_R	GGAATTTCTGAGCCACCTCTT			
c.748 -15 G>C	Exon 5	OGG1_748-15_FN	TCTTCCACAAGGGCTCAATC	205	57	ARMS[3]
		OGG1_748-15_FM	TCTTCCACAAGGGCTCAATG			
		OGG1_748-15_RCP	CTGCAGCCCACTTTCCTCAC			
p.A288V	Exon 5	OGG1_A288V_FC	GCACCCTACCACGTCCCATGC	138	56	ARMS[1]
(c.863 C>T)		OGG1_A288V_FT	GCACCCTACCACGTCCCAAGT			
		OGG1_A288V_R	GTCTGAGGGAGAGGTGACTA			
p.S326C	Exon 7	OGG1_S326F_G	TCAGTGCCGACCTGCGCCATTG	210	60	ARMS[1]
(c.977 C>G)		OGG1_S326F_C	TCAGTGCCGACCTGCGCCATTC			

		OGG1_S326_R2	GTGCCTGGCCTTTGAGGTAGTC			
c.1037+ 279 A>G	Exon 7	OGG1_Ex7_INTF OGG1_Ex7_INTRM	CAGGCACCCCCAAATCAAGC CCTGTCCCCTCCTCACCTCCC	179	55	ARMS[3]
		OGG1_Ex7_INTRN	CCTGTCCCCTCCTCACCTACT			
p.I321T	Exon 7	OGG1_Ex7b_FNI	CTGACAGCTCTGTCAGGTTAT	196	59	ARMS[3]
(c.962 T>C)		OGG1_Ex7b_FMT	CTGACAGCTCTGTCAGGTGAC			
		OGG1_Ex7b_RCP	AGGGATCCTTACTGAAGGAC			
NEIL1 **(NM_001256552.1)**						
c.434+2 T>C	Exon 2	NEI1_434+2_F_T	CAGGAGTACCAGCAGTTCAAGT	174	58	ARMS[1]
		NEI1_434+2_F_C	CAGGAGTACCAGCAGTTCATGC			
		NEI1_434+2_R	AGCTGACCCAGGACAGGTATA			
p.P208S	Exon 4	NEI1_Ex4F	GGATGAACTGCCCAAGTCT	358	55	*Sac*I digest
(c.622 C>T)		NEI1_Ex4R	AGCCTGGAAACACTATTGACA			T allele digested (102bp + 256bp)
p.R339Q	Exon 9	NEI1_R339_F_G	CTAAGAGGACTGCAACCCATCG	121	60	ARMS[1]
(c.1016 G>A)		NEI1_R339_F_A	CTAAGAGGACTGCAACCCAACA			
		NEI1_R339_R	GAGGAAAGCCCACCAGAGG			

NEIL2 **(NM_001135746.1)**						
p.R103Q	Exon2	NEI2_R103_Fnew	GTCTGTAAGGCTTGGATCTCT	225	58	ARMS[1]
(c.308 G>A)	fragment 1	NEI2_R103_RA	GGGGACGAGCTCTGCAGGCT			
		NEI2_R103_RG	GGGGACGAGCTCTGCAGCCC			
p.R164T	Exon 2	NEI2_R164_RG	ATCAGATGGCCACACCATTCCC	181	56	ARMS[1]
(c.491 G>C)	Fragment 2	NEI2_R164_RC	ATCAGATGGCCACACCATTCCG			
		NEI2_R164_F	GCGAGGATGATTCTGAGTATT			
p.P188P	Exon 3	NEI2_Ex3new_F	ATGTGGGGATGTGTGTATGT	396	60	*Bsr*I digest
(c.564 A>G)		NEI2_Ex3new_R	AACAGCACTGGCACAACTAG			A allele digested (174bp + 222bp)
c.689 -13 C>T	Exon 4	NEI2_689-13_RC	ATGATGTTCCCTGAAATAAATCG	169	56	ARMS[1]
	fragment 1	NEI2_689-13_RT	ATGATGTTCCCTGAAATAAAACA			
		NEI2_689-13_F	GAATTGTCAACTCCATTTTACA			
p.R257L	Exon 4	NEI2_R257_FG	GTCCTGAGTGCCTCGCGGCG	482	61	ARMS[2]
(c.770 G>T)	fragment 1	NEI2_R257_FT	GTCCTGAGTGCCTCGCGCCT			
		NEI2_R257_R	CACGATGGTGCAATTCAACAG			
Exon 2 - Exon 4	n/a	NEI2_LD_F NEI2_LD_R	TAAGGCTTGGATCTCTGTTCAT TGCTCTGGCTCCTCTGACA	6500	65	LD-PCR to clone region spanning R103Q and R257L

NEIL3 **(NM_018248.2)**						
p.R15R	Exon 1	NEI3_Ex1F	GCACAGCGGTATTCTCAC	315	55	*Taq*I digest
(c.45 C>A)		NEI3_Ex1R	GCCAGAGCAAAGAGTGAC			A allele digested (100bp + 215bp)
p.R38C	Exon 1	NEI3_Ex1F	GCACAGCGGTATTCTCAC	315	55	*Bst*UI digest
(c.112 C>T)		NEI3_Ex1R	GCCAGAGCAAAGAGTGAC			C allele digested (145bp + 170bp)
p.P117R	Exon 3	NEI3_Ex3F	GACTAACATTTGTTTGCTTATA	295	55	*Bss*SI digest.
(c.349 C>G)		NEI3_Ex3R	AACCAAGCTCTGCAGTAT			G allele digested (116 + 179bp)
p.D132V	Exon 3	NEI1_Ex3F	GCGGGCAGAGATCCTGTA	289	60	*Hin*FI digest.
(c.395 AC>TG)		NEI1_Ex3R	CCACCCCTGTTGTTGAGCTA			TG allele digested (109 + 115 + 71bp)
p.Q172H	Exon 3	NEI1_Ex3R	CCACCCCTGTTGTTGAGCTA	166	56	ARMS[3]
(c.516 G>C)		Q172_NEWREV	TCACCTAGCATCCGGCCTATC			
		H172_NEWREV2	TCACCTAGCATCCGGCCTATG			
c.627+41 A>G	Exon 4	NEI3_Ex4F	ATACCCTGATTAACATAAAGTG	453	55	*Sca*I digest.
		627+41_DIG_REV	GGTAGAAATGGGAAGCGGGAC			A allele digested (93 + 360bp)

c.701+34 del 7bp	Exon 5	NEI3_Ex5_FAM NEI3_Ex5R	TTTCAGCATTCTCTTTTACAGG CTCCATGCTCTTTCCACT	258	58	Fluorescent PCR products sized
p.H286R	Exon 6	H286R_FOR	AGTGCCGTAAAGCAGGACTTGCT	180	52	ARMS[3]
(c.857 A>G)		H286_REV	CAATATCTTACCATATGTCAATAT			
		R286_REV	CAATATCTTACCATATGTCAAGAC			
p.R315Q	Exon 7	NEI3_Ex7F	TGGGGTATTAATGGTTCTATA	307	55	*Dde*I digest
(c.944 G>A)		NEI3_Ex7R	CCCTCTTAGAAATGTAACAAA			A allele digested (8 + 60 + 75 + 164bp)
c.1040 -85 T>C	Exon 8	NEI3_Ex8F1new	CATGTCTGGGTTTCTGTAAGT	428	55	*Apo*I digest
		NEI3_Ex8R1new	ACAAAGCAAAACCAGATACTAG			T allele digested (24 + 111 + 118 + 175bp)
p.V424V	Exon 8	NEI3_Ex8F2new	CTGCATTTGGAACTACAACTCT	245	52	ARMS[3]
(c.1272 G>T)		V424V_NREV	GGGTGCTGTATATCATTCAAATAC			
		V424V_MREV	GGGTGCTGTATATCATTCAAAGAA			
p.R520G	Exon 9	NEI3_Ex9F	TGGGCTTTCCTCATCACTC	265	55	*Bsl*I digest
(c.1558 A>G)		NEI3_Ex9R	GCAGGGCAAGACAATAAGA			G allele digested (163+ 165bp)

MTH1 **(NM_002452.3)**						
p.V106M	Exon 4	MTH1X4V106M_F	TCCCTGGGCTGTGTGTAGAT	220	56	ARMS[4]
(c.316 G>A)		MTH1X4V106M_RN	TCTGTGCAGAAGACATGGAC			
		MTH1X4V106M_RM	GTCTGTGCAGAAGACATGTAT			
p.D142D	Exon 5	D142D_FCP	ATGAAGTTTGGGTTGCACCTC	128	59	ARMS[3]
(426 C>T)		D142D_RN	AGTGGAAACCAGTAGCTGACG			
		D142D_RM	AGTGGAAACCAGTAGCTGCCA			
p.K155I	Exon 5	MTH1X5K155I_F	TGAAGTTTGGGTTGCACCTC	164	57	ARMS[3]
(c.464 A>T)		MTH1X5K155I_RN	CTTGAAGTACCCGTGGACTT			
		MTH1X5K155I_RM	CTTGAAGTACCCGTGGAGTA			
NTH1 **(NM_002528.5)**						
p.Q90X	Exon 2	NTH_2F	ACCTGTGGCCCCACCAGAGG	373	58	*Bst*NI digest.
(c.268 C>T)		NTH_2R	GGGAGGGTGCCAGCCAAAAG			C allele digested (170 + 203bp)

MSH6 **(NG_007111.1)**						
p.P92P	Exon 2	normal	CTTGGCCCAAACCAAATCTTCT	243		
(c.276A>G)		variant	CTTGGCCCAAACCAAATCTTCC			
		common	TAGAATTTCTGTGCTTCAATATTA			
p.D180D	Exon 3	normal	GAGAGCAATGCAACGTGCAAAT	260		
(c.540T>C)		variant	GAGAGCAATGCAACGTGCATAC			
		common	CCCCATCACCCTAACATAAAT			
p.Y214Y (c.642C>T)	Exon 4 fragment 1	MSH6X4.1F MSH6X4.1R	AAATACTCTTTCCTTGCCTGGC CTTTGACAGGGCTGTTCAGG	366	60	*Sna*BI digest. C allele digested (122 + 244bp)
c.3438+14A>T	Exon 5	normal	GACTTTCTGATAACAAAACCTT	237		
		variant	GACTTTCTGATAACAAAACCTA			
		common	ATCCTTGCATTACGAAGACT			

Internal control primers used to validated ARMS assays:

[1]TSC2_Ex26F (5'-GAGCTTTGGCCCTTGGTGATA) and TSC2_Ex26R (5'-CTCGCCCACAGGAGACCTAGA), product size 338bp; [2]TSC2_Ex38F (5'-ACCAGGCAGTAGCCGAGATC) and TSC2_Ex38R (5'-CCACAAGGCCCTCCATGTC), product size 275bp; [3]AJ32 (5'-TGGCACTGAGTTGACACTCT) and AJ31 (5'-CAAATAATGTTTTCCAGAGACA), product size 315bp; [4]AJ54 (5'-CCCTGTCTGCCGTTAAATAC) and AJ55 (5'-GCTGTTTTGCACTCCTCAAT), product size 307bp.

Appendix C: Amplification and dHPLC conditions for glycosylase genes.

Gene / primer name	Primer sequence (5' to 3')	Product Size (bp)	Annealing Temp (ºC)	dHPLC conditions	
				Temp 1 (ºC)	Temp 2 (ºC)
OGG1 **(NM_002542.5)**					
MIF	CTTTGGGCGTCGACGAG	237	60	62.8	65.8
M1R	GAGGGGACAGGCTTCTCAG				
M2F1	ATTGAGTGCCAGGGTTGTCA	245	52	62.1	63
M2R1	CGGAACCCCAGTGGTGATAC				
M2F2	TGTACTAGCGGATCAAGTAT	286	52	61	/
M2R2	TGGCAAAACTGAGTCATAG				
M3F1	GTCTGGTGTTGCTTTCTCTAAC	229	60	59.6	63
M3R1	GTGATGCGGGCGATGTT				
M3F2	TCTCCAGGTGTGCGACTGC	275	60	60	63.5
M3R2	AGGAAGCCTTGAGAAGGTAACC				
M4F	GGAAGAACTTGAAGATGCCT	296	52	63.5	64
M4R	GCTCATTTCCTGCTCTCC				
M5F	CCGGCTTTGGGGCTATA	279	60	63	/
M5R	GTTTCTACCATCCCAGCCCA				
M6F	TACTTCTGTTGATGGGTCAC	153	60	59.3	61.5[a]
M6R	TGGAGGAGAGGAAACCTAG				
OGG1_Ex7F	ACCTCCCAACACTGTCACTA	489	60	55.5	61.5
OGG1_Ex7R2	TGAACCGGGAGTTTCTCTGC				
OGG1_Ex7bF	AGGCTTAGCACTTGCACTTCC	267	60	58.7	/
OGG1_Ex7bR	AGGGATCCTTACTGAAGGAC				
M8F1	CTGTGGCCCACGCACTTGTG	253	60	61.5	/
M8R1	ACGTCCTTGGTCCAGCAGTGGT				
M8F2	GAGAGGGGATTCACAAGGTG	287	60	60.5	62
M8R2	GCCATTAGCTCCAGGCTTAC				
NEIL1 **(NM_001256552.1)**					
NEI1_Ex1F1	AGCCGCTACCTCACAAAGTC	234	60	63.5	65
NEI1_Ex1R1	CGGGCTGAAGCTGAGATG				
NEI1_Ex1F2	GTGGAGAAGTCCTCTGTCAG	426	55	62.5	65
NEI1_Ex1R2	CCAAGAAGGCACTAAGAGA				
NEI1_Ex2F_NEW	AGGGAGAAGAGGAACTGTAAC	352	60	62	/
NEI1_Ex2R_NEW	GGGATCTTCAGCCTGTAGT				
NEI1_Ex3F	GCGGGCAGAGATCCTGTA	289	60	64	/
NEI1_Ex3R	CCACCCCTGTTGTTGAGCTA				
NEI1_Ex4F	GGATGAACTGCCCAAGTCT	358	55	61	62
NEI1_Ex4R	AGCCTGGAAACACTATTGACA				

NEI1_Ex5F	GAGCCTGCCCTCTGATCTC	205	60	62.5	64.5
NEI1_Ex5R	TGGGGTCTCTGCCTGTGT				
NEI1_Ex6/7F	CCAGGCTGATTCCTGAATTA	411	60	58.5	62
NEI1_Ex6/7R	GGCTCAGAAAGCAGTTCAGA				
NEI1_Ex8F	GCCCTAACCAACCTCTGA	281	60	62.5	63.5
NEI1_Ex8R	CATCCCATCCTCTCCTGTAG				
NEI1_Ex9F	TGGGCTTTCCTCATCACTC	265	55	62	/
NEI1_Ex9R	GCAGGGCAAGACAATAAGA				
NEIL2 **(NM_001135746.1)**					
NEI2_Ex1F_NEW	TGCTTGGCACCTGTTAAAG	238	56	62	/
NEI2_Ex1R_NEW	CCTTCAGAGGAGTATTAC				
NEI2_Ex2F1	TTCCCCTCTGGGTCTGT	330	55	56	62.5[b]
NEI2_Ex2R1	CTGCCAAACAAACCAAAGC				
NEI2_Ex2F2	GGCGAGGATGATTCTGAGTA	210	55	62	/
NEI2_Ex2R2	GGCCCTATGACTATGAGAGC				
NEI2_Ex3F_NEW	ATGTGGGGATGTGTGTATGT	402	60	59	60.5
NEI2_Ex3R_NEW	AACAGCACTGGCACAACTAG				
NEI2_Ex4F1	AGCTGATTTCTGCCTTGTTAA	275	60	59	62.5[a]
NEI2_Ex4R1	GGCACTGTTCTTTCTGGTAGA				
NEI2_Ex4F2	ACGTGGTGGAGTTCAGTA	328	60	63	/
NEI2_Ex4R2	GCACCTCTGACCCACACTAT				
NEIL3 **(NM_018248.2)**					
NEI3_Ex1F	GCACAGCGGTATTCTCAC	315	60	62.5	64.5
NEI3_Ex1R	GCCAGAGCAAAGAGATGAC				
NEI3_Ex2F	ACCTGGGCGTGAAGTAATA	348	55	57	/
NEI3_Ex2R	GGAGCTCTCTTAAAAACACAGT				
NEI3_Ex3F	GACTAACATTTGTTTGCTTATA	354	55	55	/
NEI3_Ex3R	TTTCATCAATTACAAACTCTT				
NEI3_Ex4F	TACCCTGATTAACATAAAGTG	382	55[c]	53.5	56.5
NEI3_Ex4R	CGTTGGTTCCTTCACAGTA				
NEI3_Ex5F	CAAGGCGTATTATTTCTTACAG	258	60	57	/
NEI3_Ex5R	CTCCATGCTCTTTCCACTA				
NEI3_Ex6F	CTTGGCAGCACTGTTTGT	389	50c	55	57
NEI3_Ex6R	TCAGGTTCTTTGGTGACATTAT				
NEI3_Ex7F	TGGGGTATTAATGGTTCTATA	307	55	52.5	56
NEI3_Ex7R	CCCTCTTAGAAATGTAACAAA				
NEI3_Ex8F1_NEW	CATGTCTGGGTTTCTGTAAGT	428	55	53	55.5
NEI3_Ex8R1_NEW	CTAGTATCTGGTTTGCTTTGT				
NEI3_Ex8F2	CTGCATTTGGAACTACAACTCT	349	55	55	56.5
NEI3_Ex8R2	CAATTATCTTAAAGACCAAGGTT				

NEI3_Ex9F	GGGGTAAAGTGGTGTGAAT	328	50	53	57
NEI3_Ex9R	CTAGGAAAGTATTGCTATTTGTTA				
NEI3_Ex10F	TGATAACAGCAGTGAAGACTATA	351	60[c]	55.5	56.5
NEI3_Ex10R	GAGGAGGACCAAACATTATAC				
MTH1 **(NM_002452.3)**					
T2F	GCAAGGACAGAGGGCTTTCTG	249	60	60.5	61.5
T2R	CCAGCAGGCCATCAACTGAT				
T3F	GCACGTCATGGCTGACTCT	246	55	62	63.5
T3R	CTGGGAAAGCCGGTTCTAT				
T4F	TCCCTGGGCTGTGTGTAGAT	298	55	64.5	/
T4R	GAGATGGGACCCGCATAGT				
T5F	TGAAGTTTGGGTTGCACCTC	281	55	61.5	63.5
T5R	AGATGGTTTGCGGCTGTTC				
NTH1 **(NM_002528.5)**					
NTH1F	TGTAGTTCTGTGCCGCCCTCTG	273	55	65.5	66.5
NTH1R	TCCAGCCTGCAGCCCCTATC				
NTH2F	ACCTGTGGCCCCACCAGAGG	373	58	62	63.5
NTH2R	GGGAGGGTGCCAGCCAAAAG				
NTH3F	GCAACAAACCAGGGTGTGTC	323	58	64.5	/
NTH3R	GAGGTCTCTCTGAGGCCACTG				
NTH4F	GGGCTGCATCCTCCCAGGTT	287	58	61.5	64.5
NTH4R	GATGTGGGGAATCCCAAGAGC				
NTH5F1	GGTGGAGTGTGCCCCTGTT	423	57[c]	63	64
NTH6R1	CCTGAAGCGTAAAGCCACTTC				

Time shifts of +0.5 min[a] and +1.0 min[b] were applied for dHPLC analysis at this temperature; [c]PCR reaction included 4% DMSO.

Appendix D: PCR Primers.

Gene/primer name	Primer Sequence (5' to 3')	Product Size (bp)	Annealing Temp. (ºC)
APEX **(NM_001244249.1)**			
APE1X1F	GGAGGCGGGGAAAGGATTTAGAG	366	60
APE1X2R	CTGCGACTTCTTCACAAACC		
APE1X3F	GTGAAGAAGTCGCAGGAACCGTAGGCTT	381	60
APE1X3R	CCTGAAGGCTAAACGGAGAA		
APE1X4F	GAATATTGTGCTGCTTGACTC	420	60
APE1X4R	GGGAAAGCAATCAAGAGGTG		
APE1X5F	TTGCTAATTCTCTATCTCTG	673	56
APE1X5R	GAGTGTTTAAAGAAGGAATGG		
MBD4 **(NM_003925.1)**			
MBD4_Ex1F	CTTTCGCAACATTCAGACCTC	297	60
MBD4_Ex1R	ACTGTCCACTCTCCCGATACC		
MBD4_Ex2F	TGAGTAGGCAGTGGAAGATAA	374	58
MBD4_Ex2R	AAGCTAAGATTCCTGCTATGC		
MBD4_Ex3F1	AATGTGGTCCAGTTCTTTTAA	399	56
MBD4_Ex3R1	ATCAACACCCTCATCTTCTTT		
MBD4_Ex3F2	TTCAAACTGGAACCTCAGGAC	361	60
MBD4_Ex3R2	AGCATCAGAAATGCAGACAGT		
MBD4_Ex3F3	GATGCTGAAAGTGAACCTGTT	496	58
MBD4_Ex3F3	TCTTGGCTCTATTTTCACATC		
MBD4_Ex4F	ATTATTTGCATCCCTCAATAT	214	56
MBD4_Ex4R	ATAGTGCATCAGAATTGAAAA		
MBD4_Ex5F	AATCAGAACAGCAAATTCTAA	298	56
MBD4_Ex5R	TGACACACTCAAAATGGACT		
MBD4_Ex6F	CCACCTGGAGTCTTGTAATCA	225	56
MBD4_Ex6R	TATGTTTTTCCTTTGGGTGTA		
MBD4_Ex7F	ATTTTGGGAGGGTGTCTTTAG	205	58
MBD4_Ex7R	CAGAGACCAAATGTGCTGAAT		
MBD4_Ex8F	GGTCTCTGCCTCTGTATCTTATGTTT	262	56
MBD4_Ex8R	TCTTAATGTGTGTGCCAATG		
MPG **(NM_001015052.2)**			
MPG_Ex1F	TCGAGTGTGTCAGGGTGTT	185	58
MPG_Ex1R	CGTCGGCAAAACTGTAATG		
MPG_Ex2F	CCTATTCGGATGCTTATTTA	378	60
MPG_Ex2R	GGGTTCAGGGACAACTG		
MPG_Ex3F	GGGCACTGTTAGGGTGAG	357	60
MPG_Ex3R	CCACCTCAGTCCTCCTAG		
MPG_Ex4F1	GCTCCACTTCCAAACTGTC	242	60
MPG_Ex4R1	CCAGCCATACAGCTTCATC		

MPG_Ex4F2	CCAGCCGTGTCCTCAAG	311	58
MPG_Ex4R2	AAAATCTTGTCTGGGCAGG		
NEIL3			
(NM_018248.2)			
NEI3_Ex1F	GCACAGCGGTATTCTCAC	315	58
NEI3_Ex1R	GCCAGAGCAAAGAGATGAC		
NEI3_Ex2F	ACCTGGGCGTGAAGTAATA	348	58
NEI3_Ex2R	GGAGCTCTCTTAAAAACACAGT		
NEI3_Ex3F	GATATACTTTCTGCCACTCAAAAATGGT	289	56
NEI3_Ex3R	TTTCATCAATTACAAACTCTT		
NEI3_Ex4F	TACCCTGATTAACATAAAGTG	381	56
NEI3_Ex4R	CGTTGGTTCCTTCACAGTA		
NEI3_Ex5F	CAAGGCGTATTATTTCTTACAG	258	56
NEI3_Ex5R	CTCCATGCTCTTTCCACTA		
NEI3_Ex6F	AGCACTGTTTTGTGGATAACAGAATT	383	58
NEI3_Ex6R	TCAGGTTCTTTGGTGACATTAT		
NEI3_Ex7F	TGGGGTATTAATGGTTCTATA	307	58
NEI3_Ex7R	CCCTCTTAGAAATGTAACAAA		
NEI3_Ex8F1	CATGTCTGGGTTTCTGTAAGT	360	58
NEI3_Ex8R1	TGAGGAGTTTCAAAACTCTCCTCCTG		
NEI3_Ex8F2	CTGCATTTGGAACTACAACTCT	349	58
NEI3_Ex8R2	CAATTATCTTAAAGACCAAGGTT		
NEI3_Ex9F	GGGGTAAAGTGGTGTGAAT	328	58
NEI3_Ex9R	CTAGGAAAGTATTGCTATTTGTTA		
NEI3_Ex10F	TGATAACAGCAGTGAAGACTATA	351	58
NEI3_Ex10R	GAGGAGGACCAAACATTATAC		
NUDT5			
(NM_014142.2)			
NUDT5x1F	GCGGCAGTTCCCTCACATAC	389	58
NUDT5x1R	CGAGCCCAAAAAAGGGAGTA		
NUDT5x2F	AAACGTGCCCACAGATTATT	389	58
NUDT5x2R	CCCAATTTTATTCAGCAACTAT		
NUDT5x3F	GGCAAGAGTTTTCGTTGTTA	527	58
NUDT5x3R	AGGTGTGGCTTCAAACTATAT		
NUDT5x4(1)F	CTGTGGAAGGTCAGTGTAGTC	557	58
NUDT5x4(1)R	CTCGCTACGAAATGGTTTAG		
NUDT5x5F	GATTTCCCGCTCCCATCAC	313	60
NUDT5x5R	AGGCCCTTCTGGCTCCAG		
NUDT5x6F	CTACCAAATGTGAGAACTAGT	500	58
NUDT5x6R	GTGGGAATACACTTCATATAG		
NUDT5x7/8F	GCTCGAGTTTGACAATGTAT	465	58
NUDT5x7/8R	CTAGGCATTTGACTTTAGTGA		
NUDT5x9F	AGCATAGGAAGTGACATATA	319	60
NUDT5x9R	AGCTAATGGCAAATCTAC		

PCNA			
(NM_002592.2)			
PCNAX1F	ACGTTCGCCCGCTCGCTCTG	421	64
PCNAX1R	GATGGCCCACGCCAGCCAATGA		
PCNAX2F	GAGAGTGGTAACCCCTTCTAA	501	60
PCNAX3R	CCCAGGAATCCCAGGTTAGA		
PCNAX4F	AGTGAGGGTGCCAAATCATT	536	58
PCNAX4R	GGTAGGATTGGGAAGTTAGGC		
PCNAX5F	TTGGCTGCTTATTAGAAAAACT	603	58
PCNAX5R	GAACTGCTTCTAAGATGCAAG		
POL λ			
(NM_001174084.1)			
POLLx1F	CTACCCCCAAAGCCTGGTCAG	365	64
POLLx1R	AGGCCCTGGACAGGCAGAGT		
POLLx2F	GGGGACTGTGAACACGTCAT	615	60
POLLx2R	CCTAAACCTCTTATAGCTGGGACTAC		
POLLx3F	GCCTTGATCGTACCACTGTAC	492	58
POLLx3R	AGAGCTAAATGGCTTCACAATA		
POLLx4.1F	GGCCTCACACCCAAGGAGA	446	60
POLLx4.1R	GCATTGATGGCCTTGGCATAG		
POLLx4.2F	CCAGAAGGCGACCAATCAC	329	60
POLLx4.2R	CTCCCAGCTTCAACAACTATCAA		
POLLx5F	TTACCCAGCCCTCATTCTATC	471	60
POLLx5R	CCCATCAGAGCACAGCATAG		
POLLx6F	TTCCCAAGTCCTGCTGAGTAC	454	60
POLLx6R	GGCCTGGAGCTTCAGTCTTA		
POLLx7F	GCTTGCCTCCTGCACAGT	482	60
POLLx7R	TGCCTCAGGACTGGAACTTC		
POLLx8.1F	AAGGGCCCAGAGAGGGTAGT	466	64
POLLx8.1R	AGGCCTAAGAGCCTGAAGACAT		
POLLx8.2F	GCCAAAACCAAGGGCATGAGTCT	300	64
POLLx8.2R	GGCCCTGCTCGCTGAGGAA		
POLLx2(1)F	TTTCAGGGTAGGGGACTGT	569	60
POLLx2(1)R	CTCCCACCTAAACCTCTTATAG		
SMUG			
(NM_001243787.1)			
SMUG_Ex1F	TGGATCCCTCCTACTCTG	420	58
SMUG_Ex1R	CCAAGCATCCACCTAGAA		
SMUG_Ex2F1	GGCCTCAGGTCTCCAGTT	314	60
SMUG_Ex2R1	GGCAGCTCAGCAGGAGTA		
SMUG_Ex3F	GAGGTCTTCTTCCATCACTGT	400	60
SMUG_Ex3R	CTTCGAGGTCTTGAATGTGTC		
TDG			

(NM_003211.4)			
TDG_Ex1F	CAGCCACTGTCTGGGTACTG	394	64
TDG_Ex1R	AGAGCAGCCCCGACCTC		
TDG_Ex2F	CTCTCCTCTGTAATCCACTCTA	329	58
TDG_Ex2R	ATCCGATGTTGAACTTTCTAA		
TDG_Ex3F	AGCTGCTAAAGTTTCTAAGTTAA	372	60
TDG_Ex3R	CAAGGACAACTGTTAAGTAAAG		
TDG_Ex4F	TCCACCACTCCTCCATAGAA	360	60
TDG_Ex4R	ACATCCCTCCATTCTCATAGAC		
TDG_Ex5F	GATCGTGCCACTACACTCTA	317	58
TDG_Ex5R	AGCTCAGCTTGAACTAGATACA		
TDG_Ex6_7F	GCTGTCTGAATTTAGCATATTATA	409	58
TDG_Ex6_7R	TCACAATGGATAGGACAAATAA		
TDG_Ex8F	ACAAATATTCTAATCTCAATGAGT	293	60
TDG_Ex8R	TATACACACACAAAATGAATAAA		
TDG_Ex9F	CGGTTTTTACAGTTCTTATG	406	58
TDG_Ex9R	ATTCCCATTCTTCAATAATTT		
TDG_Ex10F	CTGCAAAGAGCTGTGATCAT	343	58
TDG_Ex10R	AGCAAACTGAGGTTCTACTTGT		

Appendix E: PCR Primers

Amplicon	Forward Primer (5' > 3')	Reverse Primer (5' > 3')	Size (bp)
ERCC5 **(NM_000123.3)**			
ERCC5_exon_1	GCCATCTTTGTTGTGTAGGAG	CATCATCCTGCAGATGCCAC	308
ERCC5_exon_2	TTTAGGTAGATCCCATGAGAGC	TTGTACCCATGATGAACTCTC	538
ERCC5_exon_3	ACAGCAATGTTTCTAGTGGTC	AGAATCGCAGGAAATCAAGAC	253
ERCC5_exon_4	AGGTTCCTTCCTTTCTCTCGG	AGAGCATGCCTATTTCAGATGC	345
ERCC5_exon_5	TGTAAGGGGTCCTTAAAAATC	CTTTCACAGTTTGATATACCTC	208
ERCC5_exon_6	CACTTTGTTGCCTGTCACAG	CTGAGATATCGTGTAAGTATTGC	409
ERCC5_exon_7	GGGAAAGGGTGGAAATATGG	GTTTGGCACTATAGTAGTTAATCC	443
ERCC5_exon_8.1	CTAGAAGCGTATTGTCACACTG	TCTAACCACTGCACTCTCC	708
ERCC5_exon_8.2	ACGTAGCCAGCACTAATGAG	GATCTCTGAATTCCTACAGAGG	652
ERCC5_exon_9	GCTCTTGATGATTGCAGGATC	GCAACCACAAGATGTACTGC	435
ERCC5_exon_10	CAGTCAGACTAAATGCAGGC	TTAGGGACACACAGTGACC	346
ERCC5_exon_11	AACTCTGCAGGAATGAATGC	ATGCATCAAACTAGTTCCTTGG	417
ERCC5_exon_12	AGAAGCTGAAACCCACCAGG	CAGCATGACAGTTCATGCTG	748
ERCC5_exon_13	GCCATCATTATACATTGTGGC	AAAGAGTGAAAAGGAGAGCG	517
ERCC5_exon_14	CAAGAATGGGTTCTTTGGACC	GGTCTTTAACAGCTGTCAACC	237
ERCC5_exon_15	CAAGGTTGAGCTTGTTGATTTGG	ATACAAAGACCGTGCCACCAG	778
XPA **(NM_000380.3)**			
XPA_exon_1	GGCGCTCTCACTCAGAAAG	GCTTGCACGAGCCAGTCT	368
XPA_exon_2	AGGTAACATACAGGCTTACC	CATTTCCATATGCATGGCTG	366
XPA_exon_3	TCAGGCATTGCATACATGCTG	TTCTATGGCAGAACCATCGG	434
XPA_exon_4	GTGCCCCTAAGTTGCTGG	CCACACTCTGTAAGCAAAAGCCA	342
XPA_exon_5	TCCTGTGACAATACAGTCAGAG	CTTGAAGACCAACATACTGAGG	546
XPA_exon_6	CATGTACATGGCTGAAAGCTTG	ACAAGGGTTTCATTCATCTATGAAG	359
XPC **(NM_001145769.1)**			
XPC_exon_1	TGACTAGGCCTCCAACGAAG	TACGCAGGAGCTTGGATCG	453
XPC_exon_2	ATAAGCTGCACTGCCTCCAC	GATCCAATCTTCCATGGACC	371
XPC_exon_3	GCTTGAATGGAACACTAGG	TAGTGATCTGACTCCAAACAG	257
XPC_exon_4	TGATTCTGTTCAGTACAGTAGC	CAAAGTCCTCCTAAGCAGC	308
XPC_exon_5	GAGGAGAAGGAATTGCCTG	AGCACAAGCTCTTTGCACC	224
XPC_exon_6	CATGTCTTGACTTTGGCAGC	CTGTGGAAGTGACCTGAACC	325
XPC_exon_7	CTTGGCTGGAAATGAAAATTCC	GCACATGGCTGCCATTATC	257
XPC_exon_8	TTCTTAGGATAACTATGTTCTTCC	ACTCCGTGAATACCAGCTC	232
XPC_exon_9.1	CTCTAGCTGGTGACTTAACC	CTTGAAGAGCTTGAGGATGC	629
XPC_exon_9.2	GCTCTGATTTTGAGCTCTCC	CCTGACTGTGTCTTGGAGC	694
XPC_exon_10	GTCTAAGGATCATCTCCCTC	TGCTGTCCAGTCAGATGAGC	345

XPC_exon_11	ACGTTCAAGGCTGTTTGCC	GCTCATCATCACTTCTCTGC	344
XPC_exon_12&13	TGAGGAACTGGATGCCTTTG	TGAAAATTGGAGCCACCAGG	558
XPC_exon_14	CACTGTCTTCCACAAACTGG	TGTATTCAGTGCTCGCTCC	333
XPC_exon_15	ACTTGGTGTGAAGGAGAGGC	CCTTTCTGAGCTGCATCTCC	291
XPC_exon_16	GAACTTGCTGCCTCTTCATGG	TGCCTTCTCAGCAGAGAAGC	416

RPA1
(NM_002945.3)

RPA1_exon_1	AACTTCTCGGGCCAATAACTG	CGATCGAAAGAATGCTTTGAGG	319
RPA1_exon_2	ACTAGATGCTTCAGCTGACAG	AGGGTTACAATGTAGACTGGC	339
RPA1_exon_3	GCCACTTCGGTTTACGTATCC	CGTAGCCAAGCAAGTACGG	266
RPA1_exon_4	TAGCACAGGTATGAGTAGCTC	GCCATGTGTTAGCTACCTGA	320
RPA1_exon_5	GTTCATCACAATGATCATCGGG	GAATCCAAAAGGTAGAACTCCG	262
RPA1_exon_6	CGAATTCTAATCCATGGGAGTC	CATGATTCCGAAGTCTGACAC	280
RPA1_exon_7	ATTTCACTAGTGGCACCTCC	TGGTGACCCTGAAGTTCAC	336
RPA1_exon_8	TATGCGTAAGACGAGAAAGGC	AGCTGAGCCTTCTGTTTCC	305
RPA1_exon_9	TATCCCAGTGTCACTTGGG	CCTGAGACTACCATCAAATCG	273
RPA1_exon_10	TGGCAGACTAGGGGTTTCTG	ACGCCAGCAAATGATGAACGG	364
RPA1_exon_11	CCGTTCATCATTTGCTGGCGT	CCAACTCCAGGAGCTCCT	322
RPA1_exon_12	GTATGGATTCCATGTACGCTG	TTCAAGAAACACGGAAGCTGC	406
RPA1_exon_13	TAGCAGCAAGTTGCATGTGG	AATTACGGCTCTGACAGCTG	310
RPA1_exon_14	TAAGGGCAGGCTTTGAGCTG	TTGTGACCACAGTGCCTGG	337
RPA1_exon_15	TCTCCCCATCTTCTCAGTG	GTGGGGTCAGTGTTATAGAATG	245
RPA1_exon_16	TGAAACTACCCAGGAGATGC	TGCCGTAGGTGACAAACAG	256
RPA1_exon_17	TCACTGGAATGACTGAACTCTG	GTAGCTAACATGGGATCGTC	332
RPA2_exon_1&2	GTGCCAGAGAAAAGTAGCC	GTCATAGATGACTCAGGGAC	556
RPA2_exon_3	GGTGAATTCTCAGAGCAACATG	CCCTGAAACCAAGGCTACAT	451
RPA2_exon_4	AATGTAGCCTTGGTTTCAGGG	GTTCAACATCTCCACAGTTCCA	236
RPA2_exon_5	AAGAGGCTTTTACCAGCATC	TCCAAAAGCCATGACATGAATC	207
RPA2_exon_6	CCATATTTGGAATACCTCTGAGAG	GAGCCAGAGAAAGACATCACTAA	382
RPA2_exon_7	TGTCCACTCCAATGTAAGCAG	CCAGGTACTTAAACAGTCGTAG	347
RPA2_exon_8	GAGCTACGACTGTTTAAGTACC	GTACCTATTTCACCTCACAAGG	269
RPA2_exon_9	CTGTGAGAGCTTTCTATGTGC	GAAACCTACTTCCTAGAAGCC	350

GTF2H1
(NM_001142307.1)

GTF2H1_exon_2	TTGTAGAGGAATGATGCCGTG	CAACTATGACATACAAATACATCAG	379
GTF2H1_exon_3	AGGAATCCTGAATCATCTTGGG	ACTAGGAGACTATAGAGAACCC	346
GTF2H1_exon_4	GAACTGAAGGACCTCTGCTGC	AATGGTCTTAAGACACATATTGGG	379
GTF2H1_exon_5	CTTTGTCCAGTGTCCACTG	TTGTGGGTTAGGAAGGTCAC	444
GTF2H1_exon_6&7	GTAGAGTTGAGAGCTTTATGGTC	CAAGTGGAATAGGTCTACTACGC	549
GTF2H1_exon_8&9	GTTCGTGAGAAGTAATAAAAGTCCC	GAACTGAGTCACAGGAAATAAGCG	546
GTF2H1_exon_10	GAAGCTCATGGTGGGAAGAC	TGTACTTTGATGGTGACGAACTG	316
GTF2H1_exon_11	CTTATCGTTTCTTGCCTTGAGG	ATGGTTCAATCTGCTTGTGCTC	312

GTF2H1_exon_12	AATGTTAAGAAAGAGAGGTGGC	AACCAATAAAAGTTCAGCACTCC	250
GTF2H1_exon_13	ATAGCACATCCTGGTTTGGC	AGTCACTACCACAATTATCTGGC	288
GTF2H1_exon_14	TCTGTTGGTCTCTACAGCTTG	ACTACAAAGAGAAAGAAGATCAGGC	199
GTF2H1_exon_15	AGCTCCGAAACTACACTCTG	CAGGTCATAGTTCTCTCAATCTC	238

GTF2H2
(NM_001515.3)

GTF2H2_exon_2	GAATAAGATGACCTGGCATTCC	AGAAAAGACAGGGCAATAAAGCTC	242
GTF2H2_exon_3	GGGCTAGTAATCATTTTAGCTTGTC	GTCAGTTACCCTGAAACACAGC	242
GTF2H2_exon_4	TGCACGTTATGACATTCTTACTC	AGCAAGACAGAATGTTTCTCTAACC	236
GTF2H2_exon_5	AGTACTTGAAATTGGCCCTTTC	ATCCAGTGTTCTTCCCCAC	208
GTF2H2_exon_6	GAGAAGTGATGCTTGATTAACAGG	TGGGTAATTCACAAAATACATGTCC	252
GTF2H2_exon_7	TCTGCTCCAATATCATCTACAGG	TAAACAGGGCTACGAGGTTGG	220
GTF2H2_exon_8	CAGTATGAAGTGAGAGTGTTTAC	CCATATTAACACATCACTTCAGC	346
GTF2H2_exon_9	CCTATGTTTACTGTTCTTTCTGGG	ACCACAAAGATGAATCACACATAC	281
GTF2H2_exon_10	CAGCCTTACAACTTGCGATCC	GTAGCTCAGAACAGATTAAAACATCC	446
GTF2H2_exon_11	CAGTTCGTTATGTCTATGAAGGTG	TTGTGCAAATGAAAATGGAGAGC	219
GTF2H2_exon_12	AGCTTGGAATTAGTCAGTTTCATC	GATAGTCTTCATTGCCTATCTTTTG	331
GTF2H2_exon_13	GTCAAGATAGCATGCTTTCCC	TACGTGGAAGTGGAGTACTGG	341
GTF2H2_exon_14	CAATCTTCTGTTATAGCCACAGG	ATGGCATTCTTTATCTCATCTTCC	391
GTF2H2_exon_15	TTCCTGAAAGTCATGTGTAAAGAC	CCTGTTGAAAACTTATAGTACCATTG	240
GTF2H2_exon_16	TTGAATGTCATCTAGCTGTGCTG	AGAAGTCCTTTTATTCACAGTTGC	258

GTF2H3
(NM_001516.3)

GTF2H3_exon_1	ACTCCACTTCCGCAGATTAGG	TTCCATCTGGCCAGTCGTAG	228
GTF2H3_exon_2	ACTGGGTCTTCTGTTAATCTC	CTCAGTTAAACCAGAATTCTCC	279
GTF2H3_exon_3	CTTAATTATGTCAAGACCTGACATGG	CTTTTGCAACATGGAGTAGGAAC	279
GTF2H3_exon_4	CATTCTGATCATGAGTGAGCC	GTGAGAACGTGCAGGCATAC	345
GTF2H3_exon_5	CATACTTAGTTCAATAAAGCACCC	CTTCAGAGTACTCACATGTCC	263
GTF2H3_exon_6	TTAGCTAGGTAGATGCTGG	GATTTCATTTCCTGATTGTCTAC	307
GTF2H3_exon_7	GAGCCTACGTTTTCCTATGAG	CTCCCTGATATTCTGACTGAG	766
GTF2H3_exon_8	TGTGACATTGGTGTGAGAGG	CAGAATACCTCAGAGTAGAGC	227
GTF2H3_exon_9	TCTCTGTTGTGGAGTGGC	TCACTAATTGTCATAACCACCG	379
GTF2H3_exon_10	TGAGAACCTCATTCCTCTGG	CTTCCATTCCTTTCCCAGGAC	212
GTF2H3_exon_11	AAATGGAGGCCTTGGAGTTCC	GAAAAACAACACAGCTGAGAAACTG	314
GTF2H3_exon_12	GTTTTTCAGCCACCCTATTGTTTC	GTCTTCCCAACAAAGAATCTGC	305

GTF2H4
(NM_001517.4)

GTF2H4_exon_2	GAATCAGTTAGAAAGGTCAGGG	TCCAGGCTGCTACACTTTTACC	299
GTF2H4_exon_3	GATGTTTGAGAGGTAATTGAGGG	AGTGCTCCAAGGAACAGCAG	254
GTF2H4_exon_4	GTTCAGAACAGGCAGAGATGG	TGGAGACCACAAGGTTCTGG	334
GTF2H4_exon_5&6	CAGGGTTCCTTACTCTTGGC	CCCAGATACCATTTCTTAGGG	383

GTF2H4_exon_7&8	AAACGTGAGTGGACAAGTGGG	TTCTCAGTCTACTCTACTGCC	489
GTF2H4_exon_9	CAGAACGAACAGAGATGGAG	AAAGGCAACACACCACCCTG	238
GTF2H4_exon_10	CAGGCAGGAAGATGTAAGGC	CATGAGAAATGTCAGAGAGCTC	303
GTF2H4_exon_11	CTCATGACACTTGAAAGAAGGG	AGCCTTCCATAATGTGACCC	292
GTF2H4_exon_12	ACAGCTCAGATGGCTTTCCTG	ATCATGTCATCACCAGCTGCC	164
GTF2H4_exon_13	AGCTGGTGATGACATGATGG	ATGCAGACTGAACTGTCTGCC	264
GTF2H4_exon_14	CTTCACTTTCTCGTCTTCTCC	AACACCTGAGTTCTGATGCC	293
GTF2H5 **(NM_207118.2)**			
GTF2H5_exon_1	CTCGTTTCAGAGGCAGATCC	CTGTGCCACTTGTTAAAAGCG	281
GTF2H5_exon_2	GCTCAAGTCTCTGTGATGTG	CCTCTATGTCTAATCAGTCACC	336

Appendix F: Fetch2.py code.

This python code takes an input file of GeneIDs and returns information on nonsynonymous SNPs including detailed population frequency data that may be included in the annotation.

```
Fetch2.py
C:\Python25

import sys
import EntrezServer

from xml.etree import ElementTree

fxnmap = {
   'lr': 'locus-region',
   'cn': 'coding-nonsynonymous',
   'cs': 'coding-synonymous',
   'ex': 'exception',
   'in': 'intron',
   'mu': 'mrna-utr',
   're': 'reference',
   'ss': 'splice-site'
   }

usage = """Usage: python %s <locus-id> [<fxnlist.>]
where
locus-id is a LocusID from NCBI database
fxnlist is any number of the following 2-letter filters
""" % sys.argv[0]

if len(sys.argv) < 2:
   sys.stderr.write(usage)
   for key, val in fxnmap.items():
      sys.stderr.write("\t%s\t%s\n" % (key, val))
   sys.exit(0)
locus = sys.argv[1]
fc = sys.argv[2:]
if len(fc) == 0:
   fxnterm = ''
   fxnlist = []
else:
   fxnterm = ' AND (%s)' % ' OR '.join([fxnmap[x] + '[FXN_CLASS]' for x in fc])
   fxnlist = [fxnmap[x] for x in fc]

snp = EntrezServer.ncbi.Database('snp')

ctx = snp.Context()
term = '%s[LOCUS_ID]' % locus
```

```
print 'Search term:', term
sys.stdout.flush()
ctx.Search(term = term)

print 'Found', ctx.Count(), 'results'
print 'Locus;RS;Tag(s);Study;Population;Class;A;C;G;T'

ns_docsum = '{http://www.ncbi.nlm.nih.gov/SNP/docsum}(Morak et al., 2011)'
ns_geno = '{http://www.ncbi.nlm.nih.gov/SNP/geno}'

for i in range(ctx.Count()):
   ExchangeSet = ctx.Fetch(retstart=i, retmax=1, mode='text', report='XML')
   Rs = ExchangeSet.find(ns_docsum + 'Rs')
   Rs_assembly = Rs.find(ns_docsum + 'Rs_assembly')
   labels = {}
   match = False
   for Assembly in Rs_assembly.findall(ns_docsum + 'Assembly'):
      Assembly_component =  Assembly.find(ns_docsum + 'Assembly_component')
      Assembly_groupLabel \
         = Assembly.find(ns_docsum + 'Assembly_groupLabel').text
      Component = Assembly_component.find(ns_docsum + 'Component')
      Component_mapLoc = Component.find(ns_docsum + 'Component_mapLoc')
      MapLoc = Component_mapLoc.find(ns_docsum + 'MapLoc')
      MapLoc_fxnSet = MapLoc.find(ns_docsum + 'MapLoc_fxnSet')
      if MapLoc_fxnSet is None:
         continue
      label = [None, None, None]
      reflist = []
      aalist = []
      varlist = []
      vallist = []
      for FxnSet in MapLoc_fxnSet.findall(ns_docsum + 'FxnSet'):
         geneId = FxnSet.find(ns_docsum + 'FxnSet_geneId').text
         if geneId == locus:
            fxnClass = FxnSet.find(ns_docsum + 'FxnSet_fxnClass')
            value = fxnClass.attrib['value']
            if value == 'reference':
               ref = FxnSet.find(ns_docsum + 'FxnSet_residue').text
               if ref not in reflist:
                  reflist.append(ref)
               aa = int(FxnSet.find(ns_docsum + 'FxnSet_aaPosition').text) + 1
               if aa not in aalist:
                  aalist.append(aa)
            else:
               residue = FxnSet.find(ns_docsum + 'FxnSet_residue')
               if residue is None:
                  continue
               else:
                  var = residue.text
```

```
            if var not in varlist:
               varlist.append(var)
            aaPos = FxnSet.find(ns_docsum + 'FxnSet_aaPosition')
            if aaPos is None:
               continue
            else:
               aa = int(aaPos.text) + 1
            if aa not in aalist:
               aalist.append(aa)
            if value not in vallist:
               vallist.append(value)
               if value in fxnlist:
                  match = True
      if not reflist and not varlist:
         # due to it being a different gene
         continue
      reflist.sort()
      aalist.sort()
      varlist.sort()
      for val in vallist:
         if val not in fxnlist:
            break
      else:
         vallist = []
      vallist.sort()
      label = ('%s%s%s %s'
            % ('/'.join(reflist), '/'.join([str(aa) for aa in aalist]),
               '/'.join(varlist), '/'.join(vallist))).strip()
      if labels.has_key(label):
         labels[label].append(Assembly_groupLabel)
      else:
         labels[label] = [Assembly_groupLabel]
   if not match:
      continue
   tags = []
   if len(labels) == 1:
      tag = labels.keys()[0]
   else:
      for label, groups in labels.items():
         group = ','.join(groups)
         tags.append('%s (%s)' % (label, group))
      tag = ' / '.join(tags)
   if not tag:
      # This is an overlapping snip that belongs to another locus
      continue

   GenoExchange = ctx.Fetch(retstart=i, retmax=1, mode='text', report='GENXML')
   # Create the population index
   populations = {}
```

```
  for Population in GenoExchange.findall(ns_geno + 'Population'):
    popId = Population.attrib['popId']
    pop = {}
    pop['study'] = Population.attrib['handle']
    pop['label'] = Population.attrib['locPopId']
    popClass = Population.find(ns_geno + 'popClass')
    pop['class'] = popClass.attrib['self']
    pop['group'] = []
    pop['subgroup'] = []
    populations[popId] = pop
  # Retrieve the frequencies
  for SnpInfo in GenoExchange.findall(ns_geno + 'SnpInfo'):
    sys.stdout.flush()
    rsId = SnpInfo.attrib['rsId']
    SsInfoList = SnpInfo.findall(ns_geno + 'SsInfo')
    if SsInfoList:
      for SsInfo in SsInfoList:
        for ByPop in SsInfo.findall(ns_geno + 'ByPop'):
          popId = ByPop.attrib['popId']
          pop = populations[popId]
          study = pop['study']
          label = pop['label']
          Class = pop['class']
          distribution = {'A': 0, 'C': 0, 'G': 0, 'T': 0}
          for AlleleFreq in ByPop.findall(ns_geno + 'AlleleFreq'):
            allele = AlleleFreq.attrib['allele']
            freq = AlleleFreq.attrib['freq']
            distribution[allele] = freq
          sys.stdout.write('%s;%s;%s;%s;%s;%s;%s;%s;%s;%s\n'
                  % (locus, rsId, tag, study, label, Class,
                    distribution['A'], distribution['C'],
                    distribution['G'], distribution['T']))
    else:
      sys.stdout.write('%s;%s;%s;No studies\n' % (locus, rsId, tag))

print 'Done'
```

Appendix G: Assayed SNPs.

All 221 nonsynonymous SNPs assayed in the 480 EBV-transformed lymphoblastoid cell lines from unrelated healthy individuals.

Gene	rs #
ALKBH3	rs2434470
APEX1	rs1048945
APEX1	rs1130409
ATM	rs17174393
ATM	rs1800056
ATM	rs1800057
ATM	rs1800058
ATM	rs1801516
ATM	rs2234997
ATM	rs3218673[a]
ATM	rs35813135
ATR	rs2227928
ATR	rs2229032
ATR	rs2229033
ATR	rs34124242
ATRX	rs3088074
ATRX	rs45439799
BLM	rs11852361
BRCA1	rs16941
BRCA1	rs16942
BRCA1	rs1799950
BRCA1	rs1799966
BRCA1	rs28897674
BRCA1	rs28897687
BRCA1	rs4986850
BRCA1	rs4986852
BRCA1	rs799917
BRCA2	rs1046984
BRCA2	rs11571640
BRCA2	rs11571660
BRCA2	rs11571833
BRCA2	rs144848
BRCA2	rs1799944
BRCA2	rs1799954
BRCA2	rs28897708
BRCA2	rs28897727
BRCA2	rs28897729
BRCA2	rs28897731
EXO1	rs4149965
EXO1	rs735943
EXO1	rs9350
FANCA	rs11646374
FANCA	rs17233497
FANCA	rs1800282
FANCA	rs2239359
FANCA	rs7190823
FANCA	rs7195066
FANCA	rs9282681
FANCD2	rs3864017
FANCE	rs7761870
FANCE	rs9462088
FANCM	rs10138997
FANCM	rs1367580
FANCM	rs3736772
FLJ35220	rs34933300
FLJ35220	rs35549084
HEL308	rs1494961
LIG1	rs3730947
LIG1	rs3730980
LIG3	rs3744356
LIG4	rs1805388
LIG4	rs1805389
MBD4	rs10342
MBD4	rs140693
MGMT	rs12917
MGMT	rs2020893
MGMT	rs2308321
MGMT	rs2308327
MLH1	rs1799977
MLH3	rs175080
MLH3	rs175081
MLH3	rs17782839
MLH3	rs28756982
MMS19L	rs12360068
MMS19L	rs29001285
MMS19L	rs3740526
POLD1	rs1726801
POLE	rs5744751
POLE	rs5744934
POLE	rs5745066
POLG	rs2307441
POLG	rs3087374
POLI	rs3218786
POLI	rs8305
POLL	rs3730463
POLL	rs3730477
POLM	rs28382644
POLN	rs10011549
POLN	rs10018786
POLN	rs11725880
POLN	rs2353552
POLN	rs9328764
POLQ	rs1381057
POLQ	rs3218634
POLQ	rs3218649
POLQ	rs3218651
POLQ	rs487848
POLQ	rs532411
PRKDC	rs4278157
PRKDC	rs7830743
PRKDC	rs8178017
RAD1	rs1805327
RAD17	rs1045051
RAD18	rs373572
RAD23B	rs1805329
RAD51L1	rs34594234
RAD51L3	rs4796033
RAD52	rs11571463
RAD52	rs4987206
RAD52	rs4987207
RAD52	rs4987208
RAD9A	rs2066492
RDM1	rs2251660
RECQL4	rs4244612

BRCA2	rs28897743
BRCA2	rs28897758
BRCA2	rs4987047
BRCA2	rs4987117
BRIP1	rs4986764
BRIP1	rs4988350
C19orf40	rs2304103
C19orf40	rs3816032
CCNH	rs2266690
CHAF1A	rs8100525
CHAF1A	rs9352
CHEK1	rs506504
DCLRE1A	rs3750898
DCLRE1B	rs12022378
DCLRE1C	rs12768894
EME1	rs12450550
EME1	rs17714854
EME1	rs3760413
ERCC2	rs13181
ERCC2	rs1799792
ERCC2	rs1799793
ERCC4	rs1800067
ERCC5	rs1047769
ERCC5	rs17655
ERCC5	rs2227869
ERCC6	rs2228526
ERCC6	rs2228527
ERCC6	rs2228528
ERCC6	rs2228529
ERCC6	rs4253211
EXO1	rs1047840
EXO1	rs12122770
EXO1	rs1776148
EXO1	rs4149909
EXO1	rs4149910
EXO1	rs4149963

MPG	rs2308313
MSH3	rs1650697
MSH3	rs184967
MSH3	rs26279
MSH4	rs5745325
MSH4	rs5745459
MSH4	rs5745549
MSH5	rs1802127
MSH5	rs2075789
MSH5	rs28381349
MSH6	rs1042821
MSH6	rs2020912
MUS81	rs13817
MUS81	rs545500
MUTYH	rs3219484
MUTYH	rs3219489
NBN	rs1805794
NEIL1	rs5745926
NEIL3	rs13112358
NEIL3	rs13112390
NEIL3	rs1876268
NEIL3	rs2048074
NEIL3	rs34193982
NEIL3	rs7689099
OGG1	rs1052133
OGG1	rs17050550
PARP1	rs1136410
PARP1	rs3219062
PARP2	rs3093921
PARP2	rs3093926
PMS1	rs2066459
PMS2	rs17420802
PMS2	rs2228006
PMS2L3	rs17147225
PMS2L3	rs17425318
PMS2L3	rs17435215

RECQL4	rs4251691
RECQL5	rs35566780
RECQL5	rs820196
REV1	rs3087386
REV1	rs3087399
REV1	rs3087403
REV3L	rs3204953
REV3L	rs3218599
REV3L	rs458017
REV3L	rs462779
RPA1	rs5030755
RPA4	rs2642219
TDG	rs2888805
TDG	rs3953597
TDP1	rs28365054
TELO2	rs2235624
TELO2	rs2248128
TELO2	rs2667661
TP53	rs1042522
WRN	rs1346044
WRN	rs1800391
WRN	rs2230009
WRN	rs2725362
XPC	rs2227999
XPC	rs2228000
XPC	rs2228001
XRCC1	rs1799782
XRCC1	rs25487
XRCC1	rs25489
XRCC1	rs25490
XRCC1	rs25495
XRCC1	rs25496
XRCC2	rs3218536
XRCC3	rs861539
XRCC4	rs28360135

Appendix H: Common Nonsynonymous Variants

135 SNPs from Fetch.py with empirically determined MAF of 4% or greater.

Gene	***rs ID***	***AA name***	***Determined MAF***
ALKBH3	rs2434470	D228E	0.261
APEX1	rs1048945	Q51H	0.041
APEX1	rs1130409	D148E	0.475
ATM	rs1801516	D1853N	0.138
ATR	rs2227928	T211M	0.417
ATR	rs2229032	R2425Q	0.160
ATRX	rs3088074	Q929E	0.310
BLM	rs11852361	P868L	0.072
BRCA1	rs16941	E1038G	0.350
BRCA1	rs16942	K1183R	0.323
BRCA1	rs1799950	Q356R	0.063
BRCA1	rs1799966	S1613G	0.324
BRCA1	rs4986850	D693N	0.073
BRCA1	rs799917	L871P	0.336
BRCA2	rs144848	N372H	0.290
BRIP1	rs4986764	S919P	0.399
C19orf40	rs2304103	S158L	0.042
C19orf40	rs3816032	I192T	0.115
CCNH	rs2266690	V270A	0.201
CHAF1A	rs9352	V923A	0.467
DCLRE1A	rs3750898	H317D	0.237
DCLRE1B	rs12022378	H61Y	0.170
DCLRE1C	rs12768894	H243R	0.182
EME1	rs12450550	I350T	0.290
ERCC2	rs13181	K751Q	0.364
ERCC2	rs1799793	D312N	0.340
ERCC4	rs1800067	R415Q	0.077
ERCC5	rs1047769	M254V	0.040
ERCC5	rs17655	H1104D	0.213
ERCC6	rs2228526	M1097V	0.200
ERCC6	rs2228527	R1213G	0.190
ERCC6	rs2228528	G399D	0.164
ERCC6	rs2228529	Q1413R	0.190
ERCC6	rs4253211	R1230P	0.110
EXO1	rs1047840	K589E	0.384
EXO1	rs1776148	E670G	0.373
EXO1	rs4149963	T439M	0.074
EXO1	rs4149965	V458M	0.270
EXO1	rs735943	H354R	0.454
EXO1	rs9350	P757L	0.157
FANCA	rs11646374	A412V	0.110
FANCA	rs17233497	S1088F	0.110

FANCA	rs1800282	V6D	0.093
FANCA	rs2239359	G501S	0.385
FANCA	rs7190823	A266T	0.431
FANCA	rs7195066	D809G	0.340
FANCA	rs9282681	T1328A	0.110
FANCE	rs9462088	A502T	0.057
FANCM	rs10138997	S175F	0.052
FANCM	rs1367580	V878L	0.118
FANCM	rs3736772	P1812A	0.101
FLJ35220	rs34933300	R112Q	0.258
FLJ35220	rs35549084	V29I	0.050
HEL308	rs1494961	V306I	0.496
LIG4	rs1805388	T9I	0.165
LIG4	rs1805389	A3V	0.050
MBD4	rs10342	A273S/T	0.090
MGMT	rs12917	L84F	0.128
MGMT	rs2308321	I143V	0.138
MGMT	rs2308327	K178R	0.137
MLH1	rs1799977	I219V	0.325
MLH3	rs175080	P844L	0.458
MMS19L	rs12360068	A579V	0.043
MMS19L	rs3740526	G811D	0.450
MSH3	rs1650697	I79V	0.220
MSH3	rs184967	Q949R	0.156
MSH3	rs26279	A1045T	0.284
MSH4	rs5745325	A97T	0.270
MSH5	rs2075789	P29S	0.100
MSH6	rs1042821	G39E	0.170
MUTYH	rs3219484	V22M	0.079
MUTYH	rs3219489	Q324H	0.252
NBN	rs1805794	E185Q	0.324
NEIL3	rs13112358	P443L	0.210
NEIL3	rs13112390	Q471H	0.205
NEIL3	rs2048074	R381-/R	0.210
NEIL3	rs34193982	H286R	0.166
NEIL3	rs7689099	P117R	0.107
OGG1	rs1052133	P332A	0.229
PARP1	rs1136410	V762A	0.146
PARP2	rs3093926	R283Q	0.064
PMS2	rs2228006	K541E	0.155
POLD1	rs1726801	R119H	0.050
POLE	rs5744751	A252V	0.120
POLE	rs5744934	N1396S	0.137
POLG	rs2307441	E1143G	0.043
POLG	rs3087374	Q1236H	0.080
POLI	rs8305	A731T	0.306
POLL	rs3730477	R438W	0.221

POLN	rs10011549	G336S	0.101
POLN	rs10018786	M310L	0.100
POLN	rs2353552	Q121H	0.128
POLN	rs9328764	R425C	0.101
POLQ	rs1381057	R2513Q	0.306
POLQ	rs3218634	L2538V	0.068
POLQ	rs3218649	T982R	0.365
POLQ	rs3218651	H1201R	0.163
POLQ	rs487848	A581V	0.068
POLQ	rs532411	A2304V	0.068
PRKDC	rs4278157	R2899C	0.060
PRKDC	rs7830743	I3434T	0.053
PRKDC	rs8178017	M333I	0.043
RAD1	rs1805327	E281G	0.073
RAD17	rs1045051	L546R	0.316
RAD18	rs373572	Q302R	0.277
RAD23B	rs1805329	A249V	0.196
RAD51L3	rs4796033	R185Q	0.130
RDM1	rs2251660	C127W	0.148
RECQL4	rs4244612	D267E	0.390
RECQL4	rs4251691	Q1005R	0.460
RECQL5	rs820196	D480G	0.232
REV1	rs3087386	F257S	0.438
REV1	rs3087399	N373S	0.127
REV1	rs3087403	V138M	0.280
REV3L	rs3204953	V3064I	0.146
REV3L	rs458017	Y1156C	0.056
REV3L	rs462779	I1224T	0.184
RPA1	rs5030755	T351A	0.112
RPA4	rs2642219	T33A	0.306
TDG	rs2888805	V367M	0.108
TELO2	rs2235624	R146Q	0.393
TELO2	rs2248128	Q674R	0.220
TELO2	rs2667661	E7G	0.230
TP53	rs1042522	P72R	0.257
WRN	rs1346044	C1367R	0.257
WRN	rs1800391	M387I	0.075
WRN	rs2230009	V114I	0.065
WRN	rs2725362	L1074F	0.447
XPC	rs2227999	R492H	0.060
XPC	rs2228000	A499V	0.247
XPC	rs2228001	K939Q	0.398
XRCC1	rs1799782	R194W	0.062
XRCC1	rs25487	R399Q	0.365
XRCC2	rs3218536	R188H	0.087
XRCC3	rs861539	T241M	0.371

Appendix I: In silico analyses predicating functional effect of SNPs.

Functionality of 134 common (MAFs>4%) nonsynonymous SNPs in DNA repair genes.

Gene	SNPrs.no.	p.Name	Align GVGD	Polyphen	SIFT
ALKBH3	rs2434470	D228E	C35	Possibly Damaging	Damaging
APEX1	rs1130409	D148E	C0	Benign	Tolerated
APEX1	rs1048945	Q51H	C15	Possibly Damaging	Damaging
ATM	rs1801516	D1853N	C15	Possibly Damaging	Tolerated
ATR	rs2227928	M211T	C0	Benign	Tolerated
ATR	rs2229032	R2425Q	C0	Benign	Tolerated
ATRX	rs3088074	Q929E	C0	Benign	Tolerated
BLM	rs11852361	P868L	C65	Probably Damaging	Damaging
BRCA1	rs16941	E1038G	C65	Benign	Damaging
BRCA1	rs16942	K1183R	C0	Benign	Tolerated
BRCA1	rs1799950	Q356R	C0	Probably Damaging	Damaging
BRCA1	rs1799966	S1613G	C0	Benign	Damaging
BRCA1	rs4986850	D693N	C0	Benign	Damaging
BRCA1	rs799917	P871L	C0	Benign	Damaging
BRCA2	rs144848	N372H	C0	Benign	Tolerated
BRIP1	rs4986764	S919P	C0	Benign	Tolerated
C19orf40	rs2304103	S158L	C15	Benign	Tolerated
C19orf40	rs3816032	I192T	C0	Benign	Tolerated
CCNH	rs2266690	V270A	C65	Possibly Damaging	Damaging
CHAF1A	rs9352	A923V	C15	Benign	Tolerated
DCLRE1A	rs3750898	D317H	C0	Possibly Damaging	Damaging
DCLRE1B	rs12022378	H61Y	C15	Possibly Damaging	Damaging
DCLRE1C	rs12768894	H243R	C25	Probably Damaging	Damaging
EME1	rs12450550	I350T	C0	Benign	Tolerated
ERCC2	rs13181	K751Q	C0	Benign	Tolerated
ERCC2	rs1799793	D312N	C15	Benign	Tolerated
ERCC4	rs1800067	R415Q	C35	Benign	Damaging
ERCC5	rs1047769	M254V	C15	Probably Damaging	Damaging
ERCC5	rs17655	D1104H	C0	Possibly Damaging	Damaging
ERCC6	rs2228526	M1097V	C0	Benign	Tolerated
ERCC6	rs2228527	R1213G	C65	Benign	Tolerated
ERCC6	rs2228528	G399D	C0	Benign	Tolerated
ERCC6	rs2228529	Q1413R	C0	Benign	Tolerated
ERCC6	rs4253211	R1230P	C15	Probably Damaging	Tolerated
EXO1	rs1047840	E589K	C0	Benign	Tolerated
EXO1	rs1776148	E670G	C0	Benign	Tolerated
EXO1	rs4149963	T439M	C0	Possibly Damaging	Tolerated
EXO1	rs4149965	V458M	C0	Benign	Tolerated
EXO1	rs735943	H354R	C0	Benign	Tolerated

EXO1	rs9350	P757L	C65	Probably Damaging	Damaging
FANCA	rs11646374	A412V	C0	Benign	Tolerated
FANCA	rs17233497	S1088F	C65	Possibly Damaging	Damaging
FANCA	rs1800282	V6D	C15	Possibly Damaging	Tolerated
FANCA	rs2239359	G501S	C0	Possibly Damaging	Tolerated
FANCA	rs7190823	T266A	C0	Benign	Tolerated
FANCA	rs7195066	G809D	C0	Benign	Tolerated
FANCA	rs9282681	T1328A	C0	Benign	Tolerated
FANCE	rs9462088	A502T	C0	Benign	Tolerated
FANCM	rs10138997	S175F	C0	Possibly Damaging	Tolerated
FANCM	rs1367580	V878L	C0	Benign	Tolerated
FANCM	rs3736772	P1812A	C0	Possibly Damaging	Tolerated
FLJ35220	rs34933300	R112Q	C0	Benign	Tolerated
FLJ35220	rs35549084	V29I	C0	Benign	Tolerated
HEL308	rs1494961	V306I	C0	Benign	Tolerated
LIG4	rs1805388	T9I	C65	Possibly Damaging	Damaging
LIG4	rs1805389	A3V	C0	Benign	Damaging
MBD4	rs10342	A273S	C0	Benign	Tolerated
MGMT	rs12917	L84F	C15	Benign	Tolerated
MGMT	rs2308321	I143V	C0	Benign	Tolerated
MGMT	rs2308327	K209R	C0	Benign	Damaging
MLH1	rs1799977	I219V	C25	Benign	Tolerated
MLH3	rs175080	P844L	C65	Benign	Damaging
MMS19	Rs12360068	A558V	C65	Benign	Tolerated
MMS19	rs3740526	G790D	C0	Benign	Tolerated
MSH3	rs1650697	I79V	C0	Benign	Tolerated
MSH3	rs184967	Q949R	C0	Benign	Tolerated
MSH3	rs26279	A1045T	C0	Benign	Tolerated
MSH4	rs5745325	A97T	C55	Benign	Tolerated
MSH5	rs2075789	P29S	C0	Possibly Damaging	Damaging
MSH6	rs1042821	G39E	C0	Possibly Damaging	Tolerated
MUTYH	rs3219484	V22M	C15	Benign	Damaging
MUTYH	rs3219489	Q324H	C0	Possibly Damaging	Tolerated
NBN	rs1805794	E185Q	C0	Benign	Tolerated
NEIL3	rs13112358	P443L	C0	Benign	Tolerated
NEIL3	rs13112390	Q471H	C0	Probably Damaging	Tolerated
NEIL3	rs34193982	H286R	C0	Probably Damaging	Tolerated
NEIL3	rs7689099	P117R	C65	Probably Damaging	Damaging
OGG1	rs1052133	P332A	C25	Benign	Damaging
PARP1	rs1136410	V762A	C65	Benign	Tolerated
PARP2	rs3093926	R283Q	C35	Benign	Tolerated
PMS2	rs2228006	E541K	C15	Benign	Tolerated
POLD1	rs1726801	R119H	C0	Benign	Tolerated
POLE	rs5744751	A252V	C65	Benign	Tolerated

POLE	rs5744934	N1396S	C0	Benign	Damaging
POLG	rs2307441	E1143G	C65	Possibly Damaging	Damaging
POLG	rs3087374	Q1236H	C0	Benign	Damaging
POLI	rs8305	A731T	C0	Benign	Tolerated
POLL	rs3730477	R438W	C0	Benign	Damaging
POLN	rs10011549	G336S	C55	Benign	Tolerated
POLN	rs10018786	M310L	C0	Benign	Tolerated
POLN	rs2353552	Q121H	C15	Possibly Damaging	Damaging
POLN	rs9328764	R425C	C0	Probably Damaging	Tolerated
POLQ	rs1381057	Q2513R	C0	Benign	Tolerated
POLQ	rs3218634	L2538V	C25	Benign	Damaging
POLQ	rs3218649	T982R	C0	Benign	Tolerated
POLQ	rs3218651	H1201R	C0	Possibly Damaging	Damaging
POLQ	rs487848	A581V	C65	Benign	Tolerated
POLQ	rs532411	A2304V	C0	Benign	Damaging
PRKDC	rs8178017	M333I	C0	Possibly Damaging	Tolerated
PRKDC	rs4278157	R2899C	C45	Possibly Damaging	Tolerated
PRKDC	rs7830743	I3434T	C0	Benign	Tolerated
RAD1	rs1805327	E281G	C65	Possibly Damaging	Damaging
RAD17	rs1045051	L546R	C65	Probably Damaging	Damaging
RAD18	rs373572	R302Q	C0	Benign	Tolerated
RAD23B	rs1805329	A249V	C0	Benign	Tolerated
RAD51L3	rs4796033	R185Q	C0	Benign	Tolerated
RDM1	rs2251660	C127W	C0	Probably Damaging	Tolerated
RECQL4	rs4244612	E267D	C0	Benign	Tolerated
RECQL4	rs4251691	R1005Q	C0	Benign	Tolerated
RECQL5	rs820196	D480G	C0	Benign	Tolerated
REV1	rs3087386	F257S	C0	Benign	Tolerated
REV1	rs3087399	N373S	C0	Benign	Tolerated
REV1	rs3087403	V138M	C0	Benign	Tolerated
REV3L	rs3204953	V3064I	C25	Benign	Tolerated
REV3L	rs458017	Y1156C	C0	Benign	Tolerated
REV3L	rs462779	T1224I	C0	Benign	Tolerated
RPA1	rs5030755	T351A	C0	Benign	Tolerated
RPA4	rs2642219	A33T	C55	Benign	Damaging
TDG	rs2888805	V367M	C0	Benign	Tolerated
TELO2	rs2235624	Q146R	C0	Benign	Tolerated
TELO2	rs2248128	Q674R	C0	Possibly Damaging	Damaging
TELO2	rs2667661	E7G	C0	Benign	Tolerated
TP53	rs1042522	P72R	C0	Benign	Tolerated
WRN	rs1346044	C1367R	C65	Probably Damaging	Tolerated
WRN	rs1800391	M387I	C0	Benign	Tolerated
WRN	rs2230009	V114I	C0	Benign	Tolerated
WRN	rs2725362	L1074F	C0	Benign	Tolerated

XPC	rs2227999	R492H	C0	Benign	Tolerated
XPC	rs2228000	A499V	C0	Benign	Tolerated
XPC	rs2228001	Q939K	C0	Benign	Tolerated
XRCC1	rs1799782	R194W	C65	Probably Damaging	Damaging
XRCC1	rs25487	Q399R	C0	Benign	Damaging
XRCC2	rs3218536	R188H	C25	Benign	Tolerated
XRCC3	rs861539	T241M	C15	Benign	Tolerated

Appendix J: Analyses of variants (MAF>1%) and 12-week response, any toxicity and peripheral neuropathy.

rs no.	Response		Any toxicity (except peripheral neuropathy)		peripheral neuropathy	
	X^2 (df)	**p-value**	**χ^2 (df)**	**p-value**	**χ^2 (df)**	**p-value**
rs13181	0.36 (2)	0.84	0.35 (2)	0.84	6.66 (2)	0.036
rs17655	3.30 (2)	0.19	0.07 (2)	0.97	1.62 (2)	0.44
rs1799782	3.02 (2)	0.22	3.21 (2)	0.2	1.22 (2)	0.54
rs1799977	1.62 (2)	0.45	1.11 (2)	0.57	n/a	n/a
rs1800067	0.25 (2)	0.88	1.68 (2)	0.43	0.49 (2)	0.78
rs2227869	0.97 (1)	0.33	0.00 (1)	0.95	1.22 (1)	0.27
rs2228527	1.71 (2)	0.43	5.34 (2)	0.069	0.98 (2)	0.61
rs2228528	3.85 (2)	0.15	3.49 (2)	0.17	0.35 (2)	0.84
rs2228529	1.50 (2)	0.47	5.51 (2)	0.064	0.97 (2)	0.61
rs25487	9.57 (2)	0.0083	0.50 (2)	0.78	0.06 (2)	0.97
rs25489	0.04 (1)	0.85	0.05 (1)	0.81	0.35 (1)	0.56
rs861539	3.44 (2)	0.18	0.82 (2)	0.66	1.05 (2)	0.59
rs1799977	1.62 (2)	0.45	1.11 (2)	0.57	0.16 (2)	0.92
rs10011549	2.11 (2)	0.35	0.04 (2)	0.98	0.73 (2)	0.69
rs1042522	1.05 (2)	0.59	1.16 (2)	0.56	2.34 (2)	0.31
rs1045051	0.12 (2)	0.94	2.77 (2)	0.25	1.36 (2)	0.51
rs1047840	4.90 (2)	0.086	1.11 (2)	0.57	0.30 (2)	0.86
rs1048945	0.00 (1)	0.99	1.43 (2)	0.49	0.52 (1)	0.47
rs1052133	3.69 (2)	0.16	0.53 (2)	0.77	3.42 (2)	0.18
rs1130409	3.55 (2)	0.17	0.82 (2)	0.66	0.08 (2)	0.96
rs1136410	4.23 (2)	0.12	3.23 (2)	0.2	4.22 (2)	0.12
rs11571833	0.74 (1)	0.39	0.05 (1)	0.82	0.00 (1)	0.96
rs11725880	1.31 (1)	0.25	2.37 (1)	0.12	0.01 (1)	0.93
rs12022378	2.69 (2)	0.26	6.18 (2)	0.046	2.42 (2)	0.3
rs12450550	1.50 (2)	0.47	2.68 (2)	0.26	4.34 (2)	0.11
rs12768894	1.36 (2)	0.51	0.67 (2)	0.71	0.83 (2)	0.66
rs12917	4.03 (2)	0.13	3.62 (2)	0.16	2.04 (2)	0.36
rs13112390	1.44 (2)	0.49	0.07 (2)	0.97	0.20 (2)	0.91
rs1346044	0.38 (2)	0.83	0.55 (2)	0.76	3.41 (2)	0.18
rs1367580	0.61 (2)	0.74	4.53 (2)	0.1	2.39 (2)	0.3
rs1381057	4.01 (2)	0.13	0.78 (2)	0.68	0.19 (2)	0.91
rs144848	7.14 (2)	0.028	0.69 (2)	0.71	3.98 (2)	0.14
rs1494961	4.49 (2)	0.11	0.62 (2)	0.73	0.09 (2)	0.96
rs16942	0.19 (2)	0.91	5.12 (2)	0.077	1.14 (2)	0.56
rs175080	1.21 (2)	0.55	0.25 (2)	0.88	2.91 (2)	0.23
rs1776148	0.02 (2)	0.99	0.97 (2)	0.61	1.33 (2)	0.52
rs1799950	1.65 (2)	0.44	1.39 (2)	0.5	0.11 (1)	0.74
rs1799966	0.15 (2)	0.93	6.13 (2)	0.047	0.94 (2)	0.63
rs1800058	1.26 (1)	0.26	3.04 (1)	0.081	5.50 (1)	0.019

rs1800282	0.66 (2)	0.72	1.57 (2)	0.46	0.60 (1)	0.44
rs1800391	3.36 (2)	0.19	0.85 (2)	0.65	0.37 (2)	0.83
rs1801516	0.44 (2)	0.8	0.47 (2)	0.79	0.89 (2)	0.64
rs1802127	0.71 (1)	0.4	0.02 (1)	0.88	0.60 (1)	0.44
rs1805327	4.00 (2)	0.14	0.68 (2)	0.71	0.20 (2)	0.91
rs1805329	0.79 (2)	0.67	1.01 (2)	0.6	2.33 (2)	0.31
rs1805388	6.29 (2)	0.043	4.26 (2)	0.12	1.16 (2)	0.56
rs1805794	1.90 (2)	0.39	0.36 (2)	0.84	1.96 (2)	0.38
rs184967	2.14 (2)	0.34	3.31 (2)	0.19	0.25 (2)	0.88
rs1876268	0.21 (2)	0.9	0.04 (2)	0.98	0.36 (1)	0.55
rs2066459	0.24 (1)	0.62	1.31 (1)	0.25	0.56 (1)	0.45
rs2227928	0.80 (2)	0.67	0.21 (2)	0.9	0.27 (2)	0.87
rs2228006	0.11 (2)	0.94	0.89 (2)	0.64	1.76 (2)	0.41
rs2229032	0.82 (2)	0.66	1.26 (2)	0.53	0.83 (2)	0.66
rs2229033	0.42 (1)	0.52	0.03 (1)	0.85	1.08 (1)	0.3
rs2230009	4.55 (2)	0.1	0.56 (2)	0.76	0.73 (1)	0.39
rs2235624	4.51 (2)	0.1	1.65 (2)	0.44	2.03 (2)	0.36
rs2239359	1.39 (2)	0.5	1.32 (2)	0.52	0.83 (2)	0.66
rs2248128	3.04 (2)	0.22	3.02 (2)	0.22	0.80 (2)	0.67
rs2251660	1.37 (2)	0.5	0.85 (2)	0.65	1.28 (2)	0.53
rs2266690	1.18 (2)	0.55	2.37 (2)	0.31	1.45 (2)	0.48
rs2304103	0.41 (1)	0.52	0.45 (2)	0.8	0.99 (1)	0.32
rs2307441	0.13 (1)	0.71	0.63 (1)	0.43	1.84 (1)	0.17
rs2308321	0.75 (2)	0.69	3.47 (2)	0.18	1.24 (2)	0.54
rs2308327	0.37 (2)	0.83	4.61 (2)	0.1	1.08 (2)	0.58
rs2353552	2.23 (2)	0.33	0.05 (2)	0.98	2.87 (2)	0.24
rs2434470	0.56 (2)	0.76	1.55 (2)	0.46	4.49 (2)	0.11
rs26279	1.58 (2)	0.45	2.76 (2)	0.25	1.26 (2)	0.53
rs2725362	5.10 (2)	0.078	5.90 (2)	0.053	0.84 (2)	0.66
rs28360135	0.03 (1)	0.85	0.62 (1)	0.43	1.25 (1)	0.26
rs28365054	0.10 (1)	0.75	0.22 (1)	0.64	1.02 (1)	0.31
rs28756982	0.10 (1)	0.75	0.42 (1)	0.52	0.01 (1)	0.92
rs2888805	4.92 (2)	0.086	2.04 (2)	0.36	2.29 (2)	0.32
rs3087374	4.15 (2)	0.13	1.91 (2)	0.39	2.55 (2)	0.28
rs3087386	0.70 (2)	0.71	1.65 (2)	0.44	0.13 (2)	0.93
rs3087399	0.85 (2)	0.65	4.85 (2)	0.088	0.04 (2)	0.98
rs3087403	4.04 (2)	0.13	2.89 (2)	0.24	2.97 (2)	0.23
rs3093921	1.51 (1)	0.22	0.35(1)	0.55	4.86 (1)	0.027
rs3093926	3.03 (2)	0.22	0.58 (2)	0.75	2.63 (2)	0.27
rs3204953	0.54 (2)	0.77	1.69 (2)	0.43	0.11 (2)	0.95
rs3218536	0.29 (2)	0.86	4.29 (2)	0.12	0.57 (2)	0.75
rs3218599	0.02 (1)	0.89	0.12 (1)	0.73	2.07 (2)	0.36
rs3218634	1.20 (2)	0.55	2.34 (2)	0.31	0.03 (1)	0.87
rs3218649	3.02 (2)	0.22	0.03 (2)	0.98	1.18 (2)	0.55
rs3218651	1.38 (2)	0.5	0.50 (2)	0.78	2.41 (2)	0.3
rs3219484	5.31 (2)	0.07	2.57 (2)	0.28	2.19 (2)	0.34

rs3219489	1.20 (2)	0.55	2.61 (2)	0.27	0.44 (2)	0.8
rs34193982	0.38 (2)	0.83	1.67 (2)	0.43	0.23 (2)	0.89
rs34933300	3.42 (2)	0.18	2.57 (2)	0.28	0.14 (2)	0.93
rs3730477	2.83 (2)	0.24	3.98 (2)	0.14	3.93 (2)	0.14
rs373572	0.45 (2)	0.8	0.94 (2)	0.62	1.78 (2)	0.41
rs3736772	1.55 (2)	0.46	3.19 (2)	0.2	4.00 (2)	0.14
rs3750898	0.59 (2)	0.75	0.49 (2)	0.78	8.55 (2)	0.014
rs3816032	1.00 (2)	0.61	0.31 (2)	0.86	0.68 (1)	0.41
rs4149909	8.67 (1)	0.0032	3.11 (1)	0.078	0.58 (1)	0.45
rs4149963	0.04 (2)	0.98	1.61 (2)	0.45	0.38 (2)	0.83
rs4251691	0.75 (2)	0.69	0.41 (2)	0.81	1.79 (2)	0.41
rs458017	1.83 (2)	0.4	4.36 (2)	0.11	1.60 (1)	0.21
rs462779	0.45 (2)	0.8	2.35 (2)	0.31	0.97 (2)	0.62
rs487848	1.22 (2)	0.54	2.39 (2)	0.3	0.07 (1)	0.79
rs4986764	0.47 (2)	0.79	1.50 (2)	0.47	0.69 (2)	0.71
rs4986850	0.23 (2)	0.89	7.30 (2)	0.026	4.92 (2)	0.086
rs4987117	0.99 (2)	0.61	0.25 (2)	0.88	2.05 (1)	0.15
rs5030755	1.54 (2)	0.46	1.87 (2)	0.39	2.91 (2)	0.23
rs506504	2.97 (2)	0.23	0.57 (2)	0.75	2.55 (2)	0.28
rs532411	0.46 (2)	0.79	2.87 (2)	0.24	0.07 (1)	0.79
rs5744934	1.17 (2)	0.56	1.19 (2)	0.55	0.17 (2)	0.92
rs5745066	0.72 (1)	0.4	0.25 (2)	0.88	0.10 (1)	0.75
rs5745459	0.03 (1)	0.87	4.06(1)	0.044	0.80 (1)	0.37
rs5745549	0.15 (2)	0.93	1.10 (2)	0.58	0.57 (1)	0.45
rs7190823	0.32 (2)	0.85	0.84 (2)	0.66	0.20 (2)	0.91
rs735943	0.91 (2)	0.63	2.58 (2)	0.28	0.93 (2)	0.63
rs7689099	1.47 (2)	0.48	0.29 (2)	0.86	0.29 (2)	0.86
rs7830743	4.16 (2)	0.12	1.46 (2)	0.48	1.58 (2)	0.45
rs799917	0.20 (2)	0.9	5.45 (2)	0.066	0.57 (2)	0.75
rs8100525	0.04 (1)	0.85	0.35 (1)	0.56	0.00 (1)	0.95
rs8178017	1.36 (1)	0.24	0.02 (2)	0.99	0.39 (2)	0.82
rs820196	0.67 (2)	0.72	0.66 (2)	0.72	0.56 (2)	0.76
rs8305	1.02 (2)	0.6	3.27 (2)	0.2	0.81 (2)	0.67
rs9328764	2.00 (2)	0.37	0.05 (2)	0.97	0.88 (2)	0.64
rs9350	0.05 (2)	0.97	1.27 (2)	0.53	0.73 (2)	0.69
rs9352	0.58 (2)	0.75	2.83 (2)	0.24	6.06 (2)	0.048
rs2228000	0.12 (2)	0.94	1.41 (2)	0.5	2.93 (2)	0.23
rs2228001	3.27 (2)	0.19	4.37 (2)	0.11	0.16 (2)	0.92
rs34594234	1.71 (1)	0.19	0.08 (1)	0.78	2.76 (1)	0.097
rs17714854	4.74 (1)	0.029	2.06 (1)	0.15	0.65 (1)	0.42
rs17782839	3.29 (1)	0.07	0.66 (1)	0.42	0.31 (1)	0.58
rs1800056	0.27 (1)	0.6	2.37 (1)	0.12	0.00 (1)	0.96
rs7761870	1.69 (1)	0.19	5.18 (2)	0.075	0.95 (1)	0.33
rs1800057	0.03 (1)	0.87	1.78 (1)	0.18	0.54 (1)	0.46
rs3218786	2.07 (1)	0.15	3.33 (2)	0.19	0.42 (1)	0.52
rs1799944	0.50 (1)	0.48	0.07 (1)	0.79	1.95 (1)	0.16

rs12360068	0.93 (1)	0.33	0.22 (1)	0.64	0.61 (1)	0.43
rs10138997	0.29 (2)	0.86	0.42 (2)	0.81	0.71 (1)	0.4
rs11852361	3.17 (2)	0.2	0.72 (2)	0.7	0.24 (2)	0.89

n/a – not assessed.