Reliability of genomic variants across different next-generation sequencing platforms and bioinformatic processing pipelines

Susanne Gerber (sugerber@uni-mainz.de)
Johannes Gutenberg Universitat Mainz https://orcid.org/0000-0001-9513-0729
Stephan Weißbach
Johannes Gutenberg Universitat Mainz
Stanislav Jur`Evic Sys
Johannes Gutenberg Universitat Universitatsmedizin
Charlotte Hewel
Johannes Gutenberg Universitat Universitatsmedizin Humangenetics
Hristo Todorov
Johannes Gutenberg Universitat Universitatsmedizin https://orcid.org/0000-0003-2734-7701
Susann Schweiger
Johannes Gutenberg Universitat Universitatsmedizin
Jennifer Winter
Johannes Gutenberg Universitat Universitatsmedizin
Markus Pfenninger
Senckenberg Gesellschaft fur Naturforschung https://orcid.org/0000-0002-1547-7245
Ali Torkamani
Scripps Translational Science Institute https://orcid.org/0000-0003-0232-8053
Doug Evans
Scripps Translational Science Institute
Joachim Burger
Johannes Gutenberg Universitat Mainz
Karin Everschor-Sitte
Johannes Gutenberg-Universitat Mainz Fachbereich Physik Mathematik und Informatik
Helen May-Simera
Johannes Gutenberg Universitat Mainz

Research article

Keywords: Next-generation sequencing (NGS) technologies, platform-biases, Genome-wide association studies (GWAS), healthy aging, Illumina, Wellderly, Longevity, Complete Genomics, Aging, conserved loci

Posted Date: August 7th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-50691/v1

Version of Record: A version of this preprint was published on January 19th, 2021. See the published version at https://doi.org/10.1186/s12864-020-07362-8.

Abstract

Background

Next Generation Sequencing (NGS) is the fundament of various studies providing insights into questions from biology and medicine. Nevertheless, integrating data from different experimental backgrounds can introduce strong biases. In order to methodically investigate the magnitude of systematic errors, we performed a cross-sectional observational study on a genomic cohort of 99 subjects each sequenced via (i) Illumina HiSeq X, (ii) Illumina HiSeq and (iii) Complete Genomics. Consequently, we systematically analyzed the heterogeneity between the sequencing cohorts with respect to genomic annotation and common filter criteria like minimum allele frequency (MAF).

Results

The number of detected variants/variant classes per individual was highly dependent on the sequencing technology. We observed a statistically significant overrepresentation of variants uniquely called by a single platform which indicates potential systematic biases. These variants were enriched in low complexity genomic regions and simple repeats. Furthermore, estimates of allele frequency were highly discrepant for a subset of variants in pairwise comparisons between different sequencing platforms. Applying common filters – such as MAF 5% and HWE- greatly reduced the heterogeneity between cohorts but still left discrepancies of several thousand variants after filtering.

Conclusion

We provide empirical evidence of systematic heterogeneity in variant calls between alternative experimental and data analysis setups. Our results highlight the potential benefit of reprocessing genomic data with harmonized pipelines when integrating data from different studies.

Background

From sequencing, over variant calling to subsequent statistical analysis – variation can be introduced at any step of the genomics workflow. Each sequencing technology produces its own imprint of systematic biases, imposing one of the most crucial bottlenecks in genomics research. Sequencing is susceptible to high- and low-GC regions, as well as long homopolymer runs. Repetitive regions were also a main cause of uncertainty when assessing trio-samples for inconsistent mendelian errors [1]. Furthermore, Lam and colleagues demonstrated the discrepancy in sequencing accuracy between platforms on the basis of one particular individual, revealing tens of thousands of platform-specific calls [2]. Another important aspect that can contribute to variance between studies is due to heterogenous bioinformatic pipelines. To this end, several studies have been conducted. For example, O'Rawe et al., sequenced exomes and whole genomes from 15 individuals. Subsequent bioinformatic analysis with different pipelines found generally low concordance between data sets [3]. In 2015 three different studies analyzed the NA12878 sample from the 1000 Genomes Project with slightly alternate setups and pipelines, yielding slightly different

recommendations [4-6]. More recently, in 2019 Kumaran et al., and Chen et al., both re-examined whole-exome sequencing data (WES) from NA12878, although the latter also compared whole-genome sequencing (WGS) [7, 8]. Hwang et al., compared both the European NA12878 and the African NA19240 samples from the 1000 Genomes Project. They found a pipeline consisting of BWA-MEM and subsequently the Genome Analysis ToolKit with the Haplotype Caller to be sufficient to reliably detect variants in most regions, except for rare variants and difficult regions, as for example simple repeats [9]. These studies provide evidence that processing pipelines and variant calling algorithms directly contribute to the heterogeneity observed when comparing the results from different sequencing experiments.

While the previous examples already offer important insights over a multitude of sequencing platforms and bioinformatic pipelines over the years, they also highlight the inconsistencies. Most of all, a vast majority of the previous studies were conducted on a single reference individual or at most only on a small number of individuals, not allowing to achieve statistically-significant conclusions that can be generalized and translated to large cohorts - and making it difficult to apply any filtering for which a whole sequencing cohort would be needed. However, a common strategy prior to conducting Genome-Wide Association Studies (GWAS), for example, is to apply several sequencing cohort-specific filters to reduce variance, such as missingness, minor allele frequency (MAF) and departure from Hardy-Weinberg-Equilibrium filters (HWE) [10]. To the best of our knowledge, this has not been considered in studies formally comparing a multitude of sequencing platforms and pipelines for human data. Hence, quantifying the impact of these filters with respect to the actual concordance rates remains an open problem. Additionally, most analyses gravitate towards using the same references, as for example provided by the Genome In A Bottle (GIAB) Consortium [11, 12]. This is a repository of well described, widely accepted, gold standard variants. Such high-confidence variants are valuable and indispensable for benchmarking different sequencing technologies and bioinformatic pipelines. However, they also fail to represent the full spectrum of sequence cohort heterogeneity. Therefore, the assessment of further non-gold standard data sets can serve to solidify observations made on the gold standard variants and to uncover potential differences.

To this end, we we re-analyzed the vcf-files from an extensive European reference data set from a cohort of 99 individuals associated with a healthy aging phenotype [13]. All subjects were sequenced three times via different technologies, namely (1.) Illumina HiSeq X (HSX), (2.) Illumina HiSeq (MOL) and (3.) Complete Genomics (CG). In order to asses sequencing cohort heterogeneity, we determined the number of concordant and discordant variants between the three platforms, we identified the genomic regions and patterns where these variants intersect and we investigated the impact of common QC filters on sequencing cohort heterogeneity.

Results

In the current study, we re-examined the genomes of 99 individuals from the Wellderly project [13]. The wellderly phenotype describes individuals over the age of 80 that present without any known chronic

diseases and do not take any chronic medication. Each individual had already been sequenced with three different next-generation sequencing platforms and variant calling had already been performed via three different bioinformatic pipelines (Fig 1a). Therefore, the raw data in our current study were the resulting vcf files. We only considered variants, i.e. single nucleotide polymorphisms (SNP) and insertion/deletion polymorphisms (InDel) with filter tag 'PASS', as defined by the variant calling pipeline, without equalizing the 'PASS' criteria between different setups. We believe that this is the most realistic approach when comparing datasets generated by different methods. Additionally, we left-aligned the variants and split multiallelic variants in consecutive blocks to equalize the variant annotation between the different sequencing cohorts.

Throughout the study, we use the name of the sequencing platform (HSX, MOL or CG) to describe the three different experimental setups consisting of the sequencing technology and the respective mapping and variant calling algorithm. The three different genomic data sets obtained in this way are referred to as sequencing cohorts.

Our comparative analysis consisted of three different investigations. First, we determined the absolute number of variants in the respective sequencing cohort and all of their converging sets. Subsequently, we focused on the subset of variants within different genomic regions, such as introns and exons or repetitive elements. Finally, we focused on variant sites with highly differing allele frequencies between the sequencing cohorts in order to investigate systematic "miscalls" at the cohort-level.

Comparison of concordant variants between the three platforms

In the first step of the analysis, we estimated the number of variants predicted for each experimental setup and assessed the concordance between the three methods (pipeline details are summarized in Fig. 1b).

HSX was associated with the highest average number of variants, followed by MOL and then CG (Fig. 2a). Altogether, an average of 3,332,799 variants were detected per individual by all three platforms which corresponds to 79.4% of HSX variants, 88.2% of MOL variants and 89.4% of CG variants (Fig. 2a, b). We tested for statistical over- and underrepresentation of observed variant calls in the different sets of the Venn-diagram (Fig. 2b) with a Monte Carlo simulation approach. The intersection between all sequencing techniques was slightly higher than expected by chance (observed mean 3,332,789 vs. expected mean 2,921,107, $p < 0.001$, Fig 2c), increasing the confidence in these calls. However, the variants unique to each platform were highly overrepresented (MOL: 2.34, HSX: 2.98 and CG: 4.37 times more unique variants than expected by chance) at the expense of the observed number of variants called by exactly two platforms (Fig 2c). This indicated the presence of platform specific systematic sequencing errors.

Next, we investigated the concordance between different platforms more thoroughly by determining the average number and composition of variants detected by all three, exactly two, at least two, or in just one experimental setup. The proportion of InDels in the intersection of all platforms (6.84%) was lower compared to each individual platform (Fig. 2d). We observed a fairly similar pattern in the distribution of

variants found by at least two platforms where the proportion of InDels varied between ~7-9% (Fig. 2e). The analysis of variants detected under only one experimental setup revealed that HSX was associated with the highest number of unique variants, followed by CG and then MOL (Fig. 2f). Interestingly, the proportion of InDels was higher than SNPs in the case of unique variants detected by HSX only. Finally, we observed a similar pattern for the concordant variants detected by exactly two platforms (Fig. 2g). Specifically, the proportion of concordant InDels between HSX and CG was higher than the proportion of SNPs. Furthermore, the general overlap between HSX and the other two sequencers was greater than the overlap between CG and MOL (Fig. 2g).

Apart from the average number of variants per individual, we were also interested in investigating the total number of variant sites as a function of varying cohort size. For this, we randomly selected a subset of individuals while increasing the sample size from 1 to 99 and estimated the total number of variants in each resulting cohort. The results for all three platforms were similar and are shown in S. Fig 1. As expected, the number of called variants steadily increased with increasing sample size. However, the function quickly plateaued after applying the MAF filter. Consequently, increasing the sample size of the cohort above 30 individuals was not associated with a considerable increase in the number of detected variant sites.

Distribution of variants along the genome

One way of estimating the reliability of predicted variants is to correlate them to evolutionary highly conserved elements in the genome – such as regions with high PhastCons scores. These scores, ranging between zero and one, were calculated for the human genome (hg19) on the basis of 99 vertebrates, and stand for the probability of conservation of a nucleotide in the genome. A high score indicates a high rate of conservation[14]. These genomic positions are known to be under strong purifying selection, which makes the occurrence of a variant improbable. Therefore, variants detected within such conserved regions are most likely erroneous. Since no recommendations about an appropriate cut-off value have been described in the literature, we chose a threshold of PhastCons score > 0.8. This allowed us to retain conserved regions with high confidence without being overly restrictive.

Interestingly, relative proportions of PhastCons variants were virtually the same for each experimental setup (~2.4-2.6%). In the case of variants called by solely one platform, this remained the same for MOL and CG with ~1.7%, whereas HSX found 4.15% of variants in PhastCons regions. We consider these variants to be candidates for false positives, since we would expect less variation in highly conserved genomic regions. For details, see Supplemental Table S Table 1.

Additionally, we were interested in regional effects of variant calling across the genome. Therefore, we compared the genome-wide distribution of variants in particular areas, such as exonic, intergenic and intronic areas. The proportional difference in the distribution of variants between the platforms was negligible with ~2% of variants found in exonic, ~56% variants in intergenic and ~42% of variants found in intronic regions by all machines. The observed number of variants significantly differed from the expected value for all three platforms based on the relative length of exons, introns and intergenic regions

and assuming an equal distribution of variants along the genome ($p<0.0001$, chi-squared test). Notably, the observed number of variant sites in exons was lower than the expected quantity (observed mean 84867 vs. expected mean 126748 for HSX; observed mean 79071 vs. expected mean 114137 for MOL; observed mean 79551 vs. expected mean 112639 for CG).

With regard to the mean number of variants found only by one platform, the HSX-cohort had a mean number of 197865 unique variants in intergenic regions, whereas MOL had 55679 unique variants and CG had 109274 unique variants in intergenic regions. For intronic regions, a mean count of 142209 variants per sample was found by HSX only, MOL had 37160 mean unique variants and CG had 78368 mean unique variants in introns. For exons there was a mean number of 2762 of variants detected solely by HSX, whereas MOL had on average 436 unique variants and CG had 962 unique variants in exonic regions per sample. For details refer to Supplemental Table S Table 1.

Distribution of variants in repetitive genomic regions

Regions that are frequently excluded from the analysis of genomic variants are repetitive elements and GC rich regions, because they are known to accumulate sequencing errors[1]. In order to calculate the impact of such regions, we proceeded as follows (Fig. 1c): We obtained the RepeatMasker annotation for the following classes of repetitive elements, Alu elements, long interspersed nuclear elements (LINE), low complexity regions, long terminal repeats (LTR) and simple repeats. Then, we calculated the ratio of variants detected exclusively by one platform to the total number of variants detected by the same platform for each type of repetitive element. We refer to this as the observed value of the unique variant rate in the respective RepeatMasker region. The unique variant rate can generally indicate either higher sensitivity or increased false positive rate. However, we believe that this measure more likely corresponds to the false positive rate of the variant calling method in repetitive low complexity regions. In order to estimate the expected value of the unique variant rate, we randomly selected genomic regions of the same length for each type of RepeatMasker repetitive element and calculated the rate of unique variants as described above.

Fig. 3 depicts the results for all three pipelines. The expected unique variant rate was ~12.5% for HSX, ~4.7% for MOL and ~19.2% for CG. The observed unique variant rate in low complexity regions and simple repeats was considerably higher than the expected value for all three platforms. The observed unique variant rate was around 48% in low complexity regions and approximately 54% in simple repeats for both HSX and CG (Fig. 3a, 3c). These values were lower for MOL with an observed rate of 19% in low complexity regions and 25.6% in simple repeats (Fig. 3b). This finding is however likely due to the fact that MOL generally detected the lowest number of unique variants compared to the other platforms (cf. Fig. 2f). Interestingly, however, the absolute number of variants detected in low complexity regions and simple repeats was higher for MOL than both HSX and CG (Fig. 3d).

Furthermore, we examined the general length distribution of InDels and their proportion overlapping the repetitive regions defined by RepeatMasker. Notably, CG included the highest number of 1-bp insertions and deletions while MOL was the only sequencer with more 1-bp deletions than 1-bp insertions (Fig. 4a-

c). Additionally, the proportion of InDels in simple repeats and low complexity regions was well above 0.5 for all sequencers, whereas a value of ~0.2 would be expected. In contrast, InDels were less abundant than expected in LINE and LTR regions (Fig. 4d-f).

Factors with a significant impact on the concordance of called variants

We performed a Poisson regression analysis in order to investigate factors which potentially influence the concordance between the different pipelines. The response variable was the number of pipelines which detected each variant. The model included the four categorical predictors genomic region, repetitive element, MAF for the European cohort from the 1000 Genomes Project and positional conservation. Genomic regions included intergenic, introns and exons (reference category). Repetitive elements were obtained from the RepeatMasker annotation including simple repeats, LINE, LTR, low complexity repeats and non-repetitive elements (reference category). Variants were classified as common if the MAF in the 1000 Genomes Project was >5% (reference category), low if MAF was between 0.5% and 5%, rare if MAF was below 0.5% and not detected if the variant was not present in the 1000 Genomes Project. Positions were considered conserved if the PhastCons scores was above 0.8 (reference category) and non-conserved otherwise.

Results from the regression analysis for SNPs are depicted in Fig. 5a. The factors which showed the strongest negative effect on the concordance between different pipelines were simple repeats ($p<2x10^{-16}$), low complexity regions ($p<2x10^{-16}$) and novel variants not detected in the 1000 Genomes Project ($p<2x10^{-16}$). The concordance of detected variants was also significantly worse for LINEs relative to non-repetitive elements as well as variants with low and rare MAF compared to common variants. In contrast, LTRs and introns were associated with a significantly improved concordance relative to non-repetitive elements and exons, respectively. The impact of intergenic regions and positional conservation was not statistically significant.

All factors that we included in the regression analysis significantly influenced the concordance between the pipelines for InDels (Fig. 5b). Novel variants not detected in the 1000 Genomes Project were associated with the strongest negative impact on the concordance between the three pipelines. Furthermore, intergenic regions and introns relative to exons, simple repeats and low complexity regions relative to non-repetitive regions, rare variants and non-conserved genomic positions all significantly contributed to reduced concordance. Conversely, the agreement between the pipelines was significantly better for variants with a low MAF relative to common variants as well as in LINE and LTR elements compared to non-repetitive elements. Interestingly, this correlates with the observed rate of InDels in LINE and LTR regions being lower than the expected average (cf. Fig. 4d-f).

Comparison of allele frequency estimates between different platforms

In the final step of the analysis, we investigated whether the estimated allele frequency of variants was comparable between each pair of platforms for variants detected by both respective platforms. Workflow details are summarized in Fig. 1d.

Fig. 6a-c depicts the allele frequency correlation between the different experimental setups for chromosome 22, while the results for the remaining chromosomes are shown in Supplemental Figure S. Fig. 2. Generally, each pair of sequencing technologies demonstrated a strong consensus in estimated allele frequency indicated by the high density of data points around the diagonal line of the plot. However, numerous variants were associated with a very high allele frequency for one platform while simultaneously having a very low occurrence in the other platform.

In order to quantify the variants with strong disagreement in estimated allele frequency between the different setups, we created a subset with variants having an allele frequency difference >0.8 for each pair of sequencers. We chose this threshold to be able to focus on variants with a very high occurrence in one platform and a very low occurrence in the other. Such extreme discrepancies could point to systematic rather than random errors. We observed an average of 14746.26 variants with allele frequency difference >0.8 for HSX and CG, closely followed by the average for CG and MOL, whereas this value was much lower for HSX and MOL (Fig. 6d).

Next, we examined how many of the variants with highly diverging estimated allele frequency correspond to variants annotated in GWA studies which could illustrate potential problems in downstream analysis. The average number of such variants for the comparison of HSX and CG as well as CG and MOL was around 80 (Fig. 6e). Importantly, we could reduce this number to 0 by applying both the MAF and HWE filter to the data (Fig. 6f).The majority of these annotated GWAS vairants with very divergent estimated allele frequency were filtered out already after applying only the MAF filter (Fig 6g), whereas the impact of the HWE filter alone was much more moderate (Fig. 6h). Consequently, these results indicate that is seems prudent to combine both filters as neither one of them managed to remove these highly discordant variants completely when applied individually.

Discussion

The reliability of genomic variants, especially when merging multiple cohorts, is still not fully evaluated, both due to the considerable heterogeneity in laboratory protocols and variant-calling pipelines. Furthermore, multiple studies have led to conflicting estimates of accuracy and of preferred analysis pipelines for sequencing data, and challenges remain in benchmarking variant call datasets [3-6, 8, 9, 16]. In contrast to previous studies, which were only able to compare a handful of genome or exome sequences, we were given the unique opportunity to analyze a larger cohort of 99 individuals. This real-world data set allowed us to perform a practical comparison of the consistency of different sequencing and variant calling methods.

To begin with, the average number of variants consistently detected throughout all three experimental setups accounted for approximately 89% of MOL and CG variants and only ~79% of HSX variants. A Monte Carlo simulation-based statistical test revealed that the observed number of variants called by all platforms was significantly higher than expected by chance. Nevertheless, the observed number of variants unique to each method was also significantly increased relative to the expected quantity which

hints at platform-specific biases. This finding also implies that the choice of sequencing platform and the following bioinformatic processing strategy directly affect variant calling. This might in turn account for between-study heterogeneity observed in downstream analyses such as GWAS [17-19].

Upon inspection of the different subtypes of variants, we observed that all three experimental setups demonstrated a greater concordance in SNPs than in InDel detection. Conversely, the absolute and relative quantity of InDels varied strongly between the platforms. Notably, HSX detected considerably more InDels compared to both MOL and CG. Furthermore, InDels also accounted for the majority of variants unique to HSX. InDels are a class of variants that are particularly prone to amass sequencing errors. These can occur either during the PCR amplification step, the sequencing reaction, or during the alignment step, because the aligner may have difficulty placing a single insertion/deletion event in a highly repetitive stretch of the genome [20]. In line with this, we observed a strong enrichment of InDels in simple repeats and in low complexity regions, which is characteristic for this type of mutation [21]. Nevertheless, the difference in the overall number of insertions and deletions was specific to each experimental setup, thereby indicating a systematic bias rather than random error. It is important to mention that the proportion of variants detected in intronic, exonic and intergenic regions was very consistent between all three platforms. However, absolute numbers varied considerably. In exonic regions alone, which are often the most relevant regions on the functional level, the difference of unique variants for the HSX cohort still remained at about 2000 unique variants per person, which is far above the expected value of mutations a single person should have, potentially indicating reduced specificity. In line with this, Conrad and colleagues estimated that approximately 1000 mutations per diploid genome can be introduced due to somatic mutations, which is already less than the discordance we observed in exons alone [22].

Since we did not have an independently validated set of high confidence variants, we were not able to formally assess the performance of different methods in terms of for example sensitivity or specificity. One surrogate measure that we used to evaluate the false positive rate, was the proportion of variants unique to only one platform detected at highly conserved genomic positions (PhastCons score >0.8). HSX was associated with the highest proportion of such unique variants potentially indicating reduced specificity. However, it should also be mentioned, that HSX detected more unique variants in general and this was also the most recently sequenced data set. Consequently, it remains to be elucidated if HSX is more error prone or if the higher number of unique variants is an indicator of increased sensitivity.

We also evaluated the unique variant rate in repetitive genomic elements and compared it to what we would expect by chance. The observed values were much higher than the expected rates in low complexity regions and simple repeats for all platforms. Furthermore, Poisson regression revealed that the concordance of InDels between the three setups in these two types of repetitive elements was significantly worse than non-repetitive regions. Therefore, it might be reasonable to exclude variants detected in such regions from GWAS and meta-analyses as these might lead to false conclusions.

Another potential major issue that we discovered in our analysis were the large discrepancies in the estimated allele frequencies between pairs of sequencing platforms for a subset of the detected variants. This finding has direct implications for inferences drawn from GWAS and is yet another example of how the choice of sequencing and bioinformatic processing methods might account for systematic between-study heterogeneity. For instance, in a classic case-control genetic study to identify disease-associated variants, the effect size usually reported is the odds ratio which is calculated based on allele frequencies [23]. Therefore, inaccurate estimates of allele frequencies lead to reporting biased odds ratios. Moreover, variants with a MAF <0.05% as obtained from the European cohort of the 1000 Genomes Project were associated with a significantly worse concordance between the different pipelines in a Poisson regression analysis. Importantly, applying a combination of the standard filters MAF and HWE removed the variants with large discrepancies in the estimated allele frequency. However, removing variants with a MAF <5% prohibits the investigation of uncommon or very rare variants.

As previously mentioned, our experimental setup is not suitable for making a definitive statement about the reliability of individual variants. This issue could be tackled by including an artificially created synthetic reference in sequencing experiments. Such references would have to include regions with high variations or regions which are difficult to align in order to assess challenging base-calls. However, we would like to point out that the goal of our study was not to benchmark methods but to provide practical evidence that the sequencing and bioinformatic methods introduce systematic between-study variation. Another potential drawback of our study is that the used sequencing platforms have a legacy status and that the bulk of new data generated today stems from different platforms. Nevertheless, there is still a considerable number of recently published studies, which make use of older sequencing data from a wide variety of sources [24-26] and we believe this will continue to be the case. Common incentives for re-analyzing genomic cohorts include re-mapping reads to a new reference genome version [27], periodic re-analysis of disease cohorts to diagnose more patients [28] or large meta-GWAS [29], aiming to achieve statistically significant results by increasing sample sizes.

Conclusions

In our study based on a cohort of 99 subjects sequenced with three different platforms, we demonstrated a considerable discordance between the sequencing technologies and their respective bioinformatic processing pipelines. In contrast to previous studies, which have focused extensively on individuals or smaller groups, our approach using a larger cohort of 99 individuals provides a direct insight into the challenges that arise when integrating data from different sources. While variants that are uniquely detected by a single setup might point to increased sensitivity, they might also be the results of systematic errors. In agreement with previous reports, our study also highlighted the complexity of correctly calling InDels especially in tandem repeats and low complexity genomic regions. Our setup does not allow us to specifically trace the source of these discrepancies. However, it is reasonable to believe that both experimental factors such as the sequencing technology as well as the choice of data analysis method considerably contribute to heterogenous results. The ever-growing amount of available whole genome sequencing data underlies the need for reliable sequencing platforms and respective

bioinformatic processing pipelines Back in 2001, Ioannidis and colleagues suggested that meta-analyses of genetic studies would greatly benefit from including individual data instead of analyzing summary statistics [17]. At the time, this seemed unrealistic due to the huge collaborative data-sharing effort necessary to achieve such goal. Currently, however, it is common practice to make raw data publicly available. While differences in the choice of sequencing platform cannot be abolished once the data have been generated, it seems prudent to reprocess raw data in a unified manner prior to conducting a meta-analysis or generally integrating data from different sources in genomic investigations. This approach would ensure more reliable and reproducible results by removing biases originating from discrepancies in the bioinformatic processing pipelines.

Methods

Genomes investigated

A cohort of 99 subjects with the so-called “wellderly phenotype” was investigated. The wellderly phenotype refers to individuals older than 80 years who do not have any known chronic diseases and do not receive chronic medication. The subjects in our current study were sampled from a larger cohort described in Erikson et al., 2013 [13]. All individuals were sequenced three times with different sequencers: (i) Complete Genomics, (ii) Illumina HiSeq X and (iii) Illumina HiSeq with TruSeq Synthetic Long-Read DNA Library Prep Kit for long reads (Fig. 1a). All multiple nucleotide polymorphisms (MNP) were decomposed into consecutive SNP since the Illumina HiSeq data set did not encode them as such. An MNP is a variant that extends over several base pairs and has sequential bases that differ from the reference genome. We used VT decompose to split each MNP into a sequence of SNP.

Opensource tools used for analysis

BCFtools 1.9 (http://samtools.github.io/bcftools/bcftools.html) was a standard analysis tool in this study. We used BCFtools for filtering (QUAL, MAF, HWE and region-based filter), querying (i.e. creation of a bed-file) and intersecting. BCFtools stats was used to retrieve information about the vcf-files (number of variants/SNP/InDel, Ti/Tv ratio, InDel length distribution). In addition, we worked with the plugin fill-tags to update variant tags in vcf-files (i.e. AF, HWE). In order to join all individual vcf-files to a single cohort-level file, we used BCFtools merge. Since vcf is a reduced file format, which only includes the differences to the reference genome of a given genetic sequence, we assumed, that a missing variant corresponds to a reference genome type at this position.

Tabix 1.7.2 (http://www.htslib.org/doc/tabix.html) was used to create index files for gziped vcf-files and is needed by BCFtools for processing.

Bedtools 2.27.1 (https://bedtools.readthedocs.io/en/latest/) was used to intersect a bed with a vcf-file (Bedtools intersect). To pseudo-randomly assign every entry of a bed-file to a new position on the chromosome we worked with bedtools shuffle. To ensure reproducibility we chose 27111992 as a seed value.

With Vt 0.57721 decompose (https://github.com/atks/vt) we split MNP into consecutives SNP. This was necessary because Illumina HiSeq did not include MNP.

With **BigWigToWig** unversioned [downloaded: 23.01.19] (https://www.encodeproject.org/software/bigwigtowig/) bigwig-files were converted to wig-files and afterwards with Bedops 2.4.35 (https://academic.oup.com/bioinformatics/article/28/14/1919/218826) to bed-files.

We used **R 3.6.1** (https://www.r-project.org/) for basic calculations, data manipulation and as a framework for plotting using ggplot2 3.2.1 (https://ggplot2.tidyverse.org/). With dplyr 0.8.3 (https://dplyr.tidyverse.org/) we filtered variants according to their allele frequency difference (Fig. 1d). The Venn diagrams were plotted with VennDiagram 1.6.20 (https://cran.r-project.org/web/packages/VennDiagram/VennDiagram.pdf). In order to manipulate the data frames according to the needs of VennDiagramm, we employed the packages tidyverse 1.3 (https://www.tidyverse.org/), hrbrthemes 0.6.0 (https://cran.r-project.org/web/packages/hrbrthemes/index.html), tm 0.7.6 (https://cran.r-project.org/web/packages/tm/tm.pdf) and proustr 0.4.0 (https://cran.r-project.org/web/packages/proustr/index.html). Whenever possible we worked with GNU parallel 20161222 (https://www.gnu.org/software/parallel/) to minimize the run-time.

Tracks

We aimed to identify regions which differ more between the sequencers and post-sequencing algorithmns. We used several annotation tracks which can be downloaded or easily applied. If not already available, we converted the track into the bed-file format and processed our files with it.

Repeatmasker: Alu, LINE, low complexty, LTR, simple repeats (http://www.repeatmasker.org/genomes/hg19/RepeatMasker-rm405-db20140131/hg19.fa.out.gz 10.02.2019)

PhastCons 100 way: highly conserved regions with a score > 0.8 (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/PhastCons.conserved.bed 10.02.2019)

Exon, Intron, Intergenic (https://genome.uscs.edu/cgi-bin/hgTables 16.07.19)

Additionaly we created bed-files of the following features:

- Minorallelefrequency > 0.05
- Hardy-Weinberg-Equilibrium < 0.05

Minor Allele Frequency

The Minor Allele Frequency (MAF) is a commonly used filter for downstream analysis of genetic data. The individual data of all samples were merged into a cohort file. Since vcf-format is sparse and only contains positions that differ from the reference genome, we set 'missing positions' as reference. After adding the MAF-tag and filtering for 5% a bed-file with the genetic positions of the variants above the threshold was created for each sequencer. These bed-files were used to filter each individual vcf-file.

Hardy-Weinberg-Equilibrium

The Hardy-Weinberg-Equilibrium (HWE) is a theoretical measure for the derivation of variants from the expected Mendelian heritage. It has been shown that HWE is a good instrument to minimize the effects of genotype errors [30]. Here, we used an HWE-threshold of $p > 0.05$. Our approach was identical as described in the section "Minor Allele Frequency".

PhastCons – highly conserved regions

UCSC provides the PhastCons 100 way annotation, a conservation score for every genetic position [14]. The scores were created using 99 vertebrates and the human genome and they correspond to the probability of a nucleotide being conserved. We were interested in keeping highly conserved regions, therefore we chose a PhastCons score >0.8.

Variants with highly differing estimated allele frequency

As shown in Figure 1d section 1, we compared the estimated allele frequencies of each pair sequencers for all variants. We defined variants with highly differing allele frequency as such, if the difference in estimates was >0.8.

NGHRI-EBI GWAS-catalogue

The NHGRI-EBI GWAS catalogue of published genome-wide association studies (GWAS catalogue) consists of 5687 GWA studies and contains 71673 SNP-trait associations from 3567 published studies [15]. All included SNPs are associated to common traits or diseases. A study and an association must meet strict criteria to be included in the GWAS catalogue: an array based GWAS with at least 100,000 SNP and a SNP-trait association must have a p-value $< 1*10^{-5}$ [31]. In our workflow we filtered for variants with highly differing estimated allele frequency that were present in the GWAS catalogue (with matching chromosome, position and rsID). (Figure 1d, section 2).

Unique variant rate (UVR)

We defined the unique variant rate (UVR) as the number of variants, that were only found by one sequencer, divided by the number of all variants, that were found by the sequencer.

Statistical analysis

The observed and expected number of variants for the three analysis pipelines and their intersections (Venn-diagram in Fig. 2b) was statistically compared using a Monte Carlo simulation approach. Assuming the total average number of variants observed (4,500,440) as the "true" base population from which variants were called by each sequencing approach, we could infer the statistical over- or under-representation of each section of the Venn-diagram. We computed the null-distribution for the expected number of variants in each set by randomly drawing the observed number of variants for each sequencing approach and determining the respective intersections. This step was repeated 1000 times and the mean values over all runs were taken as estimates for the null distribution of the number of variants in each set of the Venn-diagram. The observed number of variants was then compared against the expected null distribution with a chi-squared test.

The Poisson regression analysis to investigate potential factors which significantly impact concordance between different sequencers was performed by fitting a generalized linear model with the glm function from the stats package in R. In order to investigate if there is an overdispersion in the response variable, we also attempted to fit negative binomial regression models with the glm.nb function from the MASS package. However, these models failed to converge suggesting that no overdispersion was present therefore the results from the Poisson model were reported.

Abbreviations

AF: Allele Frequency

CG: Complete Genomics

GIAB: Genome in a bottle

GWAS: Genome-wide Association Study

HSX: Illumina HiSeq X

HWE: Hardy-Weinberg-Equilibrium

InDel: Insertion/Deletion Polymorphism

LINE: Long Interspersed Nuclear Elements

LTR: Long Terminal Repeats

MAF: Minor Allele Frequency

MOL: Illumina HiSeq

NGS: Next Generation Sequencing

SNP: Single Nucleotide Polymorphism

UVR: Unique Variant Rate

WGS: Whole-Genome Sequencing

WES: Whole-Exome Sequencing

Declarations

Ethics approval and consent to participate

In the current study, data obtained from a previous study by Erikson et al. (2016), https://doi.org/10.1016/j.cell.2016.03.022 were re-analyzed. The original study (IRB-13-6142) was approved by the Scripps Institutional Review Board in July 2007

Consent for publication

Not applicable

Availability of data and materials

The data that support the findings of this study are not publicly available since they contain highly sensitive personal information allowing to identify the individual subjects. Data are however available from the authors upon reasonable request and in combination with an IRB Approval.

Competing interests

The authors declare that they have no conflict of interests.

Funding

SW acknowledges funding from the Emergent AI Center granted by the Carl-Zeiss-Stiftung. The work of CH was partly funded by the Boehringer Ingelheim Stiftung. SSy acknowledges funding from the Mainz Institute of Multiscale Modeling. HT acknowledges funding from the ReALity initiative. The funders had no role in the design of the research, data collection and analysis, writing of the manuscript and the decision to publish.

Authors' contributions

SW, SSy, CH and HT wrote the manuscript, performed the analysis and interpreted the data. JW, SSc, and JB contributed to writing, editing and interpreting the study. MP contributed to statistical analysis, editing and interpreting the study. AT provided the data and contributed to writing and editing. SG and CH designed and supervised the research. SG edited the manuscript and contributed to writing and interpreting the data. HMS contributed to supervising the study and to editing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

References

1. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB: **Characterizing and measuring bias in sequence data.** *Genome Biology* 2013, **14:**R51.
2. Lam HYK, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, et al: **Correction: Corrigendum: Performance comparison of whole-genome sequencing platforms.** *Nature Biotechnology* 2012, **30:**562-562.
3. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, et al: **Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing.** *Genome Medicine* 2013, **5:**28.
4. Cornish A, Guda C: **A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference.** *BioMed research international* 2015, **2015:**456479-456479.
5. Highnam G, Wang JJ, Kusler D, Zook J, Vijayan V, Leibovich N, Mittelman D: **An analytical framework for optimizing variant discovery from personal genomes.** *Nature Communications* 2015, **6:**6275.
6. Hwang S, Kim E, Lee I, Marcotte EM: **Systematic comparison of variant calling pipelines using gold standard personal exome variants.** *Scientific Reports* 2015, **5:**17875.
7. Chen J, Li X, Zhong H, Meng Y, Du H: **Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers.** *Scientific Reports* 2019, **9:**9345.
8. Kumaran M, Subramanian U, Devarajan B: **Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data.** *BMC bioinformatics* 2019, **20:**342-342.
9. Hwang K-B, Lee I-H, Li H, Won D-G, Hernandez-Ferrer C, Negron JA, Kong SW: **Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings.** *Scientific Reports* 2019, **9:**3219.
10. Kim JH: **GWAS Data Analysis.** In *Genome Data Analysis Learning Materials in Biosciences.* Singapore: Springer; 2019
11. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M: **Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls.** *Nature Biotechnology* 2014, **32:**246-251.
12. Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean CY, et al: **An open resource for accurately benchmarking small variant and reference calls.** *Nature Biotechnology* 2019, **37:**561-566.
13. Erikson GA, Bodian DL, Rueda M, Molparia B, Scott ER, Scott-Van Zeeland AA, Topol SE, Wineinger NE, Niederhuber JE, Topol EJ, Torkamani A: **Whole-Genome Sequencing of a Healthy Aging Cohort.** *Cell* 2016, **165:**1002-1011.

14. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Research* 2005, **15:**1034-1050.
15. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al: **The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.** *Nucleic acids research* 2019, **47:**D1005-D1012.
16. 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: **Best practices for benchmarking germline small-variant calls in human genomes.** *Nature Biotechnology* 2019, **37:**555-560.
17. Ioannidis JPA, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG: **Replication validity of genetic association studies.** *Nature Genetics* 2001, **29:**306-309.
18. Ioannidis JPA, Patsopoulos NA, Evangelou E: **Heterogeneity in meta-analyses of genome-wide association investigations.** *PloS one* 2007, **2:**e841-e841.
19. Pei Y-F, Tian Q, Zhang L, Deng H-W: **Exploring the Major Sources and Extent of Heterogeneity in a Genome-Wide Association Meta-Analysis.** *Annals of human genetics* 2016, **80:**113-122.
20. Narzisi G, Schatz MC: **The challenge of small-scale repeats for indel discovery.** *Frontiers in bioengineering and biotechnology* 2015, **3:**8-8.
21. Montgomery SB, Goode D, Kvikstad E, Albers CA, Zhang Z, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, et al: **The origin, evolution and functional impact of short insertion-deletion variants identified in 179 human genomes.** *Genome Research* 2013.
22. Conrad DF, Keebler JEM, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, et al: **Variation in genome-wide mutation rates within and between human families.** *Nature Genetics* 2011, **43:**712-714.
23. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT: **Basic statistical analysis in genetic case-control studies.** *Nature protocols* 2011, **6:**121-133.
24. Hamdan FF, Myers CT, Cossette P, Lemay P, Spiegelman D, Laporte AD, Nassif C, Diallo O, Monlong J, Cadieux-Dion M, et al: **High Rate of Recurrent De Novo Mutations in Developmental and Epileptic Encephalopathies.** *The American Journal of Human Genetics* 2017, **101:**664-685.
25. Ma Y, Jun GR, Zhang X, Chung J, Naj AC, Chen Y, Bellenguez C, Hamilton-Nelson K, Martin ER, Kunkle BW, et al: **Analysis of Whole-Exome Sequencing Data for Alzheimer Disease Stratified by APOE Genotype.** *JAMA Neurol* 2019, **76:**1099-1108.
26. Qiao D, Ameli A, Prokopenko D, Chen H, Kho AT, Parker MM, Morrow J, Hobbs BD, Liu Y, Beaty TH, et al: **Whole exome sequencing analysis in severe chronic obstructive pulmonary disease.** *Human molecular genetics* 2018, **27:**3801-3812.
27. Gao GF, Parker JS, Reynolds SM, Silva TC, Wang L-B, Zhou W, Akbani R, Bailey M, Balu S, Berman BP, et al: **Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data.** *Cell Systems* 2019, **9:**24-34.e10.

28. Costain G, Jobling R, Walker S, Reuter MS, Snell M, Bowdin S, Cohn RD, Dupuis L, Hewson S, Mercimek-Andrews S, et al: **Periodic reanalysis of whole-genome sequencing data enhances the diagnostic advantage over standard clinical genetic testing.** *European Journal of Human Genetics* 2018, **26:**740-744.
29. Hysi PG, Choquet H, Khawaja AP, Wojciechowski R, Tedja MS, Yin J, Simcoe MJ, Patasova K, Mahroo OA, Thai KK, et al: **Meta-analysis of 542,934 subjects of European ancestry identifies new genes and mechanisms predisposing to refractive error and myopia.** *Nature Genetics* 2020, **52:**401-407.
30. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, de Andrade M, Doheny KF, Haines JL, Hayes G, et al: **Quality control procedures for genome-wide association studies.** *Current protocols in human genetics* 2011, **Chapter 1:**Unit1.19-Unit11.19.
31. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al: **The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog).** *Nucleic acids research* 2017, **45:**D896-D901.

Figures

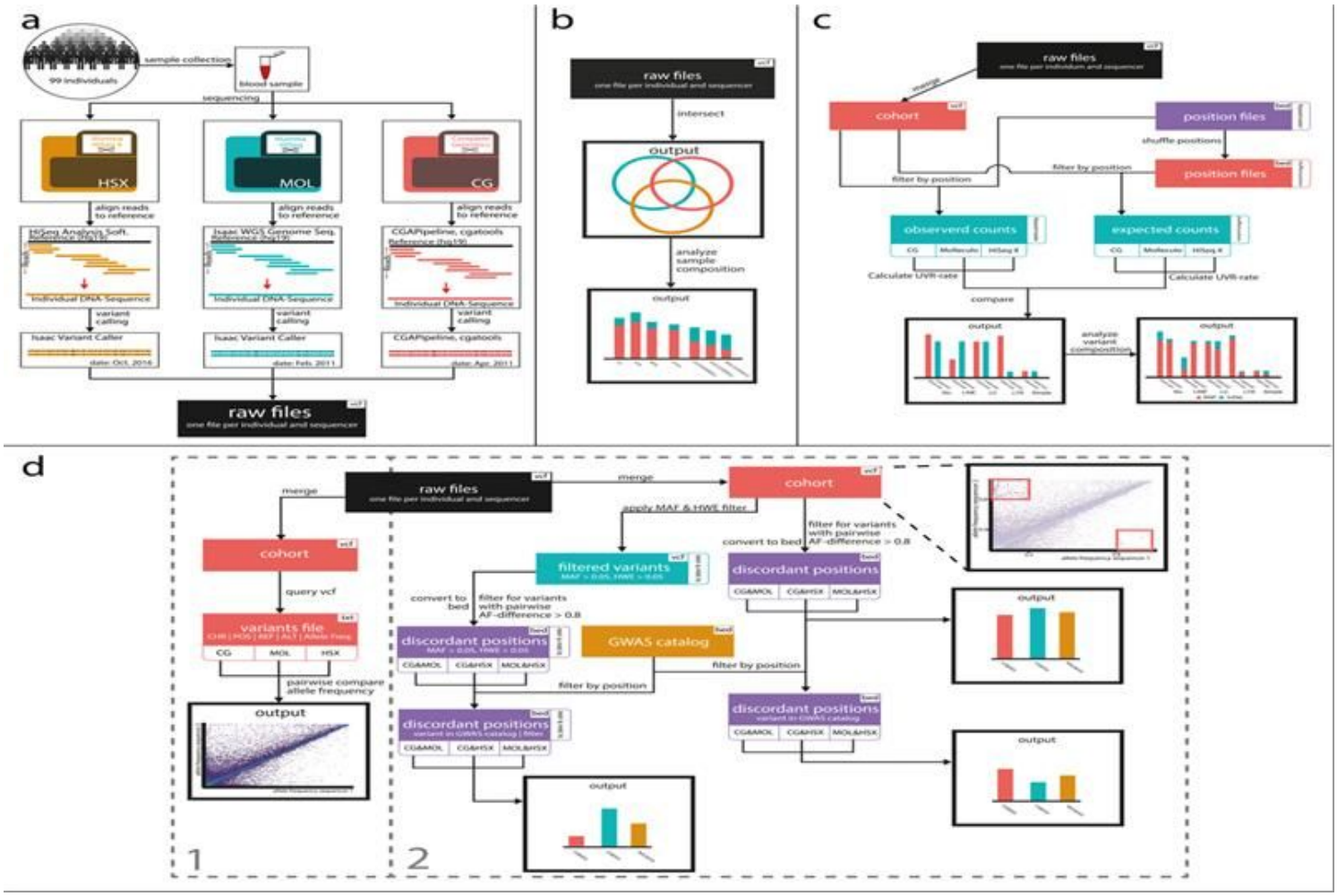


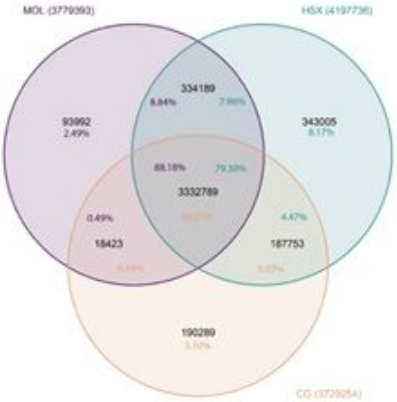
Figure 1

Graphical abstract of the study. a, Schematic overview of the workflow before our analysis. 99 individuals with the wellderly phenotype were sequenced three times with Illumina HiSeq X (HSX, Apr. 2016), Illumina HiSeq (MOL, Feb. 2011) and Complete Genomics (CG, Apr. 2011). The resulting sequencing data were processed with three different bioinformatic pipelines. For CG, the in-house software cgatools (v. 1.6.0) with the Complete Genomics Analysis pipeline (v. 2.0.22) was used, while both Illumina cohorts were aligned and called with Isaac Alignment Software and Isaac Variant Caller. In our current study, we analyzed the resulting vcf-file (three files per individual). b, We evaluated the concordance of called variants between different sequencing platforms by intersecting the respective vcf-files. The average number of concordant variants for different intersections was calculated and reported. c, Schematic overview of the work process for variants detected in repetitive genomic elements annotated with the RepeatMasker software. Annotations were converted into a bed-file. All contained sections of the bed-file were pseudo-randomly assigned to a new position on the same chromosome in order to sample random genomic regions of the same length as the respective RepeatMasker regions. The raw vcf-files were filtered with the position files. The observed and expected unique variant rate (UVR) in RepeatMasker regions was estimated by dividing the number of unique variants by all variants found by one sequencer in the respective region. d, (1) For each sequencer we created a cohort file by merging all 99 individual vcf-files into one. This cohort vcf-file was queried to obtain the information about all variants including the chromosome, position, reference allele, alternative allele and allele frequency. (2) We then quantified all variants with a difference in estimated allele frequency >0.8 between each pair of sequencers. The resulting variants were intersected with SNPs, which are included in the EMBL-EBI GWAS catalogue. To evaluate the influence of the commonly used filters Minor Allele Frequency (MAF) and Hardy-Weinberg-Equilibrium (HWE), we filtered the variant sets overlapping GWAS catalogue SNPs for MAF > 0.05 and HWE > 0.05.

a

	Illumina HiSeq X	Illumina Moleculo	Complete Genomics	Var. found by all platforms
Number of variants	4,197,735.343	3,779,392.838	3,729,254.071	3,332,788.687
Number of SNP	3,542,255.788	3,408,242.929	3,330,085.515	3,104,790.576
Number of InDel	655,479.556	371,149.909	399,168.556	227,998.111

b



c

Platform	Observed value	Expected value	Observed/Expected ratio	p-Value
HSX\(MOL&CG)	343,005.00	115,245.45	2.98	<0.001
MOL\(HSX&CG)	93,992.00	40,186.90	2.34	<0.001
CG\(HSX&MOL)	190,289.00	43,556.73	4.37	<0.001
(HSX&CG)\MOL	187,753.00	604,076.00	0.31	<0.001
(MOL&HSX)\CG	334,189.00	557,307.03	0.60	<0.001
(MOL&CG)\HSX	18,423.00	210,652.00	0.09	<0.001
MOL&HSX&CG	3,332,789.00	2,921,107.00	1.14	<0.001

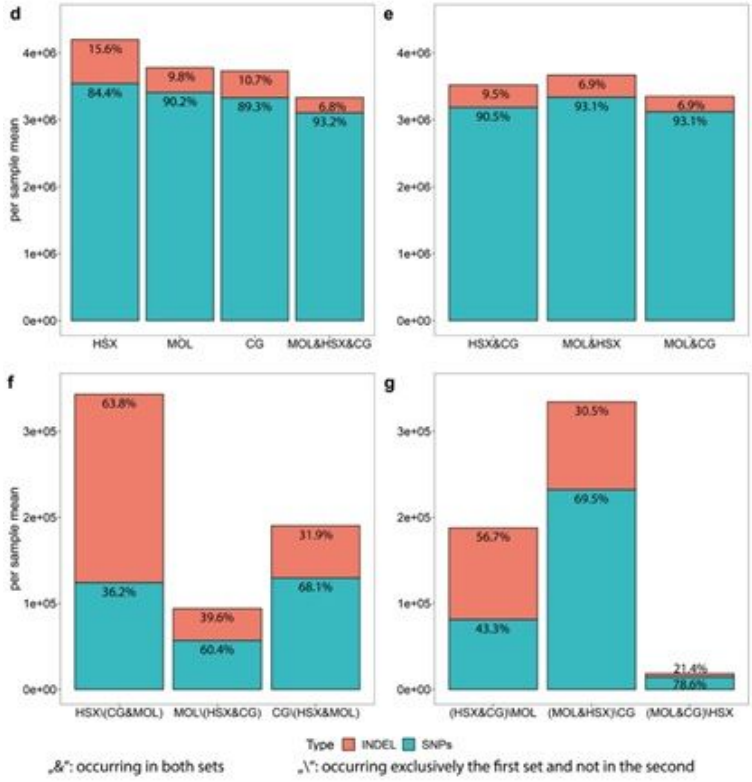


Figure 2

Composition of variants detected under different experimental setups. a, The overall average number of variants as well as the average number of single nucleotide polymorphisms (SNP) or insertions/deletions (InDel) are shown for each sequencing platform and for the intersection of all platforms. b, The Venn-diagram shows all possible intersections for the sets of variants detected for each platform. The quantity of variants in each subset is reported as absolute numbers as well as relative proportions in percent with

respect to each sequencing platform. Percentages are color-coded according to the reference platform. c, Monte Carlo simulation-based comparison of the observed and expected number of variants in the different intersections of the Venn-diagram. d, The bar plot shows the average number of variants for Illumina HiSeq X (HSX), Illumina HiSeq(MOL), Complete Genomics (CG) and variants detected by all three platforms. e, the bar plot shows the average number of variants detected by at least two platforms. f, the bar plot shows the average number of variants detected by exactly one platform. g, the bar plot shows the average number of variants detected by exactly two platforms. SNPs are colored green and InDels are colored red on the bar plots in d-g." \" indicates a logical "not", "&" corresponds to a logical "and".

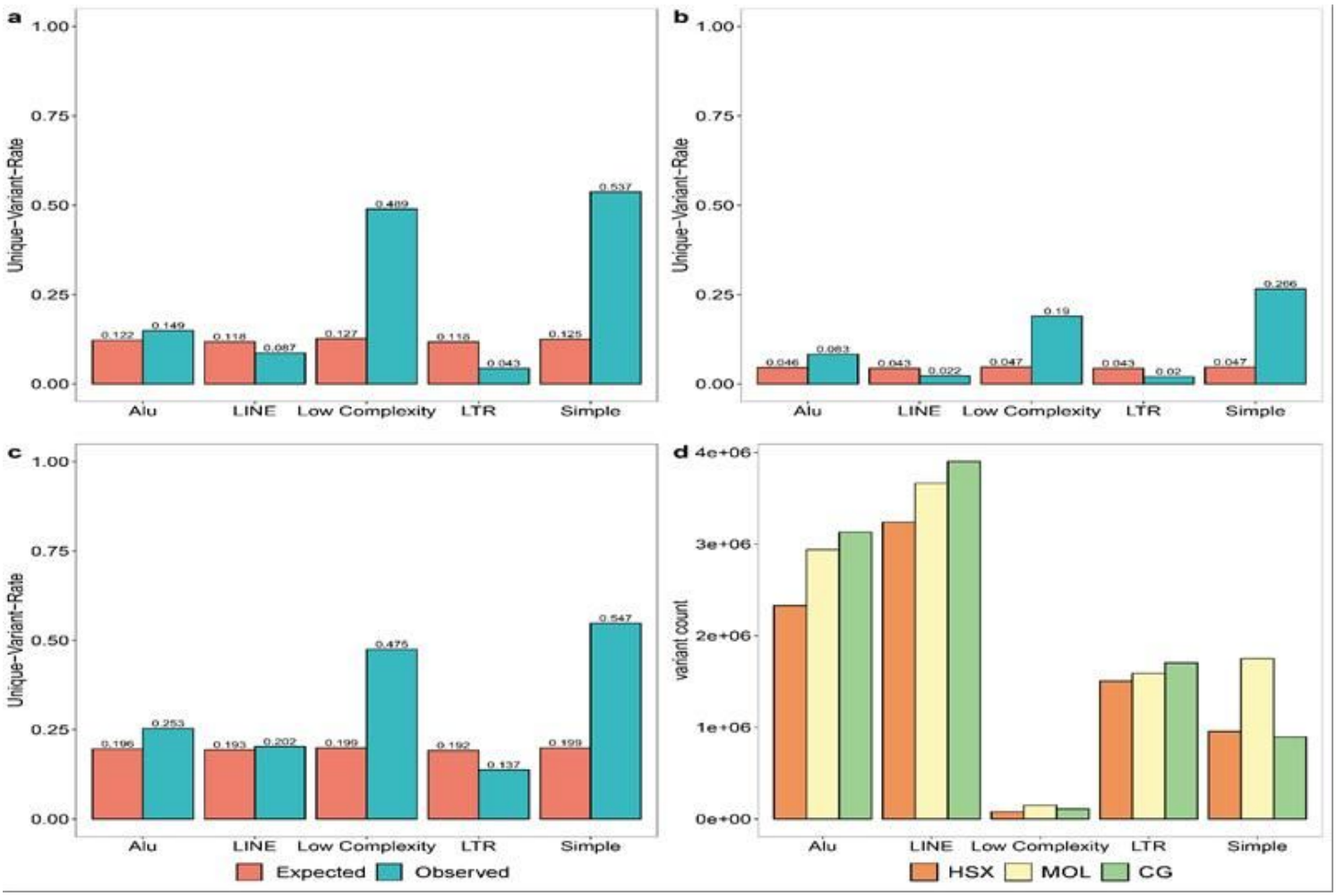


Figure 3

Unique variant rate in repetitive regions annotated with the RepeatMasker software. a, The barplot shows the observed and expected rate of unique variants in Repeatmasker regions for Illumina HiSeq X (HSX). b, The barplot shows the observed and expected rate of unique variants in Repeatmasker regions for Illumina HiSeq (MOL) c, The barplot shows the observed and expected rate of unique variants in Repeatmasker regions for Complete Genomics (CG) d, The barplot shows the average number of variants (variant count) detected for HSX, MOL and CG, respectively in each type of RepeatMasker region. LINE: long interspersed nuclear elements; LTR: long terminal repeats.

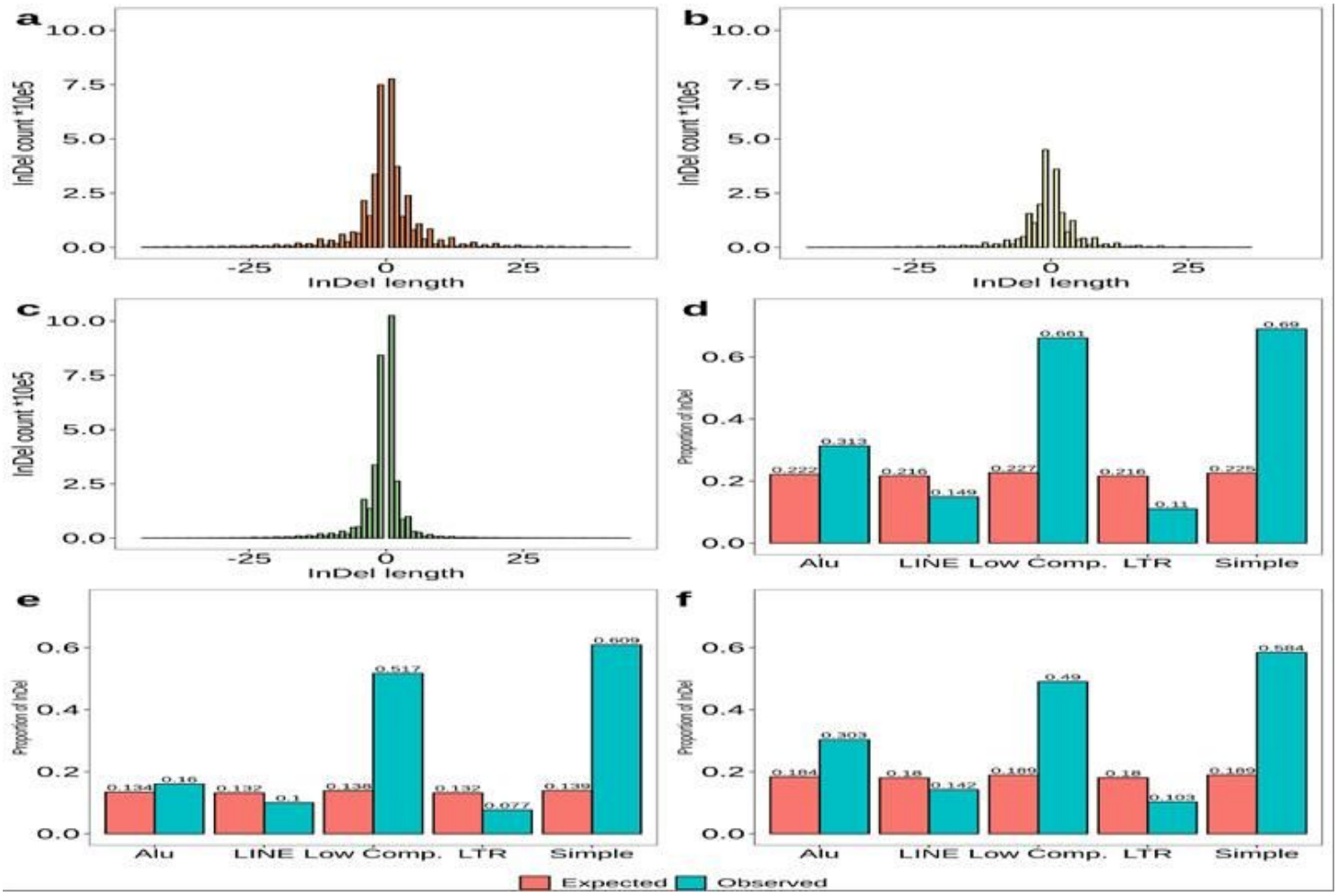


Figure 4

Insertion/deletion (InDel) length distribution and InDel proportion in RepeatMasker regions. a, The histogram shows the InDel length distribution for Illumina HiSeq X (HSX) b, The histogram shows the InDel length distribution for Illumina HiSeq (MOL). c, The histogram shows the InDel length distribution for Complete Genomics (CG). Positive values of the InDel length correspond to insertions whereas negative values represent deletions in a-c. d, Observed and expected proportion of InDels in repetitive elements annotated with the RepeatMasker software for HSX. e, Observed and expected proportion of InDels in repetitive elements annotated with the RepeatMasker software for MOL f, Observed and expected proportion of InDels in repetitive elements annotated with the RepeatMasker software for CG. LINE: long interspersed nuclear elements; Low comp.: low complexity regions; LTR: long terminal repeats.

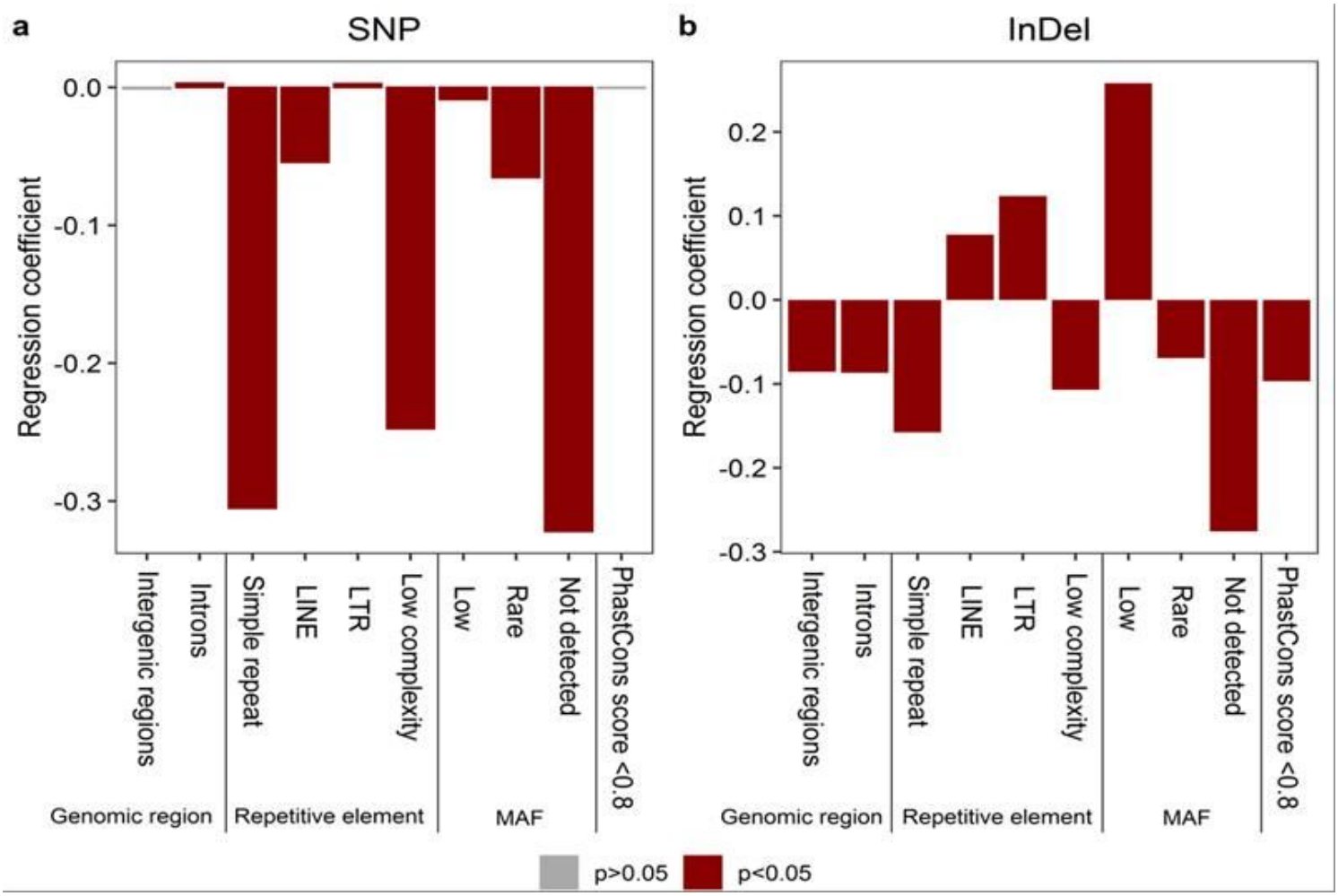


Figure 5

Regression analysis of factors impacting the concordance of variants between experimental setups. Poisson regression was performed with the four predictors genomic region, repetitive element (RepeatMasker annotation), minor allele frequency (MAF) obtained from the European cohort from the 1000 Genomes Project and positional conservation (indicated by the PhastCons score) for a, SNPs and b, InDels separately. Positive regression coefficients indicate predictors which improve the concordance between different setups, whereas negative coefficients correspond to factors which reduce the concordance. Regression coefficients that are significantly different from 0 are colored red, non-significant coefficients are colored gray.

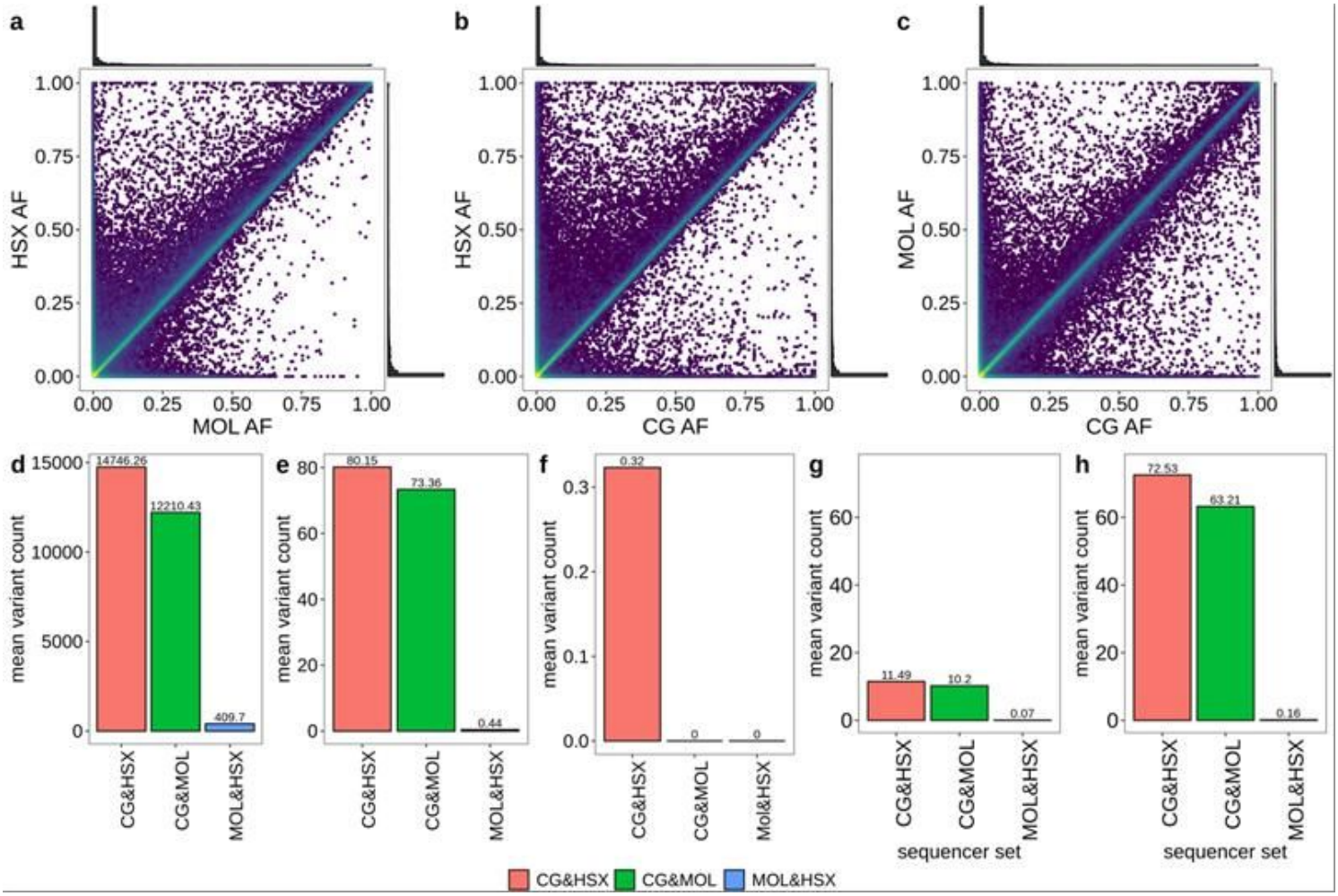


Figure 6

Comparison of allele frequency estimates between different sequencing platforms. a, Correlation plot of estimated allele frequencies (AF) for variants on chromosome 22 between Illumina HiSeq X (HSX) and Illumina HiSeq (MOL) b, Correlation plot of estimated allele frequencies for variants on chromosome 22 between HSX and Complete Genomics (CG) c, Correlation plot of estimated allele frequencies for variants on chromosome 22 between MOL and CG. The yellow/green color in a-c indicates a high density in the corresponding area of the bivariate distribution of allele frequencies for each pairwise combination of sequencing platforms. The purple color corresponds to areas with low density on the bivariate scatter plot. d, Average number of variants (mean count) having an allele frequency difference >0.8 between different platforms. e, Average number (mean count) of known genome-wide association studies (GWAS) variants having an allele frequency difference >0.8 between different platforms. f, Average number (mean count) of known GWAS variants having an allele frequency difference >0.8 between different platforms after applying the minor allele frequency (MAF) and Hardy-Weinberg-Equilibrium (HWE) filter. Known GWAS variants were retrieved from the EMBL-EBI GWAS catalogue[15]. g, Average number (mean count) of known GWAS variants having an allele frequency difference >0.8 between different platforms after applying the minor MAF filter only. h. Average number (mean count) of known GWAS variants having an

allele frequency difference >0.8 between different platforms after applying only the HWE filter. Known GWAS variants were retrieved from the EMBL-EBI GWAS catalogue [15].

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- STROBEchecklist1.doc