Identifying Polymorphisms in the Alzheimer’s Related APP Gene Using the Minion Sequencer

Keeley Brookes*, Tulsi Patel*, Gabriela Zapata-Erazo, Imelda Barber, Anne Braae, Naomi Clement, Tamar Guetta-Baranes, Sally Chappell and Kevin Morgan

Human Genetics Group, School of Life Sciences, University of Nottingham, Queen’s Medical Centre, Nottingham, UK

Abstract

The MinION is a bench top sequencer by Oxford nanopore technologies (ONT) that allows long reads of DNA sequence. Few studies have tested whether polymorphisms can be detected using this device. Several polymorphisms within the APP gene were used to test this capability. Library preparation and sequencing were performed using standard ONT protocols for samples harbouring five different mutations. Alignments to the reference sequence were analysed in MinoTour and basecalls were manually investigated using proportion of reference calls between samples to identify the variants. MinoTour’s algorithm for variant detection was unable to identify the polymorphisms due to high base calling error rate. By calculating the difference in reference basecall proportions along the amplicon, it was possible to identify the polymorphisms above a Bonferroni-corrected threshold (p<1 × 10^-4). The MinION has potential for polymorphism detection when comparing samples; however careful interpretation is needed as high base calling error rates can mask the presence of polymorphisms.

Keywords: Nanopore technology; MinION; Sequencing; Polymorphism detection; Deletion

Introduction

Sequencing of DNA samples for genetic analysis has become common practice in molecular diagnostics. Over time, the cost and duration taken to sequence DNA has reduced but at the loss of read length from up to 1000 bp in 1st generation sequencing (Sanger) to only 200 bp reads in 2nd generation platforms (e.g. Illumina). The reduction in read length requires a greater depth of coverage to enable genome assembly. In addition, the inability to produce long reads results in reduced capability to observe tandem repeat polymorphisms and determine cis haplotypes. Although costs are continually decreasing, sequencing of entire genomes is still expensive.

Now 3rd generation sequencing could potentially rival standard platforms with use of nanopore technology to sequence DNA quickly, cheaply and with read lengths likely extending more than kilobases in size. The frontrunner for this technology has been Oxford nanopore technologies (ONT) in the form of its MiniION device. The MiniION, part of ONT’s arsenal of bench top sequencers, is being tested in a small number of laboratories worldwide as part of its MiniION access programme (MAP). The device, only the size of a large memory stick, offers relatively cheap in-house sequencing with real-time data production. DNA tethered to motor proteins and adaptor sequencers is passed through a biological membrane pore and specific ion current changes for each base are detected. This allows base calls to be made using Metrichor software housed in ONT’s cloud-based server.

Several publications have documented the MiniION’s sequencing ability and whilst the majority focus on accuracy of sequencing small bacterial genomes, few have looked at the ability of the MiniION to detect polymorphisms within the genome [1]. A high error rate in base calling is observed with the device compared to current platforms, largely due to the influence of simultaneous, multiple adjacent nucleotides on the ion current, amongst other physical attributes such as the enzymes driving DNA through the pores too quickly for a current to be detected [2,3]. As a result, the presence of polymorphisms in the sequence is difficult to observe. The detection of structural variations <300 bp in size, with 500x amplicon coverage [4], However, the detection of single nucleotide polymorphisms (SNPs) appears more problematic due to the high base calling error. The human CYP2D6, HLA-A and HLA-B loci to determine cis-haplotypes using the long sequence reads producible by the MinION [5]. Error rates were too high for variant calling using conventional tools such as GATK, therefore polymorphisms were simply identified by classing variant basecalls that occurred in more than a third of total reads as a true SNP.

Algorithms that are more complex have been used to control sequencing error and identify novel polymorphisms in comparison to a reference sequence. The M13mp18 phage genome and aligned basecalls to a reference sequence with computationally generated variation in an effort to detect variants with the MinION [6]. The algorithm was able to detect variations with an optimal F-score, 97% recall and precision using only 60 times coverage in order to call substitutions at 1% frequency. The increase in frequency of substitutions along the reference sequence reduced variant detection accuracy; likely due to the difficulty in aligning the experimental data to the mutated reference sequence. Despite the high sequencing error rate observed, theoretically SNPs could be readily identified.

Similarly the PoreSeq algorithm which considers ion current information using a statistical model of the underlying physical system, a source of error generation in basecalling, to increase sequencing accuracy [7]. PoreSeq was also examined for its ability to detect variants by altering the reference sequence and computing likelihood scores of wild type and mutant sequences. When the observed likelihood score was greater for the correct base than the altered reference base, a correct call was made. PoreSeq was able to detect variants even at low sequence coverage.

In this investigation, several polymorphisms, two rare variants

*Corresponding authors: Keeley Brookes, Human Genetics Group, School of Life Sciences, University of Nottingham, Queen’s Medical Centre, Nottingham, UK, Tel: +44115 823 0141; E-mail: keeley.brookes@nottingham.ac.uk

Received March 31, 2016; Accepted May 11, 2016; Published May 13, 2016

Citation: Brookes K, Patel T, Zapata-Erazo G, Barber I, Braae A, et al. (2016) Identifying Polymorphisms in the Alzheimer’s Related APP Gene Using the Minion Sequencer. Next Generat Sequenc & Applic 3: 125. doi:10.4172/2469-9853.1000125

Copyright: © 2016 Brookes K, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
(including one novel SNP) and two common variants within the APP gene were used to test the device’s ability to detect variation despite the high reported error rate (Figure 1). Amplicons known to be wild type, heterozygous or homozygous for these polymorphisms (validated by Sanger method) were sequenced using the MinION to determine whether the polymorphisms could be detected. Variations in the amplicons were analysed using the MinoTour tool detection algorithms [8]. In addition, as an alternative to algorithm detection, the proportions of reference basecalls at each position along the amplicon were compared between samples in order to test the hypothesis that the error rate between samples would be similar and therefore any detections for deviations in the proportion of reference basecalls between samples would indicate points of genetic variation.

Results

Five polymorphisms located within the APP gene (Figure 1A) were sequenced in homozygous and heterozygous samples on the MinION using three different amplicons. Wild type samples for the respective polymorphisms were also sequenced as a comparison with all genotypes. All samples were previously Sanger sequencing confirming their genotype and indicating no DNA variations other than the polymorphisms under investigation. Primers designed to amplify these regions with standard PCR protocol can be found in Figure 1B.

MinoTour generates a number of descriptives per sequencing run, summarised in Table 1. The number of total reads obtained from the data included all reads of the template, complimentary and reads with both template and complimentary strands (2d). Error rate and distribution across the data suggests that error rate in basecalling is not significantly different between samples of the same amplicon, and are therefore comparable. Later versions of the library preparation kits used to sequence amplicons harbouring the common SNP’s rs2830088 and rs2830051 show an improved error rate in basecalling. Kolmogorov-Smirnov (KS) tests suggest that the error rates of the samples are normally distributed, with a right-handed skew. Despite the range in clustering denoted by Kurtosis scores, all show a narrow clustering of data points about the mean. Interestingly, increasing the number of reads generated does not seem to greatly improve the error rate in basecalling, indicated by a non-significant Pearson’s correlation for both number of Total Reads/2d Reads and error rates (p=0.61 and p=0.97 respectively), however this would need further specific investigation to confirm this.

Deletion variant (rs367709245) detection

Analysis using the MinoTour tool yielded no consensus variants (see methods for description) in samples containing the deletion (rs367709245) or rare SNPs (rs63750066, rs63749964). This implies

Figure 1: A) Schematic of the APP gene and location of the polymorphisms sequenced using the MinION. B) Primers designed for PCR amplification and sequencing of the APP polymorphisms (MAF from 1000 Genomes).

| Polymorphism     | Gene Location | Primer | Sequence (5’→3’) | Amplicon Size | Location of polymorphism in Amplicon | MAF | SNP |
|------------------|---------------|--------|------------------|---------------|-------------------------------------|-----|-----|
| rs367709245      | Intron 17     | Forward| CAGCAAGTTGGCGAGAAT| 481           | 385–390                             | 0.006|      |
| rs2830051        | Exon 16       | Reverse| CAGGGTGGATTGCTAATG | 481           | 129                                 | 0.002|      |
| rs2830051        | Exon 16       | Forward| CAGCAAGTTGGCGAGAAT| 481           | 128                                 | 0.002|      |
| rs2830051        | Exon 16       | Reverse| CAGGGTGGATTGCTAATG | 481           | 128                                 | 0.002|      |
| rs2830051        | Exon 16       | Forward| CAGCAAGTTGGCGAGAAT| 481           | 128                                 | 0.002|      |
| rs2830051        | Exon 16       | Reverse| CAGGGTGGATTGCTAATG | 481           | 128                                 | 0.002|      |

|          | rs367709245  | rs63750066 | rs63749964 | rs2830058 | rs2830088 | rs2830051 |
|----------|--------------|-------------|------------|-----------|-----------|-----------|
| Total Reads Generated | 12562        | 1301        | 8223       | 1194      | 2783      | 2648      |
| Total 2d Reads        | 10834        | 770         | 6779       | 969       | 2082      | 2002      |
| Average basecalls per position | 11654.5    | 680.1       | 6885.3     | 955.8     | 1910.2    | 1849.7    |
| Average % Error Rate (SD) | 20.1       | 25.6        | 22.9       | 24.5      | 21.3      | 20.2      |
that the heterozygous minor alleles, expected in 50% of the basecalls, were not observed at a high enough frequency to be called as consensus variants.

MinoTour detected four single base deletion variations in the sample containing the 6 bp deletion; none of these were within the rs367709245 deletion region under study (Table 2). However all variants coincided with a mononucleotide run in the sequence. Although the 6 bp deletion was not detected in the sample by the MinoTour algorithm, visualisation of base coverage across the amplicon indicated an increase in proportion of deletion calls in the heterozygous sample compared to wild type within the deletion region (Figure 2). This observation suggests that comparing the difference in proportion of alleles between the samples could lead to detection of polymorphisms against the background error rate, which would be similar across all samples.

The proportion of reference basecalls was calculated from the 2d reads provided for each position along the amplicon for the wild type and the rs367709245 6 bp deletion sample. The proportions were then compared between the two samples resulting in a percentage difference in reference base calls for each position. Percentage deviations ranged from 0% indicating a similar reference basecall rate between samples to 22% indicating significant deviations and therefore a potential polymorphism occurring in one of the samples. Four positions along the amplicon indicated a percentage difference of >20%, three of which were located within the deletion region of the polymorphism, the

| Reference | Reference Position | Deletion Calls | Total Reads | Percentage of Reads called as Deletion |
|-----------|--------------------|----------------|-------------|---------------------------------------|
| T         | 35                 | 6561           | 11084       | 59.3                                  |
| T         | 366                | 5604           | 11425       | 49.1                                  |
| T         | 397                | 5390           | 11341       | 47.5                                  |
| A         | 398                | 7877           | 11340       | 69.5                                  |
| A         | 399                | 5398           | 11341       | 47.6                                  |

Table 2: Potential deletion variations within the sequences amplified. The locations of these variations do not coincide with the position of the deletion polymorphism rs367709245, which resides in the heterozygous sample.

Figure 2: A) Scaled proportions of calls in the 2d reads of the 6bp deletion region colour coded for each call type at those positions. The graphic indicates that there is a higher proportion of deletion calls (black) in the heterozygous sample (right) than in the wild type (left). B) Scaled proportions of deletion calls from 2d run analysis of the wild-type (left) and heterozygous carrier of the 6bp deletion (right). Although proportions of deletion calls are generally higher in the heterozygous carrier, the patterns of deletion calls are similar between the two samples. However, within the 6bp deletion region (red) there is a substantial increase in deletion calls as would be expected.
fourth was located at position 30. The base at amplicon position 30 was set between two mononucleotide runs in the sequence and therefore likely due to ‘slippage’ which is unsurprising. In addition to the three positions of high percentage difference already mentioned, the other three positions of the 6 bp deletion also displayed high percentage differences in the proportion of reference basecalls between the wild type sample and the deletion sample. The average percentage difference of reference basecalls within the 6 bp deletion region was 18.2%, with the average of the entire amplicon at 5.4% (Figure 3A) indicating a high level of difference between the two samples in this region. Proportions of reference calls at this position were subjected to Chi-square ($\chi^2$) tests, the results indicated that the proportion of reference base calls at these positions were significantly different when corrected for largest (482 bp) amplicon (p value < 1 × 10^-4) for 5 positions out of the 6 containing the deletion. Conversely there was also an increase in deletion calls made at these positions with an average increase of 13.3% in deletion call rate with the rs367709245 heterozygous sample.

**Rare SNP (rs63750066, rs63749964) detection**

Two rare SNP variants were located within the same amplicon as the deletion and two samples heterozygous for these SNPs (rs63750066, rs63749964) were also sequenced. MinoTour algorithms identified numerous potential variants in the samples heterozygous for SNPs rs63750066 and rs63749964 (Table 3). This suggested that several SNPs existed within the amplicon sequences, however as many of these variants were also found in the wild type sample; they are likely to be false positives. Indeed sequencing of the samples using Sanger sequencing confirmed this was the case. Although the rs63750066 and rs63749964 polymorphisms were amongst these, the background of multiple potential variants renders it difficult to distinguish the correct polymorphism due to the high rate of sequencing error.

Percentage difference in homology to the reference sequence between wild type and heterozygote samples for rs63750066 was plotted along the amplicon (Figure 3B). Proportions of reference basecalls differed up to 16.4% with a single position displaying a difference of 39.9%. This position coincided with the position of the SNP (position 228), and was clearly above the average percentage difference (from wild type) for that sample (3.1%). A Chi-square test on this data indicated a highly significant signal with p value <1 × 10^-4. This difference in the proportion of reference base calls was mirrored by an increase in the proportion of the minor allele (A) occurring at this position of 16.4%, the largest increase in allele proportions across the amplicon.

![Figure 3: Graph for the percentage difference of reference base call proportions across the amplicon between the wild type and heterozygous samples for rs367709245 (A), rs63750066 (B) and rs63749964 (C). Red boxes indicate the percentage difference peaks that correspond to the location of the known polymorphisms along the amplicon.](image-url)
### Proportions of reference basecalls along the amplicon harbouring polymorphisms

|        | 216 | 3605 | 823 | 393 | 1256 | 152 | 3238 | 396 |
|--------|-----|------|-----|-----|------|-----|------|-----|
| T      | 226 |      |     |     |      |     |      |     |
| G      | 228 |      |     |     |      |     |      |     |
|        |     | 32   | 472 | 250 | 45   |     |      |     |
| C      | 242 |      |     |     |      |     |      |     |
| T      | 244 |      |     |     |      |     |      |     |
| C      | 245 |      |     |     |      |     |      |     |
| C      | 247 | 163  | 880 | 1316| 7362 |     |      |     |
| T      | 279 |      |     |     |      |     |      |     |
| C      | 281 | 175  | 103 | 3318| 7727 | 16  | 12   | 192 |
| T      | 284 | 134  | 8601| 1987| 567  | 7   | 222  | 61  |
| G      | 292 | 607  | 269 | 6188| 1801 | 35  | 24   | 361 |
| G      | 307 | 454  | 137 | 7292| 1937 |     |      |     |
| A      | 309 | 6900 | 120 | 221 | 1794 |     |      |     |
| A      | 349 |      |     |     |      |     |      |     |
| T      | 362 | 148  | 6778| 1676| 449  |     |      |     |
| T      | 365 | 116  | 3708| 1027| 970  | 7   | 222  | 61  |
| T      | 371 | 619  | 7284| 1146| 709  |     |      |     |
| T      | 373 | 191  | 6751| 1227| 869  | 17  | 263  | 73  |
| A      | 374 |      |     |     |      |     |      |     |
| T      | 397 | 206  | 4514| 625 | 606  |     |      |     |
| A      | 398 | 1580 | 480 | 639 | 764  | 129 | 30   | 41  |
| A      | 399 | 3945 | 622 | 681 | 695  | 224 | 33   | 48  |
| G      | 431 | 1132 | 496 | 8003| 921  |     |      |     |
| T      | 468 |      |     |     |      |     |      |     |

Table 3: Potential variants detected by MinoTour across the different sample sequencing reads of each amplicon. Although several possible variants were observed due to their frequency occuring 2sd from the error mean, the only two real variants are highlighted in red.

Likewise, comparison of the proportions of reference base calls between the wild type and the sample heterozygous for the rs63749964 also indicated that all but one position had a percentage difference of less than 20%. Position 241 of the amplicon displayed a percentage difference in the proportion of reference alleles of 34.6%, corresponding to the location of the SNP, against an average percentage difference of 4.7% along the amplicon (Figure 3C). There was a corresponding increase in minor allele (G) proportion of 22.3% at this position. 

#### Common SNP (rs2830088 and rs2830051) detection

Two further amplicons harbouring common SNPs were analysed using the MinION. For each amplicon a sample homozygous for the major allele, heterozygous and homozygous for the minor allele were sequenced and analysed with the MinoTour Tool and by direct sample sequencing. The MinoTour Tool algorithm was unable to detect consensus algorithms that sought to control the error rate in basecalling, such as the MinoTour Tool.

A similar result was observed for the second common SNP, rs2830051. The average percentage differences in reference basecalls between the major allele homozygote and the heterozygote and minor allele homozygote was 1.1% and 1.2% respectively. Larger percentage differences were observed for a single position (231), which corresponded to the location of the polymorphism (Figure 4B). Increasing differences in proportion of reference basecalls was seen with comparison of the heterozygote (25.3%) and of the homozygous minor allele sample (44.8%). These differences, when tested were also significant at study-wide level (p<1 × 10^{-4}). The proportion of basecalls for the minor allele of the SNP was also shown to increase to 26.8% in the heterozygote and 46.2% in the homozygous minor allele carriers (Table 5).

### Discussion

This investigation set out to determine whether polymorphisms could be detected by nanopore sequencing using the Oxford nanopore technology (ONT) MinION device. Variation detection was implemented using the MinoTour Tool algorithms and by direct sample comparison of the proportion of reference basecalls along the entire length of each respective amplicon. The hypothesis being that the error rate would be similar between samples and therefore any deviation would be indicative of a polymorphism, negating the use of complex algorithms that sought to control the error rate in basecalling, such as the MinoTour Tool.

MinoTour [8] has a user-friendly web interface that utilises the
Table 4: Potential variants called by MinoTour for the samples containing the rs2830088 (A) and rs2830051 (B) polymorphisms.

| Ref Seq | Ref Pos | rs2830088 WT n=10 | rs2830088 Het n=8 | rs2830088 Mut n=9 |
|---------|---------|------------------|------------------|------------------|
|         | A       | T                | G                | C                | Total   | A       | T           | G             | C        | Total   |
| A       | 25      | 1320             | 67               | 103              | 182       | 1672    | 1340      | 68             | 85       | 143      | 1661     |
| C       | 54      | 46               | 61               | 186              | 1514      | 1807    | 408       | 11624          | 1507     | 834      | 14373    |
| C       | 71      | 141              | 148              | 189              | 1288      | 1766    | 1090      | 1318           | 1456     | 9843     | 13707    |
| T       | 72      | 36               | 1521             | 205              | 74        | 1836    | 30        | 1520           | 108      | 162      | 1893     |
| T       | 73      | 39               | 1555             | 129              | 1511      | 1874    | 41        | 1475           | 119      | 162      | 1873     |
| T       | 76      | 54               | 1523             | 129              | 1511      | 1874    | 1100      | 1784           | 1105     | 1250     | 14639    |
| A       | 100     | 1499             | 38               | 150              | 1276      | 1814    | 1484      | 45             | 117      | 137      | 1883     |
| C       | 125     | 18               | 704              | 16               | 965       | 1828    | 128       | 10011          | 305      | 3128     | 13572    |
| C       | 157     | 82               | 145              | 79               | 1498      | 1804    | 1325      | 38             | 140      | 83       | 1800     |
| A       | 214     | 1319             | 28               | 185              | 89        | 1621    | 122       | 202            | 86       | 1051     | 13928    |
| C       | 226     | 132              | 202              | 87               | 1058      | 1479    | 122       | 202            | 86       | 1051     | 13928    |

| Ref Seq | Ref Pos | rs2830051 WT n=19 | rs2830051 Het n=20 | rs2830051 Mut n=21 |
|---------|---------|------------------|------------------|------------------|
|         | A       | T                | G                | C                | Total   | A       | T           | G             | C        | Total   |
| T       | 25      | 41               | 1025             | 131              | 236      | 1433    | 23        | 773            | 85       | 161      | 1042     |
| A       | 26      | 939              | 58               | 96               | 226      | 1319    | 26        | 939            | 58       | 96       | 226      |
| A       | 37      | 328              | 40               | 100              | 165      | 633     | 221       | 29             | 68       | 105      | 423      |
| A       | 38      | 659              | 72               | 85               | 207      | 1023    | 466       | 53             | 61       | 101      | 681      |
| T       | 50      | 86               | 914              | 212              | 204      | 1416    | 43        | 725            | 121      | 171      | 1062     |
| A       | 54      | 466              | 172              | 61               | 118      | 817     | 383       | 117            | 48       | 74       | 622      |
| A       | 68      | 596              | 108              | 85               | 110      | 878     | 463       | 96             | 48       | 69       | 676      |
| T       | 152     | 32               | 1038             | 234              | 213      | 1517    | 28        | 771            | 173      | 150      | 1122     |
| A       | 153     | 744              | 28               | 169              | 97       | 1038    | 1102      | 41             | 262      | 118      | 1523     |
| A       | 179     | 652              | 49               | 57               | 149      | 907     | 1000      | 78             | 96       | 191      | 1365     |
| C       | 200     | 181              | 264              | 26               | 1083     | 1554    | 34        | 1121           | 136      | 249      | 1540     |
| T       | 214     | 63               | 951              | 309              | 107      | 1430    | 40        | 809            | 180      | 83       | 1112     |
| A       | 215     | 974              | 25               | 241              | 256      | 1496    | 771       | 18             | 172      | 178      | 1139     |
| T       | 231     | 24               | 622              | 76               | 372      | 1094    | 79        | 277            | 128      | 907      | 1391     |
| C       | 242     | 63               | 226              | 134              | 1218     | 1641    | 54        | 1165           | 242      | 135      | 1596     |
| C       | 257     | 84               | 152              | 163              | 1002     | 1401    | 144       | 131            | 129      | 677      | 1081     |
| C       | 294     | 242              | 192              | 155              | 851      | 1440    | 259       | 175            | 174      | 926      | 1534     |
| T       | 314     | 16               | 1208             | 139              | 265      | 1628    | 571       | 154            | 168      | 239      | 1132     |
| A       | 317     | 454              | 146              | 175              | 204      | 979     | 376       | 86             | 111      | 188      | 761      |
| A       | 329     | 726              | 128              | 75               | 114      | 1043    | 523       | 84             | 48       | 70       | 725      |
| T       | 351     | 55               | 516              | 221              | 61       | 853     | 41        | 382            | 132      | 55       | 610      |
| T       | 382     | 109              | 355              | 212              | 57       | 733     | 53        | 300            | 131      | 45       | 529      |
| C       | 418     | 96               | 107              | 190              | 996      | 1389    | 49        | 97             | 124      | 785      | 1055     |
| A       | 425     | 1014             | 32               | 270              | 147      | 1463    | 808       | 28             | 163      | 97       | 1096     |

Figure 4: Graph of percentage difference for the proportion of reference base calls between major allele homozygotes and heterozygote (blue line) and minor allele homozygote (red line) samples. A) Percentage difference along the amplicon harbouring the rs2830088 SNP. Increases in percentage difference occur at the point of variation, increasing with the number of minor alleles present. B) Percentage difference along the amplicon harbouring the rs2830051 SNP. Increases in percentage difference occur at the point of variation, increasing with the number of minor alleles present. The areas surrounding the polymorphisms are enlarged and shown inset in boxes.
 However, neither of the heterozygous or homozygous minor allele samples show the expected 50% and 100% percentage of basecalls. Both show the expected increase in the percentage of calls for the minor allele across the heterozygous and homozygous minor allele samples. However, neither of the heterozygous or homozygous minor allele samples show the expected 50% and 100% percentage of base calls.

| rs2830088 % of Minor Allele (T) reads | rs2830051 % of Minor Allele (C) reads |
|--------------------------------------|--------------------------------------|
| 6.99                                 | WT (0%)                              |
| 34.38                                | Het (50%)                            |
| 60.79                                | Mut (100%)                           |

Table 5: Percentage of basecalls for the minor allele for each common SNP across the samples. Both show the expected increase in the percentage of calls for the minor allele across the heterozygous and homozygous minor allele samples. However, neither of the heterozygous or homozygous minor allele samples show the expected 50% and 100% percentage of base calls.

As an alternative, direct sample comparisons were used to account for sequencing error rate as it was found to be similar across samples. The sequencing of a confirmed wild type sample was used as a baseline to compare the proportion of reference basecalls against samples containing polymorphisms. In doing this, the base error rate in MinION sequencing was taken into consideration, providing clearer results to detect polymorphisms. Identifying positions with significant differences in the proportion of reference basecalls between the samples would be indicative of variation between samples suggesting the presence of a polymorphism. In all SNP cases, the polymorphism position displayed highly significant differences in the proportion of reference basecalls between wild type and minor allele carrying samples. The observation that the difference in proportion was significant beyond the study-wide level indicates that this analysis could be applied to de novo detection of polymorphisms in full genome sequencing of samples with unknown genotypes.

The underlying biochemistry behind the Nanopore sequencing is improving with lower average error rates and percentage difference of reference basecalls between samples. This was observed with the common polymorphism tests as these were sequenced using an updated library preparation kit (SQK-MAP006) and protocol. Our observations indicate that sequence might also play a part in the error rate as the amplicon containing the rs2830051 polymorphism had much lower error rates and discrepancies between samples than the amplicon containing the rs2830088 polymorphism, which was sequenced at the same time. In addition to this, the accuracy of the reads may also be influenced by the flowcell, as each amplicon was sequenced on a different flowcell.

What was surprising was that given a heterozygous sample and one that was homozygous for the minor allele, percentages of the minor allele did not reach the expected 50% or 100% of basecalls for the alternative allele (Table 5). For example, the minor allele (T) in the rs2830088 polymorphism was called in 34.4% of the reads in the heterozygote and 60.8% in the homozygote. Despite the proportion almost doubling as expected it is still shy of the expected proportions potentially showing a bias towards the reference sequence in basecalling. This may be due to the inadequate removal of DNA samples from the flowcells by the washing procedure. Given that the order of sequencing began with the wild type sample first, carryover would indicate a bias towards reference basecalling. Further investigation of this would prove useful.

The MinION offers the realistic vision of every lab having its own sequencer in the future. However, in its current form, although it can provide long-read analysis of genome coverage, the ability to reliably and easily detect polymorphisms is limited. There is a need to decrease the sequencing error rate before it can become a useful commodity. The MinION and its future reincarnations will only become more accurate in basecalling abilities. With reduced error rates, the possibility of identifying polymorphisms, both known and novel, will be greatly improved by alignment to a reference sequence. This investigation demonstrates that polymorphisms can be readily identified by comparing proportions of reference calls between wild type and mutant samples.

Currently the error rate is still high and creates too many false positives when detecting polymorphisms, which prevents novel SNPs from being detected against the background of spurious signals. Therefore, a highly stringent significance threshold should be used and the most significant results fully investigated and validated by an alternative approach. Although the basecalling error rate of nanopore technology might deter users from utilising it to identify polymorphisms when sequencing genomes, we demonstrate a simple way of distinguishing known polymorphisms above the background error by calculating the differences in basecalling rates and propose that potential novel variants could also be identified.

### Methods

#### PCR

Five polymorphisms within the APP gene were sequenced using three different amplicons. Primers designed for amplification via standard PCR protocol are shown in Figure 1B. DNA was extracted from human blood samples and a single sample for each genotype was used in this experiment. Amplified products were cleaned with ExoSAP and pooled to total 1ug of PCR product in a volume of 80 µl as specified in the ONT SQK-MAP005 protocol for library preparation (rs367709245, rs63750066, rs63749964) or 1ug of PCR product in 45 µl for SQK-MAP006 (rs2830088, rs2830051). All samples were previously validated using traditional Sanger sequencing to confirm genotypes of all polymorphisms and absence of other polymorphisms.

#### Library preparation and sequencing

Samples prepared with SQK-MAP005 were end-repaired using standard NEB End-repair kits (New England Biolabs), followed by dA-tailing of the blunt-ended amplicons (New England Biolabs). In SQK-MAP006 NEBNext Ultra II End-repair/dA-tailing module was used (New England Biolabs), combining both reactions into a single mix. Subsequent purifications were carried out with AMPure XP Beads (Beckman Coulter). Samples were ligated to the ONT adaptors and purified using magnetic beads (SQK-MAP005 His-tag beads; SQK-MAP006 MyOne C1 beads) prior to loading for sequencing. Each sample was run to minimum template read coverage of 1000x, ending the run when read generation had slowed to one per minute. Flowcells were flushed through with washing buffers before loading the next sample, sample order of sequencing maintained as wild type followed by heterozygote, and finally homozygote samples where applicable. A
new flowcell was used for each amplicon to prevent contamination and spurious error rates caused by non-familiar amplicons.

Analysis

Basecalling of amplicon sequences from the MinION were made in real-time with ONT software Metrichor (V1.69) and simultaneously uploaded to the MinoTour (V0.46) analysis tool for visualisation of the data [8]. Alignment analysis was performed on 2-directional (2d) reads where both template and complement strands were read to produce a consensus, resulting in greater sequence accuracy. Details for the algorithm for the alignment tool can be found in reference [8]. MinoTour was used to detect variation from the given reference sequence for each amplicon, including those that were 100% match to the reference (wild type). The tool uses two methods to detect variants; a consensus variant occurs when a non-reference base has a greater base count than the reference allele and a potential variant is called when an alternate allele to the reference occurs more frequently than 2 standard deviations (SD) from the average error rate of the sequencing run.

The 2d counts for bases and indels at every position along the amplicons were obtained from MinoTour and subjected to manual calculation. Initial exploration of error rate for each amplicon was investigated using the percentage of non-reference basecalls. In order to observe the known polymorphisms, the proportion of reference basecalls from the total (inclusive of indels) at each position was compared between wild type and variant carrying samples. Calculating the percentage difference in these proportions allowed comparisons to be made, as a greater difference at any given location would be indicative of a potential polymorphism. To verify the increased percentage difference for reference basecalls proportions at the polymorphism site in variant samples, significance of this difference was calculated using a Chi-squared ($\chi^2$) test for each position. Assuming the null hypothesis there would be no significant difference in the proportion of reference basecalls between samples. A study-wide corrected p-value for significance was calculated using the size of the largest amplicon studied (482 bp).

Acknowledgements

The work conducted was supported by Alzheimer’s Research UK. The Neuroscience Group and University of Nottingham School of Life Sciences provided studentship funding for TP. We thank Oxford Nanopore Technologies for reagents provided as part of the Early Access Programme and Matt Loose for his guidance on utilising the MinoTour programme suite for the sequencing analysis.

References

1. Loman NJ, Quinlan AR (2014) Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics 30: 3399-3401.
2. Manrao EA, Derrington IM, Laszlo AH, Langford KW, Hopper MK, et al. (2012) Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. Nat Biotechnol 30: 349-353.
3. Manrao EA, Derrington IM, Paviennok M, Niederweis M, Gundlac JH (2011) Nucleotide discrimination with DNA immobilized in the MspA nanopore. PLoS One 6: e25723.
4. Norris AL, Workman RE, Fan Y, Eshleman JR, Timp W (2016) Nanopore sequencing detects structural variants in cancer. Cancer Biology & Therapy.
5. Ammar R, Paton TA, Torti D, Shilen A, Bader GD (2015) Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes 4: 17.
6. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B (2015) Improved data analysis for the MinION nanopore sequencer. Nat Methods 12: 351-356.
7. Szalay T, Golovchenko JA (2015) De novo sequencing and variant calling with nanopores using PoreSeq. Nat Biotechnol 33: 1087-1091.
8. Camilla LC, Loose M, Tyson JR, de Cesare M, Brown BL, et al. (2015) MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Res 4: 1075.