Genomic loci susceptible to systematic sequencing bias in clinical whole genomes

Timothy M. Freeman[1], Genomics England Research Consortium[2], Dennis Wang[1,3,#], Jason Harris[4]

[1]Sheffield Institute for Translational Neuroscience, University of Sheffield, Sheffield, UK

[2]The full list of Genomics England Research Consortium members and affiliations is listed in the acknowledgements

[3]NIHR Biomedical Research Centre, University of Sheffield, UK

[4]Personalis Inc., California, USA

# corresponding author

Abstract

Background

Highly accurate next-generation sequencing (NGS) of genetic variants is key to many areas of science and medicine, such as cataloguing population genetic variation and diagnosing patients with genetic diseases. Certain genomic loci and regions can be prone to higher rates of systematic sequencing and alignment bias that pose a challenge to achieving high accuracy, resulting in false positive variant calls. Current standard practices to differentiate between loci that can and cannot be sequenced with high confidence utilise consensus between different sequencing methods as a proxy for sequencing confidence. This assumption is not accurate in cases where all sequencing pipelines have consensus on the same errors due to similar systematic biases in sequencing. Alternative methods are therefore required to identify systematic biases.

Methods

We have developed a novel statistical method based on summarising sequenced reads from whole genome clinical samples and cataloguing them in "Incremental Databases" (IncDBs) that maintain individual confidentiality. Variant statistics were analysed and catalogued for each genomic position that consistently showed systematic biases with the corresponding sequencing pipeline.

Results

We have demonstrated that systematic errors in NGS data are widespread, with persistent low-fraction alleles present at 1.26-2.43% of the human autosomal genome across three different Illumina-based pipelines, each consisting of at least 150 patient samples. We have identified a variety of genomic regions that are more or less prone to systematic biases, such as GC-rich regions (OR = 6.47-8.19) and the NIST high-confidence genomic regions (OR = 0.154-0.191). We have verified our predictions on a gold-standard reference genome and have shown that these systematic biases can lead to suspect variant calls at clinically important loci, including within introns and exons.

Conclusions

Our results recommend increased caution to minimise the effect of systematic biases in whole genome sequencing and alignment. This study supports the utility of a statistical approach to enhance quality control of clinically sequenced samples in order to flag up variant calls made at known suspect loci for further analysis or exclusion, using anonymised summary databases from which individual patients cannot be re-identified, so that results can be shared more widely.

Introduction

DNA sequencing is an imperfect process, and although error rates are low, mistakes in identifying genomic variants can still occur. While the sources of random sequencing errors are relatively well understood (Ma et al. 2019; Benjamini and

Speed 2012), identifying systematic errors in whole genomes sequenced in a clinical or commercial setting is not always possible due to restrictions in gathering information about the samples and sequencing processes. These errors could cause incorrect decisions on presence/absence of disease relevant variants in the genome and influence the decisions of the corresponding physician or patient (Goldfeder et al. 2016).

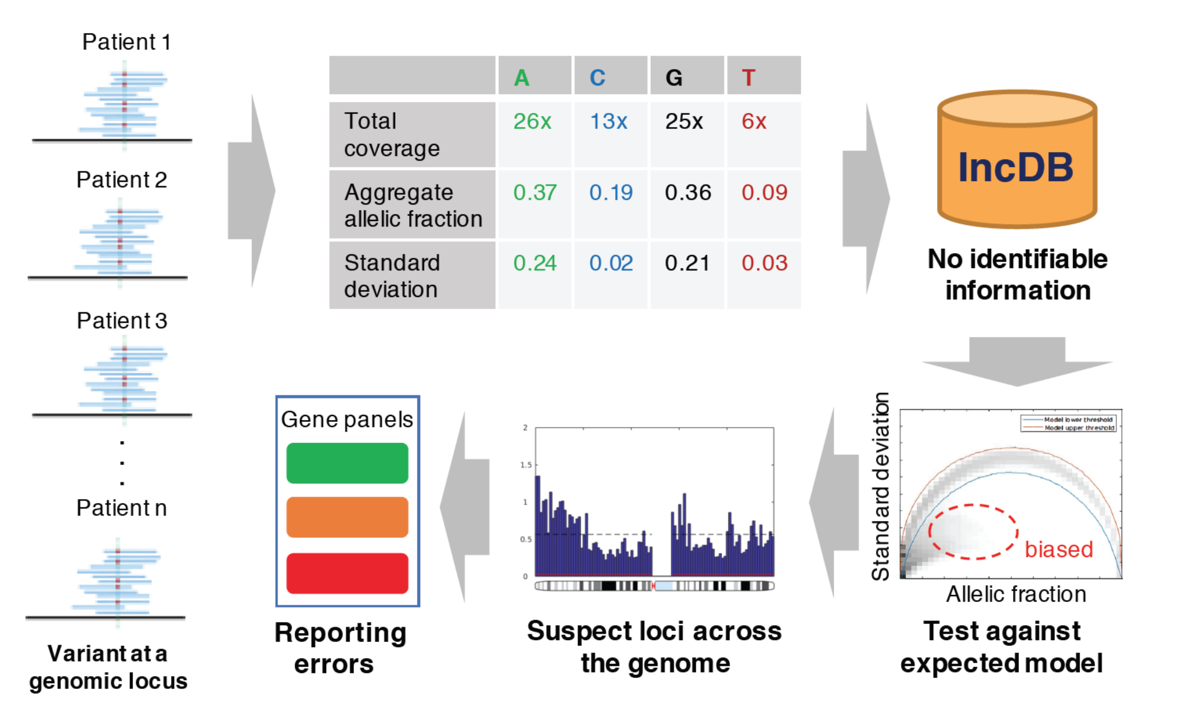
One of the major challenges to improving variant detection is that certain regions of the genome are prone to higher rates of systematic sequencing or alignment errors, which can result in the false detection of low allelic fraction variants. In the case of diploid genotype calls, true variants are expected at 50% or 100% allelic fractions, corresponding to heterozygous and homozygous loci. However, real variants often occur at low allelic fractions, such as somatic variants in tumours, and in mosaicism, where nearby cells sampled together can show genetic heterogeneity across the sample (Xu 2018). In these cases it becomes critical to be able to filter out variants that systematically exhibit a low allelic fraction across individuals, since these are unlikely to be true somatic variants.

Lists of ‘high confidence’ calls from gold-standard reference genomes are often used for quality control in clinical and commercial sequencing laboratories. The NIST GIAB consortium has proposed a list of all ‘high confidence’ positions on the genome within which sequencing accuracy is thought to be higher, developed by analysing the consensus between different sequencing technologies and variant callers for the same genomic samples to develop a ‘truth set’ of variant calls (Zook et al. 2014). However, this top-down approach cannot identify which variant calls are most likely when there is disagreement between the results from different sequencing protocols. In addition, at loci that have the same sequencing biases across all or most sequencing technologies, assuming that the consensus call is true can lead to errors being falsely called as real high-confidence variants (Krusche et al. 2019). Another drawback to this is that clinically collected samples can vary in quality and contamination and may introduce variants with low allelic fractions not seen in reference genomes. Furthermore, the number of reference genomes used may be quite small and is therefore unlikely to be representative of the diversity of clinically-sequenced genomes.

Other top-down approaches of evaluating thresholds for allelic fraction or read quality may differ depending on the variant calling pipelines used (Sandmann et al. 2017). Benchmarking these different approaches on cohorts of genomes may be insightful for research, but impractical for clinical applications and it risks leaking sensitive genetic information. Furthermore, standard variant calling quality control measures can sometimes be overly simplistic, such as fixed read depth thresholds for calling variants across the genome, which are not tailored to wide regional differences in systematic biases. An excessive reliance on high read depth for accurate variant calling increases the costs of sequencing studies, which are forced to compromise between the number of genomes they can sequence and the depth of coverage they can achieve (Consortium and The 1000 Genomes Project Consortium 2010).

Instead, a ‘bottom-up’ method that utilises large numbers of real-world clinical samples, can be trained on data from a single whole genome sequencing (WGS) pipeline, and does not rely on consensus between multiple sequencing technologies. We used the method to evaluate systematic bias in detected allelic fractions across whole genomes from three non-cancer cohorts, 150 individuals in the USA who underwent genetic testing at Personalis Inc., and two separate sets of 215 individuals in England from the 100,000 Genomes Project. Our method implements a novel technique for cataloging cohort allele fractions called an Incremental Database (IncDB). IncDBs provide statistics such as the average allelic fractions at each genomic locus and the standard deviation for these fractions across individuals (**Figure 1**), without containing any sample-specific information. This ensures that individuals remain anonymous during the quality control process of examining reads using an IncDB.

In this study we have searched the IncDBs generated for loci that persistently present a low-fraction alternate allele, across many samples. Such loci may arise from systematic sequencing or mapping errors. Because these loci may be mistaken for true biological variation, it is important to understand that such loci exist, and to utilize a catalog of these "suspect loci" when interpreting measurements of genetic variation based on NGS data. We analyse the genomic positions where systematic biases occur, in terms of chromosomal architecture and in terms of specific types of genomic regions in which they are enriched, such as GC-rich regions and Alu repeats. We also examine these biases in high confidence sequence regions and in high quality clinical genomes to demonstrate the value of our method. We propose the use of our method as a tool for researchers and clinicians to aid in differentiating between true and false positive called genomic variants with particular utility for cases where variants are expected at low allelic frequencies, such as cell-free DNA (cfDNA) sequencing in cancer patients.



	A	C	G	T
Total coverage	26x	13x	25x	6x
Aggregate allelic fraction	0.37	0.19	0.36	0.09
Standard deviation	0.24	0.02	0.21	0.03

Figure 1: Approach for detecting loci with systematic sequence bias: Variant allele frequencies are collected from a cohort of individuals at every locus and aggregated into allele-specific summary statistics. These aggregate allelic fractions and standard deviations at each genomic position is stored in the Incremental Database (IncDB), which does not contain any patient-specific information. The 99.9% confidence intervals for expected standard deviation at each allelic fraction was generated. Genomic positions where the observed standard deviation was beyond the confidence interval of the expected was considered to have “biased” allelic fractions. These positions were cataloged as “suspect loci”, and mapped to variant calls in clinically relevant genes. Prioritisation of genes for diagnostic and reporting purposes can be adjusted according to the presence of suspect loci.

Methods

Data sources

Data set 1 (Personalis Inc.):

WGS data was obtained by Personalis Inc. using Illumina HiSeq sequencing and mapped with BWA MEM to align reads against the GRCh37 reference human genome (**Table 1**). The mean depth of coverage across patients was 45x. There were 150 non-cancer individuals in the cohort, including some triplets of related individuals, recruited from hospitals in the USA and a mix of ethnicities.

Data sets 2 & 3 (100,000 Genomes Project):

WGS data was obtained by Genomics England's 100,000 Genomes Project, using Illumina HiSeq sequencing and mapped with the Illumina Isaac-aligner to align reads against the reference human genomes GRCh37 (data set 2, patients sequenced earlier in the study and mapped with an earlier version of the Isaac-aligner) and GRCh38 (data set 3, patients sequenced later in the study and mapped with a more recent version of the Isaac-aligner). The reads were aligned by Illumina with their internal workflow. Illumina workflow V2 (HiSeq Analysis Software 2) was used for GRCh37 (data set 2) and Illumina workflow V4 (NorthStar4) was used for GRCh38 (data set 3) before the output was delivered to Genomics England. The length of paired-end reads was 150bp and the mean depth of coverage across patients was 30x. Blood samples were taken from 215 distinct patients of mixed ethnicities with non-cancer neurological diseases in each cohort, recruited from hospitals in the UK. No patients were in both data sets 2 and 3.

Data set	Source	Number/type of patients	Genome build	Sequencing	Alignment
1	Personalis Inc. (USA)	150 (non-cancer)	GRCh37	Illumina HiSeqX	BWA-MEM 0.7.12
2	100,000 Genomes Project (UK)	215 (earlier sequenced neurological disorders)	GRCh37	Illumina HiSeqX	Isaac-aligner (SAAC00776.15.01.27)
3	100,000 Genomes Project (UK)	215 (later sequenced neurological disorders)	GRCh38	Illumina HiSeqX	Isaac-aligner (iSAAC-03.16.02.19)

Table 1: Sequencing protocols and data sources for all data sets used to generate IncDBs.

Incremental Database Generation

The coverage values for each allele (A, C, G, T) at every autosomal genomic locus were calculated and divided by the total coverage at the corresponding loci to get the allelic coverage fraction, x_p , for each allele at each locus in each patient, as shown in **Figure 1**. Individual IncDBs were created for each data set from the aggregate allelic fraction and standard deviation values for each allele at each locus across the entire cohort, which were calculated from x_p as described below.

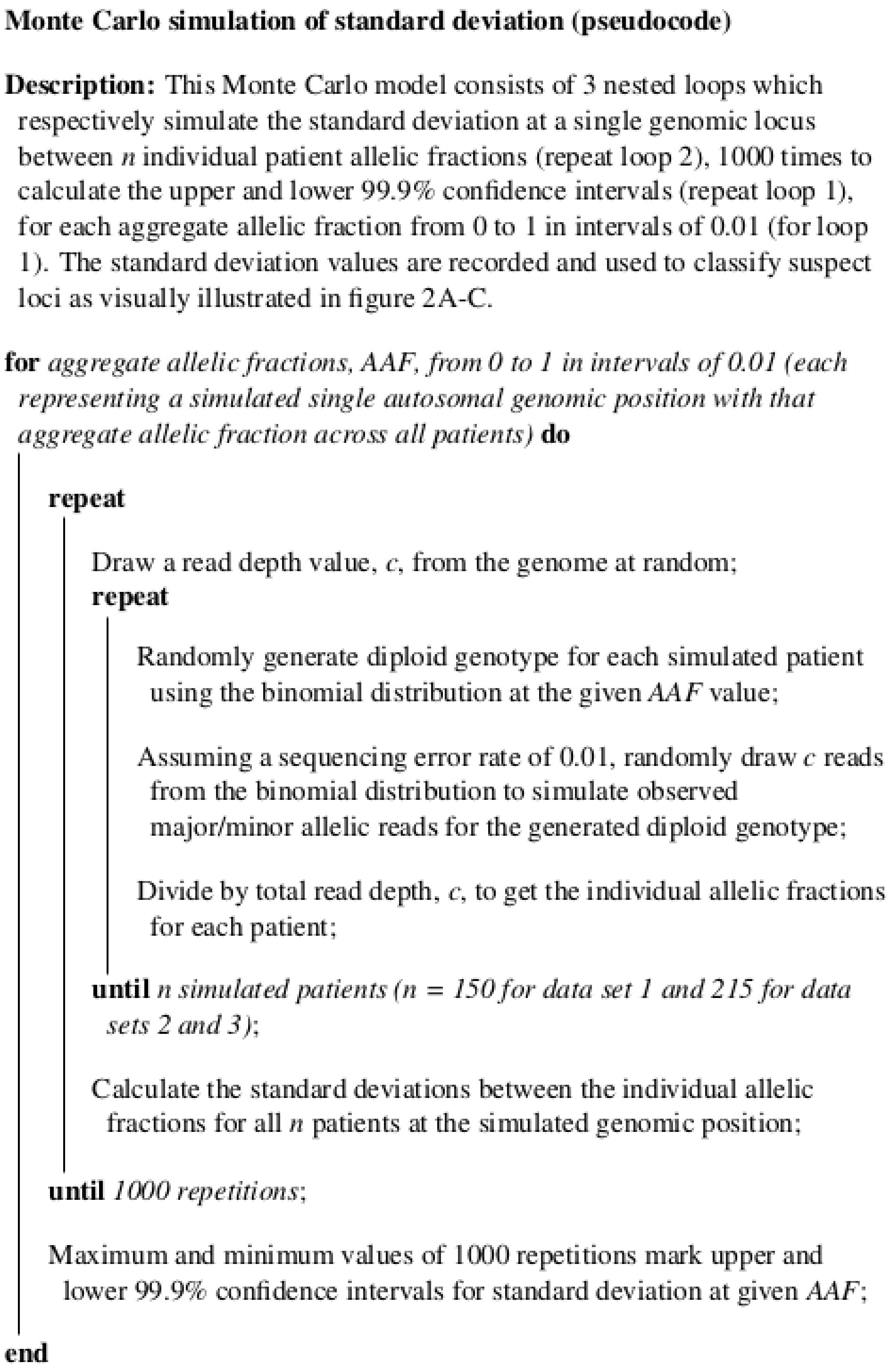
Aggregate allelic fraction = $\frac{1}{N}\sum_{p=1}^{N} x_p$

Standard deviation = $\sqrt{\frac{1}{N}\sum_{p=1}^{N}\left(x_p-\bar{x}\right)^2}$ = $\sqrt{\frac{1}{N}\sum_{p=1}^{N}\left(x_p^2\right)-\left(\frac{1}{N}\sum_{p=1}^{N}x_p\right)^2}$, where N is the number of patients, p is the patient identifier, x_p is the allelic coverage fraction for a specific allele in patient p, and $\bar{x}$ is the mean of all of the allelic coverage fractions for that same allele across all patients (aggregate allelic fraction). Notice that to compute the average allelic fraction, we do not store each individual’s x_p values, but the sum of x_p across all individuals. Similarly, we can compute the standard deviation across individuals by storing the sum of x_p , as well as the sum of x_p^2 . This approach not only removes all individual-specific genomic information, but also allows the IncDB to grow indefinitely, as more samples are sequenced and analyzed: they can simply contribute to the running sums of x_p and x_p^2 . Also note that the sum registers in the equations above do not take up more computational size on disk as the number of samples increases, so the overall IncDB file size does not increase as new samples are added.

Identifying loci affected by systematic bias (suspect loci)

For each locus in all autosomal chromosomes, the standard deviation and aggregate allelic fraction values were taken from the IncDB and plotted against each other in a density plot using MATLAB 9.6.

Monte Carlo sampling assuming diploid allele arrangements was used to generate the expected 99.9% confidence interval for the standard deviation at each aggregate allelic fraction (ranging from 0 to 1 in intervals of 0.01) with 1000 repetitions at each, see pseudo-code below. The model assumed an error rate of 0.01, corresponding to an approximation of the error rate of Illumina WGS (Wall et al. 2014). Approximately 90% of genomic reads in data set 1 had a quality score of 20 or above, corresponding to this error rate (**Figure 9**).



Monte Carlo simulation of standard deviation (pseudocode)

Description: This Monte Carlo model consists of 3 nested loops which respectively simulate the standard deviation at a single genomic locus between n individual patient allelic fractions (repeat loop 2), 1000 times to calculate the upper and lower 99.9% confidence intervals (repeat loop 1), for each aggregate allelic fraction from 0 to 1 in intervals of 0.01 (for loop 1). The standard deviation values are recorded and used to classify suspect loci as visually illustrated in figure 2A-C.

```
for aggregate allelic fractions, AAF, from 0 to 1 in intervals of 0.01 (each
  representing a simulated single autosomal genomic position with that
  aggregate allelic fraction across all patients) do
    repeat
        Draw a read depth value, $c$, from the genome at random;
        repeat
            Randomly generate diploid genotype for each simulated patient
              using the binomial distribution at the given $AAF$ value;
            Assuming a sequencing error rate of 0.01, randomly draw $c$ reads
              from the binomial distribution to simulate observed
              major/minor allelic reads for the generated diploid genotype;
            Divide by total read depth, $c$, to get the individual allelic fractions
              for each patient;
        until $n$ simulated patients ($n = 150$ for data set 1 and 215 for data
          sets 2 and 3);
        Calculate the standard deviations between the individual allelic
          fractions for all $n$ patients at the simulated genomic position;
    until 1000 repetitions;
    Maximum and minimum values of 1000 repetitions mark upper and
      lower 99.9% confidence intervals for standard deviation at given $AAF$;
end
```

The expected values for the standard deviation in the individual allelic fractions simulated from this model were used to highlight the differences between the observed and expected standard deviation distributions. Observed loci below the lower 99.9% confidence interval on the expected distribution for a given nucleotide were defined as suspect loci for that nucleotide, since they displayed excessively low standard deviation for their aggregate allelic fraction, and were not likely to be in Hardy-Weinberg equilibrium (**Figure 2**). Autosomal loci that displayed at least one suspect allele at that position were termed unique suspect loci. The total count of unique suspect loci was therefore lower than the total count of suspect loci, since some loci had multiple suspect alleles.

Analysis of regional enrichment of unique suspect loci

Histograms of suspect locus density across the chromosome were plotted in MATLAB alongside chromosome ideograms taken from the UCSC Genome Browser Downloads page for GRCh37 (data sets 1 and 2) and GRCh38 (data set 3) in order to show suspect locus density in comparison with chromosomal banding patterns.

In addition, BED files were provided by Personalis Inc. for 8 different types of genomic region which were analysed to check enrichment of unique suspect loci using a Fisher Exact test to calculate the exact significance values. A full contingency table for GC-rich regions in data set 1 (autosomal chromosomes only) is available in **Supplementary Table 1** as an example to show how this was calculated.

The regions tested were the GIAB NIST (Genome in a Bottle National Institute of Standards and Technology) high confidence regions, Alu repeats, GCgt70 (> 70% GC content) regions, NonUnique100 regions (defined as all regions where a single 100-bp read could not map uniquely), segmental duplications, small/large homopolymers, repeat masker region, introns, exons, genes, intergenic region and three neurological clinical panels (see next section for full list and details of BED files used).

Genomic region BED file sources

GIAB NIST (Genome in a Bottle National Institute of Standards and Technology) high confidence region (Xiao et al. 2014) - A selection of genomic loci covering the majority of the human genome that are considered to have high confidence calls.

GCgt70 (GC content > 70%) - Regions with greater than 70% GC content. Loci were annotated as within this region if the surrounding 100bp around each locus had greater than 70% GC content.

NonUnique100 - All regions where a single 100-bp read cannot map uniquely (so all stretches on the reference that are 100bp or longer that are repeated on the GRCh37 reference).

Segmental duplication - Long DNA sequences (> 10kb) that are found in multiple locations across the human genome as a result of duplications.

Small homopolymer - Region of DNA containing a single nucleotide (9-19 bp).

Large homopolymer - Region of DNA containing a single nucleotide (≥ 20 bp).

RepeatMasker region - A BED file containing a variety of different types of repeats (Smit, AFA & Green, P - http://www.repeatmasker.org/). The open-3-2-7 version of RepeatMasker was downloaded from the UCSC Table Browser.

Alu repeats (Hasler and Strub 2007) - The most common type of transposable element in the human genome, of which there are over one million copies. The BED file was composed of all RepeatMasker Regions downloaded from the UCSC Table Browser that were annotated as Alu repeats in the repName column.

BED files were also downloaded for genic regions, intergenic regions, exonic regions, intronic regions (03/22/19) and ClinVar short variants (06/12/19), acquired from the UCSC Table Browser.

Clinical panel BED files were also downloaded for the three most reviewed neurological clinical panels on PanelApp (for intellectual disability (10/19/18), genetic epilepsy syndromes (02/07/19) and hereditary spastic paraplegia respectively

(02/07/19)).

“All sequenced regions” referred to a BED file generated for each data set analysed containing a list of all genomic loci where the number of aligned reads was greater than zero.

Calculating allelic frequencies at suspect loci in NA12878

An indexed BAM file for NA12878 chromosome 1 was obtained from the GIAB consortium (Xiao et al. 2014). We used the Integrative Genomics Viewer (IGV) to examine NA12878’s read pileup at several data set 1 suspect locus positions at which chromosome 1 SNVs had previously been called. These positions included examples within intronic and exonic regions in the PanelApp Intellectual Disability clinical gene panel and within intergenic regions. Observing these read pileups confirmed that NA12878 exhibited low-fraction alleles at these positions. NA12878 was not part of any cohorts used to build the lncDBs in this study. Chromosome 1 SNVs in NA12878 were extracted from a VCF file corresponding to the sequencing pipeline used for data set 1. SNVs were classified as suspect if they corresponded to the same alleles at the same positions as suspect loci calculated for data set 1. The allelic frequencies for these variants were calculated from the NA12878 BAM file using SAMtools mpileup v1.9. Variants with fewer than 10 supporting reads were deemed to have insufficient coverage and were filtered out.

Analysing the proportion of gnomAD SNVs that are suspect

A list of all gnomAD variants, along with their allelic fraction and annotation as PASS-flagged or not, was obtained from a TSV file. This was filtered to only include autosomal SNVs. These were classified as suspect and non-suspect SNVs as above. Variants in gnomAD were annotated as PASS variants if they were marked this way in gnomAD v2.1.

Comparing sequencing quality between suspect and non-suspect loci

The coverage of different allelic reads across all loci/nucleotide combinations on chromosome 1 was available for data set 1 (Cov), along with the corresponding coverage of allelic reads filtered to only include reads with sequencing and mapping quality scores greater than 20 (Cov20). The filtered coverage values of allelic reads (Cov20) were divided by the corresponding unfiltered coverage values (Cov) to get the proportion of allelic reads with sequencing and mapping quality scores both greater than 20 at each locus/nucleotide combination. The cumulative distributions of these values were calculated separately for locus/nucleotide combinations that were annotated as suspect loci or non-suspect loci.

Results

Systematic biases widespread across many genomic loci

The observed relationship between standard deviation and the average allelic fraction at each genomic locus was compared to the expected distribution assuming inherited variants in Hardy-Weinberg equilibrium in **Figure 2A/B/C**. We labelled the positions which fell outside the 99.9% envelope of the Mendelian model as ‘suspect loci’; 1.26-2.43% of all autosomal loci are suspect loci for at least one allele, which we term unique suspect loci in this study. In all three datasets, the suspect loci trace a plume of low allelic fraction (up to 40%), but with much lower standard deviation across samples than a Mendelian variant with equal aggregate allele fraction would have. Single nucleotide variants (SNVs) that were called in NA12878, an individual sample separate from all of the data sets used, were analysed to check if they matched both suspect locus positions and their corresponding suspect alleles in data sets 1 and 2 (Figure 2D). The majority of NA12878 SNVs annotated as suspect were shared between data sets 1 and 2.

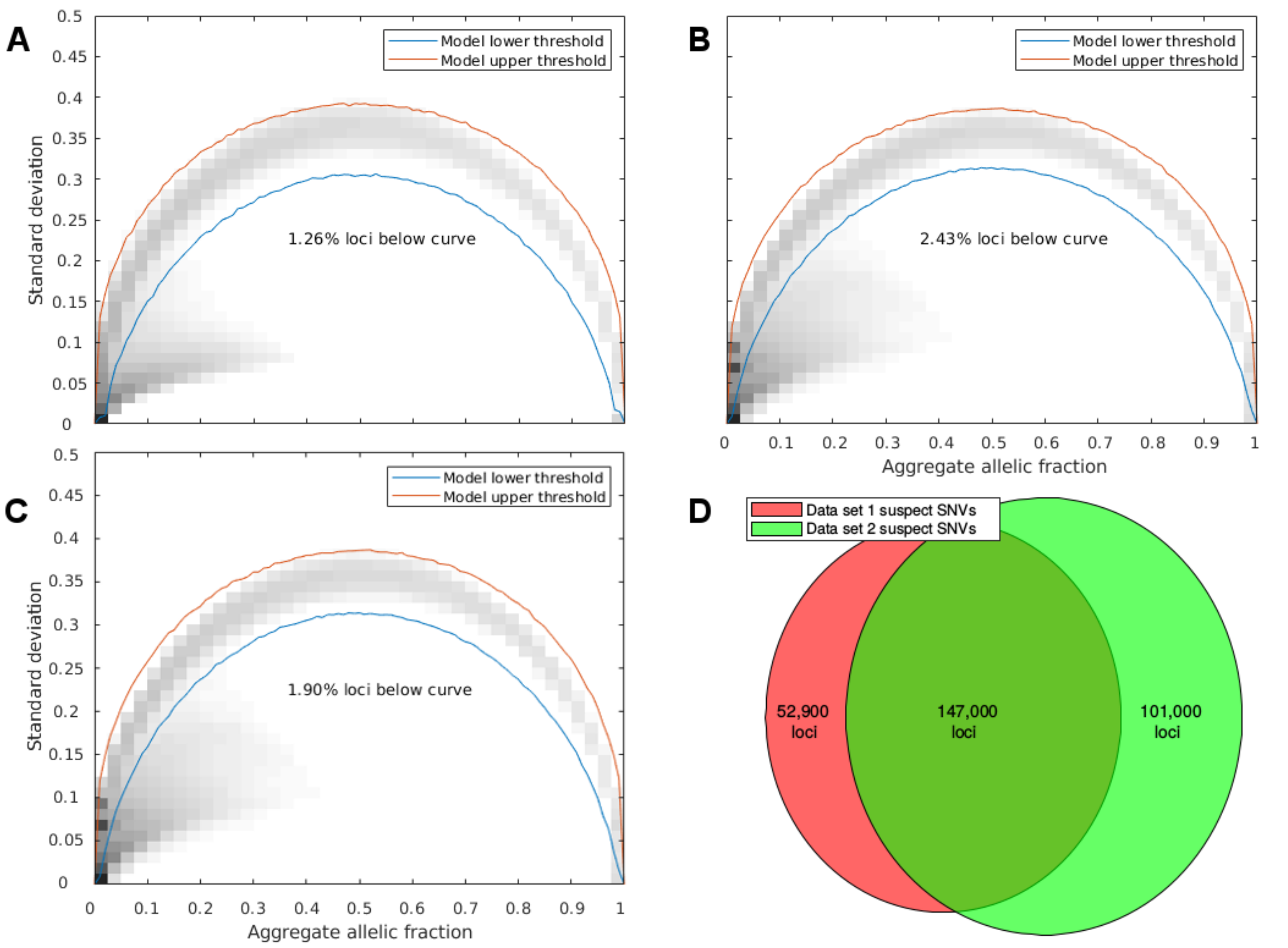


Figure 2: Identification of suspect autosomal loci/allele combinations with persistent low allelic fractions across patients. Observed and expected variant allele frequencies were estimated from three different whole genome sequenced cohorts. **A/B/C**- For all loci in autosomal chromosomes (**A**, Data set 1, **B**, Data set 2, **C**, Data set 3), the standard deviation and aggregate allelic fraction values from the Incremental Database were plotted against each other in a density plot. The darker regions have the highest concentration of loci, while lighter pixels represent combinations of standard deviation and aggregate allelic fraction that did not contain as many loci. The red lines indicate the upper and lower boundaries of the 99.9% confidence interval (shown in Supplementary Figures 1A/B respectively). Suspect loci in **A** and **B/C** were defined as the loci which occurred below this threshold. The percentage of loci with at least one suspect locus annotated (unique suspect loci) are reported below the curve on each. **D**- Venn diagram showing the overlap of suspect SNVs between data sets 1 and 2, called in NA12878 (both GRCh37). Data set 3 used the GRCh38 reference, so it was not included in the Venn diagram.

Unique suspect locus enrichment within genomic regions

Unique suspect autosomal loci were found to be present across the entire sequenced genome and only absent in unsequenced sections, although their prevalence varied across sequenced regions both locally and showing larger trends across chromosomes (**Figure 3**). We examined the distribution of unique suspect loci across different regions of the genome (**Figure 4**), and recorded the regional enrichment of unique suspect loci using odds ratios (OR)**.** All odds ratios calculated were highly statistically significant due to the very large number of genomic positions sampled, even when the odds ratios were close to 1. The 95% confidence interval lower and upper bounds were the same as the reported odds ratios to 3 significant figures in all cases. The highest/least significant p-value recorded was for the enrichment of suspect loci in the intellectual disability gene panel in data set 2 (OR=1.01, p=8.69x10^{-41}). All other p-values ranged from 10^{-322} to 10^{-79}.

GC-rich regions (OR = 8.19/6.47/7.99 for data sets 1/2/3) and Alu repeats (OR = 5.70/6.74/6.21 for data sets 1/2/3) were very strongly enriched for suspect loci. Large homopolymers (OR = 1.93/2.92/1.84 for data sets 1/2/3) and the Repeat Masker regions (OR = 1.85/2.96/2.18 for data sets 1/2/3) were mildly enriched for suspect loci. Small homopolymers on the other hand were depleted or unenriched (OR = 0.525/1.03/0.519 for data sets 1/2/3), and the NIST GIAB high confidence region was strongly depleted (OR = 0.191/0.154/0.172 for data sets 1/2/3) for suspect loci.

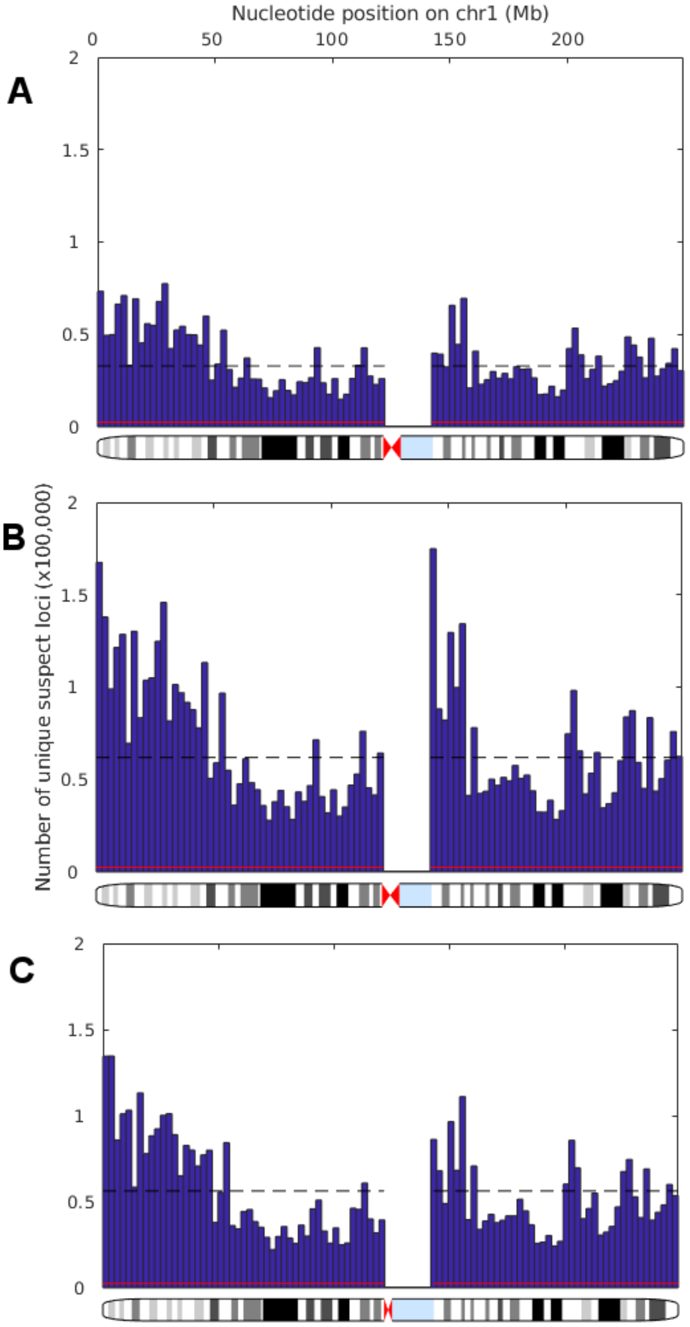


Figure 3: Variability in distribution of unique suspect loci in sequenced regions of chromosomes. Histograms show the positions of suspect loci (as defined above) across chromosome 1 (**A**: Data set 1/Personalis lncDB, **B**: Data set 2/100,000 Genomes Project GRCh37 lncDB, **C**: Data set 3/100,000 Genomes Project GRCh38 lncDB), with the number of suspect loci per bin on the y axis and the nucleotide position on the x axis (GRCh37 for **A/B**, GRCh38 for **C**). There were no suspect loci at the centromere since this could not be sequenced. The black dotted line shows the mean number of suspect loci per bin, while the red line shows the amount of suspect loci in each bin that would be expected by chance (1 per 1000 loci).

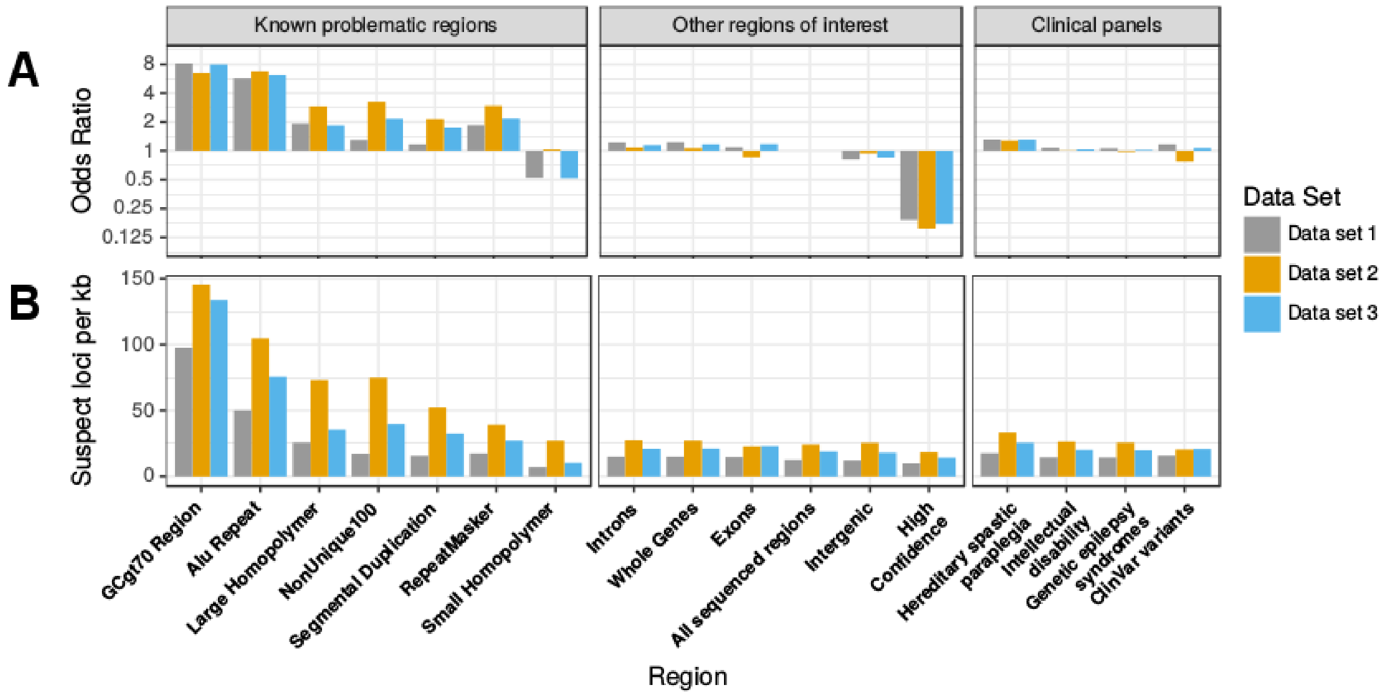


Figure 4: Enrichment of unique suspect loci in low complexity regions and depletion in high confidence regions, in three independent cohorts using Illumina sequencing. All odds ratios shown are statistically significant and equivalent to the 95% upper and lower confidence intervals to 3 significant figures.

A: Log2 scale barplot showing the odds ratios for regional enrichment of unique suspect loci across all autosomal chromosomes from three different data sets. All odds ratios were calculated and shown to be significant using Fisher's Exact Test. "All sequenced regions" had an odds ratio of 1 because it contained all of the loci tested.

B: Barplot showing the number of unique suspect loci per kb across all autosomal chromosomes from three different data sets.

While there was enrichment (**Figure 4A**) of suspect loci in repeat and homopolymer regions (Alu repeats, the Repeat masker region, Large homopolymers), the suspect loci were the most enriched in GC > 70% regions. The correlation between suspect loci and GC-rich region proportions for all autosomal chromosomes was high (Spearman's Rho = 0.822/0.922/0.760, p = $2.73 \times 10^{-6}/1.03 \times 10^{-9}/4.13 \times 10^{-5}$) for data sets 1, 2 and 3 respectively, accounting for the majority of the variability in the proportion of suspect loci in autosomal chromosomes (**Figure 5**). Even though there was wide variability in the proportions of autosomal chromosomes covered by GC > 70% regions (0.003-0.0219), there was very little variation in the overall GC content (40-42%), suggesting that this effect is a function of GC clustering specifically, rather than a sequencing issue with individual GC loci generally.

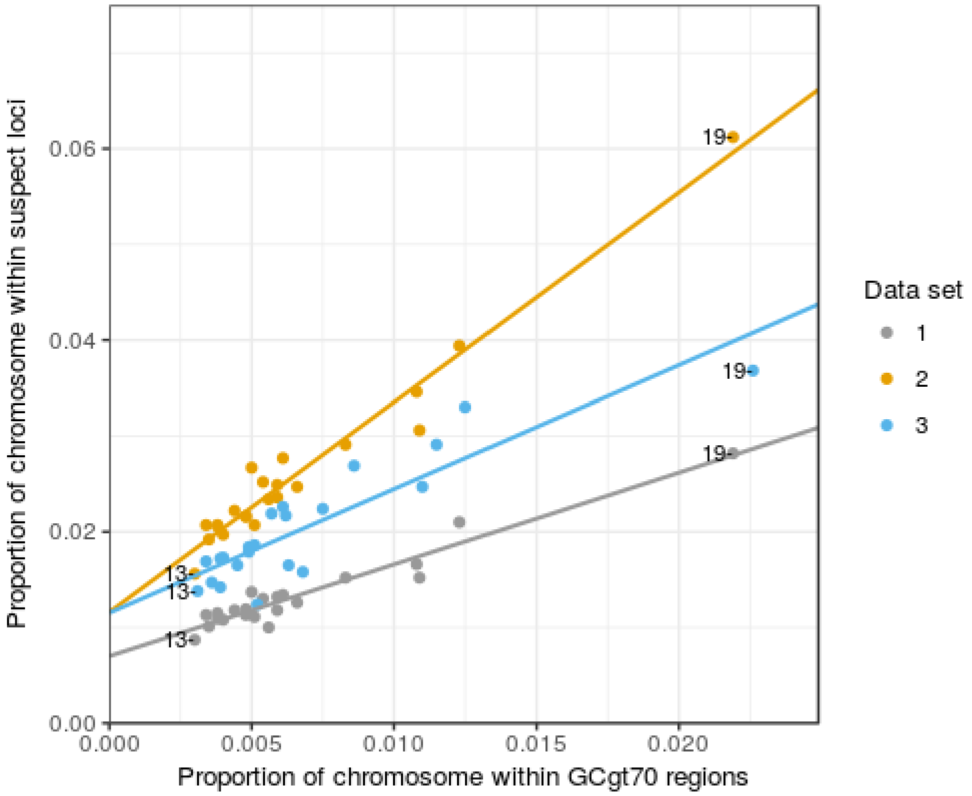


Figure 5: Proportion of chromosomal positions with at least one suspect locus vs. proportion of chromosome within GCgt70 (GC content > 70%) regions, with linear line of best fit. There is a strong linear correlation between the proportion of unique suspect loci in each chromosome and the proportion of loci in GCgt70 regions, indicating that these regions are a key contributor to systematic sequencing and/or alignment errors. Chromosomes 13 and 19 are labelled as the chromosomes with the lowest and highest pairs of values respectively in all 3 data sets, with the exception of chromosome 14, which had a slightly lower proportion of unique suspect loci than chromosome 13 in data set 3 (the GRCh38 100,000 Genomes Project lncDB).

Systematic biases confirmed in gold-standard reference sample

Our analysis based on aggregate allele statistics across more than 150 samples implies the presence of genomic loci that present a low-fraction allele in most samples. In order to confirm this prediction, select (non-unique) suspect loci were examined in the reference sample NA12878, which was not part of the training set for any of the lncDBs. NA12878 displayed the low-fraction alleles in its read pileup at each of the suspect locus positions we examined. In addition, variants called in

NA12878 were found to have far lower median allelic fraction when they occurred at suspect loci (0.300), compared to variants called elsewhere (0.579) (**Figure 6**). The Kullback-Leibler divergence value between the distributions in this figure was 0.0340. While most called variants in NA12878 conform to the expected distribution of allele fractions for heterozygous or homozygous variants, variants called at suspect loci (which comprised ~5% of all variants) show no heterozygous peak around 50% allele fraction, with the majority occurring between ~10-35%, confirming that suspect loci can yield false positive variant calls.

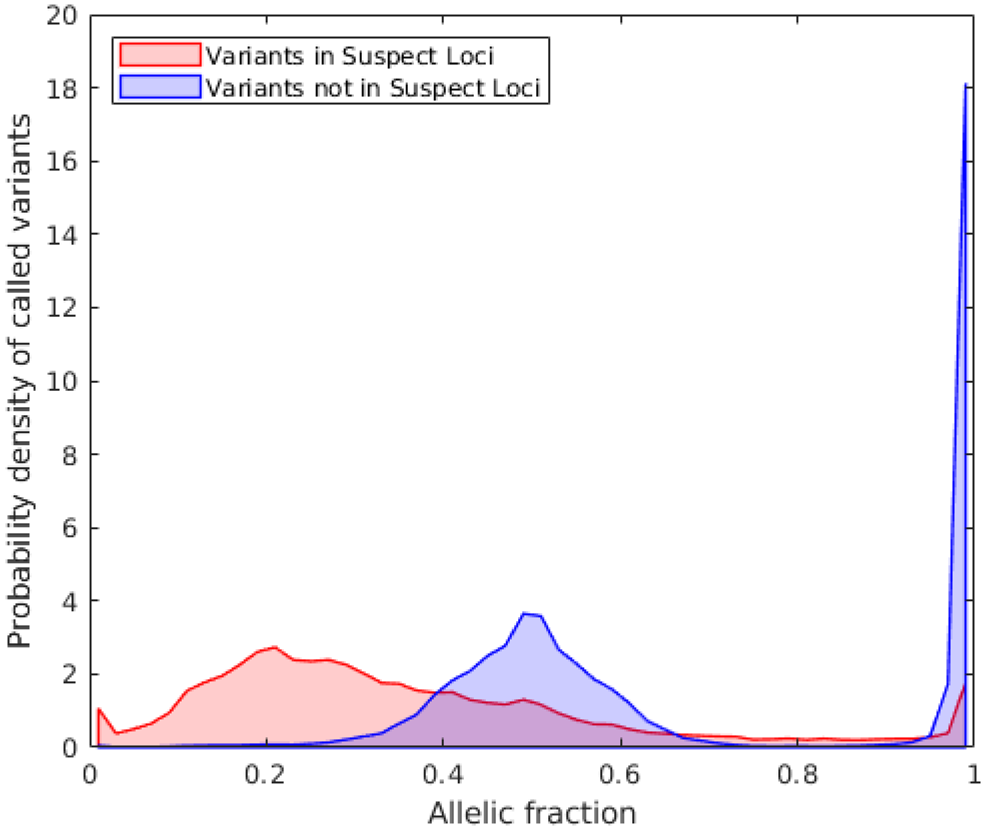


Figure 6: Distribution of allelic fractions among called SNVs on chromosome 1 in NA12878, separated by presence/absence of suspect loci at SNV in data set 1 (Personalis). Low coverage variants (<10 supporting reads) were excluded from this analysis. Area under the curve equals 1 for both curves.

In addition, individual examples of these chromosome 1 suspect variants were observed using the Integrative Genomics Viewer within the PanelApp Intellectual

Disability clinical gene panel in NA12878 (**Figure 7**). Suspect loci were examined within clinically relevant exons, introns and intergenic regions, and suspect reads were confirmed at all of these positions.

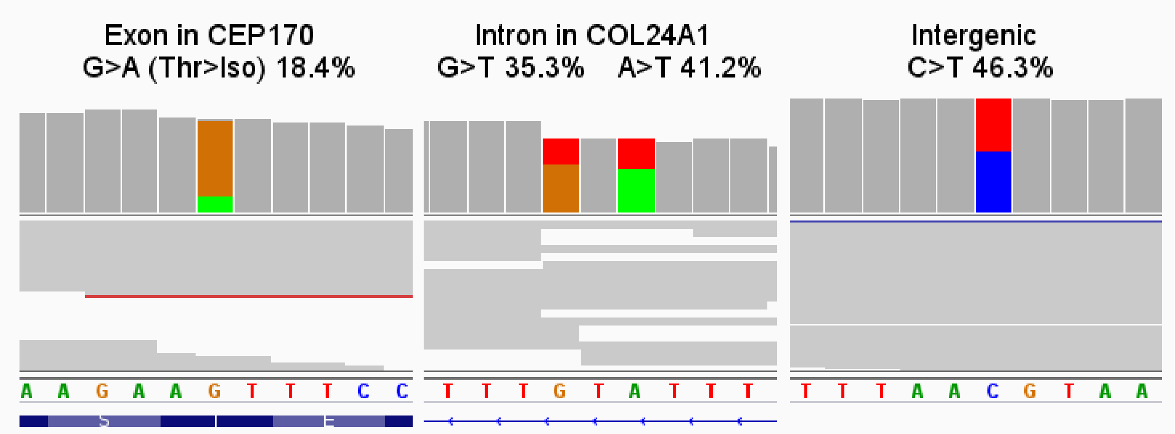


Figure 7: Cropped panels from the Integrative Genomics Viewer (Robinson et al. 2011), highlighting suspect loci from data set 1 in chromosome 1 which were called as variants separately in the GIAB pilot genome (sequenced with Illumina HiSeq, but not used as part of the patient data set to create the Incremental Database) (NA12878) (Xiao et al. 2014) predicted to have consistent low allelic fraction loci. Reads are shown in grey with coloured bands where non-reference allelic reads were observed (A=Green, C=Blue, G=Brown, T=Red). Suspect variant alleles and their respective read proportions in NA12878 are indicated above - these systematically occur at similar levels across all patients in the Incremental Databases used to identify them. Left/Middle: Suspect variant calls in exonic and intronic regions of genes in the PanelApp Intellectual Disability panel, predicted from data set 1 (Personalis IncDB) and called as variants in NA12878. Right: Suspect loci from data set 1 (Personalis IncDB) in an intergenic region of NA12878.

Discussion

The main aim of this study has been to develop and evaluate a novel statistical method to identify positions of the genome that are prone to systematic bias in genomic sequencing and alignment, using anonymised summary patient data. We developed an approach to quality control sequenced reads in the autosomal genome by cataloguing genomic positions in which there is significantly less standard deviation in patient allelic fractions than would be predicted in the absence of systematic bias, which we have labelled as 'suspect loci'. We have explored the extent to which these systematic biases occur across different genomic regions, including regions known to be difficult to sequence, higher confidence regions and clinical panels. We have also confirmed the utility of our approach in an independent gold-standard reference genome and elaborated on possible scientific and clinical applications to help resolve current issues in whole genome sequencing and alignment.

For loci unaffected by systematic bias, the standard deviation between individual allelic coverage at each locus and the measured aggregate allelic fraction, both taken from the IncDB, were expected to relate to each other in accordance with Hardy-Weinberg equilibrium. For example, variants that were fixed in the cohort were expected to show no standard deviation, while variants that were closer to a aggregate allelic fraction of 50% were expected to give higher standard deviation values. However, 1.26-2.43% of autosomal loci had significantly lower standard deviation than predicted by Hardy-Weinberg equilibrium (p=0.0005), suggesting the presence of systematic bias across numerous genomic loci. At these loci, it seems that many if not most individuals must present a low-fraction allele. Because such persistent low-fraction alleles seem to be inconsistent with our understanding of human genetics, we interpret their existence as a technological artifact: a bias or systematic error in the sequencing technology itself, or perhaps in the read mapping (**Figure 2**). The impact of these suspect loci is magnified in the context of studies looking at large numbers of genomic positions, since a small percentage of this would still correspond to a high number of genomic positions affected by systematic bias. It is therefore clear that these systematic biases are of concern and deserve further attention.

Suspect loci were widespread across all sequenced chromosomal segments, but there was variability between different types of genomic region. Despite this, there was little or no depletion of suspect loci in some regions expected to have more accurate sequencing, such as exons (OR = 1.09/0.86/1.18 in data set 1/2/3) and the clinical gene panels analysed (OR range = 0.98-1.32), suggesting that greater caution in these areas is justified.

In addition we found differences in regional systematic biases between data sets. For example, the NonUnique100 region showed much greater enrichment of suspect loci in data sets 2 (OR = 3.26) and 3 (OR = 2.17), than in data set 1 (OR = 1.30). This region is known to be prone to mapping/alignment errors, so the low enrichment in this region in data set 1 and high enrichment in other regions such as the GC-rich region suggested that the vast majority of these bias-prone loci were not likely to be the result of alignment errors. On the other hand, the higher enrichment of suspect loci in NonUnique100 in data sets 2 and 3 could indicate that the different aligners used - Isaac aligner for data set 2 and 3 rather than BWA MEM - could be the cause for the susceptibility of this region to alignment-related variant calling biases.

The selection of reference genome build used also affected the quantity of suspect loci observed. For example, data set 3 had consistently lower levels of suspect loci than data set 2, especially within small homopolymers (10.3 vs. 27.1 per kb), large homopolymers (35.5 vs. 73.1 per kb), NonUnique100 regions (39.7 vs. 74.7 per kb) and Alu repeats (75.5 vs. 104.6 per kb), which are all low-complexity regions that are associated with greater difficulty aligning reads due to their repetitive nature. These data sets both used the same sequencing pipeline with the exception that data set 2 used the earlier GRCh37 reference human genome and an earlier version of the Isaac aligner (SAAC00776.15.01.27) for read alignment while data set 3 used GRCh38 and a later version of the Isaac aligner (iSAAC-03.16.02.19). This suggested that read alignment was improved with the newer genomic build, greatly decreasing suspect locus abundance, especially in low complexity regions. Both data set 2 and 3 included data from different patients within the overall set of neurology patients in the 100,000 Genomes Project, with no patients in common between them, because no patients were sequenced on both GRCh37 and GRCh38. However, the patients were phenotypically similar, so this was considered unlikely to be a confounding factor. These types of comparisons could be used to identify the strengths and weaknesses of different sequencing/alignment protocols for benchmarking, and in order to prioritise areas for improvement in the development of new sequencing/alignment tools, and could perhaps justify the use of different parameters when sequencing in the most problematic regions, such as GC-rich areas.

There was also great variability in the distribution of suspect loci within individual genes tested (**Supplementary Data 1/2**). For example, different genes within the PanelApp gene panels examined ranged from having very low proportions of suspect loci (0% out of 2,021 total loci in LIPT2 in data set 2) to very high proportions of suspect loci (36.6% out of 3,759 total loci in NPRL2 in data set 2), suggesting that some whole genes might be particularly prone to systematic sequencing biases. Both of these genes are associated with genetic epilepsy syndromes, but LIPT2 isn't affected by systematic sequencing biases at all, while these heavily affect NPRL2. Researchers focussing on specific genes could use this information to identify how much caution they need when sequencing and calling variants in their genes of interest.

It was confirmed that the suspect loci we identify do indeed manifest as consistent low-fraction alleles in the read pileup of an individual sample that was not part of our analysis cohort. In addition, called variants that occurred at suspect loci generally had lower, more irregular allelic frequencies in NA12878. Predicted suspect loci for the same sequencing pipeline as NA12878 were confirmed in NA12878, including within introns, exons, intergenic regions and NIST GIAB 'high-confidence' regions, and within called variants in clinically relevant regions such as the PanelApp disease panels tested. This confirmed that these variants were likely false positives, and the corresponding loci were prone to systematic biases across all patients, including those outside of the IncDB data set, as we had hypothesised. This demonstrates the value of the IncDB-based method described in this study for identifying these variants, since it enables them to be analysed with increased caution and potentially filtered out of variant calling results, even from patients not used in the training set, as long as the sequencing pipeline is the same across patients.

We also examined whether our IncDB-based statistical approach added additional utility beyond pre-existing quality control processes with two additional analyses. We examined whether suspect gnomAD variants were filtered out using the pre-existing quality control processes employed in the gnomAD database (**Supplementary Figure 2**). The gnomAD database is a large database of all of the variation found across a large ethnically diverse population, taken from 125,748 exomes and 15,708 genomes (Karczewski et al. 2019). Our results revealed that suspect variants were also widespread in the gnomAD database, even after gnomAD's quality control process, across allelic frequencies. We also evaluated whether suspect loci could be identified simply by using a quality threshold (**Supplementary Figure 3**). Non-suspect loci had significantly higher proportions of high quality reads, with >90% of reads having sequencing and mapping quality scores >20 in ~90% of non-suspect loci/allele combinations vs. ~5% of suspect loci/allele combinations. This demonstrated that low quality reads were more frequent among suspect loci to a large degree, suggesting that improving sequencing and alignment quality could help with decreasing these systematic biases. However, there was still significant overlap between read quality at suspect and non-suspect loci. This showed that using a quality cutoff would therefore also filter out non-suspect loci, while not fully filtering out all suspect loci, indicating that quality filters had limited utility in the absence of the IncDB-based methods presented in this study. Our analyses therefore concluded that the existing quality control procedures already in place were not sufficient to filter out the systematic biases identified by our methods, demonstrating the additional value gained by using these methods on top of pre-existing quality control procedures.

We have demonstrated the utility of IncDBs to assess the quality of clinical whole genomes of three independent cohorts sequenced by commercial and public healthcare organisations while maintaining patient anonymity. In addition to showing the utility of this approach on whole genome Illumina sequencing, IncDBs could be applied to data from different types of sequencing platforms in the future, including specific targeted, exome-sequencing and long-read technologies such as PacBio and Oxford Nanopore. Analysis of suspect loci from different sequencing platforms would allow us to identify which loci are prone to platform-specific bias, and compare systematic biases between platforms. By doing this, the accuracy of more sequencing platforms could potentially be improved by identifying error-prone loci and the source of their associated sequencing and alignment errors, enhancing variant calling in academia, industry and healthcare, and facilitating the design and utility of genomic prognostic tools that rely on this.

Conclusions

The novel IncDB-based method explored in this study, for identifying genomic regions and loci prone to sequencing and alignment errors, has strong implications for improving variant calling accuracy by identifying likely false positive variant calls. Suspect loci have been observed to occur widely across the genome, including within clinically relevant gene panels and regions considered high-confidence. The suspect loci we identify did indeed manifest as low-fraction alleles in the read pileup of NA12878, a gold-standard reference sample separate from the IncDB training sets, confirming our findings. Identifying these systematic biases enables improvements to variant calling and has the potential to reduce errors in clinical genomic testing.

Abbreviations

BAM -	Binary Alignment Map (file format)
BED -	Browser Extensible Data (file format)
cfDNA-	Cell-free DNA
ctDNA-	Circulating tumour DNA
GIAB -	Genome in a Bottle (consortium)
gnomAD-	Genome Aggregation Database
IGV -	Integrative Genomics Viewer (software tool)
IncDB -	Incremental Database
MC -	Monte-Carlo
NGS -	Next-Generation Sequencing
NIST -	National Institute of Standards and Technology (organisation)
SD -	Standard Deviation
SNPs -	Single-Nucleotide Polymorphism
SNVs -	Single-Nucleotide Variant
WGS -	Whole-Genome Sequencing

Acknowledgements

Funding support for the analysis and methods development was from the MRC Proximity to Discovery (P2D) Research Scheme and by the NIHR Sheffield Biomedical Research Centre (BRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care (DHSC). This research was also made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support. The authors gratefully acknowledge Mark Pratt, who initially invented the concept of the Incremental Database.

Genomics England Research Consortium Members and Affiliations

Ambrose J. C.[1], Bleda M.[1], Boardman-Pretty F.[1,2], Boissiere J. M.[1], Boustred C. R.[1], Caulfield M. J.[1,2], Chan G. C.[1], Craig C. E. H.[1], Daugherty L. C.[1], de Burca A.[1], Devereau, A.[1], Elgar G.[1,2], Foulger R. E.[1], Fowler T.[1], Furió-Tarí P.[1], Hackett J. M.[1], Halai D.[1], Holman J. E.[1], Hubbard T. J. P.[1], Kasperaviciute D.[1,2], Kayikci M.[1], Lahnstein L.[1], Lawson K.[1], Leigh S. E. A.[1], Leong I. U. S.[1], Lopez F. J.[1], Maleady-Crowe F.[1], Mason J.[1], McDonagh E. M.[1,2], Moutsianas L.[1,2], Mueller M.[1,2], Need A. C.[1,2], Odhams C. A.[1], Patch C.[1,2], Perez-Gil D.[1], Polychronopoulos D.[1], Pullinger J.[1], Rahim T.[1], Rendon A.[1], Rogers T.[1], Ryten M.[1], Savage K.[1], Scott R. H.[1], Siddiq A.[1], Sieghart A.[1], Smedley D.[1,2], Smith K. R.[1,2], Sosinsky A.[1,2], Spooner W.[1], Stevens H. E.[1], Stuckey A.[1], Thomas E. R. A.[1,2], Thompson S. R.[1], Tregidgo C.[1], Tucci A.[1,2], Walsh E.[1], Watters, S. A.[1], Welland M. J.[1], Williams E.[1], Witkowska K.[1,2], Wood S. M.[1,2], Zarowiecki M.[1].

1. Genomics England, London, UK

2. William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK.

Supplementary Materials

Supplementary Figure 1 - Monte-Carlo model standard deviation vs. aggregate allelic fraction density plots.

Supplementary Figure 2 - Proportion of autosomal gnomAD SNVs annotated as suspect at different gnomAD allele frequencies in data set 1.

Supplementary Figure 3 - Proportion of allelic reads that had quality scores > 20 for both sequencing and mapping in suspect and non-suspect loci (data set 1).

Supplementary Table 1 - A Fisher Exact test contingency table for suspect locus enrichment within GC-rich regions, across all autosomal chromosomes (from data set 1)

Supplementary Data 1 - List of autosomal genes associated with hereditary spastic paraplegia from the PanelApp gene panel, ranked in descending order by suspect loci percentage. Files A, B, C refer to data sets 1, 2, 3 respectively.

Supplementary Data 2 - List of autosomal genes associated with genetic epilepsy syndromes from the PanelApp gene panel, ranked in descending order by suspect loci percentage. Files A, B, C refer to data sets 1, 2, 3 respectively.

Appendix

Websites/Tools

Genome in a Bottle - High confidence BED file (NA12878, GRCh37):
ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.2/NA12878_GIAB_highconf_CG-IllFB-IllGATKHC-Ion-Solid_ALLCHROM_v3.2_highconf.bed
Accessed 17 June 2019

SAMtools (v1.9):
http://www.htslib.org/download/
Accessed 19 June 2019

Chromosome cytoband information:
http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/cytoBandIdeo.txt.gz

Accessed 8 March 2018
http://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/cytoBandIdeo.txt.gz
Accessed 27 March 2019

UCSC Table Browser (source of exon/intron/gene/RepeatMasker/Alu repeat coordinates):
https://genome.ucsc.edu/cgi-bin/hgTables
Accessed 22 March 2019

PanelApp panels list:
https://panelapp.genomicsengland.co.uk/panels/
Accessed 7 February 2019

gnomAD:
https://macarthurlab.org/2018/10/17/gnomad-v2-1/
Accessed on 26 Feb 2019 (v2.1 GRCh37)

RepeatMasker:
http://www.repeatmasker.org/
Accessed 22 March 2019

References

Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* **40**: e72. http://dx.doi.org/10.1093/nar/gks001.

Consortium T 1000 GP, The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073. http://dx.doi.org/10.1038/nature09534.

Genome in a Bottle-High confidence BED file. (NA12878, GRCh37). ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.2/NA12878_GIAB_highconf_CG-IllFB-IllGATKHC-Ion-Solid_ALLCHROM_v3.2_highconf.bed Accessed Jun 17, 2019 (Accessed June 17, 2019).

Goldfeder RL, Priest JR, Zook JM, Grove ME, Waggott D, Wheeler MT, Salit M, Ashley EA. 2016. Medical implications of technical accuracy in genome sequencing. *Genome Medicine* **8**. http://dx.doi.org/10.1186/s13073-016-0269-0.

Hasler J, Strub K. 2007. Survey and Summary: Alu elements as regulators of gene expression. *Nucleic Acids Research* **35**: 1389–1389. http://dx.doi.org/10.1093/nar/gkm044.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. 2019. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*. https://www.biorxiv.org/content/early/2019/01/30/531210.

Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, Gonzalez-Porta M, Eberle MA, Tezak Z, Lababidi S, et al. 2019. Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol* **37**: 555–560. http://dx.doi.org/10.1038/s41587-019-0054-x.

Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, Liu Y, Chen X, Newman S, Nakitandwe J, et al. 2019. Analysis of error profiles in deep next-generation sequencing data. *Genome Biol* **20**: 50. http://dx.doi.org/10.1186/s13059-019-1659-6.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nature Biotechnology* **29**: 24–26. http://dx.doi.org/10.1038/nbt.1754.

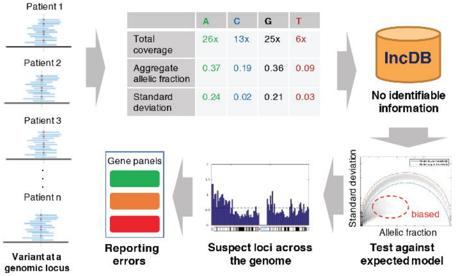
Sandmann S, de Graaf AO, Karimi M, van der Reijden BA, Hellström-Lindberg E, Jansen JH, Dugas M. 2017. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Sci Rep* **7**: 43169. http://dx.doi.org/10.1038/srep43169.

Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok P-Y, Schaefer C, Risch N. 2014. Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res* **24**: 1734–1739. http://dx.doi.org/10.1101/gr.168393.113.

Xiao C, Zook J, Trask S, Sherry S, the Genome-in-a-Bottle Consortium. 2014. Abstract 5328: GIAB: Genome reference material development resources for clinical sequencing. *Cancer Research* **74**: 5328–5328. http://dx.doi.org/10.1158/1538-7445.am2014-5328.

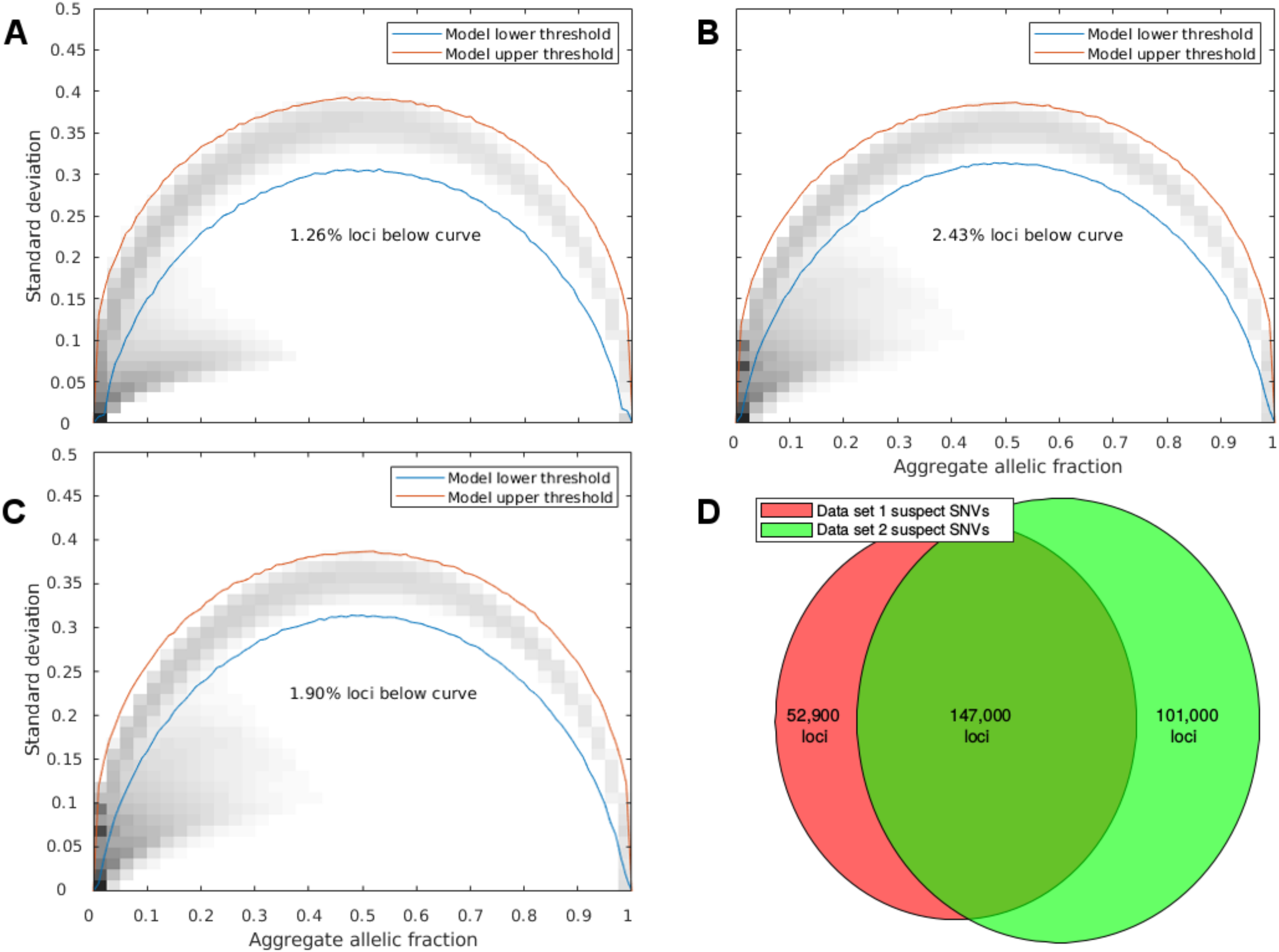
Xu C. 2018. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput Struct Biotechnol J* **16**: 15–24. http://dx.doi.org/10.1016/j.csbj.2018.01.003.

Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. 2014. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* **32**: 246–251. http://dx.doi.org/10.1038/nbt.2835.

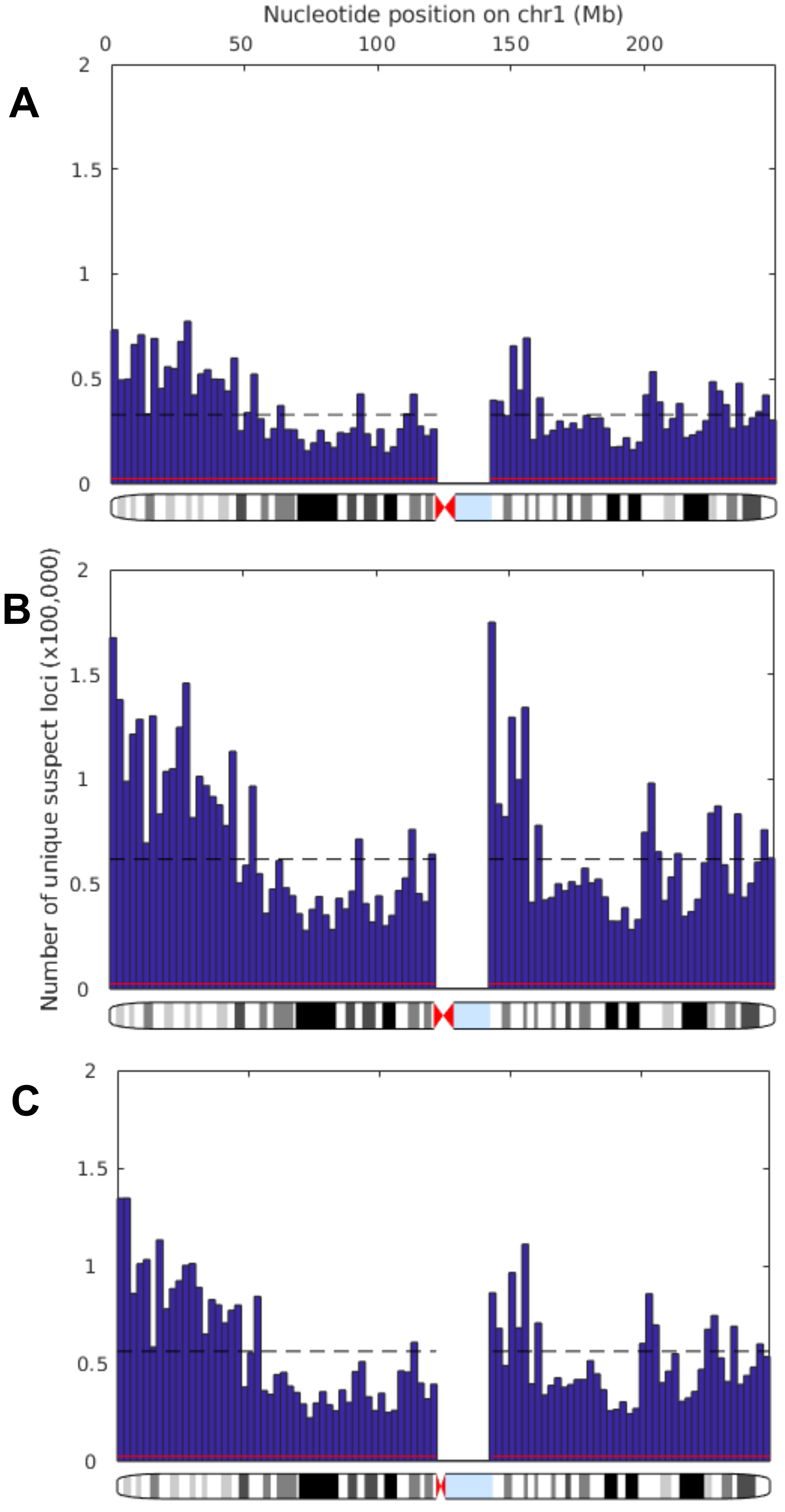


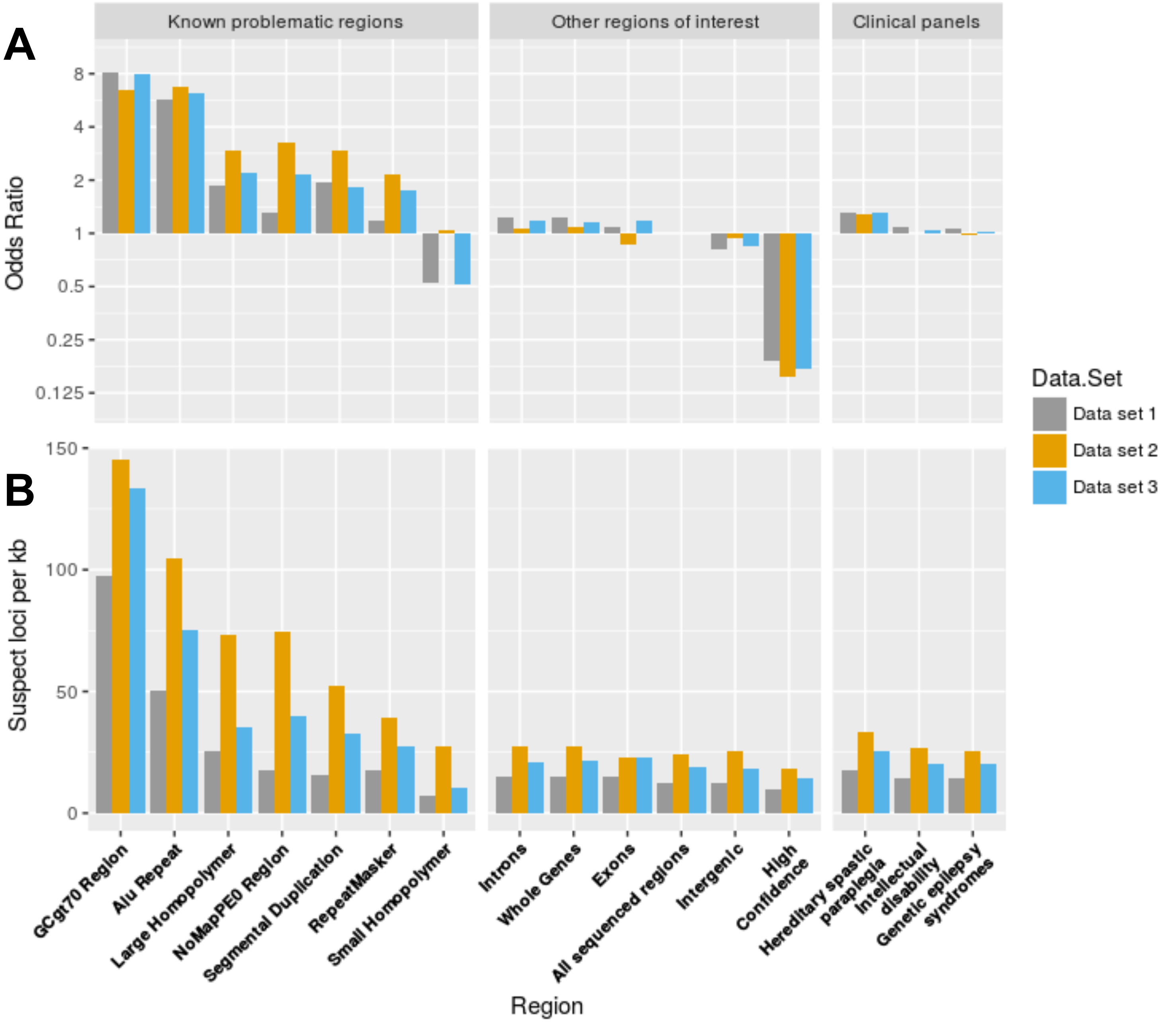
	A	C	G	T
Total coverage	26x	13x	25x	6x
Aggregate allelic fraction	0.37	0.19	0.36	0.09
Standard deviation	0.24	0.02	0.21	0.03

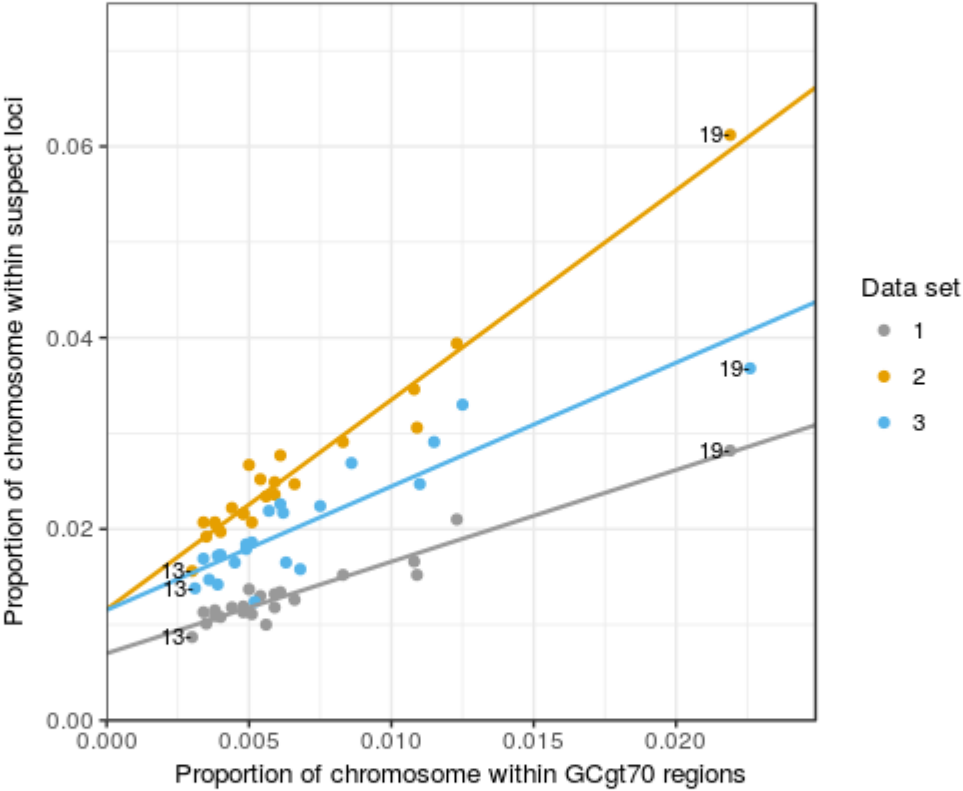
IncDB
No identifiable information
Standard deviation
Allelic fraction
biased
Test against expected model
2
1.5
1
0.5
0
Suspect loci across the genome
Gene panels
Reporting errors

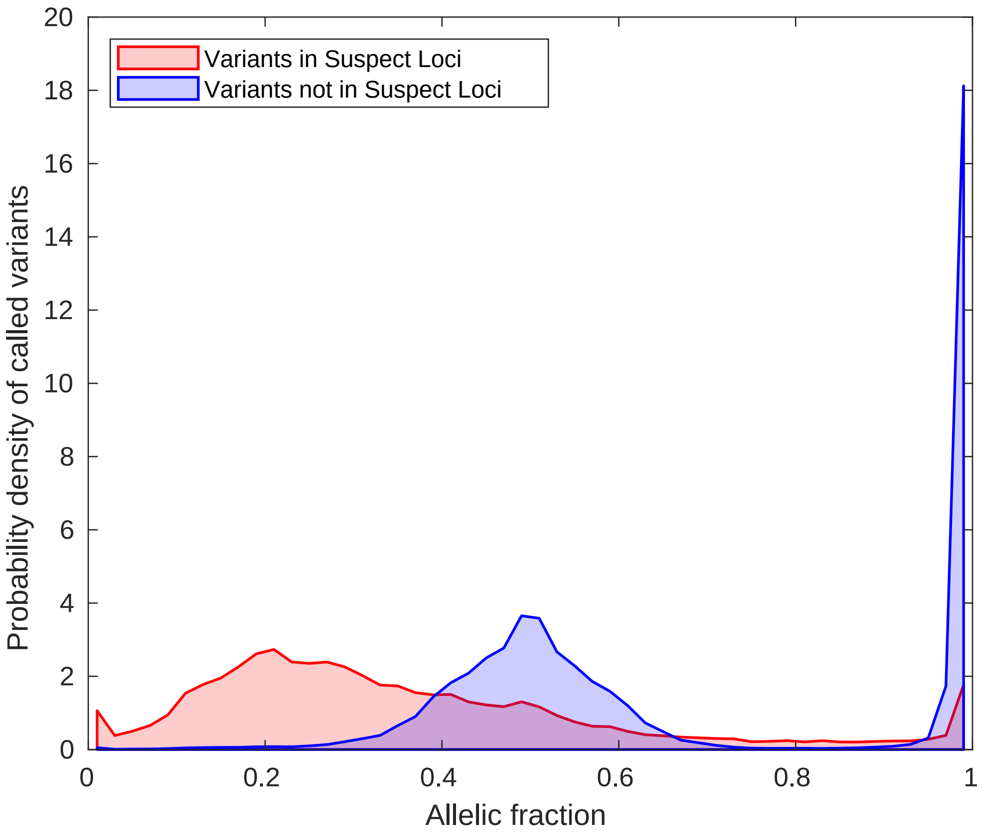


A
B
C
D
Model lower threshold
Model upper threshold
Standard deviation
Aggregate allelic fraction
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.26% loci below curve
2.43% loci below curve
1.90% loci below curve
Data set 1 suspect SNVs
Data set 2 suspect SNVs
52,900 loci
147,000 loci
101,000 loci


Nucleotide position on chr1 (Mb)
0
50
100
150
200
A
B
C
Number of unique suspect loci (x100,000)
2
1.5
1
0.5
0


A
B
Known problematic regions
Other regions of interest
Clinical panels
Odds Ratio
8
4
2
1
0.5
0.25
0.125
Suspect loci per kb
150
100
50
0
Data.Set
Data set 1
Data set 2
Data set 3
GCgt70 Region
Alu Repeat
Large Homopolymer
NoMapPE0 Region
Segmental Duplication
RepeatMasker
Small Homopolymer
Introns
Whole Genes
Exons
All sequenced regions
Intergenic
High Confidence
Hereditary spastic paraplegia
Intellectual disability
Genetic epilepsy syndromes
Region


Proportion of chromosome within suspect loci
0.00
0.02
0.04
0.06
Proportion of chromosome within GCgt70 regions
0.000
0.005
0.010
0.015
0.020
19
19
19
13
13
13
Data set
● 1
● 2
● 3


20
18
16
14
12
10
8
6
4
2
0
Probability density of called variants
0
0.2
0.4
0.6
0.8
1
Allelic fraction
Variants in Suspect Loci
Variants not in Suspect Loci