Journal of
Next Generation Sequencing & Applications

Brookes et al., Next Generat Sequenc & Applic 2016, 3:1
DOI: 10.4172/2469-9853.1000125

Research Article Open Access

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 ($p<1 \times 10^{-4}$). 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 MinION device. The MinION, part of ONT's arsenal of bench top sequencers, is being tested in a small number of laboratories worldwide as part of its MinION 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 MinION's sequencing ability and whilst the majority focus on accuracy of sequencing small bacterial genomes, few have looked at the ability of the MinION 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

(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.

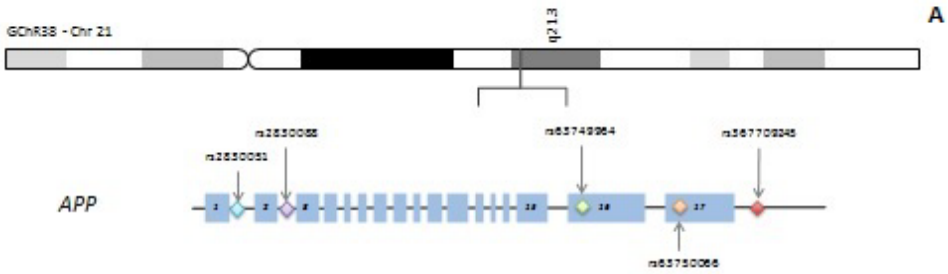
Results

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



B

Polymorphism	Gene Location	Primer	Sequence (5'->3')	Amplicon Size	Location of polymorphism in Amplicon	MAF	SNP
rs367709245 (6bp deletion)	Intron 17	Forward	CAACCAGTTGGGCAGAGAAT	482	385-390	0.0006	
		Reverse	CACGGTAAGTTGCAATGAATG				
rs63749964	Exon 17	Forward	CAACCAGTTGGGCAGAGAAT	482	241	-	T>G
		Reverse	CACGGTAAGTTGCAATGAATG				
rs63750066	Exon 16	Forward	CAACCAGTTGGGCAGAGAAT	482	228	0.0002	G>A
		Reverse	CACGGTAAGTTGCAATGAATG				
rs2830051	Intron 2	Forward	TTCCTCTCAACTTTTCTGTGAATC	451	231	0.44	T>C
		Reverse	ATGAGTGCTGCAGCTTTGTG				
rs2830088	Intron 1	Forward	ACACAGCCCTCACTCGATCT	255	125	0.48	C>T
		Reverse	AGCCTCCGACCTCTATCTCC				

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).

	Wildtype	rs367709245	rs63750066	rs63749964	rs2830088			rs2830051		
					Major	Het	Minor	Major	Het	Minor
Total Reads Generated	12562	1301	8223	1194	2783	2648	19630	5031	5028	5054
Total 2d Reads	10834	770	6779	969	2082	2002	16169	2019	2021	2109
Average basecalls per position	11654.5	680.1	6885.3	955.8	1910.2	1849.7	14973.3	1665.7	1252.4	1844.4
Average % Error Rate (SD)	20.1	25.6	22.9	24.5	21.3	20.2	20.3	16.1	15.2	17.4
	-11.6	-12.5	-12.6	-13.1	-11.1	-10.9	-11.2	-6.9	-7	-8.2

KS Test	2.2	1.4	1.7	1.8	1.1	1.7	1.6	2.1	2.8	2.7
Skewness	1.1	0.5	0.8	0.6	0.5	1.2	2.1	1.7	2	2.1
Kurtosis	2.5	0.1	0.9	0.1	1.7	4.4	14.2	5.7	6.8	7.7

Table 1: Descriptive summary of sequencing reads for the five polymorphisms tested. Two-directional (2d) reads were used for all analyses due to greater sequence accuracy. The average error rates were similar across all samples, however were reduced for the latter runs prepared using the latest library kit. Higher read generation did not appear to improve error rates. Skewness, a kurtosis and Kolmogorov-Smirnov (KS) test were performed to determine the distribution of the data and suggest that the error rates across sequencing of the same amplicons are not significantly.

		Wildtype			rs367709245			rs63750066			rs36750064		
Ref Seq	**Ref Pos**	**Deletion calls**	**Total Read**	**Percentage of Reads called as Deletion**	**Deletion Calls**	**Total Read**	**Percentage of Reads called as Deletion**	**Deletion Calls**	**Total Read**	**Percentage of Read called as Deletion**	**Deletion Calls**	**Total Read**	**Percentage of Reads called as Deletion**
T	35	6561	11064	59.3	281	599	46.9	3333	6429	51.8	431	884	48.8
T	366	5604	11425	49.1	287	659	43.6	2933	6698	43.8	413	923	44.7
T	397	5390	11341	47.5	-	-	-	3056	6639	46	407	913	44.6
A	398	7877	11340	69.5	395	654	60.4	4419	6639	66.6	593	913	65
A	399	5398	11341	47.6	288	654	44	3261	6639	49.1	440	913	48.2

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.

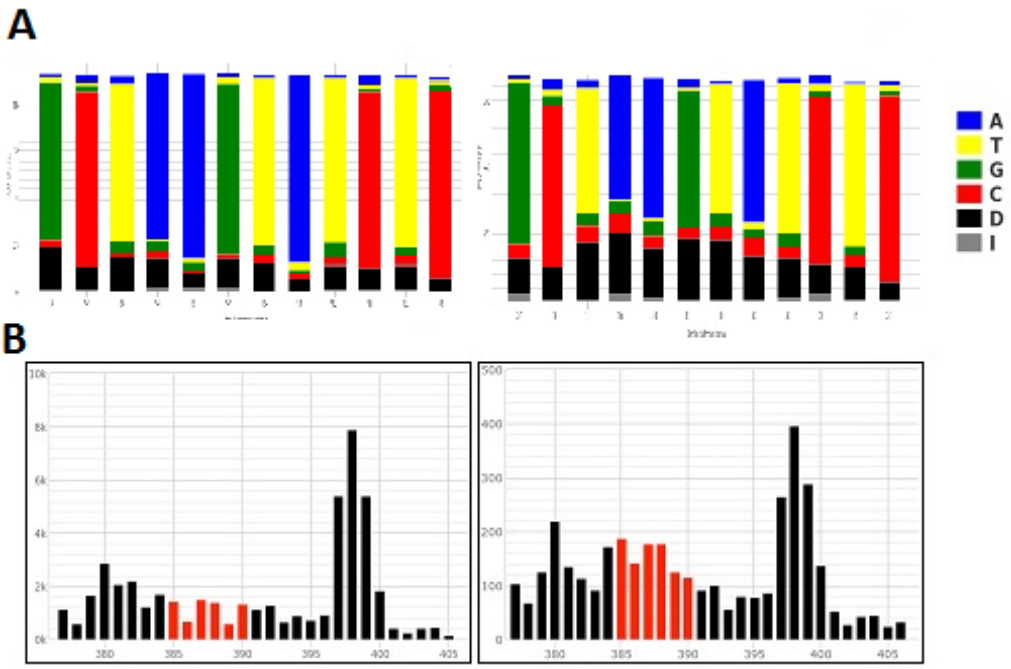


Figure 2: A) Scaled proportions of calls in the 2d reads of the 6bp deletion region colour coded for each call type at those positions. The graphic indicates that there is a higher proportion of deletion calls (black) in the heterozygous sample (right) than in the wild type (left). **B)** Scaled proportions 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.

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 sample and the rs367709245 6 bp deletion sample. The porportions 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

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 $<1 \times 10^{-4}$) 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 $<1 \times 10^{-4}$. This difference in the proportion of reference 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.

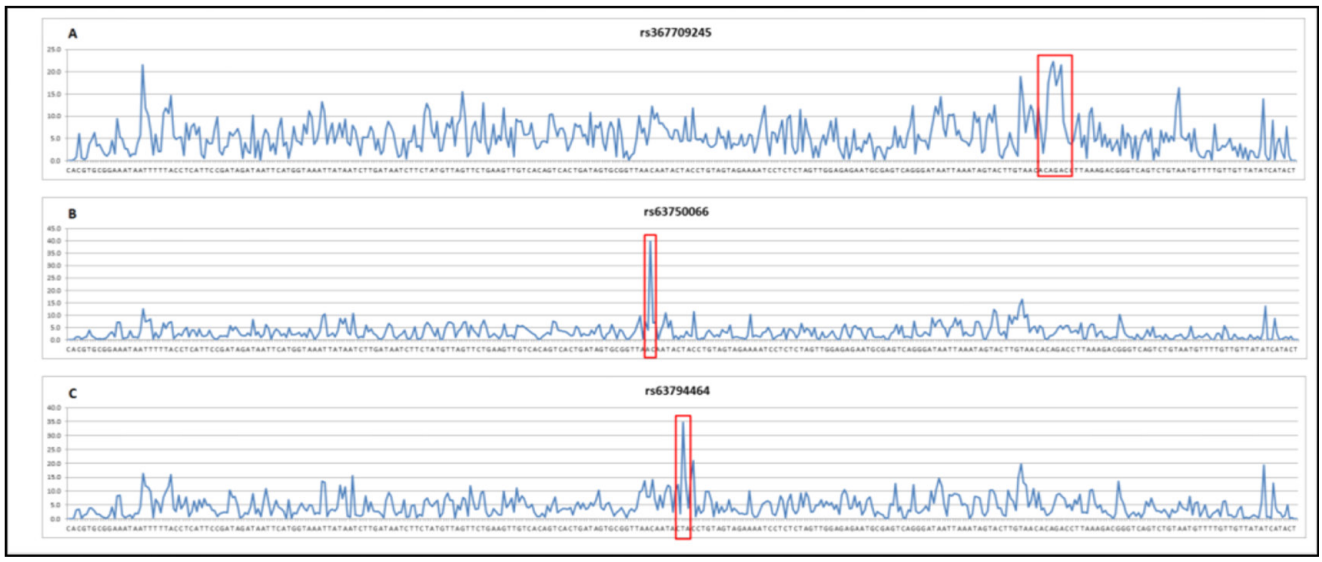


Figure 3: Graph for the percentage difference of reference base call proportions across the amplicon between the wild type and heterozygous samples for rs367709245 **(A)**, rs63750066 **(B)** and rs63749964 **(C).** Red boxes indicate the percentage difference peaks that correspond to the location of the known polymorphisms along the amplicon.

		Wildtype				rs367709245				rs63750066				rs63749964			
Ref Seq	Ref Pos	A	T	G	C	A	T	G	C	A	T	G	C	A	T	G	C
T	34	722	5798	891	236	48	304	63	31	413	3349	658	286				
T	35	200	3271	546	486	7	212	58	41	108	2098	502	388				
T	37									66	2999	666	468	9	415	103	78
A	41					320	13	42	88								
T	44													28	576	129	96
C	56	89	1370	1217	8126	19	72	99	412	96	852	890	4404	19	106	146	604
T	112					21	425	13	159	191	4018	116	1714	29	521	32	294
A	127													614	62	72	152
T	134	62	6156	459	1943	5	337	28	137								
G	190	627	243	7281	1476	42	28	361	91	357	143	3807	1261	59	23	547	175
G	201	1462	383	7648	1229	109	21	387	79	913	168	4420	779	136	21	588	118
G	202	832	459	7821	1692	63	28	428	114	487	222	4446	1144	64	45	601	167
G	205					19	444	93	77								

T	226									216	3605	823	393				
G	228									**1256**	**152**	**3238**	**396**				
T	241													**32**	**472**	**250**	**45**
C	242													88	34	143	568
T	244													60	614	170	34
C	245													37	94	67	455
C	247	163	880	1316	7362									20	80	151	571
T	279													28	616	189	76
C	281	175	103	3318	7727	16	12	192	437	128	63	1947	4547	13	12	267	619
T	284	134	8601	1987	567												
G	292	607	269	6188	1801	35	24	361	112								
G	307	454	137	7292	1937					339	92	3990	1285	43	9	561	193
A	309	6900	120	221	1794					3760	72	189	1293				
A	349									3170	327	175	747				
T	362	148	6778	1676	449					113	3129	1002	273	18	454	149	48
T	366	116	3708	1027	970	7	222	61	82	100	2413	555	697	15	312	83	100
T	367	258	5306	625	1337	22	320	31	100	353	3146	501	774	35	421	65	112
T	370	619	7284	1146	709					600	3638	767	473	68	528	116	73
T	373	191	6751	1227	869	17	263	73	88	116	2836	920	585	23	359	114	85
A	374					326	6	65	93	3655	55	665	772	461	6	120	106
T	397	206	4514	625	606												
A	398	1580	480	639	764	129	30	41	59	960	234	482	544	149	37	66	68
A	399	3945	622	681	695	224	33	48	61	1988	363	449	578	300	37	63	73
G	431	1132	496	8003	921												
T	468					49	304	94	12	574	3149	869	141	80	376	106	23

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 varaints 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 comparison of the proportion of reference basecalls between samples.

The MinoTour Tool algorithm was unable to detect consensus variants for either SNP in the heterozygous samples; however, consensus variants were detected for both homozygous minor allele samples and corresponded to the SNPs in question. In addition both homozygous and heterozygous samples yielded several potential variants (Table 4), which included the known polymorphisms present in the samples.

Proportions of reference basecalls along the amplicon harbouring the rs2830088 polymorphism were compared between the heterozygous sample and the homozygous major allele sample and between the homozygous minor allele sample and the homozygous major allele sample in order to determine the location of any sequence variation. Average percentage differences for the calls along the amplicon were 1.8% and 2% for the heterozygous and homozygous minor allele comparisons respectively. The same position (125) yielded the highest percentage difference in each comparison (Figure 4A), with the homozygous minor allele comparison displaying roughly twice the percentage difference of the heterozygous comparison (62.4% and 34.5% respectively). This coincided with the position of the SNP and displayed the largest difference across the amplicon in each comparison. Chi-square tests supported the difference giving a significant p value ($p<1 \times 10^{-4}$). Concomitantly the proportion of minor allele calls also increased to 34.4% in the heterozygote and 60.8% in the minor allele homozygote respectively.

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 ($p<1 \times 10^{-4}$). 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

A

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	T	G	C	Total
A	25	1320	67	103	182	1672	1340	68	85	143	1661	9816	595	806	1391	12608
C	54	46	61	186	1514	1807										
C	71	141	148	189	1288	1766	107	151	194	1240	1763	1090	1318	1456	9843	13707
T	72	36	1521	205	74	1836						408	11624	1507	834	14373
T	73	39	1555	129	151	1874	30	1520	108	162	1893	378	11952	1087	1344	14761
T	76	54	1523	129	151	1857	41	1475	119	162	1873	500	11784	1105	1250	14639
A	100	1499	38	150	127	1814	1484	45	117	137	1883	11206	408	1472	1234	14320
C	**125**						**18**	**704**	**16**	**965**	**1828**	**128**	**10011**	**305**	**3128**	**13572**
C	157	82	145	79	1498	1804										
A	214	1319	28	185	89	1621	1325	38	140	83	1800	10347	336	1485	766	12934
C	226	132	202	87	1058	1479	122	202	86	1051	1687	1169	1508	777	8047	11501

B

Ref Seq	Ref Position	rs2830051 WT n=19					rs2830051 Het n=20					rs2830051 Mut n=21				
		A	T	G	C	Total	A	T	G	C	Total	A	T	G	C	Total
T	25	41	1025	131	236	1433	23	773	85	161	1042	34	1121	136	249	1540
A	26	939	58	96	226	1319	649	37	64	179	929	963	77	97	230	1367
A	37	328	40	100	165	633	221	29	68	105	423	347	54	110	203	714
A	38	659	72	85	207	1023	466	53	61	101	681	738	111	69	176	1094
T	50	86	914	212	204	1416	45	725	121	171	1062	53	1114	172	247	1586
A	54	466	172	61	118	817	383	117	48	74	622	562	176	74	140	952
A	68	595	108	65	110	878	463	96	48	69	676	721	126	76	97	1020
T	152	32	1038	234	213	1517	28	771	173	150	1122	38	1125	265	228	1656
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										
T	214	63	951	309	107	1430	40	809	180	83	1112	54	1165	242	135	1596
A	215	974	25	241	256	1496	771	18	172	178	1139	1124	27	246	252	1649
T	**231**						**24**	**622**	**76**	**372**	**1094**	**79**	**277**	**128**	**907**	**1391**
C	242											63	226	134	1218	1641
C	257	84	152	163	1002	1401										
C	294	242	192	155	851	1440	144	131	129	677	1081	259	175	174	926	1534
T	314											16	1208	139	265	1628
A	317	454	146	175	204	979	376	86	111	188	761	571	154	168	239	1132
A	329	726	128	75	114	1043	523	84	48	70	725	811	135	61	110	1117
T	351	55	516	221	61	853	41	382	132	55	610	43	570	170	74	857
T	382	109	355	212	57	733	53	300	131	45	529	102	439	210	78	829
C	418	96	107	190	996	1389	49	97	124	785	1055	86	157	169	1099	1511
A	425	1014	32	270	147	1463	808	28	163	97	1096					

Table 4: Potential variants called by MinoTour for the samples containing the rs2830088 (A) and rs2830051 (B) polymorphisms.

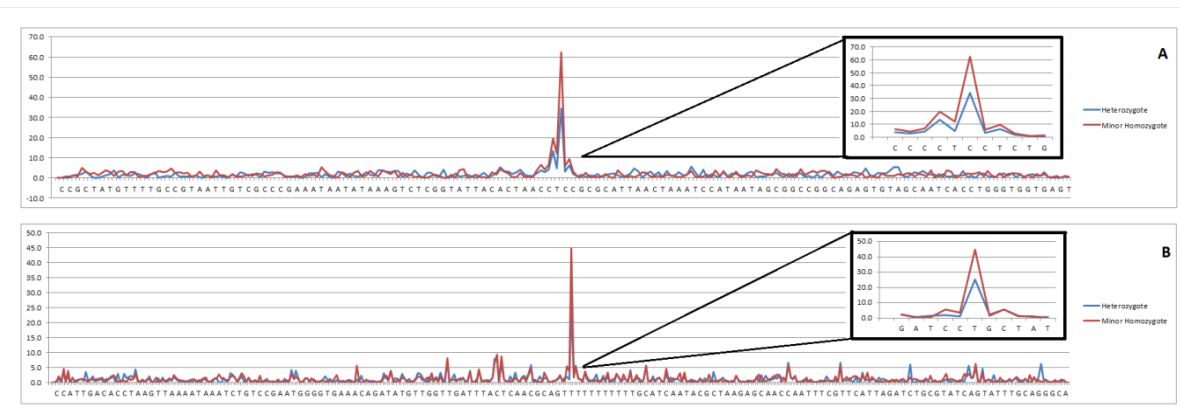


Figure 4: Graph of percentage difference for the proportion of reference base calls between major allele homozygotes and heterozygote (blue line) and minor allele homozygote (red line) samples. **A)** Percentage difference along the amplicon harbouring the rs2830088 SNP. Increases in percentage difference 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.

rs2830088 % of Minor Allele (T) reads		rs2830051 % of Minor Allele (C) reads
6.99	**WT (0%)**	1047
34.38	**Het (50%)**	26.76
60.79	**Mut (100%)**	46.23

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.

basecalls made by ONT's Metrichor server to allow users to visualise and analyse their sequence data. Like many of the programme designed to observed polymorphisms within sequences determined by nanopore technology, it aims to control the high basecalling error rate to see whether known polymorphisms could be detected in samples harbouring the variants when aligned to a reference sequence. The tool was unable to detect consensus variants for the deletion and heterozygous samples due to the high background error rate, a factor that has been identified in numerous studies [4,5]. Despite the identification of numerous potential variants including the polymorphisms of interest, against a background of false positives it was not possible to successfully detect variants using this method. However when given a minor allele homozygote sample the MinoTour Tool was able to detect the polymorphism singularly with its consensus algorithm, proving that algorithms design to align sequences to a reference can determine variants above background error rates in some instances.

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* (χ^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, Pavlenok 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, Shlien 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.