Allele Balance Bias Identifies Systematic Genotyping Errors and False Disease Associations

Francesc Muyas[1,2,6], Mattia Bosio[1,2], Anna Puig[1,2], Hana Susak[1,2], Laura Domènech-Salgado[1,2,7], Georgia Escaramis[1,2,7], Luis Zapata[1,2], German Demidov[1,2,6], Xavier Estivill[4,5], Raquel Rabionet[1,2,3,7], Stephan Ossowski[1,2,6*]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Spain. [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [3]Institut de Recerca Sant Joan de Déu; Institut de Biomedicina de la Universitat de Barcelona (IBUB); & Department of Genetics, Microbiology and Statistics, University of Barcelona, Barcelona, Spain. [4]Sidra Medicine, Doha, Qatar. [5]Women's Health Dexeus, Barcelona, Spain. [6]Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany. [7]CIBER in Epidemiology and Public Health (CIBERESP), Barcelona, Spain.

*Corresponding author, contact: stephan.ossowski@med.uni-tuebingen.de

Grant Sponsor

This work has been funded by:

1. Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2017-2021'
2. CERCA Programme / Generalitat de Catalunya
3. The "la Caixa" Foundation
4. CRG emergent translational research award
5. European Union's H2020 research and innovation programme under grant agreement No 635290 (PanCanRisk)
6. MINECO Severo Ochoa fellowship (SVP-2013-0680066)
7. PERIS program (SLT002/16/00310)

ABSTRACT

In recent years, Next Generation Sequencing (NGS) has become a cornerstone of clinical genetics and diagnostics. Many clinical applications require high precision, especially if rare events such as somatic mutations in cancer or genetic variants causing rare diseases need to be identified. Although random sequencing errors can be modeled statistically and deep sequencing minimizes their impact, systematic errors remain a problem even at high depth of coverage. Understanding their source is

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/humu.23674.

crucial to increase precision of clinical NGS applications. In this work, we studied the relation between recurrent biases in allele balance (AB), systematic errors and false positive variant calls across a large cohort of human samples analyzed by whole exome sequencing (WES). We have modeled the allele balance distribution for biallelic genotypes in 987 WES samples in order to identify positions recurrently deviating significantly from the expectation, a phenomenon we termed allele balance bias (ABB). Furthermore, we have developed a genotype callability score based on ABB for all positions of the human exome, which detects false positive variant calls that passed state-of-the-art filters. Finally, we demonstrate the use of ABB for detection of false associations proposed by rare variant association studies (RVAS). Availability: https://github.com/Francesc-Muyas/ABB.

Keywords: Genetic variant detection, allele balance, systematic NGS errors, false positive variant calls

INTRODUCTION

The rapid improvement of next generation sequencing (NGS) throughput and cost has changed biomedical research as well as clinical diagnostics of genetic diseases and cancer (Altmann et al., 2012). Numerous genome-sequencing projects catalogued millions of frequent and rare variants, some of which are associated to disease (Auton et al., 2015). NGS has facilitated the identification of novel therapeutic targets or genomic markers for clinical diagnostics and treatment, becoming the technology of choice to study the genetic causes of diseases (Hwang et al., 2015; Oleksiewicz et al., 2015; Pabinger et al., 2014).

Despite the widespread use of NGS in genetic disease studies and diagnostics, the use of short reads for identification of causal or disease-associated variants is still sensitive to technical errors and may generate false associations and diagnoses (Hardwick, Deveson, & Mercer, 2017; Lee, Abecasis, Boehnke, & Lin, 2014; Yan et al., 2016). If the studied event is rare, such as *de novo* germline mutations, the likelihood to observe false positive calls is further increased (Gómez-Romero et al., 2018; Veltman & Brunner, 2012). Moreover, rare variant association studies (RVAS) can generate

false results if genes are enriched with sequencing or alignment errors, leading to false associations to the studied disease (Hou et al., 2017; Johnston, Hu, & Cutler, 2015; Yan et al., 2016). Hence, some RVAS methods take into account error probabilities (He et al., 2015) or by pass genotype calls completely by directly modeling sequencing reads (Hu, Liao, Johnston, Allen, & Satten, 2016). The impact of false genotype calls is amplified in the study of recurrent somatic mutations in cancer, particularly, if ultra-deep sequencing is used to identify sub-clonal mutations with low minor allele frequency (Cai, Yuan, Zhang, He, & Chou, 2016). Moreover, recent benchmarking studies reported substantial disagreement between somatic SNV and indel prediction methods (Alioto et al., 2015).

Although most variant calling algorithms can deal with random sequencing errors, systematic errors have mostly been neglected in the past and thus more often lead to false positive variant calls. Several causes of errors of sequencing by synthesis-based platforms are well described, such as *crosstalk* and *dephasing* (Ledergerber & Dessimoz, 2011; Pfeiffer et al., 2018; Sleep, Schreiber, & Baumann, 2013), missed nucleotides in *low complexity regions* (H. Li, 2014), index hopping ('bleeding') (Vodák et al., 2018), and DNA damage during library preparation (Chen, Liu, Evans, & Ettwiller, 2017) caused by e.g. 8-oxo-G formation when using acoustic shearing or oxidative stress during probe hybridization (Newman et al., 2016; Park et al., 2017) and decreased coverage in regions of very *high or low CG content* (Sleep et al., 2013). Li et al. (H. Li, 2014) showed that a large fraction of systematic errors found in variant callsets were not due to sequencing errors, but erroneous re-alignments in low-complexity and repetitive regions (about 2% and 45% of the human genome, respectively), as well as the incompleteness of the reference genome with respect to the analyzed sample. While repetitive regions lead to ambiguous alignments of short reads and thus increase the likelihood of assigning a read to the wrong locus, low complexity regions also cause mis-alignment of reads at the correct position (Cordaux & Batzer, 2009; H. Li, 2014). Furthermore, many aligners tend to mis-align indels close to the end of a read, as their scoring function favors mismatches over gap openings.

Strategies for identification of systematic genotyping errors

A multitude of variant calling algorithms that apply various strategies to reduce the false positive rate (FPR) has been developed. Commonly used tools for germline variant prediction include GATK HaplotypeCaller (McKenna, 2009), Samtools mpileup (H. Li, 2011), Freebayes (Garrison & Marth, 2012) or Varscan (Koboldt et al., 2009; Koboldt DC, Larson DE, 2013). Other tools, e.g. Strelka (Saunders et al., 2012), VarScan2 and MuTect (Cibulskis et al., 2013), specialize in somatic mutation calling using tumor-normal pairs. Most of these methods apply Bayesian statistics (e.g. Bayesian classifiers) to compute genotype likelihoods (Garrison & Marth, 2012; Van der Auwera et al., 2013), or, in case of somatic mutations, the likelihood of the variant model (Cibulskis et al., 2013). Some systematic alignment issues can be addressed by haplotype-based variant calling as performed by FreeBayes (Garrison & Marth, 2012). Issues with gapped alignments around indel alleles are tackled by alignment post-processing, using either multiple-read re-alignment or local assembly (DePristo et al., 2011; Van der Auwera et al., 2013). Still, stringent post-filtration of callsets using machine learning based error models (e.g. Variant Quality Score Recalibration,VQSR (Carson et al., 2014; Van der Auwera et al., 2013)) or thresholds on various call quality statistics (e.g. genotype quality, read depth, variant allele frequency (H. Li, 2014), clustered variants, Fisher strand bias (Guo et al., 2012) is a necessity. Other strategies include removal of variants in low complexity regions as well as in repeats incompletely represented in the reference genome (typically indicated by significantly increased read coverage) (Carson et al., 2014; H. Li, 2014). However, a general issue of many post-filtration strategies is their use of hard thresholds for the various quality metrics, where small changes can dramatically influence false negative and false positive rates, or their dependence on large sample sets to be effective (e.g. VQSR) (De Summa et al., 2017; Lek et al., 2016).

Here we present a new strategy to identify systematic sequencing or alignment errors leading to false variant calls, which is based on the recurrent and significant deviation of observed to expected allele balance in a genomic position across large control cohorts. This signature, termed allele balance bias

(ABB), was found in 0.03% of all exonic positions, 4% of high confidence germline SNV calls in 987 exomes and 8% of somatic SNV calls in 200 tumor-normal pairs. We present two algorithms: one for computing ABB for all positions of the exome (or genome) using large cohorts of WES (or WGS) samples, and one for refining candidate gene lists generated by rare variant association studies (RVAS). We have trained an ABB model for the human exome optimized for the use in clinical exome diagnostics and rare variant association testing in coding genes. Finally, we provide ABB genotype callability scores for all positions of the human exome.

MATERIAL AND METHODS

Whole-Exome Sequencing and Data Analysis. We have analyzed 1197 germline samples assembled from various genetic disease and cancer studies, including case and control cohorts sequenced at the CRG-CNAG, Barcelona. Included individuals are of European ancestry. Exome capture has been performed using five different in-solution capture methods: Agilent SureSelect versions 35MB, 50MB, 71MB and V5 and Roche-Nimblegen SeqEz v3 (detailed information on samples and library preparation can be found in Supporting materials and Supp. Table S1). We removed regions showing less than an average of 10x read coverage across samples analyzed using the same kit. For variant analysis, we extended captured regions by 50 bp upstream and downstream flanking regions. Sequencing was performed on Illumina HiSeq2000 or HiSeq2500 using 2 x 100bp paired end reads (Bentley et al., 2008). Reads were aligned against the human reference genome (hg19) using BWA-MEM (H. Li, 2013; H. Li & Durbin, 2009). Alignment post-processing was performed according to GATK best practice guidelines (Van der Auwera et al., 2013), including PCR duplicate marking, Indel realignment and base quality recalibration (Bao et al., 2014). Variant calling was performed using GATK HaplotypeCaller v3.3(McKenna, 2009; Van der Auwera et al., 2013). Variants with genotype quality below 20 or Fisher strand bias (FS) in the top 10 percentile were removed. For benchmarking purposes, we generated two callsets, one with and one without applying GATK VQSR filter (tranche threshold of 99.9%). See 'ACCESION NUMBERS' section for available data.

Deviation of observed from expected allele balance. We investigated the relationship between recurrent deviation of observed from expected allele balance, systematic errors and false positive SNV calls in whole-exome sequencing (WES) data. Allele balance (AB) describes the fraction of reads supporting the alternative allele in a focal position (AB = Alternative Read Count / Total read count at focal position). When sequencing diploid species, heterozygous genotypes are expected to show an AB close to ~0.5. We modeled read distribution for heterozygous genotypes using a binomial distribution *Binomial*(D, ~0.5) (Guo et al., 2013; Nothnagel et al., 2011; O'Fallon, Wooderchak-Donahue, & Crockett, 2013). Homozygous genotypes are expected to have close to 100% of reads supporting the same allele, with the amount of deviating reads depending on the sequencing and alignment error rate and other variables. We modeled the expected read distribution for homozygous reference using zero inflated beta distribution and for homozygous alternative using one inflated beta distribution, where AB would be inside the range [0, 1] (Ospina & Ferrari, 2012). The corresponding probability density function is given by

$$\text{beinf}(y; \alpha, \gamma, \mu, \phi) = \begin{cases} \alpha(1-\gamma), & if\ y = 0, \\ \alpha\gamma, & if\ y = 1, \\ (1-\alpha) f(y; \mu, \phi), & if\ y \in (0,1), \end{cases}$$

where $f(y; \mu, \varphi)$ is the beta density function, and μ and φ are the parameters that define the shape of the beta distribution. Note that, if $y \sim$ BEINF(α,γ,μ,φ), then $P(y = 0) = \alpha(1-\gamma)$ and $P(y = 1) = \alpha\gamma$ (Ospina & Ferrari, 2009), where α is the mixture parameter and γ represents the parameter of the cumulative distribution function of a Bernoulli random variable. Here, y represents the allele balance variable. Parameters for expected AB distributions for each diploid genotype class have been estimated using post-VQSR variant calls from GATK *HaplotypeCaller* by maximum (penalized) likelihood estimation (*GAMLSS* R package). We genotyped every position of the exomes of 1197 samples by comparing observed to expected AB, obtaining one p-value for each of the three possible diploid genotypes. We assumed that the greatest p-value represents the most likely genotype of a focal sample and position. Given this genotype, we measured the deviation of observed from expected AB

(devAB equal to AB - 0 for homozygous reference, |AB - 0.5| for heterozygous and |AB – 1| for homozygous alternative).

Allele Balance Bias. Using devAB at a focal position in hundreds of samples we can identify positions showing recurrent deviation of observed from expected AB, termed Allele Balance Bias (ABB). To quantify and model ABB we processed a training cohort of 987 germline WES samples, leaving 200 germline samples for validation and testing of the model (randomly chosen from normal tissue exomes of 450 Chronic Lymphocytic Leukemia patients) (Puente et al., 2015). For Sanger validation, we used 10 independent samples, which were all obtained after ABB training (and for which ample amounts of DNA were available) (see Supporting Material and Supp. Table S1-S2 for detailed sample information). Note that the influence of somatic mutations on model training can be neglected, as training is performed only on healthy blood or normal tissue samples, in which somatic mutations are expected to be extremely rare and not recurrent across samples.

We obtained pileup files (samtools mpileup version 1.1) for each sample and collected read depth and allele counts for more than 80 Million exonic positions. We computed alternative allele fractions, most likely genotype and devAB for positions covered by at least 20 reads with base quality ≥20 (considered informative). Positions with less than 80 informative samples were excluded from further analysis. We calculated three measures of ABB strength for each position of the exome based on sample-wise devAB, termed *RdAB1, RdAB2 and RdAB3. RdAB1* represents the mean of *devAB* across all informative samples at a focal position. *RdAB2* measures the fraction of samples with a significant deviation from the expected distribution of the most likely genotype. *RdAB3* represents the arithmetic mean of $-\log_{10}$ (p-value) across all samples at a focal position.

ABB-based genotype callability model. To integrate the three measures of ABB into one genotype callability score, we trained a logistic regression model using the variant calls of 200 samples not used for defining *RdAB1, RdAB2* and *RdAB3*. This variant callset was obtained using GATK HaplotypeCaller as described above but omitting VQSR to allow the capture of an increased number

of potentially false calls for training purposes. We focused our analysis on heterozygous variants with at least 60 affected samples, resulting in 27,953 positions. To obtain the labels, we calculated the mean devAB values for all 27,953 positions, and split them into two sets using Gaussian Mixture Modeling (*mclust* R package, R version 3.2.3): non-recurrently deviated AB positions (labeled 0), and recurrently deviated AB positions (labeled 1). Two thirds of the data points (18,635 positions) were used for ABB model training (training set) 4659 positions were used for validation of the resulting ABB model and 4659 positions for final evaluation of the LR1 ABB model (test set). The logistic regression model LR1 uses *RdAB1, RdAB2, and RdAB3* as features to predict the labels obtained by Gaussian Mixture Modeling. It returns the probability of a variant site belonging to the label 1 (recurrently deviated AB positions) using the R function *glm* with family = "binomial":

$$\log(y/(1-y)) = \beta_0 + \beta_1 \cdot rdAB1 + \beta_2 \cdot rdAB2 + \beta_3 \cdot rdAB3 + \beta_4 \cdot rdAB2 \cdot rdAB3$$

with y=1 (recurrently deviated positions).

Subsequent to the estimation of the logistic regression parameters β_i, we calculated F1 scores at different probability levels and chose the maximum as optimal cutoff to assign labels. For this cutoff we calculated precision, recall, F score and false positive rate (FPR) (see formulas in Supporting Material), as well as Precision-Recall Area Under the Curve (PR-AUC) and ROC Area Under the Curve (ROC-AUC) values.

Using LR1, we calculated the probabilities to belong to the label *recurrently deviated AB* for each position of the human exome. We mapped the LR response value (probability) to precision values using the results obtained for validation and test sets. The resulting score, termed ABB genotype callability score, can be applied to estimate the callability of any position of the exome, with higher values indicating a higher likelihood of systematic errors. Based on visual inspection of the LR response to precision curve (Fig. 1D) we defined four genotype callability levels, comprising high confidence (ABB <= 0.15), medium confidence (0.15 < ABB <=0.75), low confidence (0.75 < ABB <= 0.9) and very low confidence (ABB > 0.9) positions.

Evaluation of ABB by Sanger sequencing. To benchmark the ability of ABB to identify false positive variant calls we randomly selected and Sanger sequenced 209 'suspicious' SNP calls (0.2 <= Allele Balance <= 0.35) from 10 samples not used for model training, selection or evaluation (see Supporting Material and Supp. Table S3 for more details). SNPs were sampled to similarly represent all four ABB genotype callability levels (42 high confidence, 73 medium confidence, 46 low confidence and 48 very low confidence SNPs). Additionally, 45 Sanger validations of novel disease variant candidates obtained within previous studies were included in the benchmarking. We compared false positive (FP) and failure rates (failed Sanger sequencing or ambiguous base call) between ABB bins and computed a ROC curve. Note that all variants selected for Sanger validation passed the GATK VQSR, fisher strand and minimum allele balance filters following the GATK best practice guidelines.

Relationship of ABB with other genomic features, quality measures and variant databases. Using variant calls generated by GATK-HaplotypeCaller for 10 samples (filtered with VQSR), we interrogated the correlation of ABB, fisher strand bias and transition-transversion ratio (Ti-Tv) for sites likely affected by systematic errors (ABB >= 0.9) using Wilcox test and Pearson's chi-square test, respectively. Using chi square Pearson's test we further investigated the enrichment of very low confidence variants in different public databases (dbSNP version 146, ExAC version 0.3v, 1000GP phase 3, EVS ESP6500), compared to the fraction across all informative positions of the exome. Similarly, we compared the relation of ABB with simple sequence repeats (SSRs) and tandem duplications. For that, we randomly selected a set of positions of very low (ABB >= 0.9) and high (ABB<0.15) confidence and compared the fraction of SSRs and tandem duplications using a Chi-square Pearson's test. We visualized the intersection of positions labeled un-callable by Genome in a Bottle (GIAB v3.3.2 High-Confidence regions (Zook et al., 2016)) and ABB using Venn diagrams and compared the performance in filtering false positive SNP calls on 209 sites validated by Sanger sequencing.

We interrogated the enrichment of very low confidence sites in somatic variant calls using tumor-normal paired data from 200 Chronic Lymphocytic Leukemia (CLL) patients, whose normal sample had not been used for ABB model building (germline variant positions used in model building were also excluded from enrichment analysis). Somatic SNVs were predicted using MuTect (Cibulskis et al., 2013). We measured the enrichment of high, medium, low and very low confidence sites in somatic mutation calls compared to their exome-wide expectation and the enrichment of each quality bin in Cosmic and dbSNP using in both cases Chi-Square Pearson's test.

Quality control for RVAS analysis. To test if ABB can identify false associations from rare variant association studies (RVAS) we developed Association-ABB, a method that tests if ABB can explain the difference in alternative allele counts ('burden') for a gene between cases and controls, and hence the genotype-phenotype association hypothesis can be rejected. In summary, the algorithm computes gene-wise aggregated measures of ABB in case and control cohorts in order to detect false associations arising from an uneven impact of ABB on variant calls in cases compared to controls. The algorithm takes as input the variants from the candidate genes generated by RVAS, and, for each variable position, identifies cases and control samples for which the variant caller might have missed or falsely predicted the alternative allele. Possibly "missed" alternative alleles are defined as homozygous reference calls for which the p-value within the homozygous AB zero-inflated beta distribution is less than 0.05. First, for each variant in RVAS candidate genes we test if the ratio of "called" compared to "missed" alternative genotypes is biased between cases and controls (Fisher exact test). Second, for each RVAS candidate gene, three tests are performed: 1) called-missed ratio test (similar to the variant-wise test but aggregating all rare variants per gene); 2) re-running the association test but including the "missed" calls as variants (Chi-Square Pearson's test aggregating all variants per gene); and 3) re-running the association test but removing significantly AB biased sites (Chi-Square Pearson's test). Genes with FDR lower than 0.1 in the called-missed ratio test or genes not significantly associated (FDR > 0.1) when adding "missed" variants or removing ABB variants are considered potential false positive associations. Association-ABB is available as part of the ABB

package at https://github.com/Francesc-Muyas/ABB. We have tested Association-ABB on an RVAS study for Chronic Lymphocytic Leukemia (see Supporting Material for details) in which the comparison of germline variants from 437 cases and 780 controls by SKAT-O (S. Lee, Wu, & Lin, 2012), Burden (B. Li & Leal, 2008; Madsen & Browning, 2009; Price et al., 2010), MiST (Sun, Zheng, & Hsu, 2013) and KBAC (Liu & Leal, 2010) association tests resulted in 43 CLL associated candidate genes, 10 of which were labeled as FP by Association-ABB.

RESULTS

We have developed a genotype callability score for NGS analysis based on the recurrent deviation of observed from expected allele balance, termed allele balance bias (ABB). Using an ABB model trained on 987 WES datasets we pre-computed ABB genotype callability scores for more than 81 Million positions of the human exome. We did not observe biases in genotype callability rates between kits when focusing on regions well-covered in all kits (average coverage ≥ 10, see Supporting Material and PCA in Supp. Fig. S1). To evaluate the performance of ABB on identification of systematic errors and false positive genotype calls, we used an independent set of 210 WES cases and Sanger validation. In addition, we demonstrate that ABB correlates with various measures of sequencing and alignment errors and show that public variant databases are enriched for systematic genotyping errors.

Training and Evaluation of the ABB Model

We hypothesized that systematic sequencing or alignment errors lead to recurrent deviation of allele balances in affected genomic positions across hundreds of samples. To test this hypothesis we trained, evaluated and tested a logistic regression model distinguishing positions with and without recurrently and significantly deviated AB, which integrate three measures of allele balance deviation (Methods, Fig. 1A, B, C). All coefficients of the logistic regression significantly contributed to the selected model (p-values << 0.01). We calculated F1 scores at different probability levels and chose the maximum (LR response of 0.13 at F1 of 0.91) as cutoff to assign labels. On the evaluation set of 4659

variants, LR1 showed a precision of 0.893, recall of 0.915, PR-AUC of 0.940, ROC-AUC of 0.980 and F1 of 0.904 for the optimal cutoff (Fig. 1C, Supp. Table S4). An independent test using the remaining 4,659 positions not used in any previous step showed similar performance (precision of 0.898, recall of 0.899, PR-AUC of 0.933, ROC AUC of 0.975, F1 of 0.899), demonstrating that the model was not over-fitted and can be generalized to novel datasets. Based on the correlation of LR response values and precision (Fig. 1D), probabilities obtained for each position of the exome were transformed to the precision of predicting systematic errors, which we finally use as ABB score. Higher ABB score values indicate a higher probability to obtain systematic errors in variant calling, with ABB > 0.75 considered low confidence positions (0.033% of the exome) and ABB > 0.9 considered very low confidence positions (0.025% of the exome).

ABB genotype callability filter for germline and somatic variant calling

To evaluate if the use of ABB as genotype callability filter leads to improved variant callsets, we applied an ABB very low confidence filter (ABB > 0.9) to variants predicted by GATK HaplotypeCaller with VQSR filtering. Using a callset for 10 samples not used during ABB model training, evaluation or testing, we found that 13,168 out of 346,894 (3.80%) variant sites overlapped with ABB very low confidence sites (compared to 0.025% of all exonic positions, p-value < 10^{-16}, Table 1, Supp. Fig. S2), with an average of 1,317 (3.80%) variants per sample. Surprisingly, 44.59% of known germline variants were flagged as medium confidence sites. We found that polymorphisms with ABB medium confidence are enriched for high population AF in 1000GP (mean of 26%), while polymorphic sites with ABB high confidence are mostly rare variants (population AF of 0.08%, Wilcox test p-value < 10^{-16}), reflecting that heterozygous sites are generally harder to call than homozygous sites due to a larger standard deviation of the heterozygous variant allele frequency (VAF) distribution. The distribution of allele balance across all 346,894 positions showed a 'belly' on the left of the normal distribution (AF between 0.2 and 0.35, Fig. 2A-left). Specifically, the VAF of SNVs classified as very low confidence was skewed (Fig. 2A-middle, red distribution), with a large

fraction showing VAF between 0.2 and 0.35. Application of the ABB filter resulted in a 'clean' normal distribution (Fig. 2A-right).

The transition-transversion ratio expected to be around 3 in exomes was significantly smaller for very low (ABB > 0.9) compared to high (ABB < 0.15) confidence positions (1.76 compared to 2.54, p-value < 10^{-16}). Moreover, low confidence sites showed significantly increased fisher strand bias (p-value < 10^{-16}). Furthermore, very low confidence sites were enriched for segmental duplications (27.10% of positions, p-value < 10^{-6}) and simple sequence repeats (SSRs, 16.27% of positions, p-value < 10^{-6}), compared to high confidence sites (2.50% and 0.95% of positions for segmental duplications and SRRs, respectively) (Supp. Table S5).

In order to test the applicability of ABB for improving somatic mutation callsets we generated somatic SNV calls for 200 CLL tumor-normal pairs using MuTect and obtained pre-computed ABB scores for each site (importantly, note that the ABB model is not calculated using tumor tissues, but scores are obtained from the germline-based model described above). ABB low and very low confidence positions represented 5.9% and 8.1% of the somatic mutation calls, respectively (Table 1), representing a significant increase of very low confidence positions compared to 3.8% observed for germline variant calling (p-value < 10^{-16}) and the exome-wide expectation (p-value < 10^{-16}). Interestingly, 45.38% of the very low confidence mutations were found in dbSNP. This proportion was significantly higher (p-value < 10^{-16}) than the fraction of high confidence somatic mutations in dbSNP (9.51%, Table 2), pointing at a systematic introduction of errors in dbSNP.

To demonstrate that our model is not falsely labeling real somatic mutations as systematic errors we intersected positions marked as systematic errors (ABB>=0.9) with 1896 somatic mutations validated in two studies (Papaemmanuil et al., 2014; Tarpey et al., 2013) and 341 well-known cancer driver mutation hotspots (Chang et al., 2016). We found a minimal overlap of 2/1896 and 0/341, respectively. There was no significant difference between the fraction of systematic errors identified in the whole exome (Table 1) and the set of validated somatic SNV positions (p-value = 0.1421),

demonstrating that ABB does not misclassify true somatic mutations as systematic errors more than expected by chance (Supp. Table S6).

Sanger sequencing based evaluation of ABB scores

We next evaluated if ABB scores correlate with the probability of calling false positive variants. To this end, we validated by Sanger a set of randomly selected 209 heterozygous SNPs predicted by GATK HaplotypeCaller with VQSR in 10 samples, which had AB between 0.2 and 0.35, and that were sampled equally from each of the four ABB genotype callability levels (Methods). We found that ABB genotype callability levels correlated with false positive rate (FPR) (Fig. 2C, Table 3 and Supp. Table S7). Furthermore, ABB scores were predictive of false positive calls (ROC-AUC = 0.778, Fig. 2B). Although the original variant callset produced by GATK HaplotypeCaller and VQRS could be considered high quality (tranche threshold of 99.9%), we found an FPR of 50% in the very low confidence set and 31% FPR in the low confidence set, while high and medium confidence positions showed only 0% and 15.4% FPR, respectively. Interestingly, the fraction of failed (ambiguous) Sanger sequencing experiments was significantly higher for the low confidence range when compared against high confidence range (p-value < 0.025, Table 3), indicating that low complexity regions and repeats constitute one of the underlying issues, as these also affect efficiency of Sanger sequencing (Kieleczawa, 2006).

Next, we compared the performance of ABB and the GIAB callability classifier on identification of false positive calls. Considering all exons of the autosomes (79,660,917bp included in the ABB model), GIAB classifies 75,442,680 sites as callable, leaving 4,218,237 sites as 'un-callable'. In comparison ABB classifies 46,396 sites as low or very low confidence (ABB >= 0.75, considered un-callable from here on). Of the 46,396 sites classified un-callable by ABB, 52% are classified as callable by GIAB, demonstrating that the two methods are not redundant (Supp. Fig. S3). Of the 40 GATK SNV calls confirmed as false by Sanger sequencing (out of the 209 sites evaluated by Sanger) ABB identified 30 (75%), while GIAB identified 23 (57.5%), although ABB filters substantially

fewer sites across the whole exome than GIAB (40kb vs. 4MB) (see Supporting Material and Supp. Table S8 for details). In a similar manor, we showed that both GQ and Hardy-Weinberg Equilibrium provide complementary, but not redundant sources of information for filtering of variant calls (details in Supporting Materials, Supp. Figs. S4-S6 and Supp. Tables S9 and S10).

Independent from the random Sanger evaluation we obtained validation data for disease variant candidates (*novel* mutations) prioritized in in-house analysis of various disorders (data unpublished). In each study, almost 50% of candidate variants were found to be false positives by Sanger validation. ABB labeled 11 out of 17 (64.7%) false positive calls as low or very low confidence sites, and 6 FPs as medium confidence, while all TP variants fell into the high confidence category (Supp. Table S11). We observed a large margin between ABB for TPs (average of 0.115) and FPs (average of 0.666).

ABB scores of variants in public databases

Public variant repositories differ in the way included variants are called, quality controlled and selected for integration. For instance, the 1000GP, ExAC/GnomAD and EVS databases are created in a consistent manner, using a defined pipeline for all samples (Lek et al., 2016). However, dbSNP does not dictate any specific variant prediction method or quality control procedure and contains both germline and somatic variants. Hence, we hypothesized that although all variant databases may contain systematic errors, dbSNP is specifically affected by false positives due to its inconsistent quality parameters, as previously suggested (Musumeci, 2011). We found that very low confidence positions were significantly enriched in several public variant databases (all p-values $< 10^{-16}$, see Table 4). As expected, we found the strongest enrichment of systematic errors (ABB > 0.9) in dbSNP (15.9 times more than expected). As many variant analysis pipelines use the same tools as employed for generating 1000GP, ExAC, GnomAD or EVS, one should be cautious when considering variants found in these databases as validation gold standard for variants in newly generated callsets. As we expect to see systematic errors repeatedly, this circularity issue (validation using false variants

predicted by the same tools) can lead to a 'self-fulfilling prophecy', where false variants are established as true positives in public databases and potentially influence disease studies in the future.

Filtering candidate genes from Rare Variant Association Studies

Whole exome sequencing is frequently applied to identify causal variants for genetic diseases, using rare variant association tests in large cohorts or analysis of affected families and parent-child trios. Although ABB can be used generically to filter results of variant callsets, we have in addition developed a custom algorithm, Association-ABB, for identification of cohort-specific false associations caused by systematic errors (Methods). We hypothesized that false associations can be introduced in case-control studies due to 1) a bias in systematic errors between cases and controls, leading to an uneven burden of false variant calls, or 2) copy number variants enriched in cases or controls e.g. due to biased population structure. Therefore, we re-analyzed variants in candidate genes in order to identify associations better explained by biases in the burden of systematic errors.

The Association-ABB evaluation was performed on candidate genes resulting from an RVAS for Chronic Lymphocytic Leukemia (CLL) using WES of 437 CLL normal samples from ICGC-CLL (Quesada et al., 2011) and 780 control samples. In the 43 resulting candidate genes identified by SKAT-O and Burden (see Methods), Association-ABB labeled 24 out of 739 SNPs as affected by a bias in systematic errors that were called as variants to a different extent in cases and controls ('called-missed ratio fisher test', see Methods and Supp. Table S12). In addition, these variants showed a high ABB score (average of 0.8670). We next performed a gene-wise aggregated test of biased sites and found that 10 out of 43 candidate genes were likely false associations (Supp. Table S13). In brief, we tested if the RVAS association was still significant when 1) biased sites were excluded or 2) potentially missed calls were added to the test (Methods).

One example gene, *CTDSP2*, is shown in Supp. Fig. S7. We observed that a different AB distribution in cases and controls in 8 biased positions led to an imbalanced genotyping efficiency of GATK HaplotypeCaller, explaining the significant association of this gene with CLL in the RVAS test.

Comparing called vs. potentially missed SNVs we found a significant enrichment of missed calls in the controls, i.e. positions called as homozygous reference, although more than expected reads showed the alternative allele. However, not only cases showed an enrichment of calls with significantly deviated AB in heterozygotes (AB around 0.25), but also controls that had been enriched using Agilent SureSelect, while controls prepared with NimblegenSeqEz were 'clean' (Supp. Fig. S8 and Supporting Material). We conclude that a systematic issue with few target regions of one enrichment kit introduced the false RVAS call.

The AB patterns between cases and controls in the gene CDC27 look similar to CTDSP2, as shown in Supp. Fig. S9, although this enrichment was not associated to the capture method as CTDSP2 (see Supp. Fig. S10). Moreover, literature search revealed that this gene frequently harbors false positive SNVs (Jia et al., 2012), likely caused by multiple novel retroduplications (Abyzov et al., 2013). Indeed, we found that cases with deviating AB also showed significantly increased coverage on the exons affected potentially by retroduplications (Supp. Fig. S11) (See Supporting Material for detailed explanation of this section).

Association testing when removing problematic site in CDC27 or CTDSP2 ('cleaning') or when adding potentially missed calls led to non-significant association tests. In summary, ABB identified retrotransposition as well as exome hybridization kit-related systematic errors causing false associations in an RVAS study of CLL.

DISCUSSION

In this work, we present a new genotype callability filter for exome or genome sequencing analysis, which is based on the recurrent and significant deviation of observed from expected allele balance at a genomic position across hundreds of NGS datasets. We termed the underlying phenomenon allele balance bias (ABB). Up to 4% of the positions called as germline variants and 8% of positions called as somatic mutations by state-of-the art methods show ABB scores indicative of systematic errors. We used Sanger validation of random germline calls and of disease variant candidates to show that ABB

correlates with the likelihood to identify false positive SNVs, with more than 50% FPR in the lowest genotype confidence range. Furthermore, ABB low and very low confidence positions show a low transition-transversion ratio (TiTv) (Freudenberg-hua et al., 2003; Pattnaik, Vaidyanathan, Pooja, Deepak, & Panda, 2012) and are highly enriched for low complexity regions, supporting the hypothesis that LCRs are responsible for a large fraction of systematic errors (H. Li, 2014). Nonetheless, our findings indicate that several other issues can cause systematic errors, including incomplete reference genomes and unknown CNVs or segmental duplications, among others.

Although the accuracy of variant callers has been optimized since the introduction of NGS, there are still systematic errors that cannot be identified by the current set of QC parameters. While ABB shows partial correlation with other QC measures like Fisher strand bias and LCRs, none of these parameters can identify the complete set of positions flagged by ABB, making ABB a valuable addition to the QC filter setup. Interestingly, we found that sites prone to systematic errors are highly enriched in public variant databases. As these databases are often used for benchmarking purposes this can lead to a 'fixation' of falls calls, and can skew benchmark results. dbSNP by far showed the highest enrichment of systematic errors, as suggested previously (Musumeci, 2011), demonstrating that variant callsets created consistently by a defined and reproducible pipeline and parameter setting (e.g. 1000GP, ExAC/GnomAD, EVS) are preferable. Systematic errors constitute an even bigger issue for somatic SNV calling. Even if ultra-deep sequencing is used to identify sub-clonal mutations, systematic errors, other than random errors, will still lead to false somatic SNV calls (Griffith et al., 2015). Indeed, we observed that close to 14% of somatic SNVs called by MuTect were classified as ABB low or very low confidence sites, a significantly larger fraction than observed for germline variant calling or expected on an exome-wide level. Moreover, these false positive mutations are again highly enriched in dbSNP. Considering the importance of predicted point mutations for cancer diagnostics and optimal treatment selection, removal of these FP calls is essential for the applicability of NGS in precision oncology.

Systematic false positive calls can lead to false associations of genes with disease. Using Sanger validation of disease gene candidates prioritized in previous projects we demonstrated that high-confidence ABB sites are 100% true positives, allowing to reduce the cost of Sanger validation by omitting validation of these sites. At the same time ABB identifies up to 65% of false candidates (considering low or very low confidence sites, up to 100% if also considering medium confidence as FP). We further demonstrate how systematic errors resulting in false associations can be identified by Association-ABB in a cohort specific manner. We found that in a rare variant association test for CLL around 25% of candidate gene associations were better explained by uneven burden of systematic errors in cases and controls. We further hypothesized that systematic SNV calling errors were introduced by an un-annotated CNV in at least couple of candidate genes, indirectly pointing to the real cause of the genotype-phenotype association.

The current ABB model has been built using alignments generated by bwa-mem. Hence, some systematic errors identified in our study might reflect specific alignment issues of bwa-mem and might not be observed when using e.g. bowtie2. However, bwa-mem is one of the most used aligners for human genomics, making our model directly applicable to a majority of projects using whole-exome or whole-genome sequencing of human samples. Nonetheless, one could retrain the ABB model for specific computational analysis pipelines, for other species, for whole genomes or using thousands of additional samples, a process we support by offering all scripts for generating dedicated ABB models (see Availability section).

In summary, our novel genotype callability estimator based on allele balance bias (ABB) can identify systematic variant calling errors not found by other measures and can improve the accuracy of germline and somatic variant sets as well as disease association studies in families or large cohorts.

AVAILABILITY

ABB tool is an open source package available in the GitHub repository (https://github.com/Francesc-Muyas/ABB). The pre-computed ABB score can be downloaded in https://public_docs.crg.es/sossowski/publication_data/ABB/ABB_SCORE.txt.

ACCESSION NUMBERS

Sequencing data of CLL individuals have been deposited at the European Genome-Phenome Archive (EGA, http://www.ebi.ac.uk/ega/), which is hosted by the EBI and the CRG), under accession number EGAS00000000092. Variant calls have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001003027.

ACKNOWLEDGEMENT

We acknowledge support of the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2017-2021', SEV-2016-0571; the CERCA Programme / Generalitat de Catalunya; the "la Caixa" Foundation; the CRG emergent translational research award; and the European Union's H2020 research and innovation programme under grant agreement No 635290 (PanCanRisk). LD is supported by a MINECO Severo Ochoa fellowship (SVP-2013-0680066). RR is funded through the PERIS program (SLT002/16/00310).

REFERENCES

Abyzov, A., Iskow, R., Gokcumen, O., Radke, D. W., Balasubramanian, S., Pei, B., … Gerstein, M. (2013). Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. *Genome Research*, *23*(12), 2042–52. http://doi.org/10.1101/gr.154625.113

Alioto, T. S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M. D., Hovig, E., … Gut, I. G. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications*, *6*, 10001. http://doi.org/10.1038/ncomms10001

Altmann, A., Weber, P., Bader, D., Preuß, M., Binder, E. B., & Müller-Myhsok, B. (2012). A beginners guide to SNP calling from high-Throughput DNA-sequencing data. *Human Genetics*, *131*(10), 1541–1554. http://doi.org/10.1007/s00439-012-1213-z

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. http://doi.org/10.1038/nature15393

Bao, R., Huang, L., Andrade, J., Tan, W., Kibbe, W. a, Jiang, H., & Feng, G. (2014). Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing. *Libertas Academica*, *13*, 67–82. http://doi.org/10.4137/CIN.S13779.Received

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. http://doi.org/10.1038/nature07517

Cai, L., Yuan, W., Zhang, Z., He, L., & Chou, K.-C. (2016). In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Scientific Reports*, *6*(1), 36540. http://doi.org/10.1038/srep36540

Carson, A. R., Smith, E. N., Matsui, H., Brækkan, S. K., Jepsen, K., Hansen, J.-B., & Frazer, K. a. (2014). Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics*, *15*, 125. http://doi.org/10.1186/1471-2105-15-125

Chang, M. T., Asthana, S., Gao, S. P., Lee, B. H., Chapman, J. S., Kandoth, C., … Taylor, B. S. (2016). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature Biotechnology*, *34*(2), 155–163. http://doi.org/10.1038/nbt.3391

Chen, L., Liu, P., Evans, T. C., & Ettwiller, L. M. (2017). DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science (New York, N.Y.)*, *355*(6326), 752–756. http://doi.org/10.1126/science.aai8690

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., … Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, *31*(3), 213–9. http://doi.org/10.1038/nbt.2514

Cordaux, R., & Batzer, M. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, *10*(10), 691–703. http://doi.org/10.1038/nrg2640.The

De Summa, S., Malerba, G., Pinto, R., Mori, A., Mijatovic, V., & Tommasi, S. (2017). GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. *BMC Bioinformatics*, *18*(S5), 119. http://doi.org/10.1186/s12859-017-1537-8

DePristo, M. a, Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. http://doi.org/10.1038/ng.806

Freudenberg-hua, Y., Freudenberg, J., Kluck, N., Cichon, S., Propping, P., & Nöthen, M. M. (2003). Single Nucleotide Variation Analysis in 65 Candidate Genes for CNS Disorders in a Representative Sample of the European Population Single Nucleotide Variation Analysis in 65 Candidate Genes for CNS Disorders in a Representative Sample of the European Popu. *Genome Research*, 2271–2276. http://doi.org/10.1101/gr.1299703

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv Preprint arXiv:1207.3907*, 9. http://doi.org/arXiv:1207.3907 [q-bio.GN]

Gómez-Romero, L., Palacios-Flores, K., Reyes, J., García, D., Boege, M., Dávila, G., … Palacios, R. (2018). Precise detection of de novo single nucleotide variants in human genomes. *Proceedings*

of the National Academy of Sciences of the United States of America, *115*(21), 5516–5521. http://doi.org/10.1073/pnas.1802244115

Graffelman, J., & Moreno, V. (2013). The mid p-value in exact tests for Hardy-Weinberg equilibrium. *Statistical Applications in Genetics and Molecular Biology*, *12*(4), 433–48. http://doi.org/10.1515/sagmb-2012-0039

Griffith, M., Miller, C. a., Griffith, O. L., Krysiak, K., Skidmore, Z. L., Ramu, A., … Wilson, R. K. (2015). Optimizing Cancer Genome Sequencing and Analysis. *Cell Systems*, *1*(3), 210–223. http://doi.org/10.1016/j.cels.2015.08.015

Guo, Y., Li, J., Li, C.-I., Long, J., Samuels, D. C., & Shyr, Y. (2012). The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics*, *13*(1), 666. http://doi.org/10.1186/1471-2164-13-666

Guo, Y., Samuels, D. C., Li, J., Clark, T., Li, C.-I., & Shyr, Y. (2013). Evaluation of allele frequency estimation using pooled sequencing data simulation. *TheScientificWorldJournal*, *2013*, 895496. http://doi.org/10.1155/2013/895496

Hardwick, S. A., Deveson, I. W., & Mercer, T. R. (2017). Reference standards for next-generation sequencing. *Nature Reviews Genetics*, *18*(8), 473–484. http://doi.org/10.1038/nrg.2017.44

He, L., Pitkäniemi, J., Sarin, A.-P., Salomaa, V., Sillanpää, M. J., & Ripatti, S. (2015). Hierarchical Bayesian Model for Rare Variant Association Analysis Integrating Genotype Uncertainty in Human Sequence Data. *Genetic Epidemiology*, *39*(2), 89–100. http://doi.org/10.1002/gepi.21871

Hou, L., Sun, N., Mane, S., Sayward, F., Rajeevan, N., Cheung, K.-H., … Zhao, H. (2017). Impact of genotyping errors on statistical power of association tests in genomic analyses: A case study. *Genetic Epidemiology*, *41*(2), 152–162. http://doi.org/10.1002/gepi.22027

Hu, Y.-J., Liao, P., Johnston, H. R., Allen, A. S., & Satten, G. A. (2016). Testing Rare-Variant Association without Calling Genotypes Allows for Systematic Differences in Sequencing between Cases and Controls. *PLoS Genetics*, *12*(5), e1006040. http://doi.org/10.1371/journal.pgen.1006040

Hwang, K., Lee, I., Park, J., Hambuch, T., Choi, Y., Kim, M., … Kong, S. W. (2015). Reducing false positive incidental findings with ensemble genotyping and logistic regression-based variant filtering methods, *35*(8), 936–944. http://doi.org/10.1002/humu.22587.Reducing

Jia, P., Li, F., Xia, J., Chen, H., Ji, H., Pao, W., & Zhao, Z. (2012). Consensus Rules in Variant Detection from Next-Generation Sequencing Data. *PLoS ONE*, *7*(6), e38470. http://doi.org/10.1371/journal.pone.0038470

Johnston, H. R., Hu, Y., & Cutler, D. J. (2015). Population genetics identifies challenges in analyzing rare variants. *Genetic Epidemiology*, *39*(3), 145–8. http://doi.org/10.1002/gepi.21881

Kieleczawa, J. (2006). Fundamentals of sequencing of difficult templates-An overview. *Journal of Biomolecular Techniques*, *17*(3), 207–217.

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., … Ding, L. (2009). VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, *25*(17), 2283–2285. http://doi.org/10.1093/bioinformatics/btp373

Koboldt DC, Larson DE, W. R. (2013). Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. *Curr Protoc Bioinformatics*, *44*, 15.4.1–15.4.17. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25553206

Ledergerber, C., & Dessimoz, C. (2011). Base-calling for next-generation sequencing platforms. *Briefings in Bioinformatics*, *12*(5), 489–497. http://doi.org/10.1093/bib/bbq077

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *American Journal of Human Genetics*, *95*(1), 5–23. http://doi.org/10.1016/j.ajhg.2014.06.009

Lee, S., Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, *13*(4), 762–775. http://doi.org/10.1093/biostatistics/kxs014

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … Consortium, E. A. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. http://doi.org/10.1038/nature19057

Li, B., & Leal, S. M. (2008). Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *The American Journal of Human Genetics*, *83*(3), 311–321. http://doi.org/10.1016/j.ajhg.2008.06.024

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. http://doi.org/10.1093/bioinformatics/btr509

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Preprint arXiv*, *0*(0), 3. http://doi.org/arXiv:1303.3997 [q-bio.GN]

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics (Oxford, England)*, 1–9. http://doi.org/10.1093/bioinformatics/btu356

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. http://doi.org/10.1093/bioinformatics/btp324

Liu, D. J., & Leal, S. M. (2010). A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. *PLoS Genetics*, *6*(10), e1001156. http://doi.org/10.1371/journal.pgen.1001156

Madsen, B. E., & Browning, S. R. (2009). A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genetics*, *5*(2), e1000384. http://doi.org/10.1371/journal.pgen.1000384

McKenna, S. (2009). The Genome Analysis Toolkit. *Proceedings of the International Conference on Intellectual Capital, Knowledge Management & Organizational Learning*, 254–260. http://doi.org/10.1101/gr.107524.110.20

Musumeci, L. et al. (2011). Single Nucleotide Differences (SNDs) in the dbSNP Database May Lead to Errors in Genotyping and Haplotyping Studies, *20*(4), 200–210. http://doi.org/10.1016/j.semcancer.2010.06.001.All

Newman, A. M., Lovejoy, A. F., Klass, D. M., Kurtz, D. M., Chabon, J. J., Scherer, F., … Alizadeh,

A. A. (2016). Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol*, (March), Article in press. http://doi.org/10.1038/nbt.3520

Nothnagel, M., Wolf, A., Herrmann, A., Szafranski, K., Vater, I., Brosch, M., … Krawczak, M. (2011). Statistical inference of allelic imbalance from transcriptome data. *Human Mutation*, *32*(1), 98–106. http://doi.org/10.1002/humu.21396

O'Fallon, B. D., Wooderchak-Donahue, W., & Crockett, D. K. (2013). A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data. *Bioinformatics*, *29*(11), 1361–1366. http://doi.org/10.1093/bioinformatics/btt172

Oleksiewicz, U., Tomczak, K., Woropaj, J., Markowska, M., Stępniak, P., & Shah, P. K. (2015). Review Computational characterisation of cancer molecular profiles derived using next generation sequencing. *Współczesna Onkologia*, *1A*, 78–91. http://doi.org/10.5114/wo.2014.47137

Ospina, R., & Ferrari, S. L. P. (2009). Inflated beta distributions. *Statistical Papers*, *51*(1), 111–126. http://doi.org/10.1007/s00362-008-0125-4

Ospina, R., & Ferrari, S. L. P. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics and Data Analysis*, *56*(6), 1609–1623. http://doi.org/10.1016/j.csda.2011.10.005

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., … Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, *15*(2), 256–278. http://doi.org/10.1093/bib/bbs086

Papaemmanuil, E., Rapado, I., Li, Y., Potter, N. E., Wedge, D. C., Tubio, J., … Campbell, P. J. (2014). RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nature Genetics*, *46*(2), 116–125. http://doi.org/10.1038/ng.2874

Park, G., Park, J. K., Shin, S.-H., Jeon, H.-J., Kim, N. K. D., Kim, Y. J., … Park, D. (2017). Characterization of background noise in capture-based targeted sequencing data. *Genome Biology*, *18*(1), 136. http://doi.org/10.1186/s13059-017-1275-2

Pattnaik, S., Vaidyanathan, S., Pooja, D. G., Deepak, S., & Panda, B. (2012). Customisation of the exome data analysis pipeline using a combinatorial approach. *PLoS ONE*, *7*(1). http://doi.org/10.1371/journal.pone.0030080

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L., & Mayer, G. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports*, *8*(1), 10950. http://doi.org/10.1038/s41598-018-29325-6

Price, A. L., Kryukov, G. V., de Bakker, P. I. W., Purcell, S. M., Staples, J., Wei, L.-J., & Sunyaev, S. R. (2010). Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *The American Journal of Human Genetics*, *86*(6), 832–838. http://doi.org/10.1016/j.ajhg.2010.04.005

Puente, X. S., Beà, S., Valdés-Mas, R., Villamor, N., Gutiérrez-Abril, J., Martín-Subero, J. I., … Campo, E. (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*,

526(7574), 519–524. http://doi.org/10.1038/nature14666

Quesada, V., Conde, L., Villamor, N., Ordóñez, G. R., Jares, P., Bassaganyas, L., … López-Otín, C. (2011). Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nature Genetics*, *44*(1), 47–52. http://doi.org/10.1038/ng.1032

Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., & Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, *28*(14), 1811–1817. http://doi.org/10.1093/bioinformatics/bts271

Sleep, J. a, Schreiber, A. W., & Baumann, U. (2013). Sequencing error correction without a reference genome. *BMC Bioinformatics*, *14*(1), 367. http://doi.org/10.1186/1471-2105-14-367

Sun, J., Zheng, Y., & Hsu, L. (2013). A Unified Mixed-Effects Model for Rare-Variant Association in Sequencing Studies. *Genetic Epidemiology*, *37*(4), 334–344. http://doi.org/10.1002/gepi.21717

Tarpey, P. S., Behjati, S., Cooke, S. L., Van Loo, P., Wedge, D. C., Pillay, N., … Futreal, P. A. (2013). Frequent mutation of the major cartilage collagen gene COL2A1 in chondrosarcoma. *Nature Genetics*, *45*(8), 923–926. http://doi.org/10.1038/ng.2668

Van der Auwera, G. a., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. a. (2013). *From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline*. *Current Protocols in Bioinformatics*. http://doi.org/10.1002/0471250953.bi1110s43

Veltman, J. A., & Brunner, H. G. (2012). De novo mutations in human genetic disease. *Nature Publishing Group*, *13*. http://doi.org/10.1038/nrg3241

Vodák, D., Lorenz, S., Nakken, S., Aasheim, L. B., Holte, H., Bai, B., … Hovig, E. (2018). Sample-Index Misassignment Impacts Tumour Exome Sequencing. *Scientific Reports*, *8*(1), 5307. http://doi.org/10.1038/s41598-018-23563-4

Yan, Q., Chen, R., Sutcliffe, J. S., Cook, E. H., Weeks, D. E., Li, B., & Chen, W. (2016). The impact of genotype calling errors on family-based studies. *Nature Publishing Group*. http://doi.org/10.1038/srep28323

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., … Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, *3*, 160025. http://doi.org/10.1038/sdata.2016.25

FIGURE LEGENDS

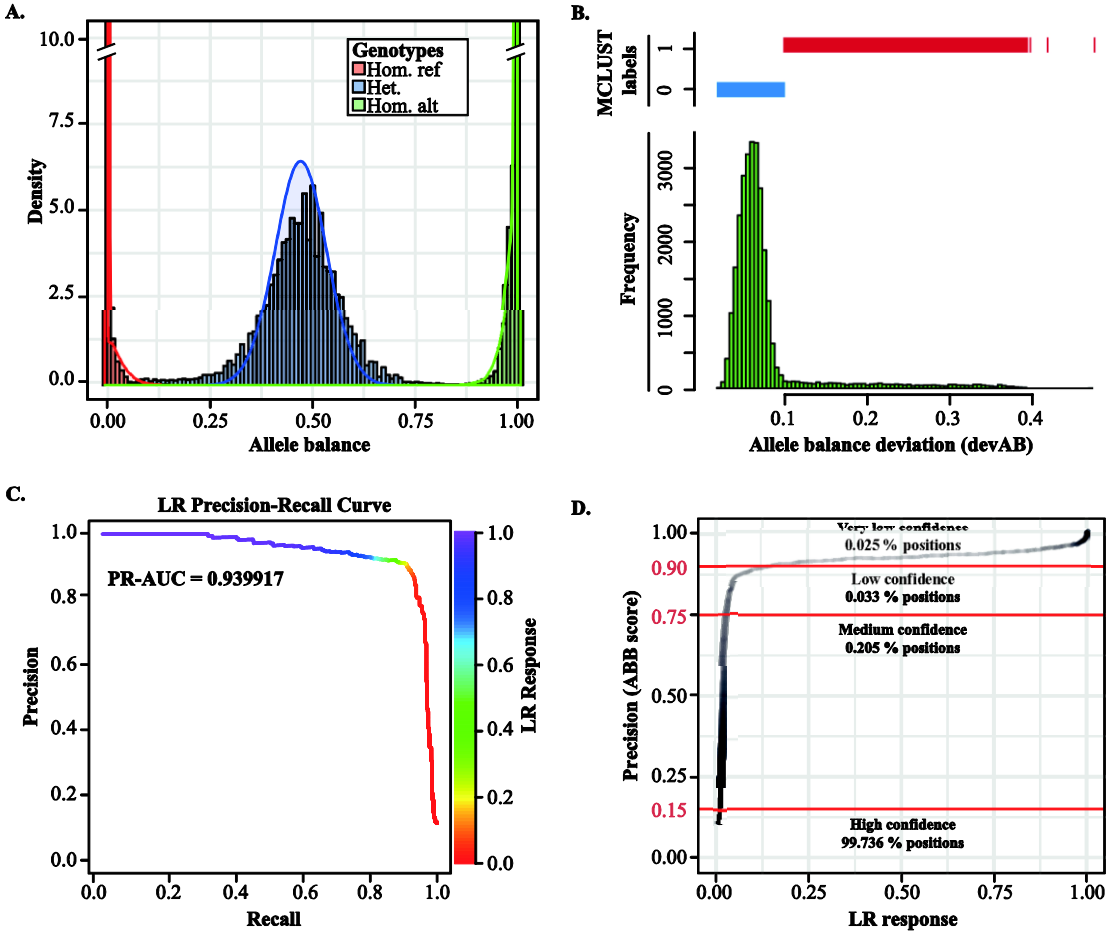

Figure 1. (A) Observed (bars) and expected (density) allele balance (AB) distributions split by genotype. (B) Gaussian mixture model of the allele balance deviation devAB, separating non-deviated (0) and deviated (1) positions. (C) Precision-Recall curves and PR-AUC for the linear regression model LR-1. The color gradient on the right shows the LR response value (probability to belong to class 1) obtained by logistic regression. (D) Correlation of LR response and precision. Precision was measured in the test and validation sets using labels defined by the GMM. Confidence levels were defined by visual inspection.

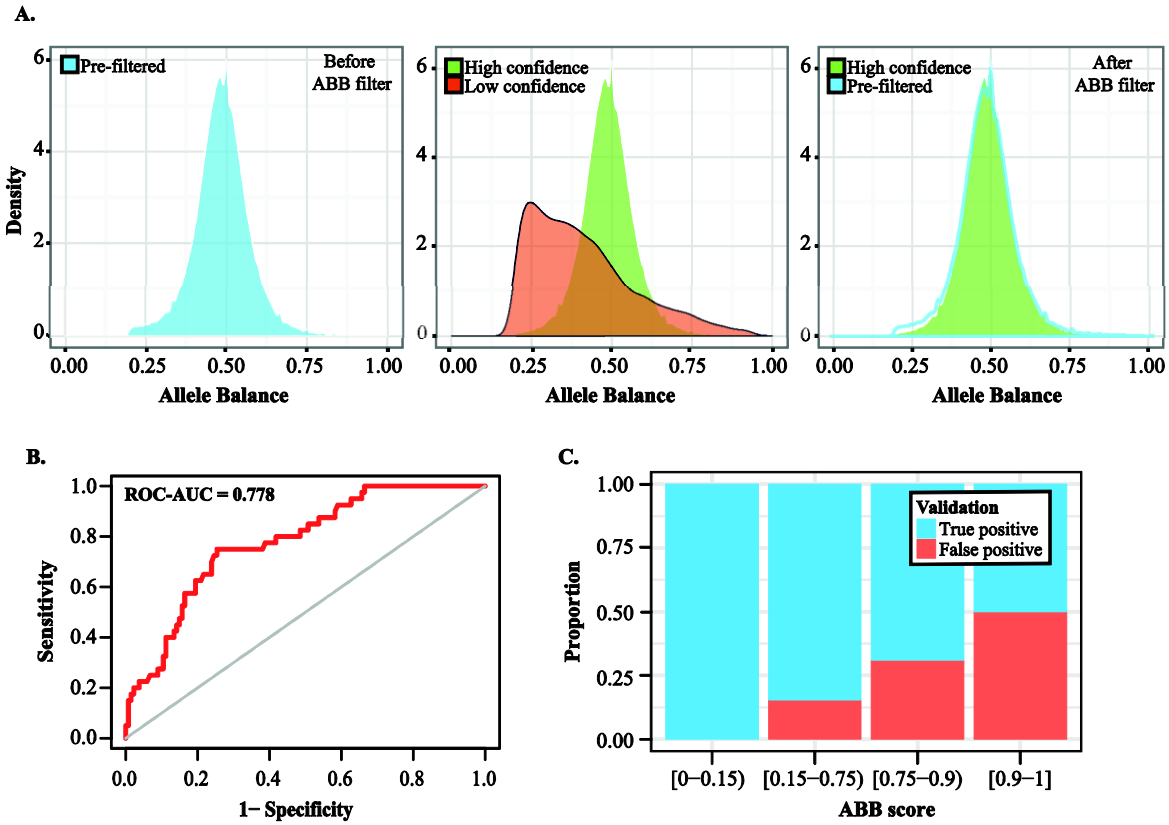

Figure 2. (A) ABB classifications of heterozygous SNPs reported by GATK Haplotype Caller. Shape of AB distribution of variants identified by GATK + VQSR (left); AB distribution of low (red) compared to high (green) confidence positions (middle); and AB distribution after ABB filtering (right). (B) ROC curve of Sanger validation results compared with ABB (AUC = 0.778). (C) Proportion of True Positive (TP) and False Positive (FP) variants in four ABB genotype callability ranges.

Table 1. Distribution of ABB genotype callability levels in the whole exome, germline SNV calls and somatic SNV calls.

ABB callability	Whole Exome	Germline SNV	Somatic SNV
High Confidence [0-0.15)	99.736%	44.955%	80.286%
Medium Confidence [0.15-0.75)	0.205%	44.585%	5.771%
Low Confidence [0.75-0.9)	0.033%	6.665%	5.865%
Very Low Confidence [0.9-1]	0.025%	3.796%	8.077%

Table 2. Enrichment of somatic SNV calls in dbSNP and Cosmic, separated by ABB callability range. Row 1 shows results for the complete call set used as baseline.

ABB callability	Novel	Cosmic	DbSNP
All SNVs [0-1]	80.89%	4.53%	14.58%
High Confidence [0-0.15)	85.60%	4.89%	9.51%
Mid Confidence [0.15-0.75)	68.08%	3.47%	28.45% ***
Low Confidence [0.75-0.9)	67.95%	4.06%	27.99% ***
Very Low Confidence [0.9-1]	52.60%	2.02% *	45.38% ***

* p-value < 10-E3

** p-value < 10E-6

*** p-value < 2E-16

Table 3. Results of Sanger validation grouped by ABB genotype callability levels. Failed Sanger sequencing experiments were ignored for the FP and TP rate calculation.

ABB callability	SNVs	TP	TP rate	FP	FP rate	Failed	Fail rate
High Confidence [0-0.15)	42	38	100.00%	0	0.00%	4	9.52%
Mid Confidence [0.15-0.75)	73	55	84.62%	10	15.38%	8	10.96%
Low Confidence [0.75-0.9)	46	20	68.97%	9	31.03%	17	36.96%
Very Low Confidence [0.9-1]	48	21	50.00%	21	50.00%	6	12.50%

Table 4. Enrichment of ABB very low confidence (VLC) positions in public variant databases. The fraction of VLC positions in the exome was used as expected value.

Database	Total positions	VLC Obs.	VLC Freq. Obs.	Ratio Obs./Exp.
Exome	81,609,944	20,725	0.03%	1
dbSNP	3,172,724	12,787	0.40%	15.87***
EVS	1,840,709	1,114	0.06%	2.38***
1000GP	2,653,982	4,690	0.18%	6.96***
EXAC	2,662,396	3,510	0.13%	5.19***

*** (P-value < 10E-16)