Is Oxford Nanopore Technology Ready for Clinical Diagnostics?

GR Taylor, Kesia Brown, Andrew Bond & Michael Yau
Viapath Clinical Genetics Labs, Guy's Hospital, London UK
email graham.r.taylor@kcl.ac.uk

Background

Bridging the "valley of death" between scientific and technological innovation and clinical implementation is a cultural challenge for many organizations, including the NHS. Nanopore sequencing is a good example of a potentially disruptive genomics technology that looks likely to converge with mainstream clinical genomics in the near future. Since the technology is packaged in a range of products from the relatively small scale (Gigabase) O.N. *Minion* to the Terabase-scale *Promethion*, service developers have the opportunity to cross the valley of death using a "rope bridge" prior to investing in major infrastructure.

Our objective is to validate diagnostic services using Oxford Nanopore's Minion in the first instance and to evaluate the cost and performance compared to existing sequencing technology in areas such as tumour DNA sequencing (and circulating tumour DNA), virology, microbiology, genetics and HLA-typing. To facilitate this we are developing R&D collaborations and securing grant funding and commercial backing.

Methods

The scale of the Minion enables direct access to small genomes, but for Gigabase genomes, enrichment in required to target the sequencing capacity to the region of interest. In the studies reported here we used long PCR to generate fragments in the range 3.5 to 16 Kb. Barcode and sequencing adapters were added by ligation. Targets were the HLA-B locus, *BRCA1, BRCA2, SMN1* and *LDLR.* The reads (2D and 1D) were called locally and converted to fastq using Poretools (Loman *et al.*). Barcode sorting used EPI2ME. Fastq read length was measured using *awk*:

```
cat bc12.fastq | awk '{if(NR%4==2) print length($1)}' | sort -n | uniq -c > bc12_length.txt
```

and plotted using R:

```
reads<-read.csv(file="bc12_length.txt", sep="", header=FALSE)
plot (reads$V2,reads$V1,type="l",xlab="bc12 (bases)",ylab="occurences",col="blue", xlim=c(0,18000), ylim=c(0,100))
```

Read mapping used either BWA, BLAT or GraphSeq. We selected a genomic region including *BRAF* and aligned reads of length 200 bases to 200 kilobases with simulated error rates of phred 10 to phred 40 to hg19 using GraphSeq, BLAT and BWA. FASTA files of length 200kbp to 200bp were generated from hg19 using samtools faidx, with starting positions of chr7: 140431813 (*BRAF*). We simulated error rates the range phred 10 to phred 40 and examined the effect of error rate and read length on the ability to identify a point mutation.

Samples

This presentation reports a series of *BRCA1* and *BRCA2* cases:

Barcode	Mutation
BC1	Heterozygous c.4478_4481delAAAG p.(Glu1493fs)
BC2	Heterozygous c.6275_6276delTT p.(Leu2092fs)
BC3	Heterozygous c.5350_5351delAA p.(Asn1784fs)
BC4	Heterozygous c.4576dupA p.(Thr1526fs)
BC5	Heterozygous c.5682C>A p.(Tyr1894Ter)
BC6	reference GiaB
BC7	Heterozygous c.1961delA p.(Lys654fs)
BC8	Heterozygous c.2475delC p.(Asp825fs)
BC9	Heterozygous c.3607C>T p.(Arg1203Ter)
BC10	Heterozygous c.3400G>T p.(Glu1134Ter)
BC11	Heterozygous c.3358_3359delGT p.(Val1120Ter)
BC12	reference GiaB

Other cases included variants in the SMN1 gene and in LDLR and HLA-B

Sequenced products

The product size distribution for the *BRCA1* and *BRCA2* series was as expected, with some small products also sequenced. Most products were unique is length and in content, indicating that sequencing errors are common.

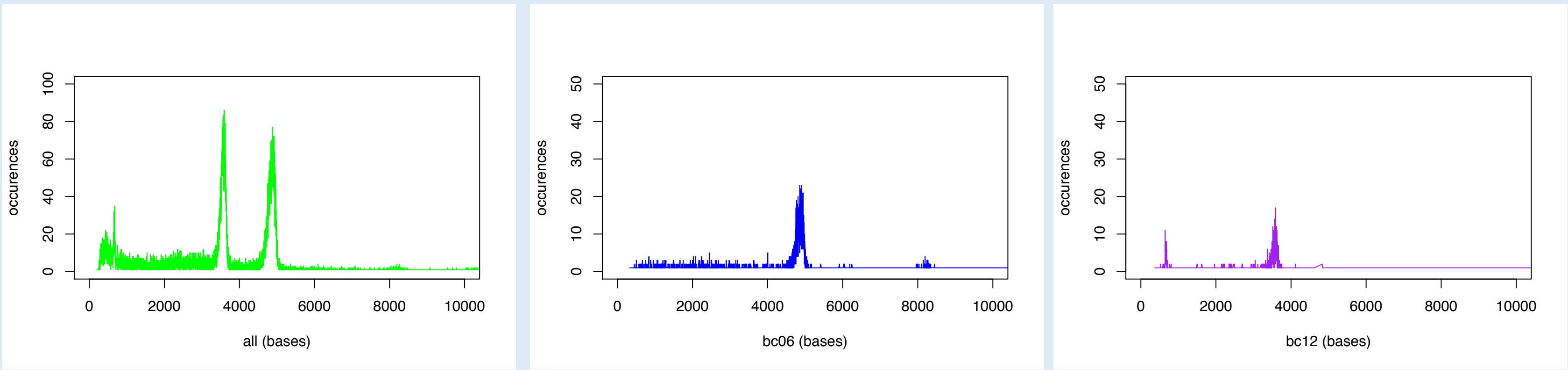


Figure 1
Read length: pooled BRCA1 & BRCA2 (green) BRCA2 (blue) and BRCA1 (purple)

Results: read mapping

Each tool successfully mapped synthetic reads with error rates of up to 10%, but the genomic indexes used by GraphMap were rather large, and so in further studies we used BWA or BLAT. Using the mpileup consensus option it was possible to call the variants in NA12878 correctly with no false positives using read depths of 500 with 2D reads (figure 2)

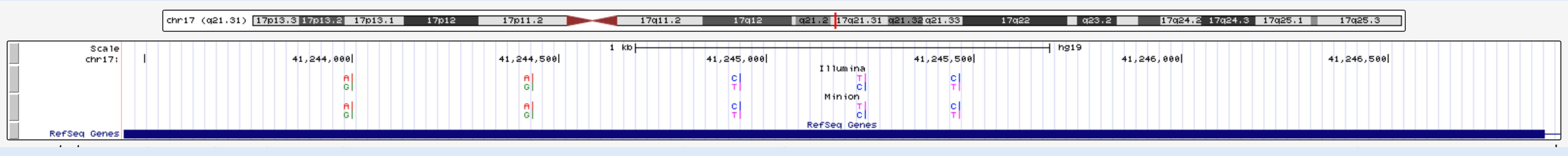


Figure 2
High sensitivity and specificty achieved using mpileup re-alignmant

Results: variant calling

Variant identification protocols are in development: in a small control series using 1D reads there is evidence or recurrent false postives caused by insertions or deletions being reported non-randomly. We are able to identify true positives, but although the accuracy of 1D reads is of the order of 95%, the errors appear to be less random that seen with 2D reads, and so consensus reads are not able to remove all of the errors.

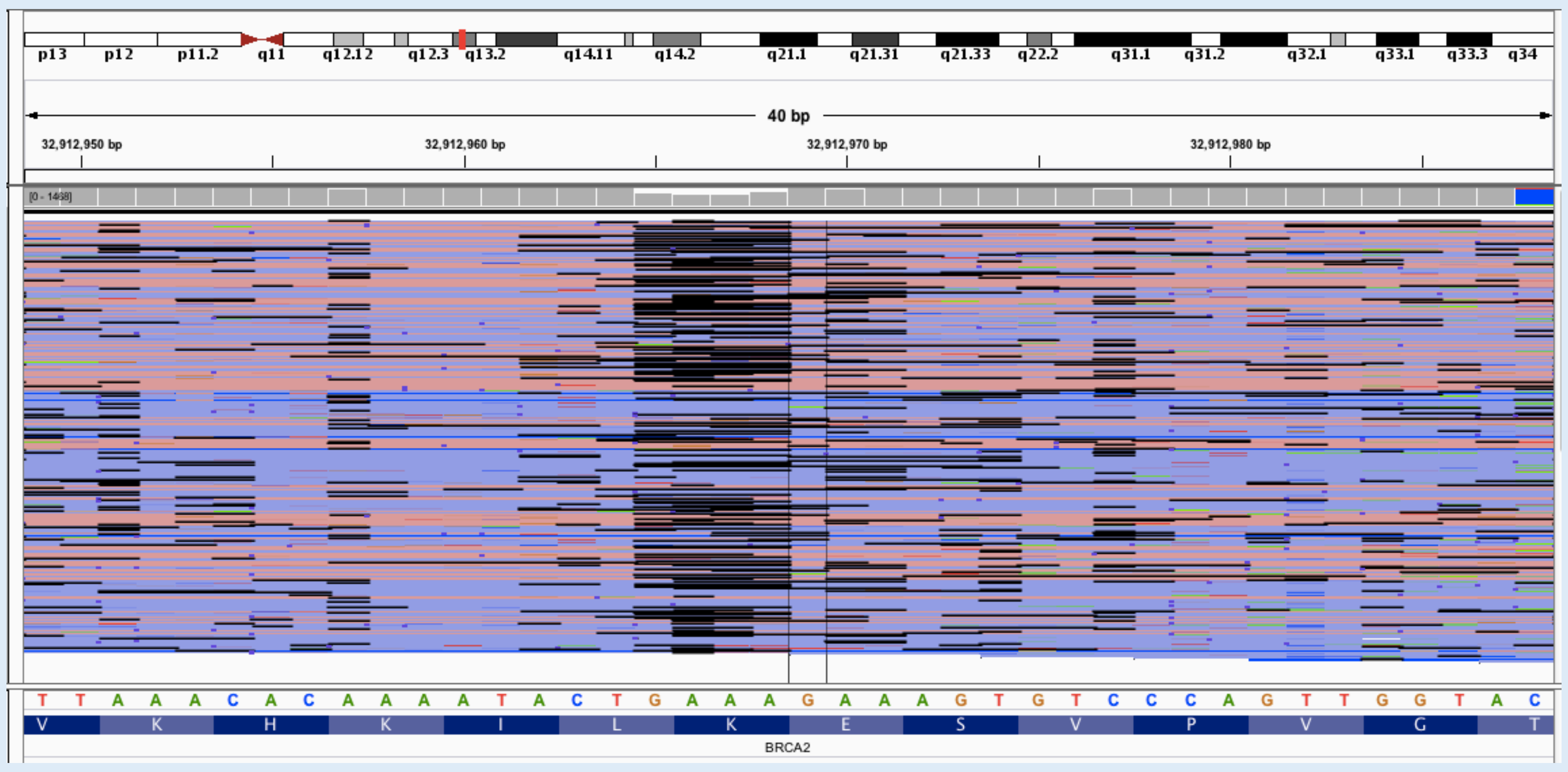


Figure 3
Visualisation of the E1493fs variant in a *BRCA2* case

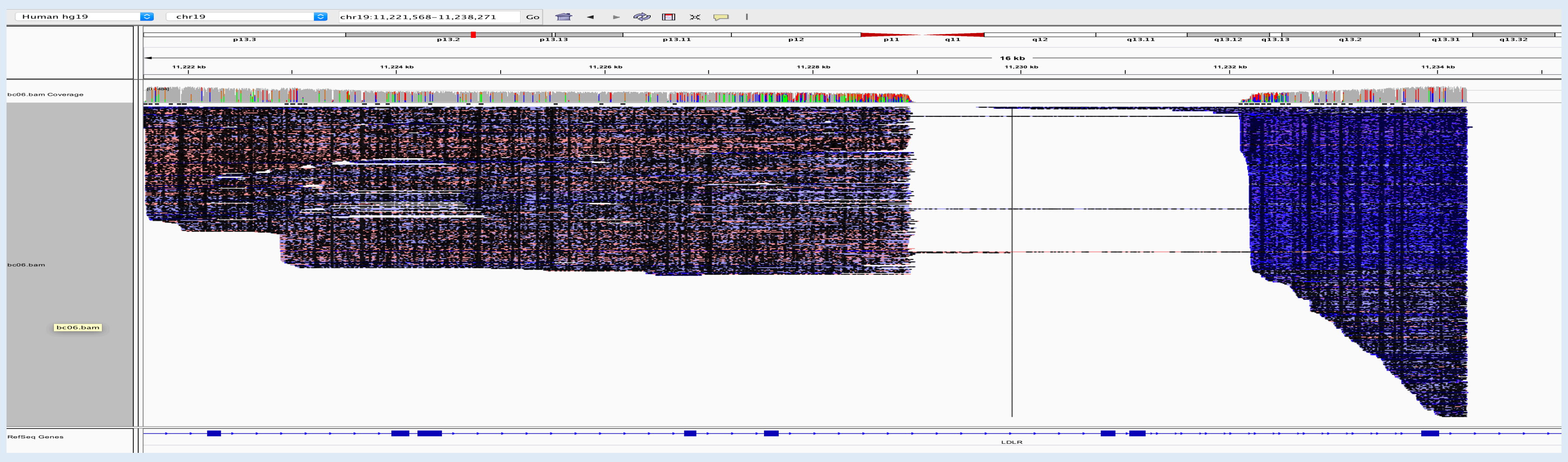


Figure 4
Exon 13-14 deletion in *LDLR*

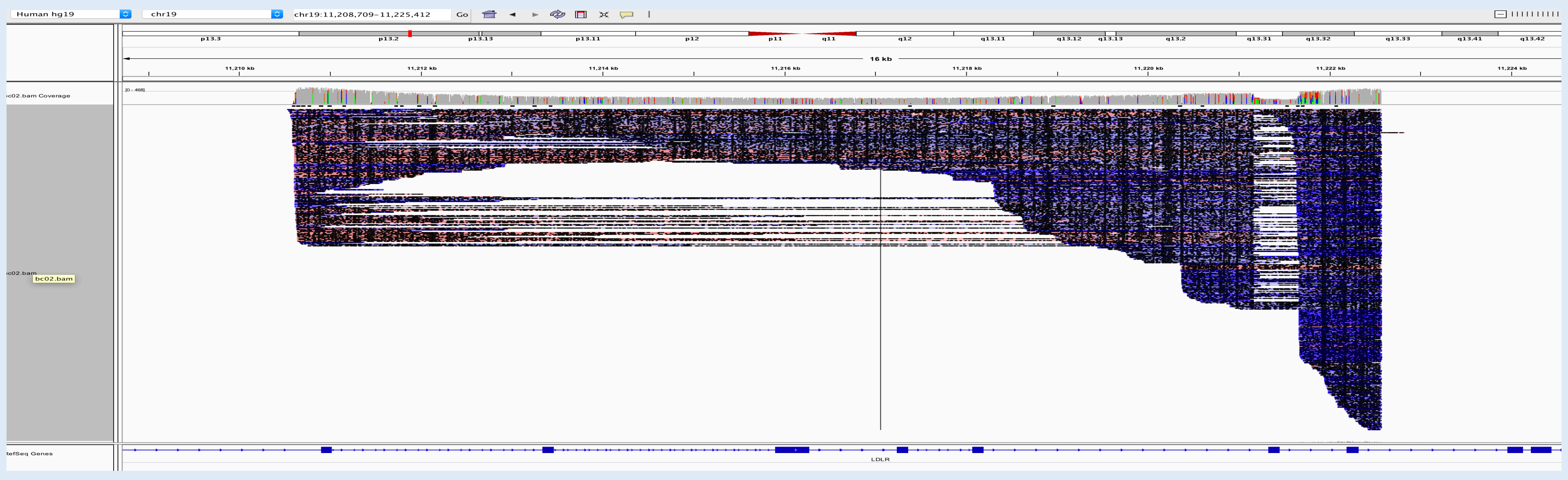


Figure 5
Exon 7 deletion in *LDLR*

Results: improving Q-scores using consensus reads

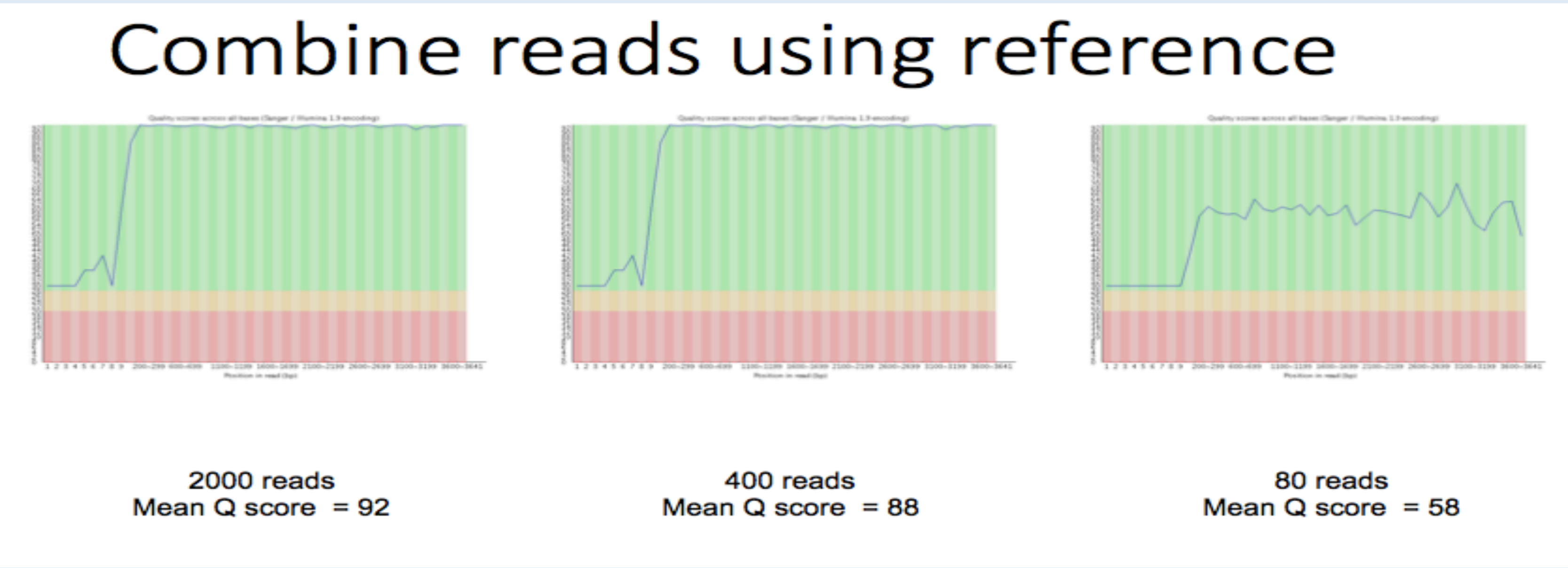


Figure 6
Increasing base calling q-score using mpiplep consenssus

Is Oxford Nanopore Technology Ready for Clinical Diagnostics?

GR Taylor, Kesia Brown, Andrew Bond & Michael Yau
Viapath Clinical Genetics Labs, Guy’s Hospital, London UK
email graham.r.taylor@kcl.ac.uk

Background

Bridging the “valley of death” between scientific and technological innovation and clinical implementation is a cultural challenge for many organizations, including the NHS. Nanopore sequencing is a good example of a potentially disruptive genomics technology that looks likely to converge with mainstream clinical genomics in the near future. Since the technology is packaged in a range of products from the relatively small scale (Gigabase) O.N. *Minion* to the Terabase-scale *Promethion*, service developers have the opportunity to cross the valley of death using a “rope bridge” prior to investing in major infrastructure.

Our objective is to validate diagnostic services using Oxford Nanopore’s Minion in the first instance and to evaluate the cost and performance compared to existing sequencing technology in areas such as tumour DNA sequencing (and circulating tumour DNA), virology, microbiology, genetics and HLA-typing. To facilitate this we are developing R&D collaborations and securing grant funding and commercial backing.

Methods

The scale of the Minion enables direct access to small genomes, but for Gigabase genomes, enrichment in required to target the sequencing capacity to the region of interest. In the studies reported here we used long PCR to generate fragments in the range 3.5 to 16 Kb. Barcode and sequencing adapters were added by ligation. Targets were the HLA-B locus, *BRCA1, BRCA2, SMN1* and *LDLR.* The reads (2D and 1D) were called locally and converted to fastq using Poretools (Loman *et al.*). Barcode sorting used EPI2ME. Fastq read length was measured using *awk*:

```
cat bc12.fastq | awk '{if(NR%4==2) print length($1)}' | sort -n | uniq -c > bc12_length.txt
```

and plotted using R:

```
reads<-read.csv(file="bc12_length.txt", sep="", header=FALSE)
plot (reads$V2,reads$V1,type="l",xlab="bc12 (bases)",ylab="occurences",col="blue", xlim=c(0,18000), ylim=c(0,100))
```

Read mapping used either BWA, BLAT or GraphSeq. We selected a genomic region including *BRAF* and and aligned reads of length 200 bases to 200 kilobases with simulated error rates of phred 10 to phred 40 to hg19 using GraphSeq, BLAT and BWA. FASTA files of length 200kbp to 200bp were generated from hg19 using samtools faidx, with starting positions of chr7: 140431813 (*BRAF*). We simulated error rates the range phred 10 to phred 40 and examined the effect of error rate and read length on the ability to identify a point mutation.

Samples

This presentation reports a series of *BRCA1* and *BRCA2* cases:

Barcode	Mutation
BC1	Heterozygous c.4478_4481delAAAG p.(Glu1493fs)
BC2	Heterozygous c.6275_6276delTT p.(Leu2092fs)
BC3	Heterozygous c.5350_5351delAA p.(Asn1784fs)
BC4	Heterozygous c.4576dupA p.(Thr1526fs)
BC5	Heterozygous c.5682C>A p.(Tyr1894Ter)
BC6	reference GiaB
BC7	Heterozygous c.1961delA p.(Lys654fs)
BC8	Heterozygous c.2475delC p.(Asp825fs)
BC9	Heterozygous c.3607C>T p.(Arg1203Ter)
BC10	Heterozygous c.3400G>T p.(Glu1134Ter)
BC11	Heterozygous c.3358_3359delGT p.(Val1120Ter)
BC12	reference GiaB

Sequenced products

The product size distribution for the *BRCA1* and *BRCA2* series was as expected, with some small products also sequenced. Most products were unique is length and in content, indicating that sequencing errors are common.

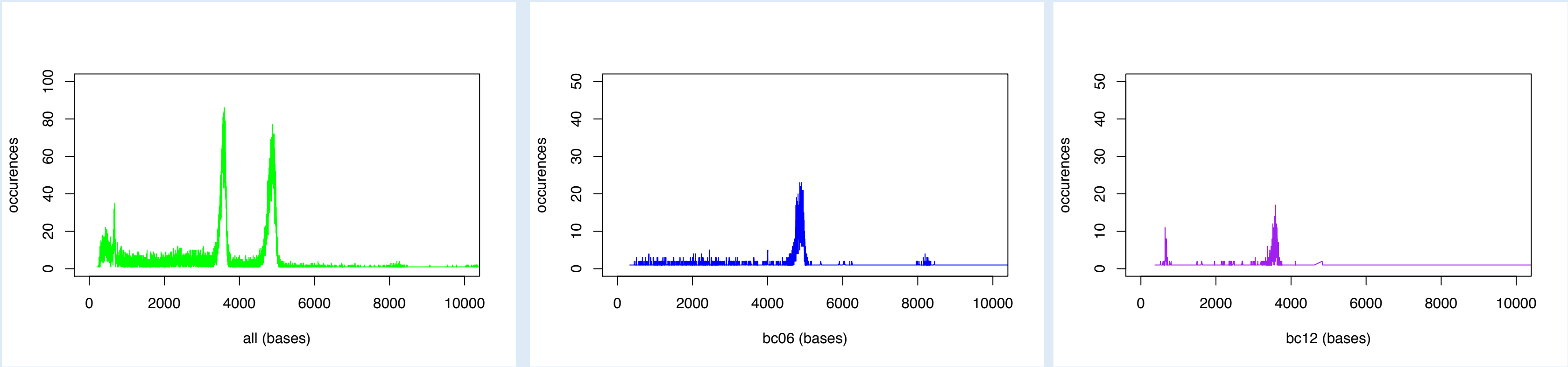


Figure 1
Read length: pooled BRCA1 & BRCA2 (green) BRCA2 (blue) and BRCA1 (purple)

Results: read mapping

Each tool successfully mapped synthetic reads with error rates of up to 10%, but the genomic indexes used by GraphMap were rather large, and so in further studies we used BWA or BLAT. Using the mpileup consensus option it was possible to call the variants in NA12878 correctly with no false positives using read depths of 500 with 2D reads (figure 2)

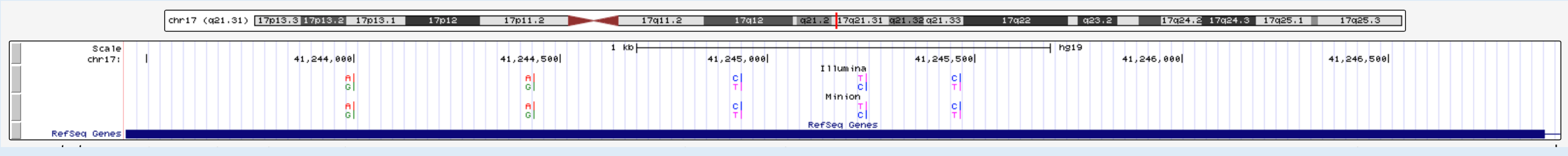


Figure 2
High sensitivity and specificty achieved using mpileup re-alignmant

Results: variant calling

Variant identification protocols are in development: in a small control series using 1D reads there is evidence or recurrent false postives caused by insertions or deletions being reported non-randomly. We are able to identify true positives, but although the accuracy of 1D reads is of the order of 95%, the errors appear to be less random that seen with 2D reads, and so consensus reads are not able to remove all of the errors.

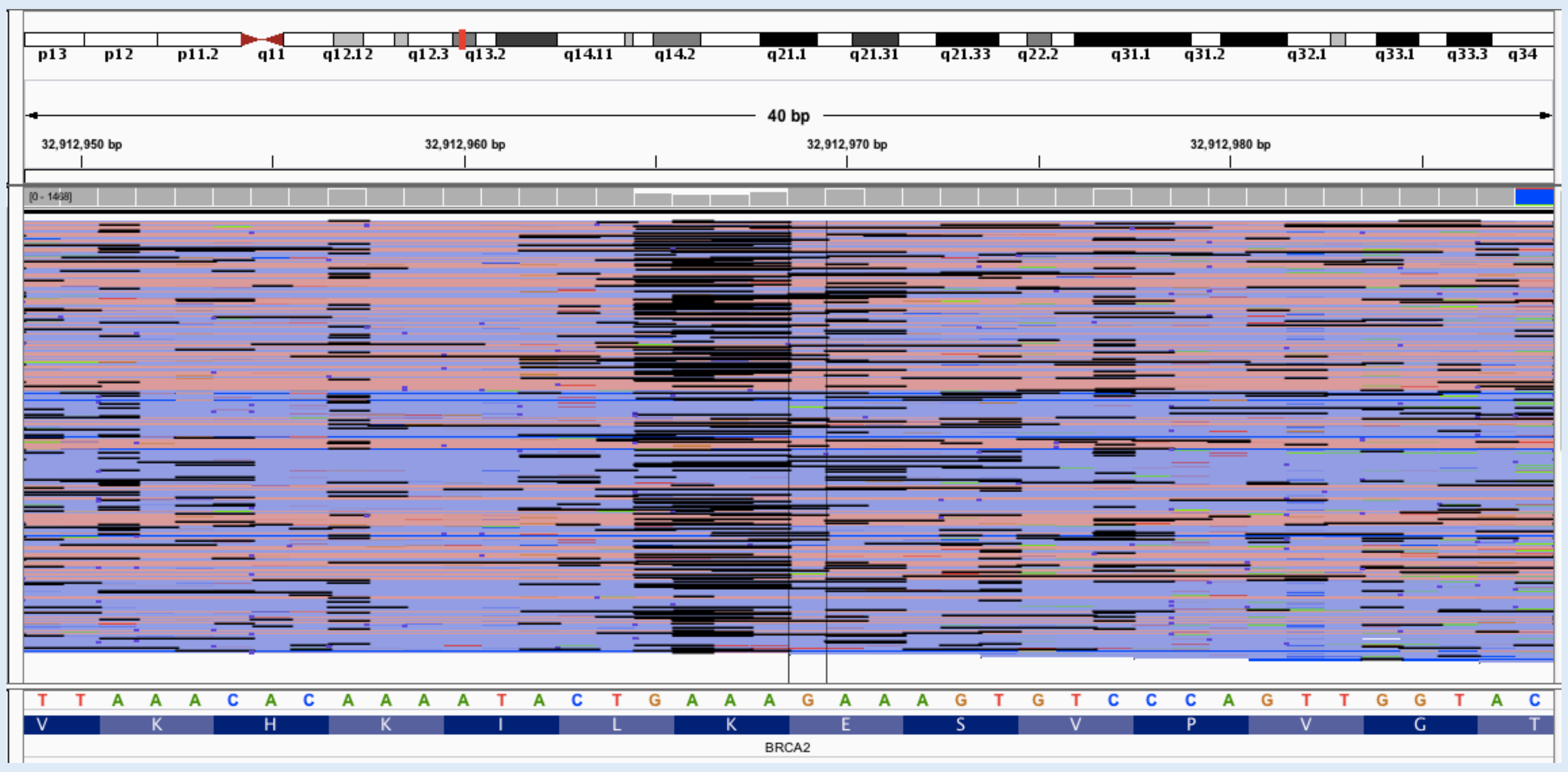


Figure 3
Visualisation of the E1493fs variant in a *BRCA2* case

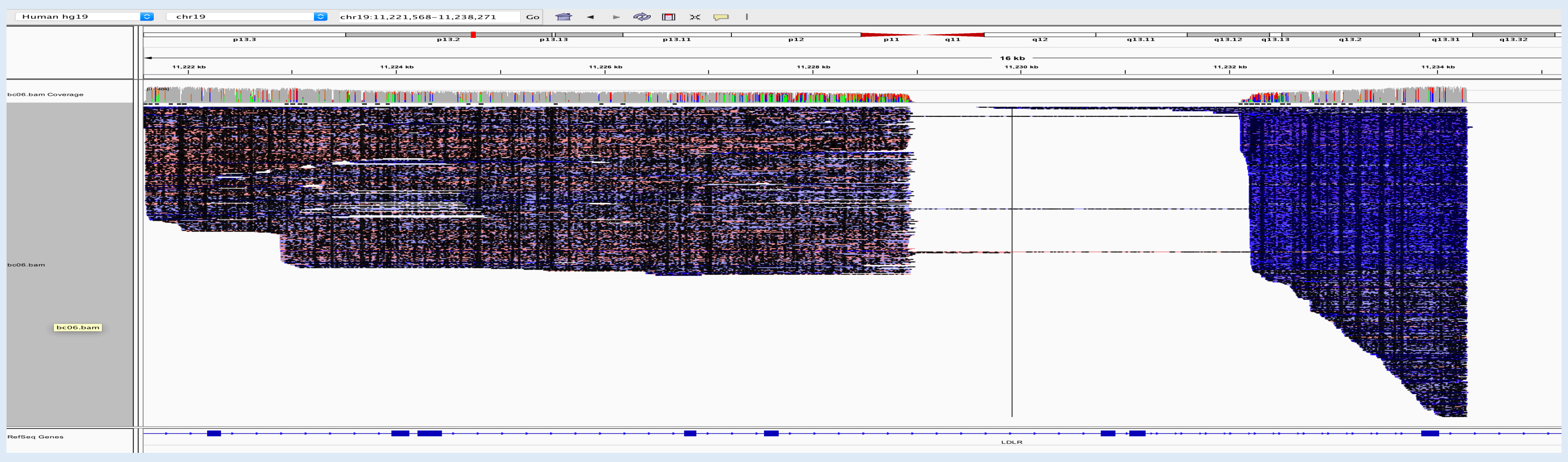


Figure 4
Exon 13-14 deletion in *LDLR*

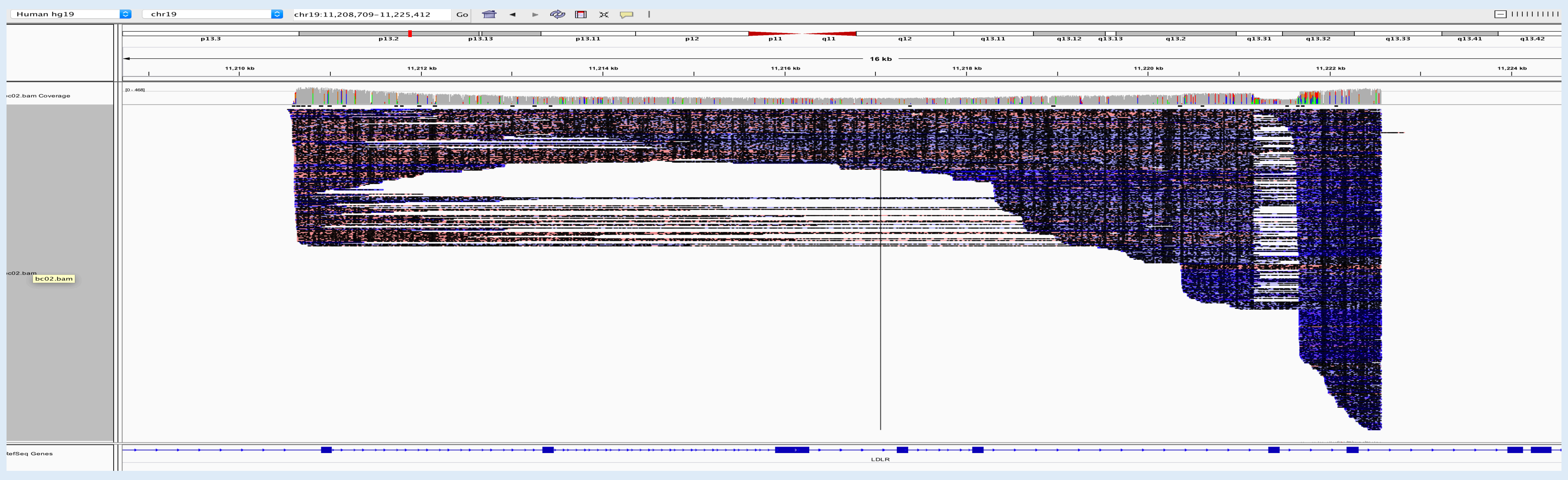
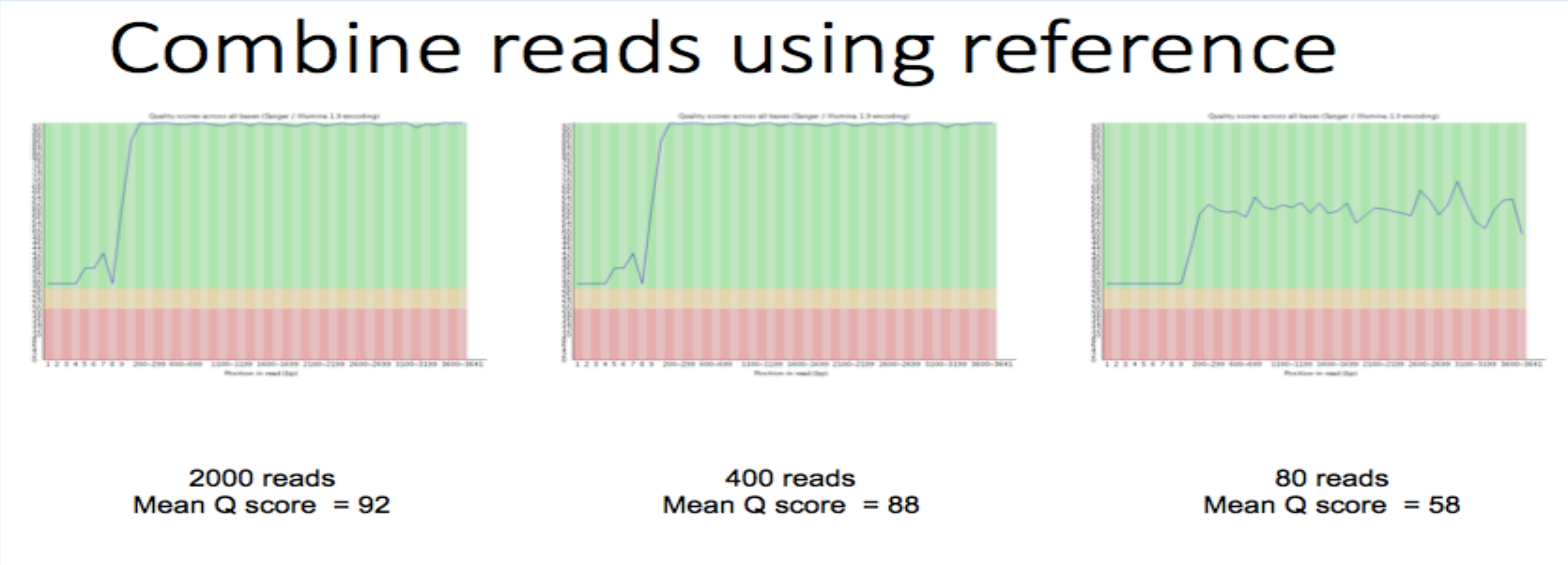


Figure 5
Exon 7 deletion in *LDLR*

Results: improving Q-scores using consensus reads



In conclusion, random error rates are tractable by either consensus alignment and oversequencing. Provided systematic errors can be avoided, as with 2D sequencing, nanopore sequencing can deliver unique tools for clinical use and point of care testing.