Int J Mol Epidemiol Genet 2012;3(4):262-275
www.ijmeg.org /ISSN:1948-1756/IJMEG1210001

Original Article

Next generation sequencing of *CLU, PICALM* and *CR1*: pitfalls and potential solutions

Jenny Lord[1], James Turton[1], Christopher Medway[1], Hui Shi[1], Kristelle Brown[1], James Lowe[2], David Mann[3], Stuart Pickering-Brown[3], Noor Kalsheker[1], Peter Passmore[4], Kevin Morgan[1]; Alzheimer's Research UK (ARUK) Consortium

[1]Human Genetics, School of Molecular Medical Sciences, Queens Medical Centre, University of Nottingham, Nottingham, UK; [2]Neuropathology, School of Molecular Medical Sciences, Queens Medical Centre, University of Nottingham, Nottingham, UK; [3]Clinical Neuroscience Research Group, Greater Manchester Neurosciences Centre, University of Manchester, Salford, UK; [4]Centre for Public Health, School of Medicine, Dentistry, and Biomedical Sciences, Queen's University Belfast, Belfast, Northern Ireland, UK

Received October 1, 2012; Accepted October 24, 2012; Epub November 15, 2012; Published November 30, 2012

Abstract: *CLU, PICALM* and *CR1* were identified as genetic risk factors for late onset Alzheimer's disease (AD) in two large genome wide association studies (GWAS) published in 2009, but the variants that convey this alteration in disease risk, and how the genes relate to AD pathology is yet to be discovered. A next generation sequencing (NGS) project was conducted targeting *CLU, CR1* and *PICALM,* in 96 AD samples (8 pools of 12), in an attempt to discover rare variants within these AD associated genes. Inclusion of repetitive regions in the design of the SureSelect capture lead to significant issues in alignment of the data, leading to poor specificity and a lower than expected depth of coverage. A strong positive correlation (0.964, p<0.001) was seen between NGS and 1000 genome project frequency estimates. Of the ~170 "novel" variants detected in the genes, seven SNPs, all of which were present in multiple sample pools, were selected for validation by Sanger sequencing. Two SNPs were successfully validated by this method, and shown to be genuine variants, while five failed validation. These spurious SNP calls occurred as a result of the presence of small indels and mononucleotide repeats, indicating such features should be regarded with caution, and validation via an independent method is important for NGS variant calls.

Keywords: Next generation sequencing, Alzheimer's disease, genes, *CLU*, *PICALM*, *CR1*

Introduction

Late onset Alzheimer's disease (AD) is a complex disorder with a strong genetic component, with heritability estimated to be up to 60-80% [1]. ApoE was the only robustly replicated genetic risk factor for AD until relatively recently [2], when the advent of the Genome Wide Association Study (GWAS) allowed researchers to test the vast majority of loci in the human genome for association with AD in a single experiment, without prior assumptions as to which might be involved. The publication of the first two sufficiently powered, large scale AD GWAS [3, 4] in 2009 identified *CLU*, *CR1* and *PICALM* as genetic risk factors for AD, and subsequent to this, another 6 genes (*BIN1, ABCA7,* the *MS4A* locus, *CD33, CD2AP* and *EPHA1* [5-7]) have been implicated and replicated through further GWAS and meta-analyses. However, it remains to be determined how these genes are involved in the pathology of AD, and which variants convey the alteration in disease risk.

In order to begin to address this issue in AD, as in numerous other disorders [8-10], extensive resources are being invested in Next Generation Sequencing (NGS) projects to discover rare, potentially causative variants at loci implicated by GWAS. Recent advances in technology have allowed scientists to generate sequence data to an unprecedented scale, with experiments that can produce millions of short sequencing reads in a single run. However, the data analysis methods needed to analyse and interpret the output from these experiments are still in their infancy, and at present there is no real "gold standard" for handling this data [11].

Various methods of target enrichment are available which allow researchers to hone in on specific genomic regions of interest [12]. Pooling of samples prior to enrichment and sequencing maximises the utility of NGS technologies, reducing the cost per sample dramatically. This enables far more subjects to be included in studies than would be feasible with individual sequencing or even indexed pooled capture. However, there is a trade off between increasing numbers of individuals within a pool, and the reliability of SNP calls. Increasing samples in a pool decreases coverage per sample, bringing the rate of a singleton SNP within a pool closer to the inherently high error rate of NGS technologies, whereas with individual sequencing, variants are present in approximately 50% of reads, at a given position. Indeed, when utilising a pooling strategy combining the DNA of 75 individuals, a validation success rate of around 20% was achieved for rare variants, compared to an almost 75% successful validation rate with pools of 12 (data unpublished). Estimating the frequencies of variants from pooled data is also reportedly unreliable, making comparisons between case and control allele frequencies for disease association testing problematic [13].

Repetitive DNA, which comprises around half of the human genome [14], and ranges from short stretches of mononucleotide repeats, to large segmental duplications, brings significant challenges for NGS projects. The main area where this presents an issue is in mapping short sequencing reads to the reference genome, since it can lead to reads having multiple potential alignment locations. Alignment programs may deal with the issue by reporting the best match only; by discarding all reads that map to multiple locations (or >n locations); or by reporting all potential alignment locations [14]. None of these methods is satisfactory, as data will be lost or aligned inaccurately, which may lead to erroneous variant calls in downstream analysis.

Given the current high error rates of NGS technologies and issues achieving accurate alignment, particularly around repetitive regions [14], validation of putative variants detected in NGS data via an independent method is important.

An NGS project was conducted targeting *CLU, CR1* and *PICALM,* in 96 AD patients, in an attempt to discover rare variants within these genes. A subset of SNPs from this NGS data were selected for validation via Sanger sequencing [15] to explore some of the potential issues mentioned above.

Materials and methods

NGS

All subjects gave informed consent to be included in the study, which was granted approval by the local Ethics Committee. 96 AD samples were obtained from two UK centres – The University of Nottingham Brain Bank and Manchester Brain Bank (48.4% female, 51.6% male; mean age at onset 70.4 years, standard deviation 11.77). ApoE alleles; ε2 - 5.7%; ε3 - 63.5%; ε4 - 30.8%). Both of these resources comprise part of the ARUK brain bank, which has been used in such projects as the GWAS that first implicated *CLU* and *PICALM* in AD risk [3]. This quantity of samples gave 80% power to detect variants with minor allele frequencies (MAF) down to 0.85%, based on the equation n=log(1-p)/log(1-MAF) where n is the number of chromosomes, p is power, and MAF is minor allele frequency. Whole genomic DNA was combined into 8 equimolar pools of 12 samples. Individual DNA samples were assessed for quality using agarose gel electrophoresis, with samples showing signs of degradation rejected from the study. The Invitrogen Quant-iT^{TM} dsDNA Broad Range Assay Kit (Life Technologies^{TM}, NY) was used to quantify the DNA concentrations to allow for accurate pooling. Enrichment of the genomic regions of interest (*CLU*, *PICALM* and *CR1,* totalling 292kb) was performed using Agilent's Sure Select XT Custom kit (Agilent Technologies, CA). SureSelect baits were designed to capture the whole gene (introns and exons), using 5x tiling, without repeat masking. Design of baits with and without the repeat masker made it possible to ascertain the proportion of each region that would not be targeted, were the repeat masker utilised; effectively the proportion of repetitive DNA in the region, although the masking method used is arguably over conservative. The region to be sequenced was also extended to include flanking regions demonstrating notable conservation across vertebrate species (areas >100bp showing at least 70% sequence identity between man, macaque, dog and mouse), up to a maximum of 5kb from the genes, using the ECR browser (http://ecrbrowser.dcode.org/) [16]. Sequencing was conducted on Illumina's GAIIx (Illumina Inc. CA), with one pool per lane on a flow cell producing 38bp single-end reads. Both of these processes were conducted by Source Bioscience (http://www.sourcebioscience.com/), following manufacturer's protocols.

Data analysis

Alignment of the sequence data was performed using BFAST v0.6.5a (Blat-like Fast Accurate Search Tool - http://bfast.sourceforge.net [17]), following the protocol in the program's manual with default settings for short single-end Illumina sequencing reads, to hg19 [18]. The quality of the alignment was assessed using SamStat [19]. SAMTools [20] v1.17 was used for basic file manipulation. Integrative Genomics Viewer [21] (IGV) v2.0 was used to visualise the aligned sequencing reads.

CRISP [22] v5, a program specifically designed for variant calling in NGS data from pooled samples, was used to identify SNPs and small indels, following the program's manual, with default settings.

Basic annotation of the polymorphisms called by CRISP was conducted using Ensembl's Variant Effect Predictor [23] (VEP, accessed November 2011), which provided information on where the variants lie in relation to the major transcripts of each gene and whether these were novel or had been documented in dbSNP.

Only variants which did not have known co-located variants (as determined by the VEP) were taken forward to further analysis. To minimise the chance of pursuing false positive variants, an additional filtering step was applied: variants which featured in less than 4% of the total number of reads in the pool in which they occurred were disregarded, since on average a singleton variant in a pool of 12 individuals (24 chromosomes) would be expected to be present in 4.2% of the reads.

The Genome Analysis Tool Kit (GATK) [24] v1.1.10 was used to perform local realignment around known indels and base quality score recalibration [25], using variant data from dbSNP 134, on the complete data for each of the genes. Both processes were run following the default protocol documented on the GATK online user guides. CRISP was then re-run on the realigned, recalibrated data.

Sanger validation

PCR primers were designed to amplify a region including at least 100bp either side of the position of interest (SNPs listed in **Table 1**) using Primer3 [26] v0.4.0 (http://frodo.wi.mit.edu/). Specificity for each primer pair was checked using UCSC's [27]; Virtual PCR function (http://genome.ucsc.edu/cgi-bin/hgPcr), and the primer binding sites were determined to be free of known polymorphisms using NGRL Manchester's SNPCheck v2.1 (https://ngrl.manchester.ac.uk/SNPCheckV2/snpcheck.htm). PCR optimisation and amplifications were completed following standard laboratory protocol (reaction mix: 1xPCR buffer (Roche Diagnostics Corp., IN); 200μM dNTPs (Fermentas); 1μM of each primer (Eurogentec Biologics, Belgium); 1 unit Taq DNA Polymerase (Roche, Diagnostics Corp., IN); plus molecular grade water up to a final volume of 30μl. Primer concentrations were halved for 8:27466924 and doubled for 1:207690803 after optimisation. Thermal cycling conditions used were 94°C for two minutes; 30 cycles of 94°C for 30 seconds, appropriate annealing temperature for 1 minute, 72°C for 1 minute; and finally 72°C for 7 minutes). Primer sequences and annealing temperatures are shown in Supplementary **Table 1**. Sequencing was conducted using PCR primers with Applied Biosystems BigDye Terminator v3.1 chemistry, run on the ABI 3130xl (Applied Biosystems, CA). Chromas Lite v2.01 (http://www.technelysium.com.au/chromas_lite.html) was used to visualise electropherograms which were assessed by eye to determine genotype. In each case, one pool of samples (12 individuals) was Sanger sequenced. The pool to be sequenced was selected based on having the highest proportion of alternative reads at the position of interest.

Tabix [28] was used to obtain variant information from 1000 genomes [29] release 20110521 and an in-house compiled Perl script was used to extract information of interest (variants and European population frequen-

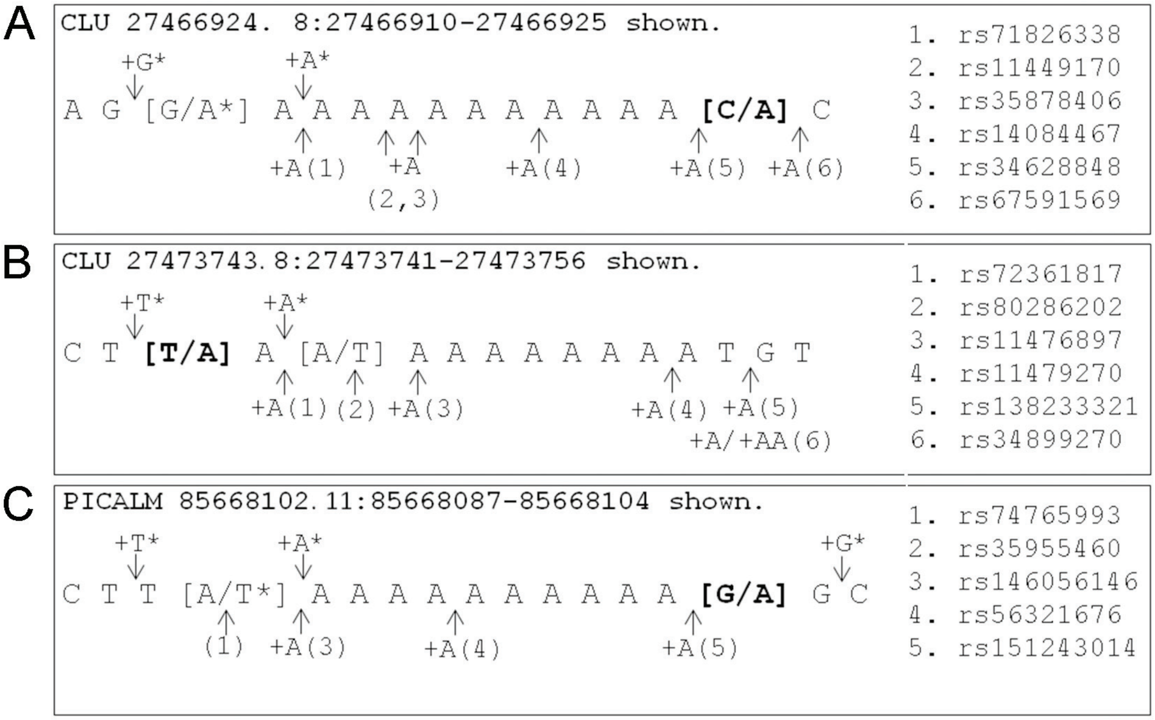

Figure 1. Sequence context of spurious SNPs called next to mononucleotide repeats, which failed validation by Sanger sequencing. The SNP shown in bold is the variant Sanger sequencing was designed to validate. Variants shown below the sequence are all present in dbSNP. A: False C/A variant call within *CLU* at position 8:27466924 (8:27466910-27466925 shown). B: Spurious T/A SNP call in *CLU* at 8:27473743 (8:27473741-27473756 shown). C: False positive SNP call at position 11:85668102 within *PICALM* (11:85668090-85668104 shown). *Other variants called by CRISP.

cy data) from this. SPSS v16 was used to calculate correlation coefficients between CRISP and 1000 genomes frequency data.

Results

Following alignment with BFAST of the ~350 million reads obtained from the NGS run, the average coverage per individual across the three genes was 17.4x (18.1x, 21.5x and 13.9x for *CLU, PICALM* and *CR1* respectively). Sequencing characteristics from the experiment are shown in **Table 2**.

The number of variants detected by CRISP in each of the genes is summarised in **Table 3**, both before and after realignment and recalibration of the data with GATK. The table also shows the number of common (MAF >5%) SNPs listed within dbSNP 134 in the targeted regions, and how many of these were detected within our data (in each case this remained the same pre- and post- GATK). As a positive control, the CRISP output was compared with variants which had a MAF between 0.85-1% in the 1000 genomes project EUR population data (i.e. the minimum MAF the study had 80% power to find). Six of the thirteen variants received particularly poor coverage in our data (<10x per individual). Of the remaining seven, four were successfully identified in our data, demonstrating the capability of this method to detect variants in this range of MAF. The final three appeared to be non-variant sites in our 96 samples.

In order to ascertain the accuracy of the frequency estimates from CRISP, MAF estimates (based on percentage of alternative reads – a surrogate for MAF) from CRISP for all of the SNPs called in the three genes were compared with MAFs from the 1000 genomes project (European population). A strong, significant positive correlation was observed between datasets (Spearman Correlation Coefficient=0.964, p=<0.001).

The seven SNPs selected for Sanger Validation were chosen on the basis of having no co-located variant (as determined by Ensembl's VEP)

Table 1. Information on SNPs for sanger sequencing validation

Gene	Chr	Coordinate	Alleles	Frequency Alternative Calls	Fold Coverage per individual	Location in gene*	rs number	Inclusion in project
CLU	8	27452179	G/T	0.07	23.69	3.5kb downstream		Targeted
	8	27452243	A/T	0.05	24.13	3.5kb downstream		Incidental
	8	27466924	C/A	0.11	17.89	Intron 2		Targeted
	8	27473743	T/A	0.19	18.11	1.5kb upstream		Targeted
PICALM	11	85668102	G/A	0.11	16.74	1.5kb downstream		Targeted
	11	85668163	G/A	0.27	19.11	1.5kb downstream	rs622110	Incidental
	11	85692077	C/T	0.05	22.90	Intron 18	rs139710547	Targeted
	11	85692181	A/C	0.63	18.46	Exon 18 (synonymous)	rs76719109	Incidental
	11	85774424	T/G	0.24	16.41	Intron 2		Targeted
	11	85774562	T/G	0.46	18.55	Intron 2	rs3016786	Incidental
CR1	1	207690803	T/C	0.07	19.57	Intron 4	rs144047769	Targeted
	1	207690871	G/C	0.19	18.40	Intron 4	rs10863358	Incidental

Information on all of the SNPs for which validation by Sanger sequencing was attempted. Coordinates stated give genomic position in hg19. *Relative to CLU transcript ENST0000031-6403, PICALM transcript ENST00000447890, CR1 transcript ENST00000367049. Distances stated are approximate.

Table 2. Sequence characteristics

Gene	Size of region targeted	Reads mapped to region (all pools)	% Reads mapped with quality >30*	Average coverage	% Target region >10x coverage per individual	% Target region >20x coverage per individual	% Repetitive DNA within targeted region
CLU	24.4kb	1115695	94.7	18.1	72.3	27.1	34
PICALM	118.3kb	6437215	93.7	21.5	82.5	49.0	34
CR1	149.2kb	5252304	67.6	13.9	53.7	29.4	48

Details of sequence characteristics from the NGS project, including information on the targeted region (region size, and the proportion of that region that would not have been targeted if the repeat masker was utilised in the experimental design); the number and quality of reads mapped; average coverage; and the proportion of the region with >10x and >20x coverage per individual. *>30 is the maximum quality score category using the SamStat tool.

Table 3. Variants detected in NGS data

Gene	Total Variants	Total SNPs (pre-GATK)	Total SNPs (post-GATK)	Total Indels (pre-GATK)	Total Indels (post-GATK)	Novel Variants*	Common (>5%) dbSNPs found (out of)
CLU	99	79	76	20	25	17	20 (20)
PICALM	541	428	414	113	116	100	117 (117)
CR1	291	253	248	38	48	34	70 (75)

Number of variants (SNPs and indels) called from the NGS data by CRISP in each of the genes, the effect GATK had on these numbers, and how many common SNPs documented within the gene regions were found in our dataset. *As determined by Ensembl's VEP.

Table 4. Validated SNPs

Chr	Coordinate	Alleles	Alternative Allele Call Frequency (CRISP - all pools)	1000 genomes MAF (EUR)	Alternative Allele Call Frequency (CRISP - sequenced pool)	Alternative Alleles (Sanger - sequenced pool)	Alternative Allele Frequency (Sanger – sequenced pool)
11	85668163	G/A	0.268	0.244	0.707	10	0.417
11	85692077	C/T	0.046	0.016	0.103	3	0.125
11	85692181	A/C	0.632	0.583	0.929	19	0.792
11	85774562	T/G	0.460	0.422	0.362	10	0.417
1	207690803	T/C	0.067	0.021	0.333	3	0.125
1	207690871	G/C	0.194	0.214	0.050	4	0.167

SNPs which were successfully validated by Sanger sequencing. The number of alternative alleles within the pool sequenced facilitated the determination of the genuine MAF within that pool (assuming Sanger results reflect true allelic counts). This was then compared to the alternative allele call frequency from the same pool in CRISP, giving a reflection of the accuracy of the CRISP frequency estimates at a much finer level than the total 96 samples allows.

Table 5. Validated indels, miscalled by CRISP as SNPs

Chr	Coordinate	Alleles	Actual Variant	Coordinate of indel	Alternative Allele Call Frequency (CRISP - all pools)	1000 genomes frequency (EUR)	rs number	Alternative Allele Call Frequency (CRISP - sequenced pool)	Alternative Alleles (Sanger - sequenced pool)	Alternative Allele Frequency (Sanger – sequenced pool)
8	27452179	G/T	T ins	27452180	0.145	0.26	rs146954978	0.232	3	0.125
8	27452243	A/T	T ins	27452242	0.099	0.17	rs35598594	0.115	1	0.042
11	85774424	T/G	TA del	85774420	0.42	Not available	rs112671434	0.455	17	0.708

Details of the indels discovered at the sites of spurious SNP calls, all of which had been previously recorded in dbSNP.

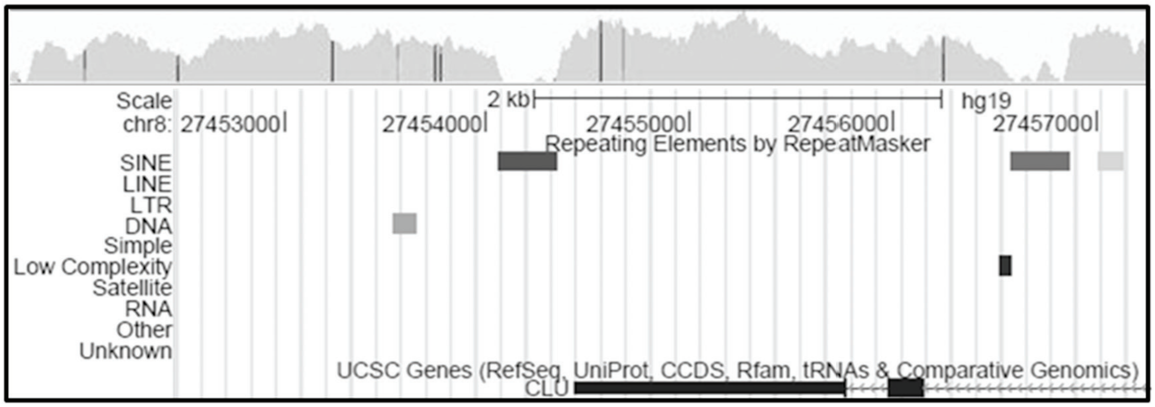

Figure 2. Depth of coverage at repetitive regions. Combined images from IGV and UCSC (http://genome.ucsc.edu) to show the drop off in coverage of NGS reads at repetitive regions in the genome. The gray graph in the top panel is from IGV. The height of the graph is proportional to the depth of coverage at a given genomic position. The gray bars in the panel below are taken from the Repeat Masker track of UCSC and show the location of repetitive DNA; with the depth of colour being proportional to the strength of the repeat sequence. The list on the left hand side explains the class of repetitive element present. The region displayed is from the 3' end of *CLU*, but is representative of the data as a whole. Vertical coloured bars within the IGV image indicate variant sites within the data, where not all base calls match the reference.

and having the alternate allele occur in >5% of sequencing reads at that position. Given the apparent commonality of these variants in our data, the fact that they had not been documented prior to the extensive resequencing efforts of the 1000 genomes project, if at all, seemed worthy of investigation. Other variants called by CRISP which fell within the regions sequenced were also considered in the analysis. Information on all of the SNPs Sanger sequenced is presented in **Table 1**.

Of the 12 putative SNPs included in Sanger sequenced regions, six were found to be genuine (**Table 4**). Reliability of frequency estimates could also be assessed once the number of actual alternative alleles in a pool was established by Sanger sequencing. Each alternative allele should contribute ~4.2% of reads to the pool total, assuming equal representation. The relationship between the number of actual alternative alleles and the proportion of NGS reads they make up is shown in **Table 4**.

Three of the remaining SNPs were not found to be present in the samples Sanger sequenced, but instead small indels were found at the suggested variant sites (all which were also called by CRISP). These variants are summarised in **Table 5**.

The final three SNPs (8:27466924, 8:27473743 and 11:85668102) were not validated by Sanger Sequencing. All three of these putative variants occurred adjacent to mononucleotide polyA repeats, with numerous other potential variants in the immediate area, called by CRISP or present in dbSNP 134 (see **Figure 1**). When CRISP was run on GATK realigned and recalibrated data, the 8:27466924 and 11:85668102 SNPs were no longer called, and 8:27473743 went from being called as present in all pools to only being present in two (including the pool Sanger Sequenced). Within each of the polyA repeats, CRISP also called a +A insertion, all of which persisted following GATK realignment and recalibration. There are also multiple +A insertions reported in dbSNP for each of the mononucleotide repeat sites, which suggests these may be genuinely variant. Due to the issues even Sanger sequencing has in dealing with mononucleotide repeats, our findings were inconclusive as to whether there were genuine variations in the number of A nucleotides present at these sites.

Discussion

NGS data

The percentage of reads which mapped to the target region was lower than expected, resulting in lower than anticipated coverage of the targeted areas. SureSelect has been reported to return 40-50% [30] of reads on target, although with the custom kits it can be lower

[31]. The figures for average coverage of the three genes are actually deceptively low – repetitive regions within the genes tended to have few reads aligned to them, lowering the calculated coverage. An optional repeat masker (based on RepBase [32] v9.11) could have been used when designing the SureSelect baits, which would have prevented strongly repetitive regions from being targeted. However, this was not used in the design of this experiment as it was deemed overly conservative (masking out 34%, 34% and 48% of *CLU, PICALM* and *CR1* respectively). However, when images from IGV and UCSC are compiled, as in **Figure 2** [18, 21, 27], it can be seen that the regions that would have been repeat masked and therefore not baited, generally had extremely low coverage, so no real information was gained by targeting these regions. Additionally, allowing these repetitive regions to be included in the bait design will have reduced the specificity of the capture, as non-target, similar DNA will also have been pulled down, reducing the recovery of the true target region, and limiting the number of reads which could be uniquely mapped.

For CR1, around 40kb of the ~150kb region targeted received effectively no coverage, reducing the average for the whole region. This gap in the sequencing is as a result of the nature of the *CR1* gene. *CR1* encodes complement receptor 1, a membrane glycoprotein and the main receptor for complement proteins C3b and C4b. There are four different CR1 isoforms, encoded by CR1-A (the F allele), CR1-B (the S-allele), CR1-C and CR1-D, with allele frequencies in Caucasian populations of 0.83, 0.15, 0.01 and <0.01 [33] respectively. The proteins differ in the number of C3b binding sites present, and the different alleles are thought to have arisen from unequal crossover events involving a stretch of DNA encoding this. CR1-C which encodes the smallest isoform has a single C3b binding site, and thus a single copy of this stretch of DNA, but alleles encoding the larger isoforms with multiple C3b binding sites will have multiple copies of this region, making accurate alignment of reads virtually impossible. A recent study [34] which used multiplex amplicon quantification to distinguish F- and S-alleles found an association between the S-allele (with an extra C3b binding site) and increased AD risk, however, with our methodology it was not possible to determine genotypes for this polymorphism within sample pools, let alone within individual samples.

When repetitive DNA comprises such a large proportion (~50% [14]) of the human genome, these regions cannot simply be ignored, but nor can they be accurately sequenced, given the current technology and data analysis methodologies available. The *CR1* example above demonstrates that repetitive DNA can be biologically important and disease relevant, but whether this is a common phenomenon or an isolated example remains to be seen.

This study had 80% power to detect variants with a MAF down to 0.85%. This figure is a higher frequency than the multiple individual variants Bettens *et al*. found in their study, however significantly more samples (several thousand) were included in their resequencing approach targeting *CLU* exons [35]. The study here documented used less samples, so would not be expected to find so many extremely rare variants, but has strengths in its coverage of full exonic, intronic and potential regulatory regions. That said, CRISP did detect several variants (rs185685560, rs188050008, rs186928661 and rs1834226) which all have MAFs as low as 0.13% in the 1000 genomes EUR population data. Of the 13 variants with MAFs from 0.85-1% in the 1000 genomes data for these genes, nine were not identified in our samples. Although it is likely some of these rare variants were simply not present in our sample set, some were likely missed due to poor coverage of the regions. This limitation arose through the inclusion of repetitive regions in the study design, creating issues during alignment, which could have been improved via the use of paired-end rather than single-end reads.

Sanger validation

Overall, the total number of variants called by CRISP in the three genes was 761 SNPs and 171 indels, of which over 150 were without co-located variants. All of the SNPs listed in dbSNP 134 with CEU frequencies >5% within the targeted regions were found in our data (see **Table 3**), with the exception of five in *CR1*, likely due to the coverage issues mentioned above. This indicates a low false negative rate, for common SNPs at least. The strong positive correlation between 1000 genomes and CRISP frequency estimates for all CRISP called SNPs indicates that even from pooled data, frequency estimates can be considered reasonably accurate, although it should also be noted that as only AD patients were included in this sequencing project, deviance from a perfect correlation could be indicative of disease association. A recent paper [13] commented that frequency estimates from pooled samples were not generally accurate enough to use in meaningful association testing between case and control groups, however, the strong correlation between CRISP and 1000 genomes frequencies in our data suggests otherwise.

However, comparison of the actual number of alleles within a pool and the CRISP frequency does not support this. Sanger sequencing allowed the determination of the exact number of alternative alleles within a pool for the validated SNPs. Frequency estimates from pooled data are based on the assumption that each allele contributes equally to the total number of reads. If this assumption is incorrect, frequency estimations will not be accurate. For the majority of the variants considered in this study, there is a discrepancy between the number of alternative alleles within the Sanger sequenced pools, and the frequency estimation from CRISP, which could indicate the assumption is invalid and alleles are not equally represented. This could occur as a result of inaccurate pooling, resulting in DNA from certain subjects being over or under represented. Alternatively, it may reflect inherent biases in the target enrichment or NGS processes if DNA from certain individuals is captured or sequenced to a lesser extent. The more samples included in a study, the more accurate estimations of frequency will become [36], so when the full 96 are considered, frequency estimates improve, but this does not mean samples are being equally represented, which should be acknowledged during analysis. It is possible that while keeping a small number of samples per pool ensures accurate variant calls, it actually reduces the reliability of frequency estimates for those pools.

With so many potential variants arising from even relatively small scale resequencing projects, it is crucial to minimise the amount of false positive variants, hence why validation via an independent methodology is important. Half of the putative SNPs in the Sanger sequenced regions were validated as being genuine SNPs, but this included only two which were deliberately targeted for validation, and almost all of the variants which were incidentally sequenced as they fell within the amplicons to be sequenced. Bansal *et al.* [37] estimated a false positive rate of <1% using CRISP, which is significantly lower than our 50%. However, our rate would be expected to be higher than this, since many of the SNPs selected for validation were in possession of unusual characteristics (e.g. being present in all pools sequenced).

The location of these variants relative to the major transcripts of their respective genes is given in **Table 1**. The majority are deeply intronic, so are unlikely to be affecting splicing activity. One variant (11:85668163) falls around 1.5kb downstream of *PICALM*, where it is not likely to be affecting the gene's function or regulation. None of the variants show a high degree of conservation. The other SNP, at 11:85692181, is a synonymous exonic change, which whilst not affecting the primary sequence of the protein, could be having an effect on splicing regulatory elements or mRNA structure. Functional studies would be needed to clarify the effects of any of these variants.

The remainder of the potential SNPs sequenced did not turn out to be genuine. These constitute false positives. However, three of these transpired to be next to genuinely variant sites of small indels; two +T insertions and a –TA deletion, all of which had been previously documented and were identified by CRISP.

The final three putative SNPs which were not validated were each next to a string of polyA mononucleotide repeats. These repeats are too small for the issue to be solved by masking repeats in the design of SureSelect baits. Unlike longer repeats, short stretches of mononucleotide repeats do not make alignment of reads impossible, as enough unique sequence is present, even in short 38bp reads to allow mapping. It is likely that these are genuinely variant sites, since CRISP calls +A insertions within each of them, and all have multiple rs number +A insertions falling within the repeat region. However, the problems presented by repetitive DNA even to the relatively robust Sanger sequencing meant it was not possible to tell whether these sites were polymorphic in our samples, or whether apparent variance was simply an artefact of slippage during Sanger sequencing or PCR amplification [38].

It was hoped that running CRISP on data which had been realigned around indels and had had base quality scores recalibrated using GATK would reduce these false positive SNP calls, and for two of the variants next to mononucleotide repeats this was the case, but not for the other one, or for the three spurious "SNPs" called next to indel sites. From this, it can be said that GATK is useful in clearing up some false positive calls, but certainly does not work for all, and a more reliable and less time consuming approach to avoiding these false positives might be to be aware of documented indels and mononucleotide repeats which could cause potential issues, and regard any SNPs called close to these sites with caution.

Acknowledgements

We thank the Alzheimer's Research UK (ARUK) consortium for collection of the samples used, and to the patients and families whose participation made this work possible. We also thank ARUK and the Big Lottery Fund for financial support of this project, and ARUK for funding the PhD studentship of Jenny Lord.

The Alzheimer's Research UK consortium: Peter Passmore, Bernadette McGuinness, Janet Johnston, Stephen Todd, Queen's University Belfast, UK; Reinhard Heun (now at Royal Derby Hospital), Heike Kölsch, University of Bonn, Germany; Patrick G. Kehoe, University of Bristol, UK; Nigel M. Hooper, Emma R.L.C. Vardy (now at University of Newcastle), University of Leeds, UK; David M. Mann, University of Manchester, UK; Kristelle Brown, Noor Kalsheker, Kevin Morgan, University of Nottingham, UK; A. David Smith, Gordon Wilcock, Donald Warden, University of Oxford (OPTIMA), UK; Clive Holmes, University of Southampton, UK.

Disclosure statement

There are no actual or potential conflicts of interest to disclose.

Address correspondence to: Dr. Kevin Morgan, Human Genomics and Molecular Genetics, Institute of Genetics, School of Molecular Medical Sciences, A Floor, West Block, Room 1306, Queens Medical Centre, Nottingham NG7 2UH, United Kingdom. Tel: 0115 8230724; E-mail: kevin.morgan@nottingham.ac.uk

References

[1] Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, Fiske A and Pedersen NL. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry 2006; 63: 168-174.

[2] Pericak-Vance MA, Bebout JL, Gaskell PC Jr, Yamaoka LH, Hung WY, Alberts MJ, Walker AP, Bartlett RJ, Haynes CA, Welsh KA and et al. Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. Am J Hum Genet 1991; 48: 1034-1050.

[3] Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, Pahwa JS, Moskvina V, Dowzell K, Williams A, Jones N, Thomas C, Stretton A, Morgan AR, Lovestone S, Powell J, Proitsi P, Lupton MK, Brayne C, Rubinsztein DC, Gill M, Lawlor B, Lynch A, Morgan K, Brown KS, Passmore PA, Craig D, McGuinness B, Todd S, Holmes C, Mann D, Smith AD, Love S, Kehoe PG, Hardy J, Mead S, Fox N, Rossor M, Collinge J, Maier W, Jessen F, Schurmann B, van den Bussche H, Heuser I, Kornhuber J, Wiltfang J, Dichgans M, Frolich L, Hampel H, Hull M, Rujescu D, Goate AM, Kauwe JS, Cruchaga C, Nowotny P, Morris JC, Mayo K, Sleegers K, Bettens K, Engelborghs S, De Deyn PP, Van Broeckhoven C, Livingston G, Bass NJ, Gurling H, McQuillin A, Gwilliam R, Deloukas P, Al-Chalabi A, Shaw CE, Tsolaki M, Singleton AB, Guerreiro R, Muhleisen TW, Nothen MM, Moebus S, Jockel KH, Klopp N, Wichmann HE, Carrasquillo MM, Pankratz VS, Younkin SG, Holmans PA, O'Donovan M, Owen MJ and Williams J. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet 2009; 41: 1088-1093.

[4] Lambert JC, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, Combarros O, Zelenika D, Bullido MJ, Tavernier B, Letenneur L, Bettens K, Berr C, Pasquier F, Fievet N, Barberger-Gateau P, Engelborghs S, De Deyn P, Mateo I, Franck A, Helisalmi S, Porcellini E, Hanon O, de Pancorbo MM, Lendon C, Dufouil C, Jaillard C, Leveillard T, Alvarez V, Bosco P, Mancuso M, Panza F, Nacmias B, Bossu P, Piccardi P, Annoni G, Seripa D, Galimberti D, Hannequin D, Licastro F, Soininen H, Ritchie K, Blanche H, Dartigues JF, Tzourio C, Gut I, Van Broeckhoven C, Alperovitch A, Lathrop M and Amouyel P. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet 2009; 41: 1094-1099.

[5] Seshadri S, Fitzpatrick AL, Ikram MA, DeStefano AL, Gudnason V, Boada M, Bis JC, Smith AV, Carassquillo MM, Lambert JC, Harold D, Schrijvers EM, Ramirez-Lorca R, Debette S, Longstreth WT Jr, Janssens AC, Pankratz VS, Dartigues JF, Hollingworth P, Aspelund T, Hernandez I, Beiser A, Kuller LH, Koudstaal PJ, Dickson DW, Tzourio C, Abraham R, Antunez C, Du Y, Rotter JI, Aulchenko YS, Harris TB, Petersen RC, Berr C, Owen MJ, Lopez-Arrieta J, Varadarajan BN, Becker JT, Rivadeneira F, Nalls MA, Graff-Radford NR, Campion D, Auerbach S, Rice K, Hofman A, Jonsson PV, Schmidt H, Lathrop M, Mosley TH, Au R, Psaty BM, Uitterlinden AG, Farrer LA, Lumley T, Ruiz A, Williams J, Amouyel P, Younkin SG, Wolf PA, Launer LJ, Lopez OL, van Duijn CM and Breteler MM. Genome-wide analysis of genetic loci associated with Alzheimer disease. JAMA 2010; 303: 1832-1840.

[6] Hollingworth P, Harold D, Sims R, Gerrish A, Lambert JC, Carrasquillo MM, Abraham R, Hamshere ML, Pahwa JS, Moskvina V, Dowzell K, Jones N, Stretton A, Thomas C, Richards A, Ivanov D, Widdowson C, Chapman J, Lovestone S, Powell J, Proitsi P, Lupton MK, Brayne C, Rubinsztein DC, Gill M, Lawlor B, Lynch A, Brown KS, Passmore PA, Craig D, McGuinness B, Todd S, Holmes C, Mann D, Smith AD, Beaumont H, Warden D, Wilcock G, Love S, Kehoe PG, Hooper NM, Vardy ER, Hardy J, Mead S, Fox NC, Rossor M, Collinge J, Maier W, Jessen F, Ruther E, Schurmann B, Heun R, Kolsch H, van den Bussche H, Heuser I, Kornhuber J, Wiltfang J, Dichgans M, Frolich L, Hampel H, Gallacher J, Hull M, Rujescu D, Giegling I, Goate AM, Kauwe JS, Cruchaga C, Nowotny P, Morris JC, Mayo K, Sleegers K, Bettens K, Engelborghs S, De Deyn PP, Van Broeckhoven C, Livingston G, Bass NJ, Gurling H, McQuillin A, Gwilliam R, Deloukas P, Al-Chalabi A, Shaw CE, Tsolaki M, Singleton AB, Guerreiro R, Muhleisen TW, Nothen MM, Moebus S, Jockel KH, Klopp N, Wichmann HE, Pankratz VS, Sando SB, Aasly JO, Barcikowska M, Wszolek ZK, Dickson DW, Graff-Radford NR, Petersen RC, van Duijn CM, Breteler MM, Ikram MA, Destefano AL, Fitzpatrick AL, Lopez O, Launer LJ, Seshadri S, Berr C, Campion D, Epelbaum J, Dartigues JF, Tzourio C, Alperovitch A, Lathrop M, Feulner TM, Friedrich P, Riehle C, Krawczak M, Schreiber S, Mayhaus M, Nicolhaus S, Wagenpfeil S, Steinberg S, Stefansson H, Stefansson K, Snaedal J, Bjornsson S, Jonsson PV, Chouraki V, Genier-Boley B, Hiltunen M, Soininen H, Combarros O, Zelenika D, Delepine M, Bullido MJ, Pasquier F, Mateo I, Frank-Garcia A, Porcellini E, Hanon O, Coto E, Alvarez V, Bosco P, Siciliano G, Mancuso M, Panza F, Solfrizzi V, Nacmias B, Sorbi S, Bossu P, Piccardi P, Arosio B, Annoni G, Seripa D, Pilotto A, Scarpini E, Galimberti D, Brice A, Hannequin D, Licastro F, Jones L, Holmans PA, Jonsson T, Riemenschneider M, Morgan K, Younkin SG, Owen MJ, O'Donovan M, Amouyel P and Williams J. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat Genet 2011; 43: 429-435.

[7] Naj AC, Jun G, Beecham GW, Wang LS, Vardarajan BN, Buros J, Gallins PJ, Buxbaum JD, Jarvik GP, Crane PK, Larson EB, Bird TD, Boeve BF, Graff-Radford NR, De Jager PL, Evans D, Schneider JA, Carrasquillo MM, Ertekin-Taner N, Younkin SG, Cruchaga C, Kauwe JS, Nowotny P, Kramer P, Hardy J, Huentelman MJ, Myers AJ, Barmada MM, Demirci FY, Baldwin CT, Green RC, Rogaeva E, George-Hyslop PS, Arnold SE, Barber R, Beach T, Bigio EH, Bowen JD, Boxer A, Burke JR, Cairns NJ, Carlson CS, Carney RM, Carroll SL, Chui HC, Clark DG, Corneveaux J, Cotman CW, Cummings JL, Decarli C, Dekosky ST, Diaz-Arrastia R, Dick M, Dickson DW, Ellis WG, Faber KM, Fallon KB, Farlow MR, Ferris S, Frosch MP, Galasko DR, Ganguli M, Gearing M, Geschwind DH, Ghetti B, Gilbert JR, Gilman S, Giordani B, Glass JD, Growdon JH, Hamilton RL, Harrell LE, Head E, Honig LS, Hulette CM, Hyman BT, Jicha GA, Jin LW, Johnson N, Karlawish J, Karydas A, Kaye JA, Kim R, Koo EH, Kowall NW, Lah JJ, Levey AI, Lieberman AP, Lopez OL, Mack WJ, Marson DC, Martiniuk F, Mash DC, Masliah E, McCormick WC, McCurry SM, McDavid AN, McKee AC, Mesulam M, Miller BL, Miller CA, Miller JW, Parisi JE, Perl DP, Peskind E, Petersen RC, Poon WW, Quinn JF, Rajbhandary RA, Raskind M, Reisberg B, Ringman JM, Roberson ED, Rosenberg RN, Sano M, Schneider LS, Seeley W, Shelanski ML, Slifer MA, Smith CD, Sonnen JA, Spina S, Stern RA, Tanzi RE, Trojanowski JQ, Troncoso JC, Van Deerlin VM, Vinters HV, Vonsattel JP, Weintraub S, Welsh-Bohmer KA, Williamson J, Woltjer RL, Cantwell LB, Dombroski BA, Beekly D, Lunetta KL, Martin ER, Kamboh MI, Saykin AJ, Reiman EM, Bennett DA, Morris JC, Montine TJ, Goate AM, Blacker D, Tsuang DW, Hakonarson H, Kukull WA, Foroud TM, Haines JL, Mayeux R, Pericak-Vance MA, Farrer LA and Schellenberg GD. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet 2011; 43: 436-441.

[8] Nejentsev S, Walker N, Riches D, Egholm M and Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 2009; 324: 387-389.

[9] Johansen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, Ban MR, Martins RA, Kennedy BA, Hassell RG, Visser ME, Schwartz SM, Voight BF, Elosua R, Salomaa V, O'Donnell CJ, Dallinga-Thie GM, Anand SS, Yusuf S, Huff MW, Kathiresan S and Hegele RA. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nat Genet 2010; 42: 684-687.

[10] Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, Boucher G, Ripke S, Ellinghaus D, Burtt N, Fennell T, Kirby A, Latiano A, Goyette P, Green T, Halfvarson J, Haritunians T, Korn JM, Kuruvilla F, Lagace C, Neale B, Lo KS, Schumm P, Torkvist L, Dubinsky MC, Brant SR, Silverberg MS, Duerr RH, Altshuler D, Gabriel S, Lettre G, Franke A, D'Amato M, McGovern DP, Cho JH, Rioux JD, Xavier RJ and Daly MJ. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet 2011; 43: 1066-1073.

[11] Nielsen R, Paul JS, Albrechtsen A and Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 2011; 12: 443-451.

[12] Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J and Turner DJ. Target-enrichment strategies for next-generation sequencing. Nat Methods 2010; 7: 111-118.

[13] Day-Williams AG, McLay K, Drury E, Edkins S, Coffey AJ, Palotie A and Zeggini E. An evaluation of different target enrichment methods in pooled sequencing designs for complex disease association studies. PLoS ONE 2011; 6: e26279.

[14] Treangen TJ and Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 2012; 13: 36-46.

[15] Sanger F, Nicklen S and Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977; 74: 5463-5467.

[16] Ovcharenko I, Nobrega MA, Loots GG and Stubbs L. ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. Nucleic Acids Res 2004; 32: W280-286.

[17] Homer N, Merriman B and Nelson SF. BFAST: an alignment tool for large scale genome resequencing. PLoS ONE 2009; 4: e7767.

[18] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 2001; 409: 860-921.

[19] Lassmann T, Hayashizaki Y and Daub CO. SAMStat: monitoring biases in next generation sequencing data. Bioinformatics 2011; 27: 130-131.

[20] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25: 2078-2079.

[21] Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G and Mesirov JP. Integrative genomics viewer. Nat Biotechnol 2011; 29: 24-26.

[22] Bansal V. A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics 2010; 26: i318-324.

[23] McLaren W, Pritchard B, Rios D, Chen Y, Flicek P and Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 2010; 26: 2069-2070.

[24] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M and DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010; 20: 1297-1303.

[25] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D and Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011; 43: 491-498.

[26] Rozen S and Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 2000; 132: 365-386.

[27] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM and Haussler D. The human genome browser at UCSC. Genome Res 2002; 12: 996-1006.

[28] Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 2011; 27: 718-719.

[29] Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, De La Vega FM, Donnelly P, Egholm M, Flicek P, Gabriel SB, Gibbs RA, Knoppers BM, Lander ES, Lehrach H, Mardis ER, McVean GA, Nickerson DA, Peltonen L, Schafer AJ, Sherry ST, Wang J, Wilson R, Gibbs RA, Deiros D, Metzker M, Muzny D, Reid J, Wheeler D, Wang J, Li J, Jian M, Li G, Li R, Liang H, Tian G, Wang B, Wang J, Wang W, Yang H, Zhang X, Zheng H, Lander ES, Altshuler DL, Ambrogio L, Bloom T, Cibulskis K, Fennell TJ, Gabriel SB, Jaffe DB, Shefler E, Sougnez CL, Bentley DR, Gormley N, Humphray S, Kingsbury Z, Kokko-Gonzales P, Stone J, McKernan KJ, Costa GL, Ichikawa JK, Lee CC, Sudbrak R, Lehrach H, Borodina TA, Dahl A, Davydov AN, Marquardt P, Mertes F, Nietfeld W, Rosenstiel P, Schreiber S, Soldatov AV, Timmermann B, Tolzmann M, Egholm M, Affourtit J, Ashworth D, Attiya S, Bachorski M, Buglione E, Burke A, Caprio A, Celone C, Clark S, Conners D, Desany B, Gu L, Guccione L, Kao K, Kebbler J, Knowlton J, Labrecque M, McDade L, Mealmaker C, Minderman M, Nawrocki A, Niazi F, Pareja K, Ramenani R, Riches D, Song W, Turcotte C, Wang S, Mardis ER, Wilson RK, Dooling D, Fulton L, Fulton R, Weinstock G, Durbin RM, Burton J, Carter DM, Churcher C, Coffey A, Cox A, Palotie A, Quail M, Skelly T, Stalker J, Swerdlow HP, Turner D, De Witte A, Giles S, Gibbs RA, Wheeler D, Bainbridge M, Challis D, Sabo A, Yu F, Yu J, Wang J, Fang X, Guo X, Li R, Li Y, Luo R, Tai S, Wu H, Zheng H, Zheng X, Zhou Y, Li G, Wang J, Yang H, Marth GT, Garrison EP, Huang W, Indap A, Kural D, Lee WP, Leong WF, Quinlan AR, Stewart C, Stromberg MP, Ward AN, Wu J, Lee C, Mills RE, Shi X, Daly MJ, DePristo MA, Altshuler DL, Ball AD, Banks E, Bloom T, Browning BL, Cibulskis K, Fennell TJ, Garimella KV, Grossman SR, Handsaker RE, Hanna M, Hartl C, Jaffe DB, Kernytsky AM, Korn JM, Li H, Maguire JR, McCarroll SA, McKenna A, Nemesh JC, Philippakis AA, Poplin RE, Price A, Rivas MA, Sabeti PC, Schaffner SF, Shefler E, Shlyakhter IA, Cooper DN, Ball EV, Mort M, Phillips AD, Stenson PD, Sebat J, Makarov V, Ye K, Yoon SC, Bustamante CD, Clark AG, Boyko A, Degenhardt J, Gravel S, Gutenkunst RN, Kaganovich M, Keinan A, Lacroute P, Ma X, Reynolds A, Clarke L, Flicek P, Cunningham F, Herrero J, Keenen S, Kulesha E, Leinonen R, McLaren WM, Radhakrishnan R, Smith RE, Zalunin V, Zheng-Bradley X, Korbel JO, Stütz AM, Humphray S, Bauer M, Cheetham RK, Cox T, Eberle M, James T, Kahn S, Murray L, Chakravarti A, Ye K, De La Vega FM, Fu Y, Hyland FC, Manning JM, McLaughlin SF, Peckham HE, Sakarya O, Sun YA, Tsung EF, Batzer MA, Konkel MK, Walker JA, Sudbrak R, Albrecht MW, Amstislavskiy VS, Herwig R, Parkhomchuk DV, Sherry ST, Agarwala R, Khouri HM, Morgulis AO, Paschall JE, Phan LD, Rotmistrovsky KE, Sanders RD, Shumway MF, Xiao C, McVean GA, Auton A, Iqbal Z, Lunter G, Marchini JL, Moutsianas L, Myers S, Tumian A, Desany B, Knight J, Winer R, Craig DW, Beckstrom-Sternberg SM, Christoforides A, Kurdoglu AA, Pearson JV, Sinari SA, Tembe WD, Haussler D, Hinrichs AS, Katzman SJ, Kern A, Kuhn RM, Przeworski M, Hernandez RD, Howie

B, Kelley JL, Melton SC, Abecasis GR, Li Y, Anderson P, Blackwell T, Chen W, Cookson WO, Ding J, Kang HM, Lathrop M, Liang L, Moffatt MF, Scheet P, Sidore C, Snyder M, Zhan X, Zöllner S, Awadalla P, Casals F, Idaghdour Y, Keebler J, Stone EA, Zilversmit M, Jorde L, Xing J, Eichler EE, Aksay G, Alkan C, Hajirasouliha I, Hormozdiari F, Kidd JM, Sahinalp SC, Sudmant PH, Mardis ER, Chen K, Chinwalla A, Ding L, Koboldt DC, McLellan MD, Dooling D, Weinstock G, Wallis JW, Wendl MC, Zhang Q, Durbin RM, Albers CA, Ayub Q, Balasubramaniam S, Barrett JC, Carter DM, Chen Y, Conrad DF, Danecek P, Dermitzakis ET, Hu M, Huang N, Hurles ME, Jin H, Jostins L, Keane TM, Le SQ, Lindsay S, Long Q, MacArthur DG, Montgomery SB, Parts L, Stalker J, Tyler-Smith C, Walter K, Zhang Y, Gerstein MB, Snyder M, Abyzov A, Balasubramanian S, Bjornson R, Du J, Grubert F, Habegger L, Haraksingh R, Jee J, Khurana E, Lam HY, Leng J, Mu XJ, Urban AE, Zhang Z, Li Y, Luo R, Marth GT, Garrison EP, Kural D, Quinlan AR, Stewart C, Stromberg MP, Ward AN, Wu J, Lee C, Mills RE, Shi X, McCarroll SA, Banks E, DePristo MA, Handsaker RE, Hartl C, Korn JM, Li H, Nemesh JC, Sebat J, Makarov V, Ye K, Yoon SC, Degenhardt J, Kaganovich M, Clarke L, Smith RE, Zheng-Bradley X, Korbel JO, Humphray S, Cheetham RK, Eberle M, Kahn S, Murray L, Ye K, De La Vega FM, Fu Y, Peckham HE, Sun YA, Batzer MA, Konkel MK, Walker JA, Xiao C, Iqbal Z, Desany B, Blackwell T, Snyder M, Xing J, Eichler EE, Aksay G, Alkan C, Hajirasouliha I, Hormozdiari F, Kidd JM, Chen K, Chinwalla A, Ding L, McLellan MD, Wallis JW, Hurles ME, Conrad DF, Walter K, Zhang Y, Gerstein MB, Snyder M, Abyzov A, Du J, Grubert F, Haraksingh R, Jee J, Khurana E, Lam HY, Leng J, Mu XJ, Urban AE, Zhang Z, Gibbs RA, Bainbridge M, Challis D, Coafra C, Dinh H, Kovar C, Lee S, Muzny D, Nazareth L, Reid J, Sabo A, Yu F, Yu J, Marth GT, Garrison EP, Indap A, Leong WF, Quinlan AR, Stewart C, Ward AN, Wu J, Cibulskis K, Fennell TJ, Gabriel SB, Garimella KV, Hartl C, Shefler E, Sougnez CL, Wilkinson J, Clark AG, Gravel S, Grubert F, Clarke L, Flicek P, Smith RE, Zheng-Bradley X, Sherry ST, Khouri HM, Paschall JE, Shumway MF, Xiao C, McVean GA, Katzman SJ, Abecasis GR, Mardis ER, Dooling D, Fulton L, Fulton R, Koboldt DC, Durbin RM, Balasubramaniam S, Coffey A, Keane TM, MacArthur DG, Palotie A, Scott C, Stalker J, Tyler-Smith C, Gerstein MB, Balasubramanian S, Chakravarti A, Knoppers BM, Abecasis GR, Bustamante CD, Gharani N, Gibbs RA, Jorde L, Kaye JS, Kent A, Li T, McGuire AL, McVean GA, Ossorio PN, Rotimi CN, Su Y, Toji LH, Tyler-Smith C, Brooks LD, Felsenfeld AL, McEwen JE, Abdallah A, Juenger CR, Clemm NC, Collins FS, Duncanson A, Green ED, Guyer MS, Peterson JL, Schafer AJ, Xue Y, Cartwright RA. A map of human genome variation from population-scale sequencing. Nature 2010; 467: 1061-1073.

[30] Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES and Nusbaum C. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 2009; 27: 182-189.

[31] Kenny EM, Cormican P, Gilks WP, Gates AS, O'Dushlaine CT, Pinto C, Corvin AP, Gill M and Morris DW. Multiplex Target Enrichment Using DNA Indexing for Ultra-High Throughput SNP Detection. DNA Res 2011 Feb; 18: 31-8. Epub 2010 Dec 16.

[32] Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O and Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 2005; 110: 462-467.

[33] Krych-Goldberg M and Atkinson JP. Structure-function relationships of complement receptor type 1. Immunol Rev 2001; 180: 112-122.

[34] Brouwers N, Van Cauwenberghe C, Engelborghs S, Lambert JC, Bettens K, Le Bastard N, Pasquier F, Montoya AG, Peeters K, Mattheijssens M, Vandenberghe R, De Deyn PP, Cruts M, Amouyel P, Sleegers K and Van Broeckhoven C. Alzheimer risk associated with a copy number variation in the complement receptor 1 increasing C3b/C4b binding sites. Mol Psychiatry 2012; 17: 223-233.

[35] Bettens K, Brouwers N, Engelborghs S, Lambert JC, Rogaeva E, Vandenberghe R, Le Bastard N, Pasquier F, Vermeulen S, Van Dongen J, Mattheijssens M, Peeters K, Mayeux R, St George-Hyslop P, Amouyel P, De Deyn PP, Sleegers K and Van Broeckhoven C. Both common variations and rare non-synonymous substitutions and small insertion/deletions in CLU are associated with increased Alzheimer risk. Mol Neurodegener 2012; 7: 3.

[36] Ingman M and Gyllensten U. SNP frequency estimation using massively parallel sequencing of pooled DNA. Eur J Hum Genet 2009; 17: 383-386.

[37] Bansal V, Tewhey R, Leproust EM and Schork NJ. Efficient and cost effective population resequencing by pooling and in-solution hybridization. PLoS ONE 2011; 6: e18353.

[38] Clarke LA, Rebelo CS, Goncalves J, Boavida MG and Jordan P. PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. Mol Pathol 2001; 54: 351-353.