Trends in Immunology

CellPress OPEN ACCESS

Forum

Potential impact of long-read sequencing on complement-mediated diseases

Sarah M. Carpanini [1,2,*] and Rebecca Sims[3,*]

The complement genes harbour genetic variants that affect numerous diseases; however, these genes are notoriously repeat-heavy, and these repeat regions are largely unexplored for disease-relevant genetic variation. Elucidating these 'dark' regions is now possible using long-read sequencing (LRS), enabling identification of novel disease-relevant genetic variants.

The complement system and LRS

The complement system is a critical branch of the innate immune system and the first line of defence against invading pathogens. Genetic variants in the complement system genes influence a wide array of human diseases, including inflammatory disorders, eye disorders, renal disorders, susceptibility to infection, and neurodegeneration [1]. However, many complement-relevant gene regions – notably those residing within the regulators of complement activation (RCA) gene cluster on human chromosome 1q32 – are highly repetitive in nature and difficult to align to the reference genome, creating gaps not sequenced by short-read sequencing (SRS); these are referred to as 'dark' [2].

More than 50% of the human genome is made up of repeat elements where specific gene regions are repeated tens or hundreds of times [3]. Indeed, over 6000 genes contain repeat elements and are classified as 'dark' [4]. Dark regions are understudied areas of the genome that have historically been inaccessible to microarray and SRS technologies. SRS requires DNA to be fragmented into 50–300 bp reads and aligned to a reference genome. Some dark regions are 'dark by depth', often because the sequence is chemically difficult to sequence (e.g., GC-rich) [5,6], while others are 'dark by mapping quality' due to the ambiguous nature of aligning short reads to repetitive regions of genomic DNA [4].

LRS is a transformative technology that enables the comprehensive analysis of full-length transcripts, structural variants, and genome architecture [7] (Box 1). LRS provides higher resolution of transcript diversity, complex genetic rearrangements, and regulatory elements compared with previous methodologies. There are currently two widely available LRS platforms – Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) – which generate reads ranging from 1000 to 30 000+ bp. Both offer enhanced resolution in identifying large-scale insertions, deletions, duplications, inversions, and translocations, and, importantly, enable researchers to differentiate between maternal and paternal ancestry.

Complement genes and the dark genome

Given the repeat-heavy nature of the complement system genes, we reasoned that LRS technology would be particularly applicable to these gene loci, and we decided to determine whether complement genes are enriched for dark regions compared with the rest of the genome. A published dark gene list compiled from an analysis of Illumina whole-genome SRS data from ten unrelated males identified 36 794 dark regions in 3804 protein-coding genes that were dark either by depth or mapping quality [4]. We identified complement system genes defined as in our previous publication [7], and compared the proportion of complement system genes within and not within the dark gene list with the whole-genome proportion. We found that 21.4% of complement genes (n = 12/56) were present in the list of genes harbouring 'dark' regions compared with a genome average of 19% (n = 3804/20 000). Although this difference in gene ratios is not statistically significant (X^2 = 0.2158, P = 0.6423), there is a trend towards an enrichment of dark regions within complement system genes compared with the whole genome. Thus, complement system genes may harbour an excess of regions yet unexplored for genetic variation that may be relevant to disease.

Exploration of hypothesis

To investigate this possibility, we utilised a list of medically relevant dark genes compiled by PacBio and Twist BioScience (Twist Alliance Dark Genes Panel BED File | Twist Bioscience) [8] and tested for enrichment of complement system genes. The list of medically relevant genes was previously extracted from OMIM, HGMD and ClinVar and contained 40 of the 56 genes on our complement system gene list [9]. Comparison of the proportion of complement system genes that were dark (8/40) to the rest of the genome (370/4773) identified a highly significant enrichment of medically relevant complement system genes that contain dark regions (X^2 = 8.219, P = 0.0041).

Our findings strengthen the hypothesis that yet unexplored genetic variation within the complement system genes could be relevant to disease, and demonstrate that application of LRS has the potential to identify large structural variations that commonly have a much stronger disease impact than single nucleotide polymorphisms (SNPs). We know that numerous complement genes harbour disease-relevant structural variants which have been identified through techniques such as multiplex ligation-dependent probe amplification (MLPA) [10]. However, these

Box 1. Comparison of SRS and LRS technologies

LRS is a technology that allows the reading of DNA/RNA fragments of 1000–30 000+ bp, compared with the 50–300 bp achievable with SRS. Unlike SRS, LRS allows the sequencing of a single molecule, generating longer read length for better sequence assembly.

SRS technologies fragment and then amplify the DNA/RNA strands before computer programs are used to assemble the derived reads into a continuous sequence, aligned against a reference genome. Sequencing and/or assembling complex and repetitive regions of the genome can be challenging and can result in sequencing gaps (Figure IA). LRS methodologies vary in their chemistry, but all use longer DNA fragments and negate the need to amplify the sequence: thus making read assembly and reference genome alignment less arduous, and reducing gaps in sequencing (Figure IB).

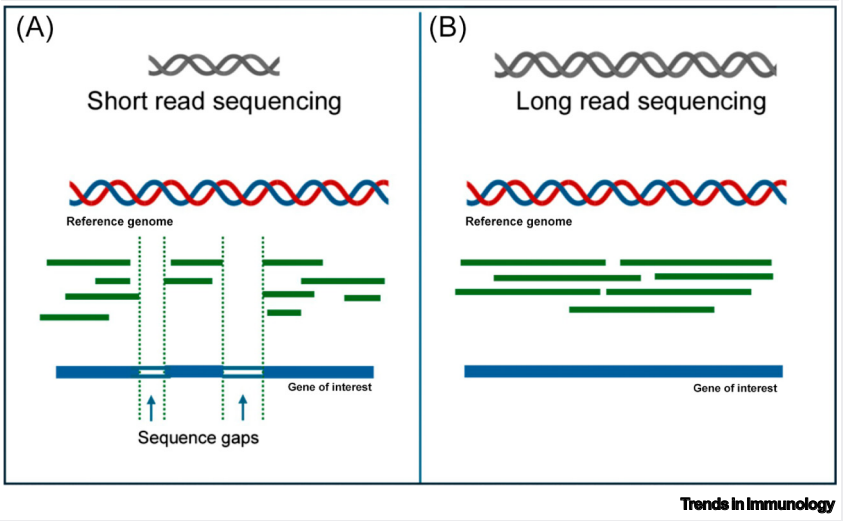

Figure I. Comparison of short-read sequencing (SRS) (A) and long-read sequencing (LRS) (B). Sequencing reads are represented by green lines. The figure shows diagrammatic visualisation of the sequencing gaps remaining in complex and repetitive regions when using SRS. Figure created using BioRender.

methods are time-consuming, limited to known variants, and have a low throughput. LRS allows us for the first time to test for structural variants and elucidate the dark regions of the genome at scale.

Complement receptor 1

Of particular interest to us is the *complement receptor 1* (*CR1*) gene which is known to have a statistically significant impact on risk for Alzheimer's disease (AD) and is present in the medically relevant dark gene panel. CR1 is a key regulator of the complement pathway, it binds complement activation products C3b and C4b, and has cofactor activity for inactivation of C3b by the serine protease factor I. *CR1* is known to harbour large genomic variation, including a copy number variant (CNV) composed of an 18 kb variable tandem repeat or long homologous repeat (LHR). The *CR1* CNV produces four co-dominant alleles encoding proteins with different number of LHRs. Each LHR encodes a C3b/C4b binding site; thus, longer CR1 molecules have an additional LHR which increases the number of C3b/C4b binding sites and thus gain of function [11].

The published dark gene list identified that 26.5% of the *CR1* gene is dark [4]. Visualisation of *CR1* gene coverage from SRS data generated by the Multi-'omics Atlas Project[i] highlights these sequencing gaps within the *CR1* gene body (Figure 1A). Comparison with the sequencing coverage achievable using ONT LRS[ii] shows that these gaps are accessible when using this superior technology (Figure 1B).

Concluding remarks

While the development of LRS technology is not novel, its application in complex disease is in its infancy. Historically, LRS was expensive, and it had reduced per-base accuracy and increased systematic errors. However, both PacBio and ONT platforms have increased their base-calling accuracy in recent years, making LRS comparable with SRS. LRS still has an increased cost per base, a lower throughput, increased input DNA/RNA requirements, and a significantly higher computational burden compared with SRS [12]. However, LRS offers enhanced resolution in identifying large structural variation (including retrotransposons), and it allows long-range haplotyping and precise allelic phasing, enabling researchers to differentiate between maternal and paternal lineage [13]. Key to the clinical utility of LRS is the ability to expand the use of epigenomic and RNA sequencing in genomic screening which is currently being explored in the diagnosis and treatment management of complex rare disease and cancer.

While the complement genome has yet to be explored via LRS in relation to disease, structural variants in other disease-relevant genes have been identified via LRS [14]. Once the current limitations – such as costs, complexity, and amount of data, availability, and lack of clinical validation in complement genes – are overcome, we expect that LRS has the potential to become a transformative technology and standard of care for genomic analysis, allowing a widespread implementation of precision medicine [13]. With falling experimental costs, we expect a surge in the use of LRS, demonstrating its potential to identify new variants in complement system

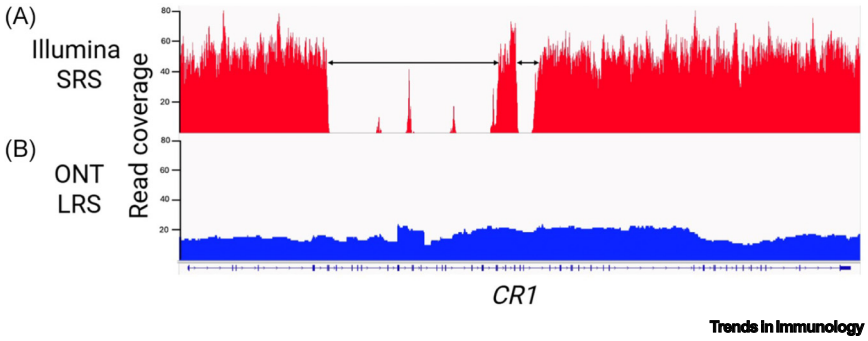

Figure 1. Comparison of short-read sequencing (SRS) and long-read sequencing (LRS) of complement receptor 1 (*CR1*). Arrows represent dark regions that appear as sequencing gaps when using SRS technology. Data from (A) Illumina NovaSeq 6000 SRS (data available via www.synapse.org/Synapse:syn36812517/wiki/619350) and (B) Oxford Nanopore technologies (ONT) LRS (data freely available at https://epi2me.nanoporetech.com/giab-2025.01/). Figure created using BioRender.

genes that are already known to impact the aetiology of a range of diseases [1]. We propose that the use of LRS has the potential to significantly deepen our understanding of complement-mediated diseases and to accelerate the development of better tests and therapeutic interventions.

Acknowledgments

The authors would like to thank Samuel Keat for producing Figure 1. The authors acknowledge support from Alzheimer's Research UK, BRACE, UKDRI, and the Moondance Dementia Laboratory.

Declaration of interests

The authors declare no competing interests.

Resources

[i]https://www.synapse.org/Synapse:syn36812517/wiki/619350

[ii]https://epi2me.nanoporetech.com/giab-2025.01/

[1]UK Dementia Research Institute at Cardiff University, Maindy Road, CF24 4HQ Cardiff, UK
[2]Division of Infection and Immunity, School of Medicine, Cardiff University, Henry Wellcome Building, Heath Park, Cardiff, Wales CF14 4XN, UK
[3]Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neuroscience, School of Medicine, Cardiff University, Cardiff, UK

*Correspondence:
CarpaniniS@Cardiff.ac.uk (S.M. Carpanini) and SimsRC@Cardiff.ac.uk (R. Sims).

https://doi.org/10.1016/j.it.2025.09.006

References

1. de Cordoba, S.R. *et al.* (2012) Complement dysregulation and disease: from genes and proteins to diagnostics and drugs. *Immunobiology* 217, 1034–1046
2. Hourcade, D. *et al.* (1989) The regulators of complement activation (RCA) gene cluster. *Adv. Immunol.* 45, 381–416
3. Hannan, A.J. (2018) Tandem repeats and repeatomes: delving deeper into the 'dark matter' of genomes. *EBioMedicine* 31, 3–4
4. Ebbert, M.T.W. *et al.* (2019) Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol.* 20, 97
5. Ebbert, M.T.W. *et al.* (2018) Long-read sequencing across the C9orf72 'GGGGCC' repeat expansion: implications for clinical use and genetic discovery efforts in human disease. *Mol. Neurodegener.* 13, 46
6. Zheng-Bradley, X. *et al.* (2017) Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *Gigascience* 6, gix038
7. Carpanini, S.M. *et al.* (2021) The impact of complement genes on the risk of late-onset Alzheimer's disease. *Genes (Basel)* 12, 443
8. Wagner, J. *et al.* (2022) Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat. Biotechnol.* 40, 672–680
9. Mandelker, D. *et al.* (2016) Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet. Med.* 18, 1282–1289
10. Garcia-Fernandez, J. *et al.* (2021) Detection of genetic rearrangements in the regulators of complement activation RCA cluster by high-throughput sequencing and MLPA. *Methods Mol. Biol.* 2227, 159–178
11. Brouwers, N. *et al.* (2012) Alzheimer risk associated with a copy number variation in the complement receptor 1 increasing C3b/C4b binding sites. *Mol. Psychiatry* 17, 223–233
12. Moustakli, E. *et al.* (2025) Long-read sequencing and structural variant detection: unlocking the hidden genome in rare genetic disorders. *Diagnostics (Basel)* 15, 1803
13. Oehler, J.B. *et al.* (2023) The application of long-read sequencing in clinical settings. *Hum. Genomics* 17, 73
14. de la Morena-Barrio, B. *et al.* (2022) Long-read sequencing identifies the first retrotransposon insertion and resolves structural variants causing antithrombin deficiency. *Thromb. Haemost.* 122, 1369–1378