Limitations in next-generation sequencing-based genotyping of breast cancer polygenic risk score loci

Alexandra Baumann[1], Christian Ruckert[2], Christoph Meier[3], Tim Hutschenreiter[1], Robert Remy[4], Benedikt Schnur[5], Marvin Döbel[6], Rudel Christian Nkouamedjo Fankep[4], Dariush Skowronek[7], Oliver Kutz[1], Norbert Arnold[8], Anna-Lena Katzke[5], Michael Forster[8], Anna-Lena Kobiela[4], Katharina Thiedig[9], Andreas Zimmer[10], Julia Ritter[11], Bernhard H.F. Weber[3,12], Ellen Honisch[13], Karl Hackmann[1], Bioinformatics Working Group of the German Consortium for Hereditary Breast & Ovarian Cancer[*], Gunnar Schmidt[5], Marc Sturm[6] & Corinna Ernst[4]

[1]Institute for Clinical Genetics, University Hospital Carl Gustav Carus at TU Dresden, Dresden, Germany; ERN GENTURIS, Hereditary Cancer Syndrome Center Dresden, Germany; National Center for Tumor Diseases Dresden (NCT/UCC), Germany: German Cancer Research Center (DKFZ), Heidelberg, Germany; Faculty of Medicine and University Hospital Carl Gustav Carus at TU Dresden, Dresden, Germany; German Cancer Consortium (DKTK), Dresden, Germany; German Cancer Research Center (DKFZ), Heidelberg, Germany; Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

[2]Institute of Human Genetics, University of Münster, Münster, Germany

[3]Institute of Human Genetics, University of Regensburg, Regensburg, Germany

[4]Center for Familial Breast and Ovarian Cancer, Center for Integrated Oncology (CIO), Medical Faculty, University of Cologne and University Hospital Cologne, Cologne, Germany

[5]Department of Human Genetics, Hannover Medical School (MHH), Hannover, Germany

[6]Institute of Medical Genetics and Applied Genomics, University Hospital Tübingen, Tübingen, Germany

[7]Department of Human Genetics, University Medicine Greifswald and Interfaculty Institute of Genetics and Functional Genomics, University of Greifswald, Greifswald, Germany

[8]Department of Gynecology and Obstetrics, Institute of Clinical Chemistry Institute of Clinical Molecular Biology, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany

[9]Division of Gynaecology and Obstetrics, Klinikum rechts der Isar der Technischen Universität München, München, Germany

[10]Institute for Human Genetics, Medical Center University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany

[11]Department of Human Genetics, Labor Berlin – Charité Vivantes GmbH, Berlin, Germany

[12]Institute of Clinical Human Genetics, University Hospital Regensburg, Regensburg, Germany

[13]Department of Gynaecology and Obstetrics, University Hospital Düsseldorf, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany

[*]A list of authors and their affiliations appears at the end of the paper.

Abstract

Considering polygenic risk scores (PRSs) in individual risk prediction is increasingly becoming the standard in genetic testing for hereditary breast cancer (BC). To calculate individual BC risks, the Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) with inclusion of the BCAC 313 or the BRIDGES 306 BC PRS is commonly used. Meaningful incorporation of PRSs relies on reproducing the allele frequencies (AFs), and hence, the distribution of PRS values, expected by the algorithm. Here, the 324 loci of the BCAC 313 and the BRIDGES 306 BC PRS were examined in population-specific database gnomAD and in real-world data sets of five centers of the German Consortium for Hereditary Breast and Ovarian Cancer (GC-HBOC), to determine whether these expected AFs are achieved with next-generation sequencing-based genotyping. Four PRS loci were non-existent in gnomAD v3.1.2 non-Finnish Europeans, further 24 loci showed noticeably deviating AFs. In real-world data, between 16 and up to 22 loci were reported with noticeably deviating AFs, and were shown to have effects on final risk prediction. Deviations depended on sequencing approach, variant caller and calling mode (forced versus unforced) employed. Therefore, this study demonstrates the necessity to apply quality assurance not only in terms of sequencing coverage but also observed AFs in a sufficiently large sample, when implementing PRSs in a routine diagnostic setting. Furthermore, future PRS design should be guided by reproducibility of expected AFs in addition to the observed effect sizes.

Keywords: polygenic risk score, breast cancer, next-generation sequencing

Introduction

The German Consortium for Hereditary Breast and Ovarian Cancer (GC-HBOC) is a consortium of interdisciplinary university centers specialized in providing counseling, genetic testing and healthcare for individuals at risk for familial breast and ovarian cancer (BC/OC). Clinical management of women found to be at increased risk for BC/OC, due to inherited pathogenic variants in established BC/OC risk genes or a strong family history of cancer, demands for accurate and age-dependent risk estimates. Numerous studies demonstrated that the effects of BC susceptibility loci, i.e., common single nucleotide variants (SNVs) and short indels, which individually contribute only slightly to individual BC risks, but whose effects can be summed up to polygenic risk scores (PRSs) which can achieve a clinically relevant degree of BC risk discrimination [1 – 3]. As the contribution of the PRS to BC risks has also been confirmed for carriers of a pathogenic variant in moderate- to high-penetrant BC risk genes [4 – 7], inclusion of PRSs in individual BC risk prediction is increasingly becoming standard in GC-HBOC centers [8].

The Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA), which is implemented in the CE-marked CanRisk web interface, provides (since v5) the straightforward inclusion of genetic germline test results, cancer family history, non-genetic risk factors and PRSs in a comprehensive model [9, 10]. It is therefore widely applied for individual BC risk prediction in routine diagnostics of the GC-HBOC centers. The CanRisk web interface allows the specification of individual PRSs either as manual input (including specification of the square root of the proportion of the overall polygenic variance explained) or, for a given set of PRSs, via upload of a VCF file with the genotype or dosage information per locus to consider. Whichever method is chosen, genotyping is the responsibility of the user. For PRSs for which VCF upload is supported, CanRisk provides specifications for incorporated loci, each including the variant (chromosome, genomic position for hg19, reference and effect allele), log odds ratio (i.e., effect size) and expected AF [11]. The given alleles and AFs arise from high-throughput genotyping using one of two arrays, iCOGS13 or OncoArray [2].

In the GC-HBOC centers, the BCAC 313 BC PRS, and its modified version, the BRIDGES 306 BC PRS [12],

are the preferred PRS variant sets employed for BC risk prediction. The genetic germline testing and genotyping of PRS loci is based on next-generation sequencing (NGS), e.g., using the TruRisk™ or further specifically adapted multigene panels, whole-exome or whole-genome sequencing (WGS). The BRIDGES 306 BC PRS excludes loci of the original BCAC 313 BC PRS that were found not to be appropriately designable using NGS, some of which were replaced by corresponding loci in linkage disequilibrium [12]. The assessment of designability was mainly based on sufficient read coverage for diagnostic purposes when using a multigene panel approach and mapping to human reference hg19. With the implementation of BC PRS analysis in routine diagnostics and the establishment of corresponding bioinformatic workflows, further technical challenges besides insufficient coverage were identified, e.g., missing variant calls or variant calling resulted in deviating alleles. Studies systematically assessing and comparing quality and pitfalls of germline genotyping using either arrays or NGS approaches, are rare and mainly date from the early days of the establishment of NGS in clinical diagnostics [13 – 16]. Hence, it cannot be excluded that the conclusions drawn (which were also contradictory with regard to NGS or array being the more reliable and preferable approach) were based on now predominantly outdated technologies. Nevertheless, it is well-known that accuracy of NGS tend to be hampered in genomic regions of low complexity, i.e., homopolymer runs, tandem repeats and strongly biased GC contents, among others [17 – 19]. In the Genome Aggregation Database (gnomAD), the largest and most widely used population-specific variant database, variants located in so-called low-complexity regions are flagged, to indicate that reported AFs may be erroneous [20, 21].

In this study, the Bioinformatics Working Group of the GC-HBOC conducted a systematic evaluation across GC-HBOC centers to develop a detailed, locus-wise assessment of technical pitfalls and possible sources of error in NGS-based PRS genotyping. A three-stage approach was followed. First, the AF of PRS variants were compared to the gnomAD AF for the European general population and it was checked if the variants can be converted to the hg38 reference genome. Second, PRS variant AFs in real-world data sets provided by participating GC-HBOC centers were compared to the AFs expected by CanRisk. Third, possible workarounds for use in clinical diagnostics, i.e., usage of alternative alleles and proxys, were identified. The presented

results are of relevance beyond diagnostics for BC risk prediction, as they demonstrate principle difficulties in NGS-based PRS computation, especially for PRSs developed based on array data. Furthermore, the results underline the necessity of a comprehensive technical evaluation of PRS variant genotyping in clinical use, as the predictive ability of an individual PRS crucially depends on the assumptions made about the underlying AFs.

Materials and Methods

Evaluation of expected allele frequencies & convertibility to hg38

Two BC PRS variant sets were considered, namely of the BCAC 313 and the BRIDGES 306 BC PRS. Of the two sets, 295 loci are identical, 18 loci are unique to BCAC 313 BC PRS, and further 11 loci are unique to the BRIDGES 306 BC PRS, resulting in a total number of N = 324 variants to be considered. Expected AFs were extracted from the corresponding PRS specification files at the CanRisk knowledge base [11]. Additionally, AFs for non-Finnish Europeans (NFEs) were obtained from the gnomAD v3.1.2 database[1], which are based on more than 33,000 WGS samples mapped to the hg38 reference sequence. For conversion of the hg19-based PRS variants from CanRisk to hg38, the gnomAD liftover feature was used.

Besides AFs, gnomAD flags and warnings indicating possible technical artifacts were retrieved and recorded. These included localization within low-complexity regions, low-quality sites (i.e., sites that are covered in less than 50% of considered samples [20]) and sites not passing the allele-specific GATK Variant Quality Score Recalibration (VQSR) filter.

Determination of deviating allele frequencies

To determine PRS variants with considerably deviating AFs, thresholds had to be defined dependent on sample sizes and variances observed. Therefore, individual thresholds per data set were determined, using an

[1] https://gnomad.broadinstitute.org

elbow of the curve method. The absolute differences between observed and expected AFs were sorted in descending order, and the absolute difference referring to the point with the largest Euclidean distance to the imaginary line between thought points (0, 1) and (N + 1, 0), were chosen as threshold, i.e., all observed absolute differences greater than this threshold were determined as noticeably deviating. Corresponding curves are shown in Supplementary Figures 1 to 6. If the same set of samples was processed with two different variant callers, the smaller threshold was applied in each case, to facilitate comparing variant caller performance.

Real-world data collection

Genotyping results for either BCAC 313 or BRIDGES 306 BC PRS loci in a sample of at least 100 individuals of European ancestry were requested from GC-HBOC centers. Participating centers submitted observed AFs per locus as well as fractions of samples that did not meet required quality criteria (e.g., with regard to minimum read depth). Furthermore, details on sequencing approaches and bioinformatic analysis workflows for PRS genotyping were systematically recorded.

In total, five GC-HBOC centers provided data, namely the Institute of Medical Genetics and Applied Genomics (IMGAG), University Hospital Tübingen, the Institute for Clinical Genetics (ICG), University Hospital Carl Gustav Carus Dresden, the Institute of Human Genetics at the University of Münster (IHG-M), the Center for Familial Breast and Ovarian Cancer (CFBOC), University Hospital Cologne, and the Institute of Human Genetics at the University of Regensburg (IHG-R). Each center provided two NGS-based data sets. An overview on data characteristics is given in Table 1. A more detailed description of sample compositions, sequencing approaches and bioinformatic analyses can be found in Supplementary Methods.

Assessment of effects of deviating allele frequencies on estimated breast cancer risks

Effects of noticeably deviating AFs of PRS loci on CanRisk-based estimated BC risks, rely on a multitude of factors, such as the number and combination of affected loci, and additional risk factors such as results of

germline testing of established BC/OC risk genes, BC/OC family history, non-genetic risk factors and current age. In principle, the effect of the PRS on BC risk is expected to be decreased in carriers of a pathogenic germline variant in a BC risk gene with moderate or high penetrance, and furthermore, its effect is expected to decrease with age [10]. In order to get an estimate of expected biases in predicted BC risks due to potentially erroneous PRS genotyping, estimates of 10 year and remaining lifetime risks, i.e., cumulative risks of primary BC until age of 80 years, were calculated using the CanRisk web interface for imaginary cancer-unaffected women of three different ages, namely 20, 40, and 60 years, without any further information than (artificial) PRS.

To simulate different scenarios, artificial VCF files were constructed with an average PRS (50th percentile) by using two times the expected CanRisk AF, i.e., expected dosage. For each data set, for loci showing noticeably deviating AFs, expected dosages were replaced by two times the observed AF in the data set. Dates of birth were set to January 1 in 2003, 1983, and 1963, to simulate 20, 40, and 60 years of age at time of risk computation, which were performed in October 2023, using CanRisk v2.3.5.

Elaboration of workarounds

Potential solutions for improving genotyping performance with respect to expected AFs could be (besides improving the calling itself) the consideration of alternative alleles or proxys. Details on the identification of potential variants to substitute for this purpose are given in Supplementary Methods. Alternative variants in gnomAD v.3.1.2 with an AF matching the expected CanRisk AF, were further evaluated using the IMGAG freebayes data, as this (i) was the largest data set in the study (n=1410), and (ii) the only WGS-based data set, which allowed genotyping of the entire set of putative proxys.

Results

Missing loci & convertibility to hg38

For four BC PRS loci, no variants were listed at the specified genomic position in gnomAD v2.1.1, namely rs572022984, rs113778879, rs73754909, and rs79461387. gnomAD v3.1.2 also reported no variants for three of these four loci for corresponding loci in hg38 as defined by dbSNP [22] (Supplementary Table 1). Locus rs572022984 was listed, but with an overall allele count of zero in NFE samples (Table 2).

For two loci, conversion to hg38 resulted in a change in alleles, namely for rs143384623 (hg19: 1-145604302-C-CT; hg38: 1-145830798-C-CA) and rs550057 (hg19: 9-136146597-C-T; hg38: 9-133271182-T-C). For rs143384623, the change of the alternative allele from CT to CA did not result in a noticeable shift in AFs observed in gnomAD NFE samples (5142/13304 (0.39) in v2.1.1 versus 24316/64610 (0.38) in v3.1.2, two-sided Fisher's exact test p = 0.14). For rs550057, the observed AFs appeared exactly opposite, i.e., 3786/14828 (0.26) for allele T in gnomAD v2.1.1 and 49878/67552 (0.74) for allele C in gnomAD v3.1.2. Therefore, 1 – 49878/67552 was assumed as the gnomAD v3.1.2 effect AF at this bi-allelic site.

Allele frequencies & technical artifacts reported in gnomAD v3.1.2

For 39 of the 320 PRS loci listed with AF>0 in gnomAD v3.1.2, at least one observation of technical artifacts was reported: 38 loci were flagged as being located in low complexity regions, three as being localized at a low-quality site, and one failed the allele-specific VQSR filter (Supplementary Table 1).

Due to the absolute difference threshold 0.016 (Supplementary Figure 1), 24 loci were determined as showing deviating AFs compared to CanRisk (Figure 1, Table 2). Absolute differences ranged from 0.03 to 0.71, and for 21 out of these 24 loci (87.5%), technical artifacts were reported in gnomAD v3.1.2.

Evaluation of real-world next-generation sequencing outcome

All 50 PRS loci for which a noticeably deviating AF was observed in at least one of the data sets provided by the five participating GC-HBOC centers are listed in Table 3.

For the IMGAG DRAGEN data, 0.052 was calculated as threshold to determine noticeably deviating AFs (Supplementary Figure 2), resulting in 18 loci affected (Table 3, Figure 2). Of these, 16 were previously also

identified as missing or showing noticeably deviating AFs in gnomAD v3.1.2. The exceptions were rs62485509 and rs9931038. For IMGAG freebayes data, 0.036 was calculated as threshold (Supplementary Figure 2), resulting in 16 loci from the BCAC 313 BC PRS determined as showing a noticeably deviating AF. Of these, 11 loci were also identified as showing deviating AF in IMGAG DRAGEN data, and all but rs12406858 and rs11268668 were previously identified as missing or showing deviating AFs in gnomAD v3.1.2.

Considering genotyping data provided by ICG, 23 of the overall 324 PRS loci did not meet the minimum quality criteria (read depth ≥20) in more than 25% of samples and were discarded (Supplementary Table 2). Additionally, GATK reported read depth <20 for >25% of samples for rs143384623. For 266 of the remaining 300 PRS loci (88.67%), forced genotyping with GATK and freebayes resulted in observation of identical AFs. For both ICG GATK and freebayes data, 0.053 was calculated as threshold to determine noticeably deviating AFs (Supplementary Figure 3). Using this threshold, 19 loci showed noticeably deviating AFs in each dataset (including two loci exclusive for BCAC 313 BC PRS), with an overlap of 13 (Table 3, Figure 2).

The IHG-M provided GATK- and DRAGEN-based BRIDGES 306 BC PRS genotyping data of 593 samples. Locus rs138179519 did not meet the quality criteria, and additionally rs774021038 using DRAGEN. Of the remaining 304 loci, 252 (82.89%) showed identical AFs (Supplementary Table 2). Using a threshold of 0.046 (Supplementary Figure 4), resulted in 22 loci showing deviating AFs in GATK data, respectively 16 loci in DRAGEN data, with an overlap of 11 loci.

For the CFBOC data based on 416 samples, a threshold of 0.046 was calculated (Supplementary Figure 5). The loci of the BRIDGES 306 BC PRS were considered, 243 (79.41%) of which showed identical AFs for both callers applied (Supplementary Table 2). Overall 23 loci (all of which are included also in the BCAC 313 BC PRS) showed deviating AFs: 16 loci in GATK and 17 loci in freebayes data, with an overlap of 10 loci.

The IHG-R provided GATK- and CLC-based BRIDGES 306 BC PRS genotyping data of 251 samples (Supplementary Methods). Four loci did not meet the quality criteria in both settings, and additional four in the CLC setting. Of the remaining 298 loci, 228 (76.51%) showed identical AFs (Supplementary Table 2). Using a threshold of 0.063 (Supplementary Figure 6), resulted in 23 loci showing noticeably deviating AFs in GATK

193 data, respectively 19 loci in CLC data, with an overlap of 10 loci.

194 In summary, for five loci, deviating AFs were reported in all GC-HBOC real-world settings examined, namely
195 for rs56097627, rs113778879, rs57589542, rs3988353, and rs3057314. Further three loci, namely
196 rs574103382, rs73754909, and rs57920543, were reported with deviating AFs in all settings except for one
197 (Table 3).

198 However, there were also 13 loci that were conspicuous in a single setting exclusively, namely four in IHG-R
199 GATK data (rs1511243, rs4880038, rs1027113, rs1111207), three in IHG-R CLC-data (rs10975870, rs11049431,
200 rs144767203), two each in ICG freebayes data (rs147399132, rs199504893) and IHG-M GATK data
201 (rs143384623, rs66987842), and one each in IMGAG DRAGEN (rs9931038) and IMGAG freebayes data
202 (rs12406858). Another 6 loci (rs34207738, rs10074269, rs55941023, rs851984, rs9421410, rs35054928) showed
203 AF deviations in only one center, but these were concordant.

204 Considering the loci non-existent in gnomAD v3.1.2, rs113778879 was not observed with expected AF in any GC-
205 HBOC center, and rs73754909 only with forced DRAGEN calling in IHG-M data. For rs79461387, expected AFs
206 were reported when using freebayes or forced DRAGEN calling only. Of note, rs572022984 with zero allele count
207 in gnomAD v3.1.2 NFEs and an expected AF of 0.0364 in CanRisk, was consistently not observed at all or with a
208 maximum AF of 0.005 (Supplementary Table 2).

209 Five loci showing aberrant AFs in gnomAD v3.1.2 NFEs (Table 2) were not reported with deviating AF by any of
210 the participating GC-HBOC centers, namely rs78425380, rs62331150, rs60954078, rs10862899, and rs112855987.

211

212 **Implications on risk prediction**

213 Without further information and assuming a standardized PRS at the 50th percentile, the estimated 10 year
214 risks of developing primary BC of cancer-unaffected women of 20, 40, and 60 years of age were 0.1%, 1.5%,
215 and 3.4% according to CanRisk (Supplementary Table 3). Percentiles of PRSs from artificial VCF files with
216 aberrant dosages (see Methods) ranged from 47.5% (IHG-R CLC, BRIDGES 306) up to 55.3% (ICG freebayes,
217 BCAC 313). The risk of 0.1% for a 20 year old woman was concordantly unchanged in all scenarios including

218 artificial PRSs. For a 40 year old woman, estimated 10 year risks were increased by 0.1% in seven scenarios,
219 and for a 60 year old woman by up to 0.2% in nine scenarios.
220 Estimated remaining lifetime risks of developing primary BC assuming an average PRS (50th percentile) of
221 cancer-unaffected women aged 20, 40, and 60 years are 11.3%, 10.9%, and 7.1% according to CanRisk
222 (Supplementary Table 3). When using PRSs from artificial VCF files with aberrant dosages, estimated lifetime
223 risks ranged from 11.1% up to 11.9% for a 20 year old woman, from 10.6% up to 11.4% for a 40 year old
224 woman, and from 7.0% up to 7.4% for a 60 year old woman, whereby the lowest estimates were obtained
225 with the BRIDGES 306 BC PRS based on IHG-R CLC data with 19 artificial dosages imputed, and the highest
226 with the BCAC 313 BC PRS based on ICG freebayes data with also 19 artificial dosages imputed.
227
228 **Consideration of alternative alleles and loci in linkage disequilibrium**
229 For 20 PRS loci showing noticeably deviating AFs in at least one real-world NGS data set, alternative alleles or
230 overlapping variants with minimum AF 0.01 in NFEs were reported in gnomAD v3.1.2 (Supplementary Table
231 4). For rs73754909 and rs79461387, both SNVs and non-existent in gnomAD v3.1.2, deletions were reported
232 with comparable AFs to the ones expected by CanRisk. For both deletions, the adjacent downstream
233 nucleotide of the reference sequence was identical to the substituted nucleotide of the expected effect allele
234 (Figure 3). For rs113778879, which is also an SNV not contained in gnomAD v3.1.2, a similar observation
235 could be made (Supplementary Figure 7), but the reported AF exceeds the expected one by more than 0.1
236 (0.5762 versus 0.6818).
237 For 29 out of the 50 loci showing noticeable deviating AFs in at least one real-world data set, proxys in 1000G
238 GRCh37 microarray data, 1000G GRCh38 High Coverage WGS data, or TOPMED European data could be
239 identified (Supplementary Table 5). For rs73754909, rs79461387, and rs113778879, LDpair based on GRCh38
240 reported the same alternative alleles as gnomAD v3.1.2 (Supplementary Table 4), where the original PRS loci
241 are non-existent.
242 Proxys and alternative alleles showing AFs in gnomAD v3.1.2 comparable to expected CanRisk AFs, i.e., an

absolute deviation <0.016, were considered as possible workarounds for improved PRS genotyping, and further evaluated with respect to observed AFs in IMGAG freebayes data (Table 4). For 20 of these 22 PRS loci, absolute differences between expected and observed AFs in IMGAG freebayes data remained below the previously defined IMGAG freebayes-specific threshold of 0.036. The exceptions were the substitutions of rs12406858 and rs79461387. The latter is noteworthy because the original PRS locus, which is an SNV, is correctly called by freebayes in forced and unforced mode (Table 3), whereas GATK HaplotypeCaller seems to call an overlapping deletion of sequence GAG. Also noteworthy are the potential replacements of rs73754909 and rs111833376, as both variants are consistently called with noticeably deviating AFs in real-world data sets.

Discussion

This study describes the systematic evaluation of NGS-based PRS genotyping in real-world data sets of five GC-HBOC centers. The observed AFs of PRS loci in individuals with European descent were employed as quality criterion, as the reproducibility of expected AFs of the PRS loci, and hence, the assumptions made about the overall PRS distribution, are an essential prerequisite for a correct risk calculation. In each setting under consideration, at least 14 out of 313 BCAC BC PRS loci, respectively 306 BRIDGES BC PRS loci, showed noticeably deviating AFs. These deviations were dependent on sequencing technology, variant caller and calling mode and can be expected to affect final BC risk calculations of the BOADICEA model implemented in CanRisk. Therefore, this study demonstrates the necessity to apply quality assurance not only in terms of sequencing coverage but also in terms of observed AFs in a sufficiently large cohort, when implementing PRSs in a routine diagnostic setting.

The presented results also point to potential solutions for improving genotyping performance with respect to the achievement of expected AFs for several loci, these primarily include the use of alternative variant callers or consideration of proxy variants. The use of certain variant callers resulted consistently in noticeable deviating AFs, which were not observed for other callers. This concerned e.g. rs62485509 when using

DRAGEN, and rs11268668 when using freebayes (Table 3). In each setting under investigation considering identical samples, the number of loci whose AFs match the expected AFs could be increased by variant-specific selection of the variant caller.

Comparison to large-scale population-specific data, such as gnomAD and 1000G High Coverage WGS, indicates that several PRS loci do not appear or appear with different alleles in NGS than in array-based genotyping. Here, four loci have been identified for which the use of alternative alleles could lead to the achievement of the intended, originally array-based determined AF, if NGS-based genotyping does not do so (Table 4). Two of these loci were absent in gnomAD v3.1.2 NFEs, which was also true for rs113778879 and rs572022984. As potential workaround for rs113778879, which is an SNV, an overlapping 5bp deletion was identified, but the observed AF exceeds the expected one by more than 0.1 (Supplementary Table 4). gnomAD SV v2.1 [23] reports a 1,370bp deletion starting at the same genomic position as rs572022984, namely DEL_2_27095, with an AF of 0.0417 in Europeans. However, genotyping of structural variants requires adapted variant calling approaches and therefore might be unfeasible within the scope of PRS genotyping in a routine diagnostic setting.

If no workarounds are available for loci showing noticeably deviating AFs, only imputation of the expected dosage according to CanRisk remains. This leads to smaller errors than omitting the locus from PRS calculation or setting the genotype to 0/0. However, each imputation causes a shift towards the mean PRS, and therefore imputations are meaningful only up to a certain extent.

PRSs for calculating individual BC risks will continue to evolve. For example, currently the Confluence Project[2] aims to develop multi-ancestry PRSs. In addition, PRSs become also more and more relevant for diagnostics of other diseases with a genetic component [24,25]. The presented results underline that it would facilitate the implementation in clinical routine and thus also increase the reliability of genetic diagnostics if the design of future PRSs would be guided by the reproducibility of the expected AFs in addition to the observed effect sizes. A straightforward strategy to achieve this could be to ensure comparability of AFs in large-scale

[2] https://confluence.cancer.gov

population databases, favorably based on different genotyping approaches, prior to including a locus in a PRS. This study has limitations. Larger sample sizes may have resulted in more accurate estimators of AFs. Furthermore, there was a strong enrichment for samples derived from individuals with familial BC/OC, which may have resulted in deviating AFs due to genetic load rather than technical artifacts. The genetic background could explain, e.g., the aberrant (but concordant) AFs of rs851984 in ICG data and of rs35054928 in CFBOC data. Finally, no statement can be made about whether the described AF deviations would persist when using arrays for genotyping, since corresponding analyses are not (yet) performed in any of the GC-HBOC centers.

Data availability

All data generated or analyzed during this study are included in this published article [and its supplementary files].

References

1. Lakeman IM, Hilbers FS, Rodrìguez-Girondo M, A. Lee A, Vreeswijk MP, Hollestelle A, Seynaeve C, Meijers-Heijboer H, Oosterwijk JC, Hoogerbrugge N, et al. Addition of a 161-SNP polygenic risk score to family history-based risk prediction: impact on clinical management in non-*BRCA1/2* breast cancer families. *J Med Genet.* 2019; 56:581–589.
2. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, Tyrer JP, Chen TH, Wang Q, Bolla MK, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am J Hum Genet.* 2019; 104:21–34.
3. Shieh Y, Hu D, Ma L, Huntsman S, Gard CC, Leung JW, Tice JA, Vachon CM, Cummings SR, Kerlikowske K, et al. Breast cancer risk prediction using a clinical risk model and polygenic risk score. *Breast Cancer Res Treat.* 2016; 159:513–525.
4. Borde J, Ernst C, Wappenschmidt B, Niederacher D, Weber-Lassalle K, Schmidt G, Hauke J, Quante AS,

316 Weber-Lassalle N, Horváth J, et al. Performance of breast cancer polygenic risk scores in 760 female
317 *CHEK2* germline mutation carriers. *J Natl Cancer Inst.* 2021; 113:893–899.
318 5. Borde J, Laitman Y, Blümcke B, Niederacher D, Weber-Lassalle K, Sutter C, Rump A, Arnold N, Wang-
319 Gohrke S, Horváth J, et al. Polygenic risk scores indicate extreme ages at onset of breast cancer in
320 female *BRCA1/2* pathogenic variant carriers. *BMC Cancer.* 2022; 22:1–9.
321 6. Gallagher S, Hughes E, Wagner S, Tshiaba P, Rosenthal E, Roa BB, Kurian AW, Domchek SM, Garber J,
322 Lancaster J, et al. Association of a polygenic risk score with breast cancer among women carriers of
323 high-and moderate-risk breast cancer genes. *JAMA Netw Open.* 2020; 3:e208501–e208501.
324 7. Kuchenbaecker KB, McGuffog L, Barrowdale L, Lee A, Soucy P, Healey S, Dennis J, Lush M, Robson
325 M, Spurdle AB, et al. Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction
326 in *BRCA1* and *BRCA2* mutation carriers. *J Natl Cancer Inst.* 2017; 109:djw302.
327 8. Stiller S, Drukewitz S, Lehmann K, Hentschel J, Strehlow V. Clinical impact of polygenic risk score for
328 breast cancer risk prediction in 382 individuals with hereditary breast and ovarian cancer syndrome.
329 *Cancers (Basel).* 2023; 15:3938.
330 9. Carver T, Hartley S, Lee A, Cunningham AP, Archer S, Babb de Villiers C, Roberts J, Ruston R, Walter
331 FM, Tischkowitz M, et al. CanRisk tool – a web interface for the prediction of breast and ovarian cancer
332 risk and the likelihood of carrying genetic pathogenic variants. *Cancer Epidemiol Biomarkers Prev.*
333 2021; 30:469–473.
334 10. Lee A, Mavaddat N, Wilcox AN, Cunningham AP, Carver T, Hartley S, Babb de Villiers C, Izquierdo A,
335 Simard J, Schmidt MK, et al. BOADICEA: a comprehensive breast cancer risk prediction model
336 incorporating genetic and nongenetic risk factors. *Genet Med.* 2019; 21:1708–1718.
337 11. Carver T. Canrisk konwledgebase. https://canrisk.atlassian.net/wiki/spaces/FAQS/
338 pages/35979266/What+variants+are+used+in+the+PRS, 2022. Accessed: 2022-11-30.
339 12. Mavaddat N, Ficorella L, Carver T, Lee A, Cunningham AP, Lush M, Dennis J, Tischkowitz M, Downes K,
340 Hu D, et al. Incorporating Alternative Polygenic Risk Scores into the BOADICEA Breast Cancer Risk

Prediction Model. *Cancer Epidemiol Biomarkers.* 2023; 32:422–427.

13. Kiialainen A, Karlberg O, Ahlford A, Sigurdsson S, Lindblad-Toh K, Syvänen AC. Performance of microarray and liquid based capture methods for target enrichment for massively parallel sequencing and SNP discovery. *PLoS One.* 2011; 6:e16486.
14. Sulonen AM, Ellonen P, Almusa H, Lepistö M, Eldfors S, Hannula S, Miettinen T, Tyynismaa H, Salo P, Heckman C, et al. Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol.* 2011; 12:1–18.
15. Teer JK, Bonnycastle LL, Chines PS, Hansen NF, Aoyama N, Swift AJ, Abaan HO, Albert TJ, Margulies EH, Green ED, et al. Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Res.* 2010; 20:1420– 1431.
16. Yi M, Zhao Y, Jia L, He M, Kebebew E, Stephens RM. Performance comparison of SNP detection tools with illumina exome sequencing data – an assessment using both family pedigree information and sample-matched SNP array data. *Nucleic Acids Res.* 2014; 42:e101– e101.
17. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics.* 2014; 30:2843–2851.
18. Reis AL, Deveson IW, Madala BS, Wong T, Barker C, Xu J, Lennon N, Tong W, Mercer TR. Using synthetic chromosome controls to evaluate the sequencing of difficult regions within the human genome. *Genome Biol.* 2022; 23:1–24.
19. Stoler N, Nekrutenko A. Sequencing error profiles of illumina sequencing instruments. *NAR Genom Bioinform.* 2021; 3:lqab019.
20. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, Genome Aggregation Database Consortium, Rehm HL, MacArthur DG, O'Donnell-Luria A. Variant interpretation using population databases: Lessons from gnomAD. *Hum Mutat.* 2022; 43:1012–1030.
21. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. The mutational constraint spectrum quantified from variation in 141,456

366 humans. *Nature.* 2020; 581:434–443.

367 22. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI
368 database of genetic variation. *Nucleic Acids Res.* 2001; 29:308–311.

369 23. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera AV, Lowther C, Gauthier LD,
370 Wang H, et al. A structural variation reference for medical and population genetics. *Nature.* 2020;
371 581:444–451.

372 24. Adeyemo A, Balaconis MK, Darnes DR, Fatumo S, Moreno PG, Hodonsky CJ, Inouye M, Kanai M,
373 Kato K, Knoppers BM, et al. Responsible use of polygenic risk scores in the clinic: potential benefits,
374 risks and gaps. *Nat Med.* 2021; 27:1876–1884.

375 25. Sugrue LP, Desikan RS. What are polygenic scores and why are they important? *JAMA.* 2019;
376 321:1820–1821.

377

378 **Acknowledgements.** We thank the coordinator of the GC-HBOC, Rita K. Schmutzler, and all GC-HBOC
379 center directors for their support of the GC-HBOC Bioinformatics Working Group. Further, we thank Joe Dennis
380 for helpful comments.

381

382 **The Bioinformatics Working Group of the German Consortium for Hereditary Breast & Ovarian Cancer.**

383 Norbert Arnold[13], Alexandra Baumann[14], Marvin Döbel[15], Stephan Drukewitz[16], Christoph Engel[17], Corinna
384 Ernst[18], Rudel Christian Nkouamedjo Fankep[18], Michael Forster[13], Peter Frommolt[19], Eva Groß[20], Karl
385 Hackmann[14], Johannes Helmuth[21], Ellen Honisch[22], Tim Hutschenreiter[14], Anna-Lena Katzke[23], Anna-
386 Lena Kobiela[18], Zarah Kowalzyk[14], Oliver Kutz[14], Christoph Meier[24,25], Maximilian Radtke[16], Juliane
387 Ramser[26], Robert Remy[18], Julia Ritter[21], Christian Ruckert[27], Gunnar Schmidt[23], Benedikt Schnur[23], Dariush
388 Skowronek[28], Marc Sturm[15], Katharina Thiedig[26], Steffen Uebe[29], Shan Wang-Gohrke[30], Andreas Zimmer[31]

389 [13]Department of Gynecology and Obstetrics, Institute of Clinical Chemistry Institute of Clinical Molecular
390 Biology, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany. [14]Institute for Clinical Genetics,

University Hospital Carl Gustav Carus at TU Dresden, Dresden, Germany; ERN GENTURIS, Hereditary Cancer Syndrome Center Dresden, Germany; National Center for Tumor Diseases Dresden (NCT/UCC), Germany: German Cancer Research Center (DKFZ), Heidelberg, Germany; Faculty of Medicine and University Hospital Carl Gustav Carus at TU Dresden, Dresden, Germany; German Cancer Consortium (DKTK), Dresden, Germany; German Cancer Re- search Center (DKFZ), Heidelberg, Germany; Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. [15]Institute of Medical Genetics and Applied Genomics, University Hospital Tübingen, Tübingen, Germany. [16]Institute of Human Genetics, University of Leipzig Medical Center, Leipzig, Germany. [17]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany. [18]Center for Familial Breast and Ovarian Cancer, Center for Integrated Oncology (CIO), Medical Faculty, University of Cologne and University Hospital Cologne, Cologne, Germany. [19]Institute for Human Genetics, University Hospital Hamburg-Eppendorf, Hamburg, Germany. [20]Department of Obstetrics and Gynecology, Ludwig-Maximilians-University of Munich, Munich, Germany. [21]Department of Human Genetics, Labor Berlin – Charité Vivantes GmbH, Berlin, Germany. [22]Department of Gynaecology and Obstetrics, University Hospital Düsseldorf, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany. [23]Department of Human Genetics, Hannover Medical School (MHH), Hannover, Germany. [24]Institute of Human Genetics, University of Regensburg, Regensburg, Germany. [25]Institute of Clinical Human Genetics, University Hospital Regensburg, Regensburg, Germany. [26]Division of Gynaecology and Obstetrics, Klinikum rechts der Isar der Technischen Universität München, München, Germany. [27]Institute of Human Genetics, University of Münster, Münster, Germany. [28]Department of Human Genetics, University Medicine Greifswald and Interfaculty Institute of Genetics and Functional Genomics, University of Greifswald, Greifswald, Germany. [29]Institute of Human Genetics, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany. [30]Department of Gynaecology and Obstetrics, University Hospital Ulm, Ulm, Germany. [31]Institute for Human Genetics, Medical Center University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany.

Author contributions. Conceptualization: GS, MS and CE. Methodology: All authors. Data analysis: AB, CR, CM, RR, MS and CE. Editing and review of manuscript: All authors. Final manuscript review: All authors.

Funding. RR and RF received funding from the German Federal Ministry of Health within the genomDE initiative. MD received funding from the German Cancer Aid (https://www.krebshilfe.de/) in the HerediVar project.

Competing interests. The authors declare no competing interest.

Ethical approval. IMGAG: The use of aggregate statistics of human subject genetics data was approved by the ethics committee of the Medical Faculty of the University of Tübingen, Germany (Genome+, ClinicalTrial.gov-Nr: NCT04315727; #066/2021BO2 for retrospective data analysis). ICG, IHG-M, CFBOC, IHG-R: Written informed consent was obtained from all patients and ethical approval was granted by the ethics committee of the Technische Universität Dresden, ethics committee of the Medical Association Westfalen-Lippe, ethics committee of the Medical Faculty of the University of Cologne (19-1360_4), the ethics committee of the University of Regensburg (21-2192-103).

Figure legends

Figure 1: Comparison of variant effect allele frequencies (AFs) specified by CanRisk and observed in gnomAD v3.1.2 non-Finnish European samples for 320 variants incorporated in BCAC 313 or BRIDGES 306 breast cancer polygenic risk scores. Extremely deviating AFs with absolute difference > 0.016 are indicated by red markers.

Figure 2: Comparison of effect allele frequencies (AFs) specified by CanRisk and observed in 10 real-world data sets for 320 loci incorporated in BCAC 313 or BRIDGES 306 breast cancer polygenic risk scores. Data was provided by the Institute of Medical Genetics and Applied Genomics (IMGAG) at University Hospital Tübingen, Institute for Clinical Genetics (ICG) at University Hospital Carl Gustav Carus Dresden, by the Institute of Human Genetics at the University of Münster (IHG-M), by the Center for Familial Breast and Ovarian Cancer (CFBOC) at University Hospital Cologne, and by the Institute of Human Genetics at the University of Regensburg (IHG-R).

Figure 3: Sequences of reference, expected effect allele and potential alternative allele of polygenic risk score loci rs73754909 and rs79461387 (hg19-based). Both alternative alleles are deletions with the adjacent downstream nucleotide identical to the expected substituted one.

Table legends

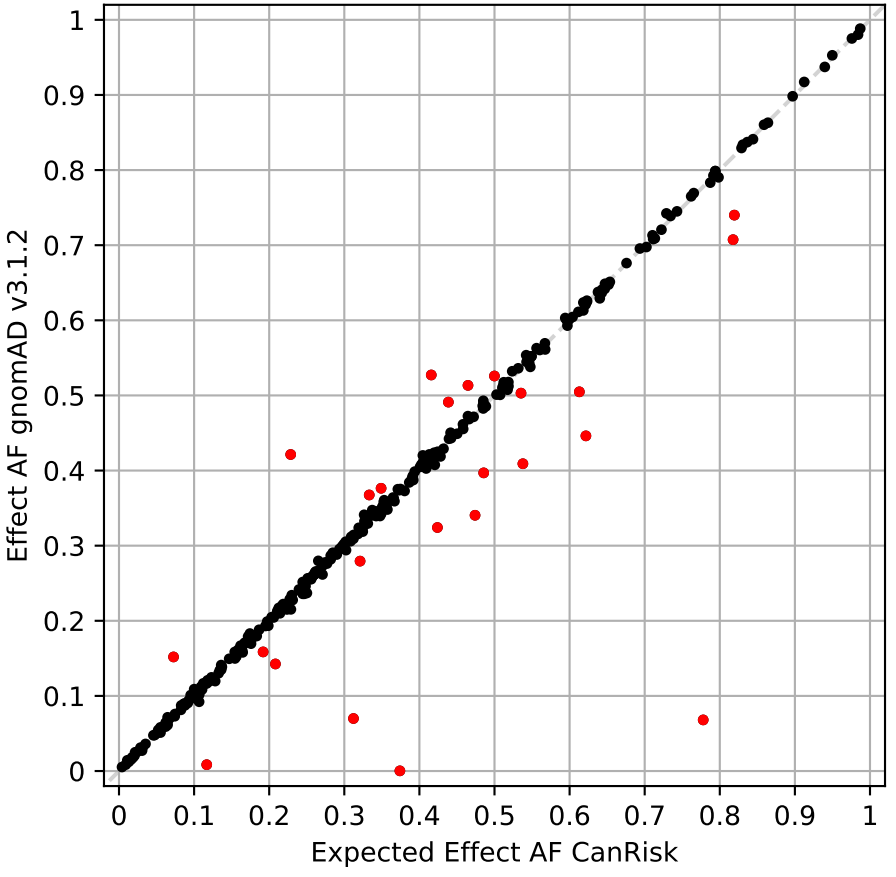
Table 1: Characteristics of data sets provided by participating centers of the German Consortium for Hereditary Breast & Ovarian Cancer (GC-HBOC), namely the Institute of Medical Genetics and Applied Genomics (IMGAG), University Hospital Tübingen, the Institute for Clinical Genetics (ICG), University Hospital Carl Gustav Carus Dresden, the Institute of Human Genetics at the University of Münster (IHG-M), the Center for Familial Breast and Ovarian Cancer (CFBOC), University Hospital Cologne, and the Institute of Human Genetics at the University of Regensburg (IHG-R). Each center provided two data sets. BC/OC: Breast/ovarian cancer; DP: Sequencing depth; PRS: Polygenic risk score.

Table 2: Characteristics of loci incorporated in the BCAC 313 or BRIDGES 306 breast cancer PRSs that were either not included in the gnomAD v3.1.2 database or reported with extremely deviating allele frequency compared to CanRisk. Log odds ratios (ORs) are identical for BCAC 313 and BRIDGES 306, but missing values indicate loci not included in the corresponding PRS. Entries in the Comment column refer to technical artifacts reported in gnomAD. LCR: low complexity region; LQS: low-quality site (in <50% of samples covered); VQSR: failed allele-specific GATK Variant Quality Score Recalibration (VQSR) filter.

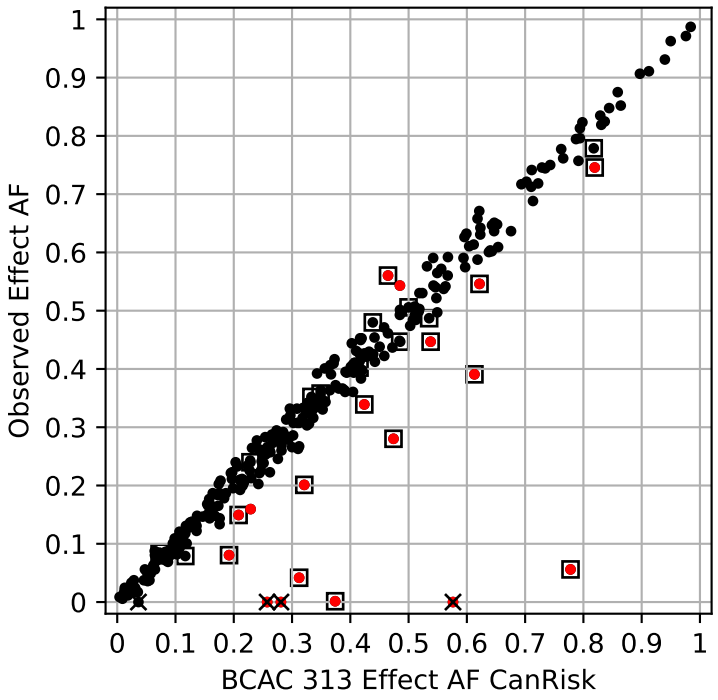
Table 3: Summary of polygenic risk score genotyping results with noticeably deviating allele frequencies (AFs) of centers of the German Consortium for Hereditary Breast and Ovarian Cancer. Noticeably deviating AFs are shown in bold. Loci (hg19-based) of rs11268668 and rs57589542 are 1-204502514-T-TTCTGAAACAGGG (hg19) and 6-152022664-CAAAAAAA-C (hg19), respectively. WGS: Whole-genome sequencing. MGP: Multi-gene panel sequencing. FB: freebayes.

Table 4: Potential solutions for improving polygenic risk score (PRS) genotyping performance with respect to the achievement of allele frequencies (AFs) expected by CanRisk, using alternative alleles or proxys. Resulting AFs were investigated based on gnomAD v3.1.2 non-Finnish European data and genotyping results of 1410

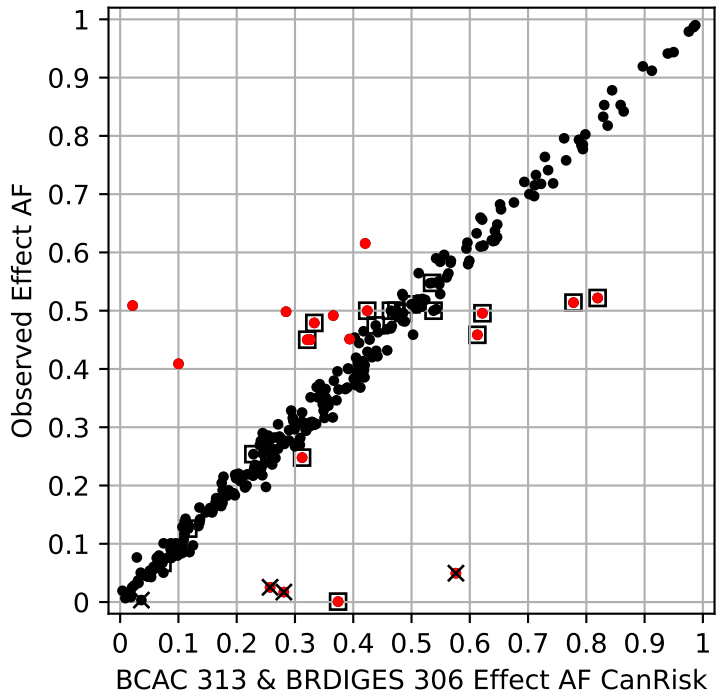
European whole-genome sequencing (WGS) samples using (unforced) freebayes, provided by the Institute of Medical Genetics and Applied Genomics (IMGAG) at University Hospital Tübingen.


Effect AF gnomAD v3.1.2
Expected Effect AF CanRisk
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

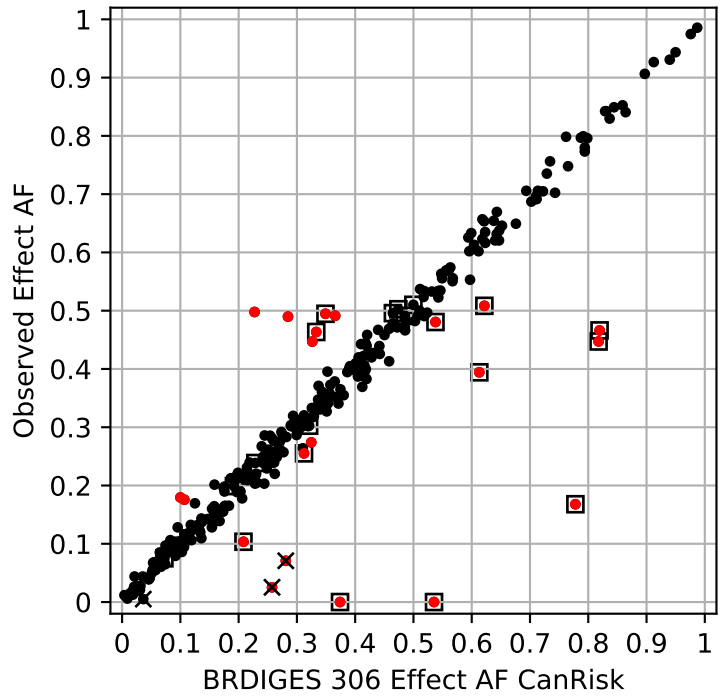
IMGAG DRAGEN (WGS, N=348)



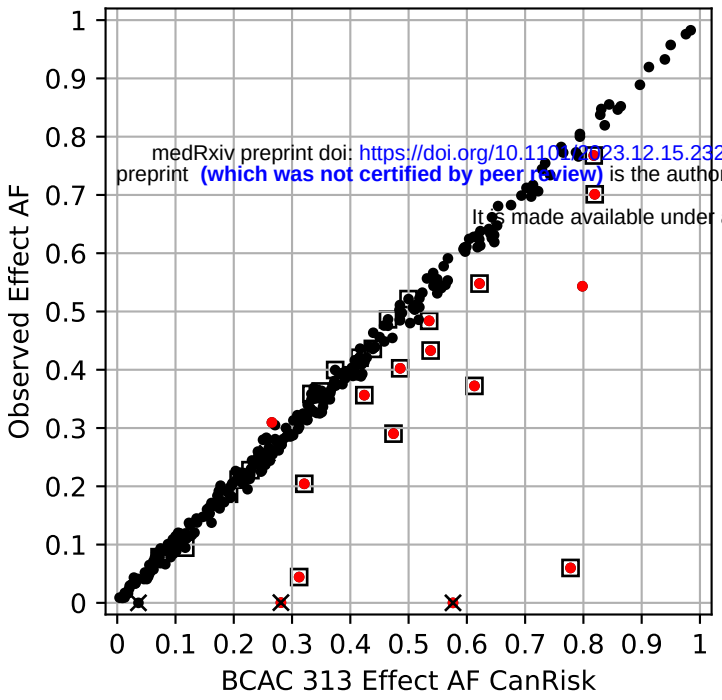
ICG GATK (Twist Panel, N=595)



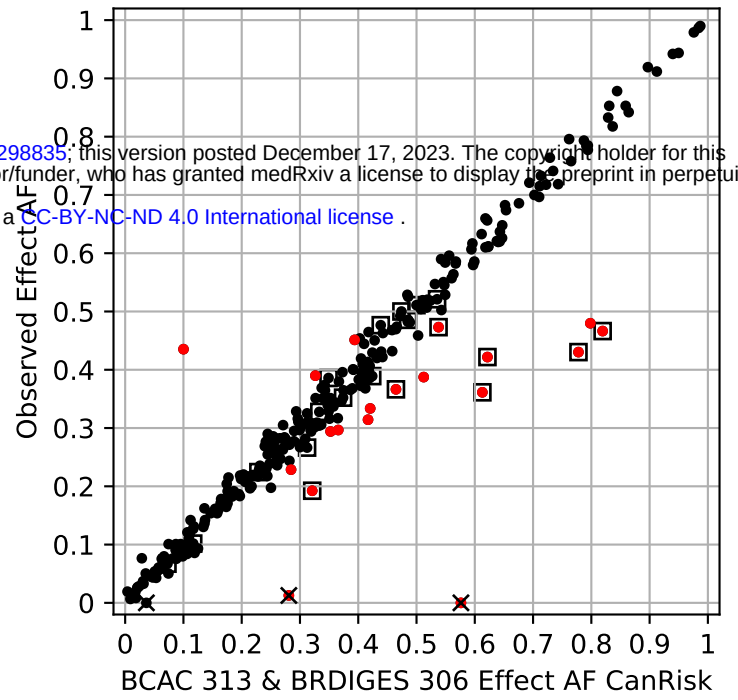
IHG-M GATK (Twist Panel, N=593)



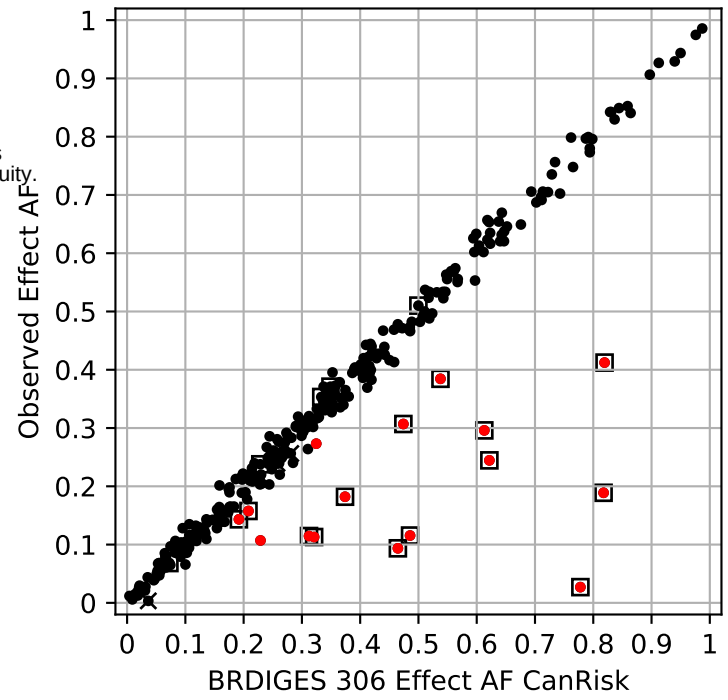
IMGAG freebayes (WGS, N=1410)



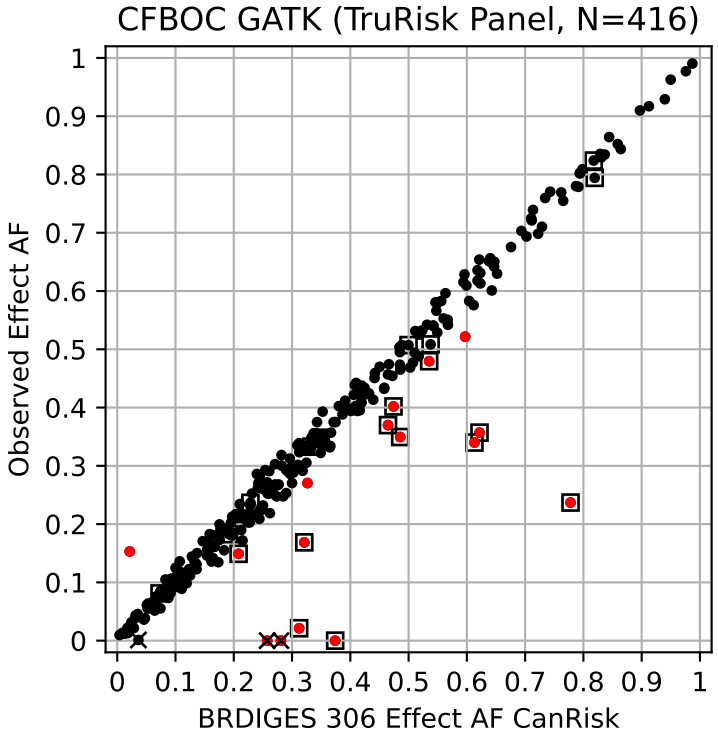
ICG freebayes (Twist Panel, N=595)



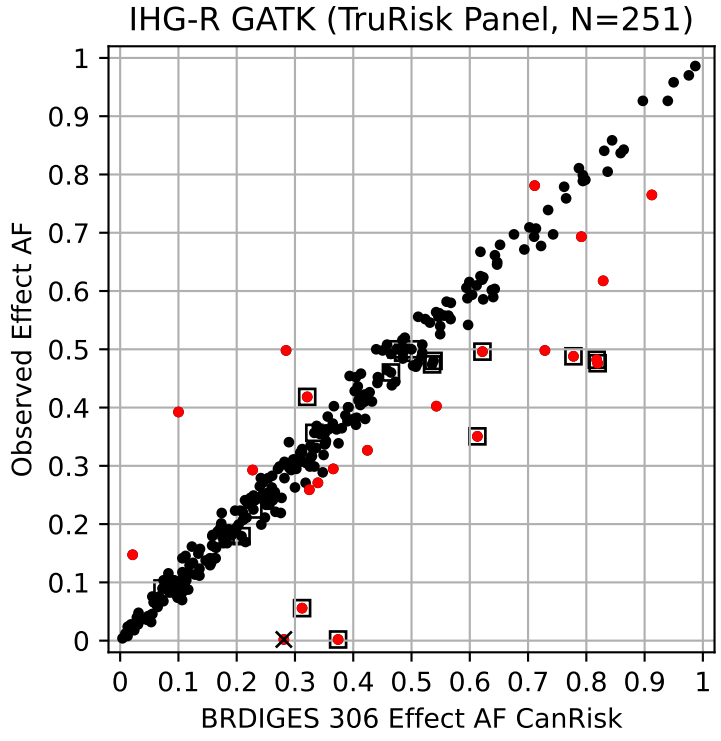
IHG-M DRAGEN (Twist Panel, N=593)

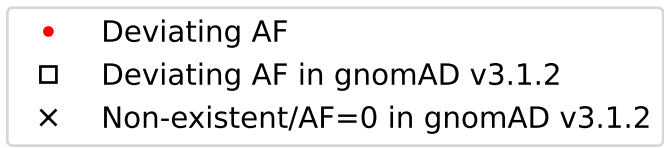


CFBOC GATK (TruRisk Panel, N=416)

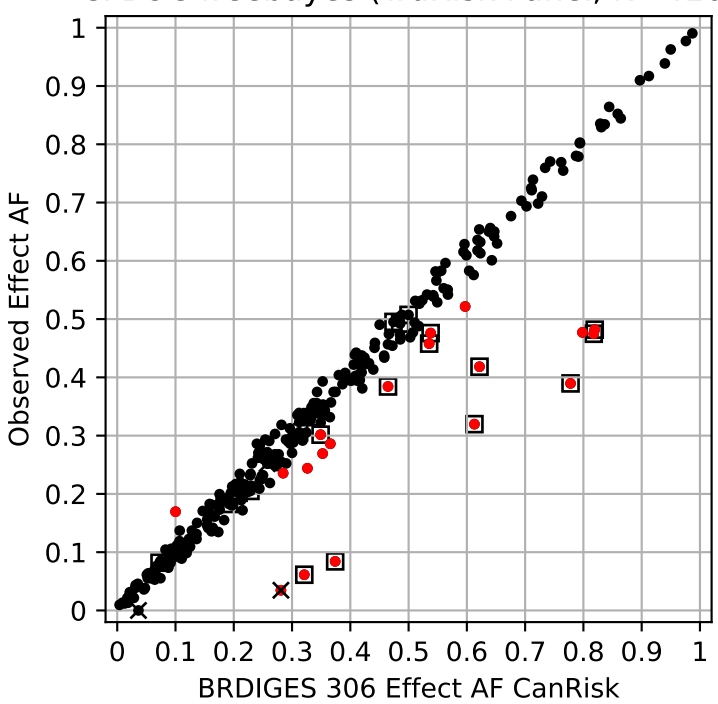


IHG-R GATK (TruRisk Panel, N=251)

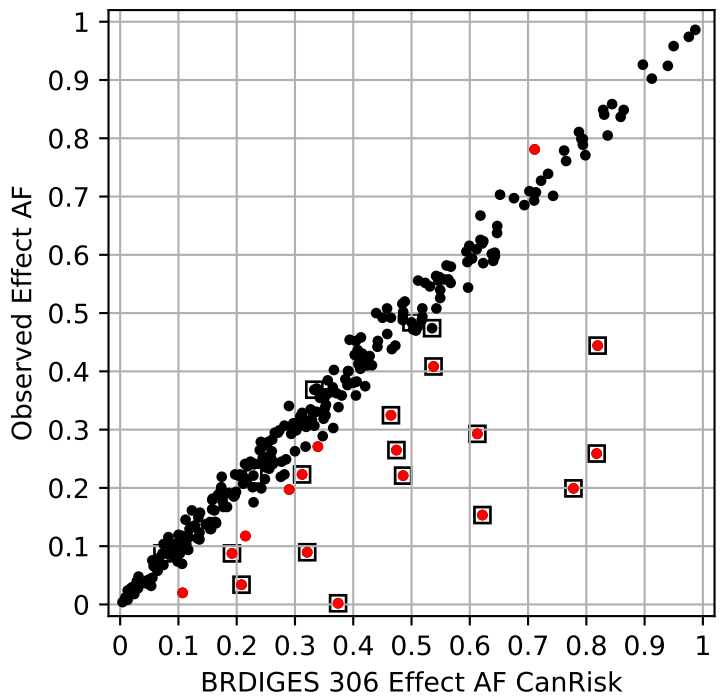




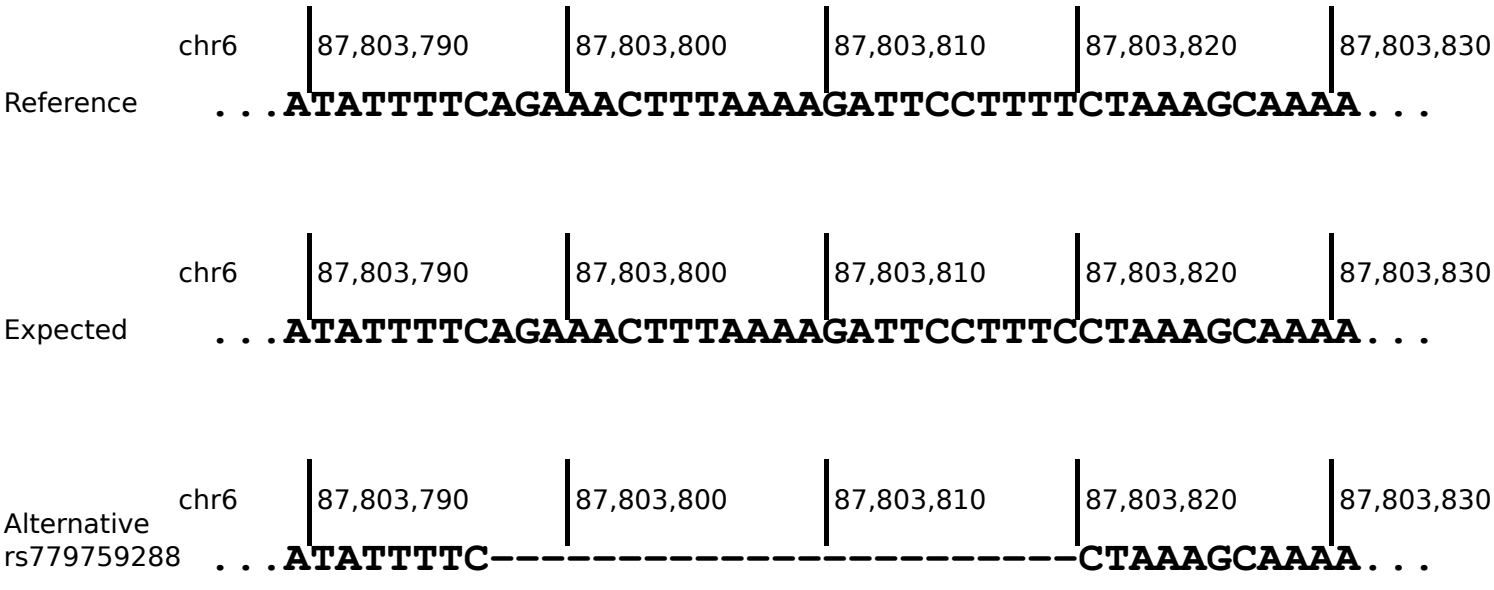
CFBOC freebayes (TruRisk Panel, N=416)



IHG-R CLC (TruRisk Panel, N=251)



rs73754909 (hg19: 6-87803819-T-C)



rs79461387 (hg19: 17-29168077-G-T)

		IMGAG	ICG	IHG-M	CFBOC	IHG-R
Sample size	Set 1	348	595	593	416	251
	Set 2	1410				
Testing indication		various	cancer-related	familial BC/OC	familial BC/OC	familial BC/OC
Considered PRS		BCAC 313	BCAC 313 BRIDGES 306	BRIDGES 306	BRIDGES 306	BRIDGES 306
NGS approach		WGS	Twist Custom Cancer Panel	Twist Custom Panel	Agilent TruRisk v3	Agilent TruRisk v3
Reference		hg38	hg19	hg19	hg19	hg38
Variant caller	Set 1	DRAGEN v4.0.3	freebayes v1.3.6	DRAGEN v4.2.4	freebayes v1.3.6	CLC LightSpeed v23.0.2
	Set 2	freebayes v1.3.6	GATK v4.2.6 HaplotypeCaller	GATK v4.4.0 HaplotypeCaller	GATK v4.3.2 HaplotypeCaller	GATK v4.2.6 HaplotypeCaller
Calling mode	Set 1	unforced	forced	forced	forced	unforced
	Set 2					forced
Quality filter		DP ≥ 15	DP ≥ 20	DP ≥ 20	DP ≥ 30	DP ≥ 10

		CanRisk			gnomAD v3.1.2	
		log OR				
rs ID	Locus (hg19)	BCAC	BRIDGDES	AF	AF	Comment
rs56168262	1-51467096-CT-C	0.0374	0.0374	0.4856	0.3969	LCR
rs56097627	1-110198129-CAAA-C	0.0458	0.0458	0.7779	0.0681	LCR
rs143384623	1-145604302-C-CT	-0.0399	-0.0399	0.3490	0.3764	LCR
rs78425380	2-10138983-T-C	0.0603		0.1168	0.0085	LCR,LQS
rs553796823	2-39699510-C-CT	-0.0402	-0.0402	0.4647	0.5134	LCR
rs572022984	2-217955896-GA-G	-0.2016	-0.2016	0.0364	allele count zero	
rs774021038	4-84370124-TA-T	-0.0464	-0.0464	0.5353	0.5030	LCR
rs147404208	4-92594859-TTCTTTC-T	-0.0407		0.4386	0.4911	LCR
rs62331150	4-106069013-G-T	0.0471	0.0471	0.2286	0.4214	LQS
rs113778879	5-58241712-C-T	-0.0434		0.5762	not listed in gnomAD	
rs543824204	6-20537845-CA-C	-0.0391	-0.0391	0.4741	0.3405	LCR
rs574103382	6-82263549-AAT-A	0.0477		0.4240	0.3242	LCR
rs73754909	6-87803819-T-C	0.0383	0.0383	0.2809	not listed in gnomAD	
rs60954078	6-151955914-A-G	0.1449	0.1449	0.0726	0.1519	LCR
rs57589542	6-152022664-CAAAAAAA-C	0.0137	0.0137	0.6130	0.5048	LCR
rs10644978	7-91459189-A-ATT	0.0452	0.0452	0.3332	0.3675	LCR
rs111963714	7-99948655-T-G	0.0420	0.0420	0.2083	0.1425	
rs5887960	7-139943702-CT-C	0.0582	0.0582	0.5378	0.4091	LCR
rs3988353	8-17787610-CT-C	-0.0377	-0.0377	0.6217	0.4462	LCR,VQSR
rs3057314	9-21964882-CAAAA-C	0.0550	0.0550	0.3210	0.2794	LCR
rs2384736	10-38523626-C-A	0.0404	0.0404	0.3740	0.0003	LCR,LQS
rs111833376	10-71335574-C-T	-0.0404	0.3122	0.3122	0.0699	LCR
rs140936696	10-95292187-CAA-C	-0.0512	-0.0512	0.8177	0.7074	LCR
rs10862899	12-85004551-C-T	0.0348	0.0348	0.4999	0.5259	
rs57920543	16-4008542-CAAAAA-C	-0.0329	-0.0329	0.8194	0.7400	LCR
rs79461387	17-29168077-G-T	-0.0568	-0.0568	0.2573	not listed in gnomAD	
rs2668667	17-44283858-G-A	-0.0540	-0.0540	0.1919	0.1586	
rs112855987	22-45319953-G-A	-0.0134		0.4158	0.5272	LCR

rs ID	Locus (hg19)	log OR			Allele Frequencies										gnomAD
					IMGAG (WGS)		ICG (MGP)		IHG-M (MGP)		CFBOC (MGP)		IHG-R (MGP)		deviating
		BCAC	BRIDGES	expected	DRAGEN	FB	GATK	FB	GATK	DRAGEN	GATK	FB	GATK	CLC	AF
rs56168262	1-51467096-CT-C	0.0458	0.0458	0.4856	0.4468	**0.4025**	0.4961	0.4857	0.4911	**0.1155**	**0.3494**	0.4923	0.4940	**0.2211**	yes
rs56097627	1-110198129-CAAA-C	0.0374	0.0374	0.7779	**0.0560**	**0.0599**	**0.5139**	**0.4303**	**0.1679**	**0.0270**	**0.237**	**0.3896**	**0.4880**	**0.1992**	yes
rs12406858	1-118141492-A-C	0.0452	0.0452	0.2654	0.2874	**0.3096**	0.2824	0.2824	0.2487	0.2487	0.2656	0.2668	0.2550	0.2948	no
rs143384623	1-145604302-C-CT	-0.0399	-0.0399	0.3490	0.3578	0.3628	-	0.3859	**0.4947**	0.3710	0.3482	0.3017	-	-	yes
rs11463354	1-172328767-T-TA	-0.0435	-0.0435	0.3264	0.3305	0.3142	**0.4503**	**0.3899**	**0.4469**	0.3187	**0.2704**	**0.2440**	0.3127	0.3167	no
rs11268668	see caption	-0.0321	-0.0321	0.7983	0.8233	**0.5433**	0.8025	**0.4798**	0.7960	0.7960	0.8089	**0.4772**	0.7908	0.7709	no
rs553796823	2-39699510-C-CT	-0.0420	-0.0420	0.4647	**0.5603**	0.4865	0.5000	**0.3664**	0.4958	**0.0936**	**0.3699**	**0.3843**	0.4602	**0.3247**	yes
rs11693806	2-218292158-C-G	-0.0757	-0.0757	0.7289	0.7457	0.7443	0.7639	0.7639	0.7352	0.7352	0.7103	0.7103	**0.4980**	-	no
rs371314787	3-49709912-C-CT	-0.0367	-0.0367	0.2847	0.2917	0.2809	**0.4983**	**0.2286**	**0.4898**	0.2403	0.2476	0.2356	**0.4980**	0.2490	no
rs34207738	3-141112859-CTT-C	0.0551	0.0551	0.4205	0.3980	0.3929	**0.6154**	**0.3336**	0.4585	0.4300	0.4246	0.3811	0.3805	0.3745	no
rs774021038	4-84370124-TA-T	-0.0464	-0.0464	0.5353	0.4871	**0.4840**	0.5477	0.5210	**0.0000**	-	**0.4796**	**0.4579**	0.4741	0.4741	yes
rs147399132	4-126752992-A-AAT	-0.0377		0.5123	0.4842	0.5092	0.5644	**0.3874**							no
rs199562199	5-52679539-C-CA	0.0571	0.0571	0.1001	0.1049	0.1142	**0.4089**	**0.4353**	**0.1795**	0.0658	0.1250	**0.1695**	**0.3924**	0.1195	no
rs113803968	5-55662540-C-CT	-0.0458	-0.0458	0.3657	0.4066	0.3603	**0.4916**	**0.2966**	**0.4913**	0.3356	0.3341	**0.2861**	**0.2948**	0.3028	no
rs113778879	5-58241712-C-T	-0.0434		0.5762	**0.0000**	**0.0000**	**0.0495**	**0.0000**							-
rs10074269	5-169591460-T-C	0.0412	0.0412	0.3393	0.3463	0.3507	0.3513	0.3513	0.3305	0.3305	0.3389	0.3389	**0.2709**	**0.2709**	no
rs543824204	6-20537845-CA-C	-0.0391	-0.0391	0.4741	**0.2802**	**0.2904**	0.5000	0.5000	0.5025	**0.3069**	**0.4019**	0.4952	0.5000	**0.2649**	yes
rs574103382	6-82263549-AAT-A	0.0477		0.4240	**0.3391**	**0.3564**	**0.5000**	0.3891							yes
rs73754909	6-87803819-T-C	0.0383	0.0383	0.2809	**0.0000**	**0.0004**	**0.0171**	**0.0127**	**0.0708**	0.2563	**0.0000**	**0.3470**	**0.0020**	-	-
rs55941023	6-130341728-C-CT	0.0472	0.0472	0.7113	0.7414	0.7099	0.7151	0.7151	0.6914	0.6914	0.7212	0.7212	**0.7809**	**0.7809**	no
rs57589542	see caption	0.0137	0.0137	0.6130	**0.3908**	**0.3723**	**0.4586**	**0.3613**	**0.3943**	**0.2960**	**0.3399**	**0.3197**	**0.3506**	**0.2928**	yes
rs851984	6-152023191-G-A	0.0626	0.0626	0.3938	0.3937	0.3957	**0.4513**	**0.4513**	0.3997	0.3997	0.3978	0.3978	0.4542	0.4542	no
rs10644978	7-91459189-A-ATT	0.0452	0.0452	0.3332	0.3520	0.3585	**0.4790**	0.3277	**0.4635**	0.3533	0.3329	0.3317	0.3566	0.3685	yes
rs111963714	7-99948655-T-G	0.0420	0.0420	0.2083	**0.1494**	0.2113	-	-	**0.1035**	**0.1577**	**0.1490**	0.2019	0.1793	**0.0339**	yes
rs5887960	7-139943702-CT-C	0.0582	0.0582	0.5378	**0.4468**	**0.4330**	0.5000	**0.4731**	**0.4806**	**0.3844**	0.5084	**0.4760**	0.4801	**0.4084**	yes
rs62485509	7-144048902-G-T	-0.0563	-0.0563	0.2289	**0.1595**	0.2255	0.2210	0.2210	0.2175	**0.1071**	0.2344	0.2344	0.2490	0.1753	no
rs3988353	8-17787610-CT-C	-0.0377	-0.0377	0.6217	**0.546**	**0.5479**	**0.4956**	**0.4218**	**0.5083**	**0.2445**	**0.3570**	**0.4183**	**0.4960**	**0.1534**	yes
rs1511243	8-76230943-A-G	0.0755	0.0755	0.8289	0.8348	0.8376	0.8328	0.8328	0.8423	0.8423	0.8353	0.8353	**0.6175**	0.8486	no
rs10975870	9-6880263-A-G	0.0348	0.0348	0.2900	0.3132	0.2848	0.2950	0.2950	0.3019	0.3019	0.2531	0.2523	0.2968	**0.1972**	no
rs3057314	9-21964882-CAAAA-C	0.0550	0.0550	0.3210	**0.2011**	**0.2043**	**0.4502**	**0.1924**	**0.3026**	**0.1130**	**0.1687**	**0.0613**	**0.4183**	**0.0896**	yes

rs4880038	9-36928288-T-C	0.0249	0.0249	0.5427	0.5431	0.544	0.5025	0.5025	0.5228	0.5228	0.5397	0.5385	**0.4024**	0.5080	no
rs542275778	10-22477776-ACC-A	0.1687	0.1687	0.0214	0.0187	0.0294	**0.5088**	0.0269	0.0439	0.0295	**0.1531**	0.0313	**0.1474**	0.0299	no
rs2384736	10-38523626-C-A	0.0404	0.0404	0.3740	**0.0014**	0.3996	**0.0008**	0.3521	**0.0000**	**0.1821**	**0.0000**	**0.0842**	**0.0020**	**0.0020**	yes
rs111833376	10-71335574-C-T	-0.4040	-0.4040	0.3122	**0.0417**	**0.0443**	**0.2479**	0.2664	**0.2552**	**0.1147**	**0.0213**	0.3059	**0.0558**	**0.2231**	yes
rs140936696	10-95292187-CAA-C	-0.0512	-0.0512	0.8177	0.7677	**0.7742**	-	-	**0.4472**	**0.1889**	0.8239	**0.4748**	**0.4821**	**0.2590**	yes
rs9421410	10-123095209-G-A	-0.0538	-0.0538	0.3246	0.3247	0.3170	0.3076	0.3076	**0.2740**	**0.2732**	0.3053	0.3053	**0.2590**	-	no
rs35054928	10-123340431-GC-G	-0.2408	-0.2408	0.5971	0.5747	0.6028	0.5798	0.5798	0.5531	0.5531	**0.5216**	**0.5216**	0.5418	0.5438	no
rs199504893	11-108267402-C-CA	-0.0022		0.4168	0.4526	0.4362	0.3879	**0.3143**							no
rs11049431	12-28347382-C-T	-0.0521	-0.0521	0.2151	0.1997	0.2053	-	-	0.2091	0.2082	0.1719	0.1719	0.1693	**0.1175**	no
rs1027113	12-29140260-G-A	0.0647	0.0647	0.9124	0.9109	0.9195	0.9118	0.9118	0.9266	0.9266	0.9171	0.9171	**0.7649**	0.9024	no
rs144767203	15-100905819-A-C	-0.0608	-0.0608	0.1072	0.0934	0.1043	-	-	0.1098	0.1037	0.1361	0.1370	0.0837	**0.0199**	no
rs57920543	16-4008542-CAAAAA-C	0.0550	0.0550	0.8194	**0.7457**	**0.7011**	**0.5218**	**0.4664**	**0.4662**	**0.4123**	0.7942	**0.4189**	**0.4761**	**0.4442**	yes
rs12709163	16-6963972-C-G	0.0354	0.0354	0.7915	0.7572	0.7660	0.7840	0.7840	0.7993	0.7993	0.7788	0.7788	**0.6932**	0.7988	no
rs9931038	16-85145977-T-C	-0.0211	-0.0211	0.4851	**0.5431**	0.5110	0.4824	0.4824	0.4730	0.4730	0.4736	0.4736	0.4940	0.4940	no
rs79461387	17-29168077-G-T	-0.0568	-0.0568	0.2573	**0.0000**	0.2567	**0.0253**	0.2613	**0.0253**	0.2395	**0.0000**	0.2519	-	-	-
rs71363517	17-43212339-C-CT	0.0438	0.0438	0.2273	0.2256	0.2128	-	-	**0.4978**	0.2142	0.2110	0.2028	**0.2928**	0.2012	no
rs2668667	17-44283858-G-A	-0.0540	-0.0540	0.1919	**0.0805**	0.1872	-	-	0.1990	**0.1433**	0.1827	0.1827	0.1813	**0.0876**	yes
rs1111207	18-24125857-T-C	0.0346	0.0346	0.4243	0.4267	0.4135	0.4294	0.4294	0.4283	0.4283	0.4339	0.4339	**0.3267**	0.4104	no
rs140702307	19-19517054-C-CGGGCG	0.0437	0.0437	0.3525	0.3405	0.3504	0.3655	**0.2941**	0.3499	0.3533	0.3413	**0.2692**	0.3386	0.3247	no
rs66987842	22-40904707-CT-C	0.1148	0.1148	0.1068	0.1207	0.1195	0.1210	0.1288	**0.1757**	0.1349	0.1142	0.1166	0.1414	0.1016	no

rs ID	Locus (hg19)	Expected AF	Workaround	gnomAD AF	TMGAC FB AF
rs12406858	1-118141492-A-C	0.2654	Proxy rs1966228	0.2622	0.3064
rs11693806	2-218292158-C-G	0.7289	Proxy rs3821098	0.7422	0.7443
rs34207738	3-141112859-CTT-C	0.4205	Summing up the AFs of deletions of two and three thymines	0.4344	0.4167
rs10074269	5-169591460-T-C	0.3393	Proxy rs4562056	0.3414	0.3511
rs73754909	6-87803819-T-C	0.2809	Alternative allele rs77846138*	0.2846	0.2578
			Proxy rs12664322	0.2849	0.2791
rs55941023	6-130341728-C-CT	0.7113	Proxy rs1415700	0.7049	0.7050
			Proxy rs11390217	0.7058	0.7046
rs851984	6-152023191-G-A	0.3938	Proxy rs851983	0.3993	0.3965
rs1511243	8-76230943-A-G	0.8289	Proxy rs6472903	0.8294	0.8376
rs10975870	9-6880263-A-G	0.29	Proxy rs12380608	0.2863	0.2840
			Proxy rs10975887	0.2849	0.2837
rs4880038	9-36928288-T-C	0.5427	Proxy rs4880039	0.5449	0.5436
			Proxy rs7032313	0.5446	0.5440
rs542275778	10-22477776-ACC-A	0.0214	Proxy rs112287594	0.0185	0.0270
rs111833376	10-71335574-C-T	0.3122	Summing up AFs of rs111833376 and rs753981427†	0.3200	0.3163
			Proxy rs12769661	0.2984	0.2929
rs9421410	10-123095209-G-A	0.3246	Proxy rs7913694	0.3142	0.3110
			Proxy rs35098964	0.3139	0.3099
rs35054928	10-123340431-GC-G	0.5971	Proxy rs2981579	0.5908	0.5996
rs11049431	12-28347382-C-T	0.2151	Proxy rs11049519	0.2142	0.2039
rs144767203	15-100905819-A-C	0.1072	Proxy rs58855876	0.1078	0.1043
			Proxy rs113438754	0.1078	0.1043
rs12709163	16-6963972-C-G	0.7915	Proxy rs1492386	0.7951	0.7684
rs9931038	16-85145977-T-C	0.4851	Proxy rs60296580	0.4903	0.5082
rs79461387	17-29168077-G-T	0.2573	Alternative allele rs550458309‡	0.2719	0.0000
rs2668667	17-44283858-G-A	0.1919	Proxy rs2532237	0.1860	0.1957
			Proxy rs150290194	0.1765	0.1858
rs1111207	18-24125857-T-C	0.4243	Proxy rs1111208	0.4249	0.4135
rs66987842	22-40904707-CT-C	0.1068	Proxy rs6001949	0.1003	0.1195

*6-87094100-CAGAAACTTTAAAAGATTCCTTTT-C (hg19)

†10-71335572-TCC-T (hg19)

‡17-29168076-AGAG-A (hg19)