

# <sup>1</sup> Legacy Data Confounds Genomics Studies

<sup>3</sup> Luke Anderson-Trocmé<sup>1,2</sup>, Rick Farouni<sup>1,2</sup>, Mathieu Bourgey<sup>1,2</sup>, Yoichiro  
<sup>4</sup> Kamatani<sup>3</sup>, Koichiro Higasa<sup>3</sup>, Jeong-Sun Seo<sup>4,5</sup>, Changhoon Kim<sup>4</sup>, Fumihiko  
<sup>5</sup> Matsuda<sup>3</sup>, Simon Gravel<sup>1,2</sup>

\*For correspondence:

simon.gravel@mcgill.ca (SG)

<sup>6</sup> <sup>1</sup>Department of Human Genetics, McGill University, Montreal, QC H3A 0G1, Canada;

<sup>7</sup> <sup>2</sup>McGill University and Genome Quebec Innovation Centre, Montreal, QC H3A 0G1,

<sup>8</sup> Canada; <sup>3</sup>Center for Genomic Medicine, Graduate School of Medicine, Kyoto University,

<sup>9</sup> Kyoto 606-8501, Japan; <sup>4</sup>Bioinformatics Institute, Macrogen Inc., Seoul, 08511, Republic of

<sup>10</sup> Korea; <sup>5</sup>Precision Medicine Center, Seoul National University Bundang Hospital,

<sup>11</sup> Seongnam, 13605, Republic of Korea

<sup>12</sup> 

---

<sup>13</sup> **Abstract** Recent reports have identified differences in the mutational spectra across human  
<sup>14</sup> populations. While some of these reports have been replicated in other cohorts, most have been  
<sup>15</sup> reported only in the 1000 Genomes Project (1kGP) data. While investigating an intriguing putative  
<sup>16</sup> population stratification within the Japanese population, we identified a previously unreported  
<sup>17</sup> batch effect leading to spurious mutation calls in the 1kGP data and to the apparent population  
<sup>18</sup> stratification. Because the 1kGP data is used extensively, we find that the batch effects also lead to  
<sup>19</sup> incorrect imputation by leading imputation servers and \*some questionable suspicious\* GWAS  
<sup>20</sup> associations. Lower-quality data from the early phases of the 1kGP thus continues to \*impact  
<sup>21</sup> contaminate\* modern studies in hidden ways. \*As technologies and analytical methods improve, it  
<sup>22</sup> is important that we renew and update these data resources accordingly\* It may be time to retire  
<sup>23</sup> or upgrade such legacy sequencing data.

<sup>24</sup> **Key words :** Batch Effect, Mutational Signature, Statistical Genetics, Population Genetics, Reference  
<sup>25</sup> Cohorts, Imputation

<sup>26</sup> 

---

## <sup>27</sup> Introduction

### <sup>28</sup> Batch Effects in Aging Reference Cohort Data

<sup>29</sup> The last 5 years have seen a drastic increase in the amount and quality of human genome sequence  
<sup>30</sup> data. Reference cohorts such as the International HapMap Project (*International HapMap Consortium, 2005*), the 1000 Genomes Project (1kGP)(*1000 Genomes Project Consortium, 2010, 2012; Consortium et al., 2015*), and the Simons Diversity project (*Mallick et al., 2016*), for example, have  
<sup>31</sup> made thousands of genome sequences publicly available for population and medical genetic analy-  
<sup>32</sup> ses. Many more genomes are available indirectly through servers providing imputation services  
<sup>33</sup> (*McCarthy et al., 2016*) or summary statistics for variant frequency estimation (*Lek et al., 2016*).

<sup>34</sup> The first genomes in the 1kGP were sequenced 10 years ago (*van Dijk et al., 2014*). Since  
<sup>35</sup> then, sequencing platforms have rapidly improved. The second phase of the 1kGP implemented  
<sup>36</sup> multiple technological and analytical improvements over its earlier phases (*1000 Genomes Project*  
<sup>37</sup> *Consortium, 2012; Consortium et al., 2015*), leading to heterogeneous sample preparations and  
<sup>38</sup> data quality over the course of the project.

41 Yet, because of the extraordinary value of freely available data, early data from the 1kGP is still  
42 widely used to impute untyped variants, to estimate allele frequencies, and to answer a wide range  
43 of medical and evolutionary questions. This raises the question of whether and how such legacy  
44 data should be included in contemporary analyses alongside more recent cohorts. Here we point  
45 out how large and previously unreported batch effects in the early phases of the 1kGP still lead to  
46 incorrect genetic conclusions through population genetic analyses and spurious GWAS associations  
47 as a result of imputation using the 1kGP as a reference.

#### 48 **Mutational Signatures**

49 Different mutagenic processes may preferentially affect different DNA motifs. Certain mutagens  
50 in tobacco smoke, for example, have been shown to preferentially bind to certain genomic motifs  
51 leading to an excess of G to T transversions (*Pfeifer et al., 2002; Pleasance et al., 2010*). Thus,  
52 exposure of populations to different mutational processes can be inferred by considering the DNA  
53 context of polymorphism in search of *signatures* of different mutational processes (*Alexandrov*  
54 *et al., 2013; Shiraishi et al., 2015*). Such genome-wide mutational signatures have been used as  
55 diagnostic tools for cancers (e.g., *Alexandrov et al. (2013); Shiraishi et al. (2015)*).

56 In addition to somatic mutational signatures, there has been recent interest in population  
57 variation in germline mutational signatures which can be revealed in large sequencing panels.  
58 In 2015, Harris reported 50% more TCC → TTC mutations in European populations compared  
59 to African populations, and this was replicated in a different cohort in 2017 (*Harris, 2015; Harris*  
60 *and Pritchard, 2017; Mathieson and Reich, 2017*). Strong population enrichments of a mutational  
61 signature suggests important genetic or environmental differences in the history of each population  
62 (*Harris, 2015; Harris and Pritchard, 2017*). Harris and Pritchard further identified distinct mutational  
63 spectra across a range of populations, which were further examined in a recent publication by  
64 Aikens et al. (*Harris and Pritchard, 2017; Aikens et al., 2019*).

65 In particular, the latter two studies identified a heterogeneous mutational signature within 1kGP  
66 Japanese individuals. This heterogeneity is intriguing because differences in germline signatures  
67 accumulate over many generations. A systematic difference within the Japanese population would  
68 suggest sustained environmental or genetic differences across sub-populations within Japan with  
69 little to no gene flow. \*Given this signature had not been reproduced in another dataset\*, we  
70 therefore decided to follow up on this observation, by using a newly sequenced dataset of Japanese  
71 individuals from Nagahama. \*Our initial goal was to explore the population structure that might be  
72 responsible for such an intriguing mutational signature.\*

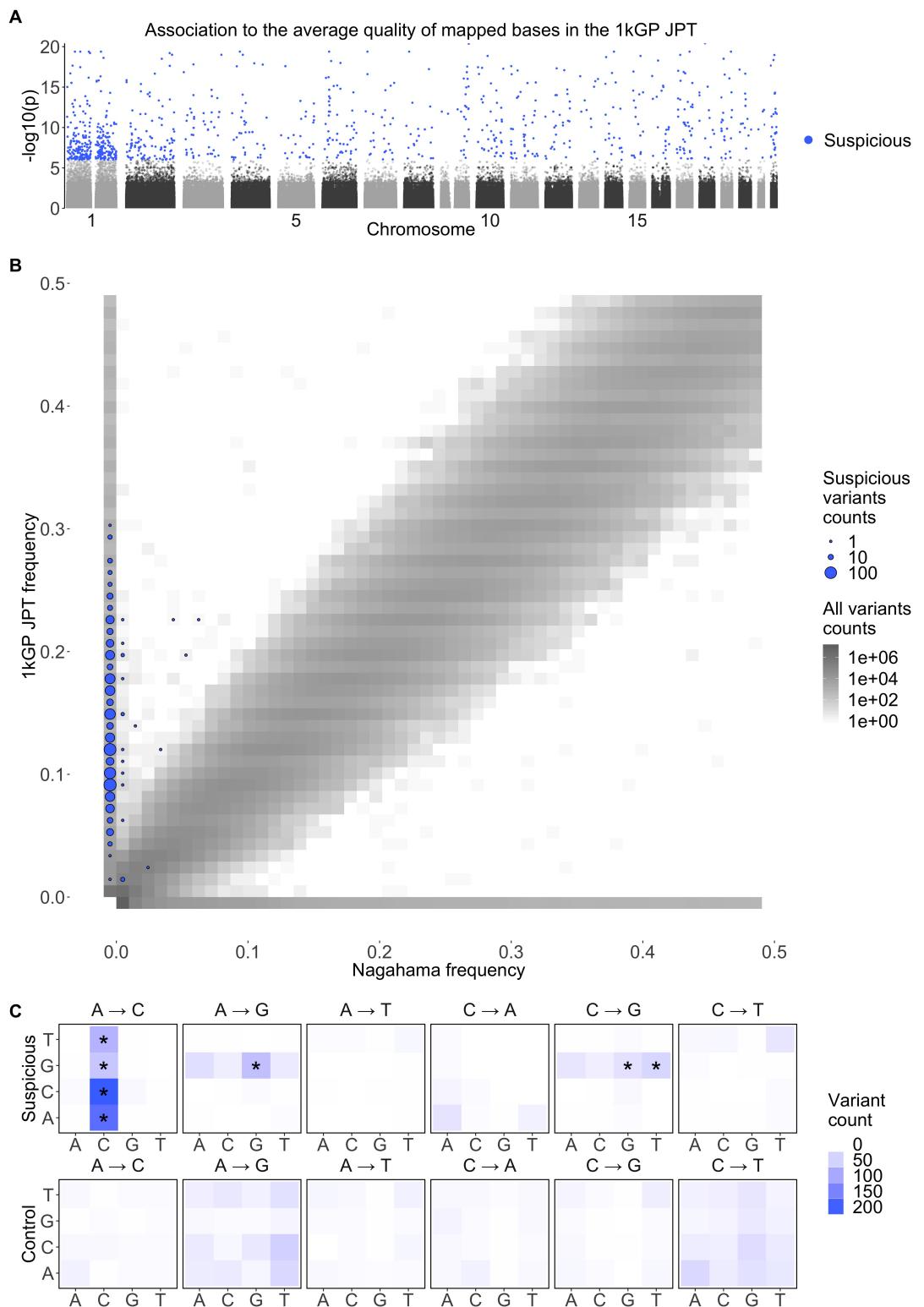
73 While we were unable to reproduce the mutational heterogeneity within the Japanese population,  
74 we could trace back the source of the discrepancy to a technical artefact in the 1kGP data. In addition  
75 to creating biases in mutational signatures, this artefact leads to spurious imputation results which  
76 have found their way in a number of recent publications.

77 The results section is organized as follows. We first attempt to reproduce the original signal and  
78 identify problematic variants in the JPT cohort from the 1kGP. Next, we expand our analysis to the  
79 other populations in the 1kGP and identify lists of variants that show evidence for technical bias.  
80 Finally, we investigate how these variants have impacted modern genomics analyses.

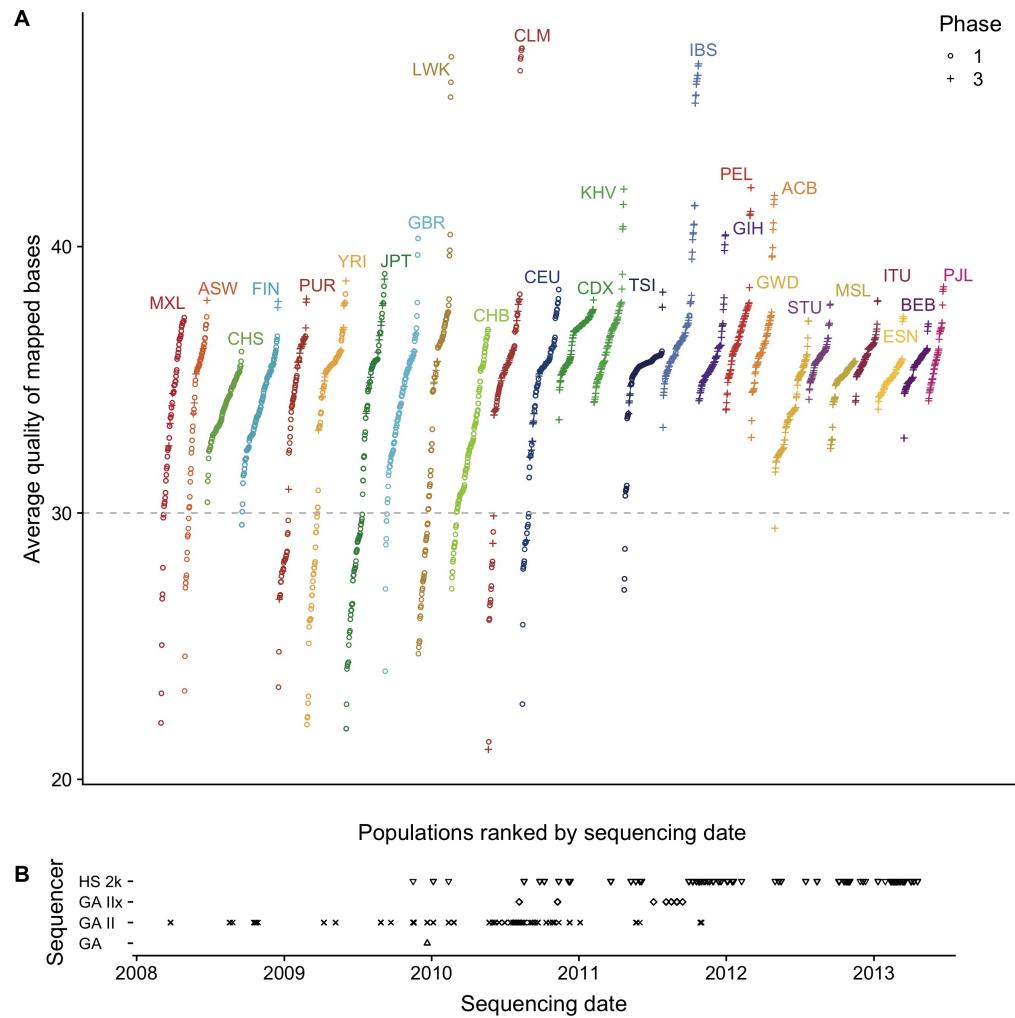
## 81 **Results**

### 82 **A peculiar mutational signature in Japan**

83 Harris and Pritchard reported an excess of a 3-mer substitution patterns \*AC→\*CC in a portion  
84 of the Japanese individuals in the 1kGP (*Harris and Pritchard, 2017*). While trying to follow up on  
85 this observation in a larger and more recent Japanese cohort from Nagahama, we did not find this  
86 particular signature. When comparing the allele frequencies between the Japanese individuals from  
87 the 1kGP and this larger dataset, we observed a number of single nucleotide polymorphisms (SNPs)  
88 private to one of the two groups (Figure 1). Given the similarity of the two populations, this strongly



**Figure 1.** Suspicious mutations carried by individuals with low quality data have distinct mutational profiles, reproduce poorly across studies, and are distributed across the genome. **A** Genome wide association of the average quality of mapped bases  $Q$  for the 104 Japanese individuals included in the 1000 Genomes Project. This GWAS identified 587  $p < 10^{-8}$  and 1034  $p < 10^{-6}$  SNPs that were associated to the average  $Q$  of SNPs mapped for an individual. The same analysis was performed independently for each of the populations in the 1000 Genomes Project. **B** Joint frequency spectrum plot of the Japanese from the 1000 Genomes Project and a more recent Japanese dataset from Nagahama. The size of blue dots are proportional to the number of variants that associate with  $Q$  in the JPT. **C** Mutation spectrum of the 1034 variants that associated with  $Q$  in the JPT ( $p < 10^{-6}$ ), compared to a random set of SNPs. The majority of the variants with significant associations to  $Q$  have the  $*AC \rightarrow *CC$  mutational pattern. There is also a slight enrichment in  $GA \rightarrow GG$  and  $GC \rightarrow GG$  mutations. These three enrichments can be summarized as  $G** \rightarrow GG*$ . Stars (\*) indicate significant difference from the mutational spectrum of a set of control SNPs.



**Figure 2.** Sampling and sequencing technologies over time in the 1000 Genomes Project. **A** The average quality of mapped bases  $Q$  for each individual per population included in the 1000 Genomes Project. Populations are ranked by mean sequencing date (the earliest sequencing date was used for individuals with multiple dates). The shape indicates whether the individual belongs to Phase 1 or Phase 3 of the 1000 Genomes project. **\*B** The sequencing technologies used over time. A change in technology coincides with an increase in overall quality of data produced.\*

suggests a technical difference rather than a population structure effect. These mismatches were maintained despite only considering sites that satisfied strict quality masks and Hardy-Weinberg equilibrium in both cohorts.

When mismatch sites are removed from the 1kGP data, the  $*AC \rightarrow *CC$  signal disappears (Figure 1). To identify possible technical reasons for the difference, we performed regressions of the prevalence of the  $*AC \rightarrow *CC$  mutational signature against different individual-level quality metrics provided by the 1kGP (see Figure S18). The average quality of mapped bases  $Q$  per individual stood out as a strong correlate : Individuals with low  $Q$  show elevated rates of the signature. Thus, sequences called from low- $Q$  data contain variants that reproduce poorly across studies and exhibit a particular mutational signature.

To identify SNPs that are likely to reproduce poorly across cohorts without having access to a second cohort, we performed an association study in the JPT for SNPs that associate strongly with low  $Q$  (Figure 1). Traditionally, genome wide association studies use genotypes as the independent variable. Here we perform a **\*genotype conditional association test (GCAT)**, where **\*** genotypes are

now the dependent variable that we attempt to predict using the continuous variable  $Q$  as the independent variable (Song et al., 2015). We use logistic regression of the genotypes on  $Q$  and identify 587 SNPs with  $p < 10^{-8}$  and 1034 SNPs with  $p < 10^{-6}$ . While identifying putative low-quality SNPs to exclude, using a higher  $p$ -value threshold increases the stringency of the filtering (i.e., excluding SNPs with  $p < 10^{-6}$  is more stringent than excluding SNPs with  $p < 10^{-8}$ ). The variants that are associated to  $Q$  have a significant enrichment in \*AC→\*CC mutations, GA\*→GG\*, and GC\*→GG\* mutations (Figure 1C). These three enrichments can be summarized as an excess of G\*\*→GG\* in individuals with low  $Q$ . \*Statistical significance of these enrichment is computed using a chi-squared test as described by Harris 2017 (Harris and Pritchard, 2017). \*

Thus, this mutational signal is heavily enriched in  $Q$ -associated SNPs, but residual signal remains in non-significant SNPs, presumably because many rare alleles found in individuals with low  $Q$  remain unidentifiable using association techniques \*due to lack of power (Supplementary Figure S2)\*. The removal of individuals with  $Q$  below 30 successfully removes the \*AC→\*CC signal, however other signals identified by Harris and Pritchard appear unchanged (Figure ??). For population genetic analyses sensitive to the accumulation of rare variants, the removal of individuals with low  $Q$  appears preferable to filtering specific low-quality SNPs. For other analyses where quality of imputation matters, identifying  $Q$ -associated variants may be preferable.

## Identifying suspicious variants in the 1000 Genomes Project

The distribution of  $Q$  across 1kGP populations shows that many populations have distributions of  $Q$  scores comparable to that of the JPT, especially populations sequenced in the phase 1 of the project: sequencing done in the early phases of the 1kGP was more variable and overall tended to include lower quality sequencing data (Figure 2 and Figure S3). This variability could result from evolving sequence platform and protocols or variation between sequencing centres. By 2011, older sequencing technologies were phased out, and methods became more consistent, resulting in higher and more uniform quality.

We therefore performed the same reverse GWAS approach in all populations independently, and similarly identified  $Q$ -associated SNPs in 23 of the 26 populations in the 1kGP, with the phase 1 populations being most affected, with on average four times as many significantly associated sites compared to the phase 3 populations. Over 812 variants were independently associated to low  $Q$  in at least two populations with  $p < 10^{-6}$  in each (Figure S4).

To build a test statistic to represent the association across all populations simultaneously, we performed a simple logistic regression predicting genotype based on  $Q$  with the logistic factor analysis (LFA) as an offset to account for population structure or Genotype-Conditional Association Test (GCAT) as proposed by (Song et al., 2015). We also considered two alternative approaches to account for confounders, namely using the leading five principal components, and using population membership as covariates. These models were broadly consistent (See Figure S1).

This method identifies a total of 24,390 variants associated to  $Q$  distributed across the genome with 15,270 passing the 1kGP strict mask filter (Figures S13,S14, S15, and S16). Most analyses below focus on the 15,270 variants satisfying the strict mask, since these variants are unlikely to be filtered by standard pipelines. To account for the large number of tests, we used a two-stage Benjamini & Hochberg step-up FDR-controlling procedure to adjust the p-values using a nominal Type-I error rate  $\alpha = 0.01$  (Benjamini et al., 2006). We tested SNPs, INDELs and repetitive regions separately as they may have different error rates (Table 1). Lists of  $Q$ -associated variants and individuals with low  $Q$  are provided in Supplementary Data.

$Q$ -associated variants are distributed across the genome, with chromosome 1 showing an excess of such variants, and other chromosomes being relatively uniform (Figure S6). At a 10kb scale, we also see rather uniform distribution with a small number of regions showing an enrichment for such variants (Figure S7). An outlying 10kb region in chromosome 17 (bases : 22,020,000 to 22,030,000) has 35  $Q$ -associated variants. Distribution of association statistics in this region is provided in Figure S8. By contrast, variants that do not pass the 1kGP strict mask are more unevenly

distributed across the genome(Figure S7). \*The mutational 3-mer substitution patterns of this list of variants is similar to the signature identified in the single population test of the 1kGP JPT in that there is an enrichment in \*AC→\*CC. However, another enrichment profile is significantly enriched : TAT→TTT, TCT→TTT and ACA→AAA (reverse complement : TGT→TTT). These three enrichments can be summarized as T\*T→TTT.\*

	Repeat	Non-Repeat	Total
SNP	3,369 0.53‰	11,059 0.56‰	14,428
INDEL	181 0.3‰	657 0.66‰	838
Total	3,550	11,716	15,270

**Table 1.** Number of statistically significant variants passing the 1000 Genomes Project strict mask per category. Variants that are flagged by the 1000 Genomes Project nested repeat mask file were analyzed separately for FDR calculation. SNPs and INDELS were also analyzed separately. A total of 15,270 are statistically significantly associated to  $Q$ . \*The grey text is the proportion of significant associations per category.\* The number of variants included in the analysis for SNPs, SNPs in repeat regions, INDELS and INDELS in repeat regions are 19,846,786, 6,312,620, 1,770,315 and 586,342 respectively.

### 158 Cell line or technical artifact

159 In 2017, Lan et al. resequenced 83 Han Chinese individuals from the 1kGP (*Lan et al., 2017*). To  
160 assess consistency between the two datasets, we consider consistency of genotype calls for  $Q$ -  
161 associated variants that are predicted to be polymorphic in these 83 individuals according to the  
162 1kGP. Among the 296 such variants that were  $Q$ -associated in the CHB or CHS, only 6 are present in  
163 the resequenced data (Figure S11). This is more than our nominal false discovery rate of \* $\alpha = 0.01$ \*  
164 of the sites. Thus a small number of variants associated to  $Q$  are present in the population but with  
165 somewhat biased genotypes.

166 We did a similar analysis using all variants identified in the GCAT model (rather than only variants  
167 significantly associated to  $Q$  within the CHB and CHS). Of the 15,270  $Q$ -associated variants identified  
168 globally, 6,307 are polymorphic in the 1kGP for the 83 resequenced individuals (See Figure S9).  
169 From this subset, only 1,139 (or 18%) are present in the resequenced data. The allele frequencies of  
170 these variants are nearly identical between datasets suggesting that among these 83 individuals,  
171 these variants are properly genotyped in the 1kGP. \*In the GCAT test, a variant may be significantly  
172 associated to  $Q$  in some populations, but not necessarily all populations. Thus, the 1,139 variants  
173 present in the resequenced Han might have stronger associations to  $Q$  in individuals from other  
174 populations.\* There are 5 \*variants\* that show differing frequencies between both datasets that  
175 are likely explained by biased genotypes. The vast majority of polymorphisms associated with  $Q$  are  
176 not present at all in the resequencing dataset, supporting sequencing rather than cell line artifacts.

177 Among the 15,270  $Q$ -associated variants, 613 are present on Illumina's Omni 2.5 chip (See Figure  
178 S17). These are likely among the small number of variants that are present in the data but exhibit  
179 biased genotyping in 1kGP.

### 180 Suspicious variants impact modern genomics analyses

181 State of the art imputation servers use a combination of many databases including some that  
182 are not freely available. From the perspective of researchers, they act as black-box imputation  
183 machines that take observed genotypes as input and return imputed genotypes.

184 To investigate whether suspicious calls from the 1kGP are imputed into genotyping studies, we  
185 submitted genotype data for the first two chromosomes of the 1kGP genotype data to the Michigan  
186 Imputation Server. We found that all of the variants associated with  $Q$  were imputed back in the  
187 samples. This suggests that the imputation reference panel still includes individuals with low  $Q$ ,  
188 and the dubious variants will be imputed in individuals who most closely match the low-quality  
189 individual. \*These  $Q$ -associated variants could also compromise the imputation of real variants

190 that happen to be close by, causing not only improper imputation, but also false negatives for real  
191 variants.\*

192 We searched the literature for any GWAS that might have reported these  $Q$ -associated variants  
193 as being significantly associated with some biological trait, even though there is no particular  
194 reason for these variants to be associated with phenotypes. The NHGRI-EBI Catalog of published  
195 genome-wide association studies identified seventeen recent publications that had reported these  
196 variants as close to or above the genome-wide significant threshold (Table 2). \*Some of these  
197 publications noted that these variants were not flanked by linkage disequilibrium peaks (*Mandise*  
198 *et al., 2017*). Indeed, we expect a large number of these variants to be filtered out by some quality  
199 controls like the Hardy-Weinberg equilibrium test. However, the fact that some of these variants  
200 are not removed and that the great majority of these variants are missing from higher quality  
201 datasets means that these  $Q$ -associated variants should be flagged for removal to avoid spurious  
202 association.\*

203 Eleven of these studies included the 1kGP in their reference panel for imputation (*Xu et al.,*  
204 *2012; Lutz et al., 2015; Park et al., 2015; Astle et al., 2016; Herold et al., 2016; Suhre et al., 2017;*  
205 *López-Mejías et al., 2017; Tian et al., 2017; Spracklen et al., 2017; Nagy et al., 2017; Gao et al.,*  
206 *2018*) and another used the 1kGP sequence data and cell lines directly (*Mandise et al., 2017*). One  
207 study used an in-house reference panel for imputation (*Nishida et al., 2018*), two studies genotyped  
208 individuals and imputed the data using the HapMap II as a reference database for imputation (*Kraja*  
209 *et al., 2011; Ebejer et al., 2013*) and two studies used genotyping chip data (*Yucesoy et al., 2015;*  
210 *Ellinghaus et al., 2016*).

211 These articles used a variety of strict quality filters, including Hardy-Weinberg equilibrium test,  
212 deviations in expected allele frequency and sequencing data quality thresholds. They also removed  
213 rare alleles and alleles with high degrees of missingness. Despite using state-of-the-art quality  
214 controls, these variants managed not only to be imputed onto real genotype data, but they also  
215 reached genome wide significance for association with biological traits.

216 These associations are not necessarily incorrect – a weak but significant bias in imputation  
217 may still result in a correct associations. To distinguish between variants with weak but significant  
218 association with  $Q$  from variants with strong biases, we distinguished between variants where  
219 the allele frequency difference between individuals with low- and high- $Q$  is larger than a factor of  
220 two (which naturally separates two clusters of variants on Figure S9). The majority (92.7%) of the  
221  $Q$ -associated variants are strongly biased in that they are more than twice as frequent in individuals  
222 with low- $Q$  compared to high- $Q$  data. By contrast, most  $Q$ -associated variants reported in the GWAS  
223 catalogue had weak bias (See Figure S10), with three exceptions. One study reports associations  
224 with seven  $Q$ -associated variants that we find to be highly biased (*Mandise et al., 2017*). That study  
225 considered copy number of Epstein-Barr virus (EBV) sequence in the 1kGP as a phenotype. \*It is  
226 unlikely that EBV copy number associations to these seven  $Q$ -associated variants could be explained  
227 by population structure as the data quality of an individual is also correlated with sequencing date.  
228 This suggests that there may be issues with how EBV is mapped to the genomes of individuals with  
229 low quality sequence data.\* Thus the phenotype in that study is likely confounded by the same  
230 technical artefacts that lead to biased SNP calling.

## 231 Discussion

232 The variants identified in this study are likely to be technical artifacts from legacy technologies.  
233 Different sequencing technologies will have different error profiles. A report comparing the Genome  
234 Analyzer II (GAII) to the Illumina HiSeq found that the GAII had much higher rates of reads below a  
235 quality score of 30 (*Minoche et al., 2011*) with, for instance, different patterns of quality decrease  
236 along reads. Differences in read quality and error profiles in turn require different calling pipelines.

237 To pinpoint the precise technical source of the discrepancy would require further forensic  
238 inquiries into the details of the heterogeneous sample preparation and data processing pipelines

Pubmed ID	Disease/Trait	rsID	GWAS	<i>Q</i>
			-log <sub>10</sub> p	-log <sub>10</sub> p (adjusted)
28654678	EBV copy number in lymphoblastoid cell lines	rs201761909	5.7	78.11
		rs201130852	5.05	72.28
		rs201255786	5.7	68.97
		rs200655768	6.52	66.67
		rs184202621	5.52	60.45
		rs80274284	6	56.15
		rs200699422	5.3	7.43
23527680	ADHD <sup>†</sup>	rs6057648	5.4	20.5
28928442	Cold sores	rs201471471	6.52	7.87
26053186	HMPMA <sup>‡</sup> levels in smokers	rs60136336	5.7	2.25
28270201	HDL cholesterol	rs453755	7.52	5.29
23023329	Prostate cancer	rs103294	*15.3	4.32
28334899	HDL cholesterol	rs103294	*29.3	4.32
28240269	Blood protein levels	rs103294	*72.7	4.32
27863252	High light scatter reticulocyte count	rs3794738	*13.15	3.73
29534301	Response to hepatitis B vaccine	rs9273062	*9.7	3.36
21386085	Metabolic syndrome	rs301	*10.52	3.02
26830138	Alzheimer disease and age of onset	rs77894924	6.7	2.77
29617998	Intraocular pressure	rs4963156	*22.4	2.52
28698626	Immunoglobulin A vasculitis	rs11015915	5.05	2.45
26974007	Chronic inflammatory diseases	rs3124998	*8.05	2.33
26634245	Post bronchodilator FEV1/FVC ratio	rs451000	6	2.28
		rs443874	5.3	2.26
		rs400942	6	2.2
25918132	Diisocyanate-induced asthma	rs76780579	6	2.09

**Table 2.** Recent publications that reported *Q*-associated variants as close to or above the genome-wide significant threshold. The variants reaching genome wide significance have a star (\*). The black text colour indicates that this variant is twice as frequent in individuals with *Q* < 30, grey text colour indicates that these variants are less than twice as frequent in individuals with *Q* < 30 (See Figure S10). <sup>†</sup> Attention deficit hyperactivity disorder. <sup>‡</sup> 3-hydroxy-1-methylpropylmercapturic acid.

used throughout the 1kGP. Given the progress in sequencing and calling that occurred since the early phases of the 1kGP (Figure 2), it is likely that the source of these biases is not longer being actively introduced in recent sequence data.

However, because the 1kGP data is widely used as a reference database, these variants are still being imputed onto new genotype data and can then impact association studies for a variety of phenotypes. Even though significant association of a variant with a quality metric is not in itself an indication that the variant is spurious, we would recommend to carefully examine GWAS

246 associations for such variants, e.g. by repeating the analysis without the 1kGP as part of the  
247 imputation panel.

248 For analyses where individual variants cannot be examined individually (mutation profiles,  
249 distributions of allele frequencies, polygenic risk scores), we would recommend to simply discard the  
250  $Q$ -associated SNPs or the individuals with  $Q < 30$  (lists of such variants and sample IDs are provided  
251 in the Supplementary Data). We also recommend that imputation servers discard individuals with  
252 low  $Q$  (or at least provide the option of performing the imputation without). Given the value of freely  
253 accessible data, resequencing individuals with low  $Q$  would also likely be a worthwhile investment  
254 for the community.

## 255 Conclusion

256 On a technical front, we were surprised that strong association between variants and technical  
257 covariates in the 1kGP project had not been identified before. The genome-wide logistic regression  
258 analysis of genotype on quality metric is straightforward, and should probably be a standard in  
259 a variety of -omics studies. The logistic factor analysis is more computationally demanding but  
260 produces more robust results (*Song et al., 2015*). Both approaches produce comparable results.

261 More generally, to improve the quality of genomic reference datasets, we can proceed by  
262 addition of new and better data and by better curation of existing data. Given rapid technological  
263 progress, the focus of genomic research is naturally on the data generation side. However, cleaning  
264 up existing databases is also important to avoid generating spurious results. The present findings  
265 suggest that a substantial fraction of data from the final release of the 1kGP project is overdue for  
266 retirement or re-sequencing.

## 267 Methods

### 268 Code and data availability

269 Since this data is primarily performed using publicly available data, we provide fully reproducible and  
270 publicly available on [GitHub](#). This repository includes scripts used for data download, processing,  
271 analysis and plotting.

### 272 Metadata

273 The metadata used in this analysis was compiled from each of the index files from the 1kGP file  
274 system. Average quality of mapped bases  $Q$  per sample was obtained from the BAS files associated  
275 with each alignment file. Each BAS file has metadata regarding each sequencing event for each  
276 sample. If a sample was sequenced more than once, we took the average of each  $Q$  score from  
277 each sequencing instance. The submission dates and sequencing centres for each sample in the  
278 analysis was available in the sequence index files.

### 279 Quality Controls

280 For the mutation spectrum analysis, we reproduced the quality control and data filtering pipelines  
281 used by Harris et al. as they applied the current state of the art quality thresholds to remove  
282 questionable sequences for detecting population level differences. Several mask files were applied  
283 to remove regions of the genome that might be lower quality, or might have very different mutation  
284 rates or base pair complexity compared to the rest of the genome. The 1kGP strict mask was used  
285 to remove low quality regions of the genome, highly conserved regions were removed using the  
286 phastCons100way mask file and highly repetitive regions were removed using the NestedRepeats  
287 mask file from RepeatMasker. Furthermore, only sites with missingness below 0.01, MAF less than  
288 0.1, and MAF greater than 0.9 were considered. In total, 7,786,023 diallelic autosomal variants  
289 passed our quality controls for the mutation spectrum analysis. We calculated the mutation  
290 spectrum of base pair triplets for the list of significant variants for the JPT population using a similar  
291 method as described in (*Harris and Pritchard, 2017*).

292 For the reverse GWAS, the only filtration used was the application of a minor and major  
 293 allele frequency cutoff of 0.000599 (removing singletons, doubletons and tripletons) resulting in  
 294 a total of S=28,516,063 variants included in the test. We also used the NestedRepeats mask file  
 295 to flag variants inside repetitive regions as these were analyzed separately for false discovery  
 296 rate estimation. Variants flagged by the 1kGP strict mask are included in the association test and  
 297 included in the FDR adjustment. These variants are only removed after the FDR and excluded from  
 298 downstream discussion of error patterns, since most population genetics analyses use the strict  
 299 mask as a filter, and we expect to find problematic variants in filtered regions.

300 **Testing the association of quality to genotype**

When conducting a statistical analysis of population genetics data, we must account for population structure. In a typical GWAS, we are interested in modelling the phenotype as a function of the genotype. Here we have the opposite situation, where the quantitative variable ( $Q$ ) is used as an explanatory variable. So we consider models where the genotype  $y$  is a function of an expected frequency  $\pi_{si}$ , based on population structure, and  $Q$ . The null model is

$$y_{si} \mid \pi_{si} \sim \text{Binomial}(2, \pi_{si}). \quad (1)$$

301 The expected frequency for a SNP  $s$  and individual  $i$  can be estimated using principal component  
 302 analysis, categorical population labels, or logistic factor analysis (**Song et al., 2015**). The alternative  
 303 model then takes in  $Q$  as a covariate:

$$y_{si} \mid q_i, \mathbf{h}^{(i)} \sim \text{Binomial}\left(2, \text{logit}^{-1}\left(\text{logit}(\pi_{si}) + \beta_s q_i\right)\right). \quad (2)$$

304 Under the null hypothesis the slope coefficient  $\beta_s$  is zero and Model (2) reduces to Model (1).  
 305  $\beta_s$  denotes the association to average quality of mapped bases  $Q$  to genotype  $y_s$ . To test the null  
 306 hypothesis, we use the generalized likelihood ratio test statistic, whose deviance is a measure of  
 307 the marginal importance of adding  $Q$  in the model. The deviance test statistic under the null model  
 308 is approximately chi-square distributed with one degrees of freedom.

309 We run a total of  $S$  regressions, where  $S$  is the total number of genomic loci. Given the large  
 310 number of tests, the large proportion of expected null hypotheses and the positive dependencies  
 311 across the genome, we used the two-stage Benjamini & Hochberg step-up FDR-controlling proce-  
 312 dure to adjust the  $p$ -values (**Benjamini et al., 2006**). By using a nominal Type-I error rate  $\alpha = 0.01$ , a  
 313 total of 15,270 variants were found to be statistically significance. See Supplementary Data for a list  
 314 of variants and adjusted  $p$ -values.

315 **Individual-specific allele frequency**

Examples of models that are widely used to account population structure include the Balding-Nichols model (**Balding and Nichols, 1995**), and the Pritchard-Stephens-Donnelly model (**Pritchard et al., 2000**). These and several other similar models used in GWAS studies can be understood in terms of the following matrix factorization.

$$\mathbf{L} = \mathbf{A}\mathbf{H} \quad (3)$$

where the  $i^{\text{th}}$  column ( $\mathbf{h}^{(i)}$ ) of the  $K \times I$  matrix  $\mathbf{H}$  encodes the population structure of the  $i^{\text{th}}$  individual and the  $s^{\text{th}}$  row of the  $S \times K$  matrix  $\mathbf{A}$  determines how that structure is manifested in SNP  $s$ . When Hardy-Weinberg equilibrium holds, observed genotype can be assumed to be generated by the following Binomial model.

$$y_{si} \mid \pi_{si} \sim \text{Binomial}(2, \pi_{si}) \quad (4)$$

316 for  $s = 1 \dots S$  and  $i = i, \dots, I$ , where  $y_{si} \in \{0, 1, 2\}$  and  $\text{logit}(\pi_{si})$  is the  $(s, i)$  element of the matrix  $\mathbf{L}$   
 317 such that  $\pi_{si}$  is the individual-specific allele frequency.

To test whether quality is associated to genotype while adjusting for population structure, we performed the Genotype-Conditional Association Test (GCAT) proposed by (*Song et al., 2015*). The GCAT is a regression approach that assumes the following model.

$$y_{si} \mid q_i, \mathbf{h}^{(i)} \sim \text{Binomial}\left(2, \text{logit}^{-1}\left(\sum_{k=0}^K a_{sk} h_{ki} + \beta_s q_i\right)\right) \quad (5)$$

for  $s = 1 \dots S$  and  $i = i, \dots, I$  ( $S = 28,516,063$  and  $I = 2,504$ ) and where  $\hat{h}_{0i} = 1$  so that  $a_{s0}$  is the intercept term and  $\text{logit}(\pi_{si}) = \sum_{k=0}^K a_{sk} h_{ki}$ . The vectors  $\mathbf{h}^i$  of the matrix  $\mathbf{H}$  are unobserved but can be estimated using Logistic Factor Analysis (LFA) (*Song et al., 2015*) and are therefore used directly in the model. We approximated the population structure using  $K = 5$  latent components from a subsampled genotype matrix consisting of  $M = 2,306,130$  SNPs (we picked SNPs from the 1kGP OMNI 2.5). To avoid possible biases in computing PCA from the biased variants, we considered the genotype matrix  $L$  obtained by downsampling 1kGP variants the positions from the OMNI 2.5M chip.

### 326 Imputation

Using the Michigan Imputation Server, we imputed the genotype data from 1kGP for chromosomes 1 and 2. We used the genotyped data from the 1kGP Omni 2.5M chip genotype data. The VCF file returned from the server was then downloaded and used to search for the number of significant variants successfully imputed.

### 331 Acknowledgments

We would like to thank Kelly Harris for sharing her mutation spectrum scripts. We would also like to thank the members of the Gravel lab for their help with coding and useful discussions.

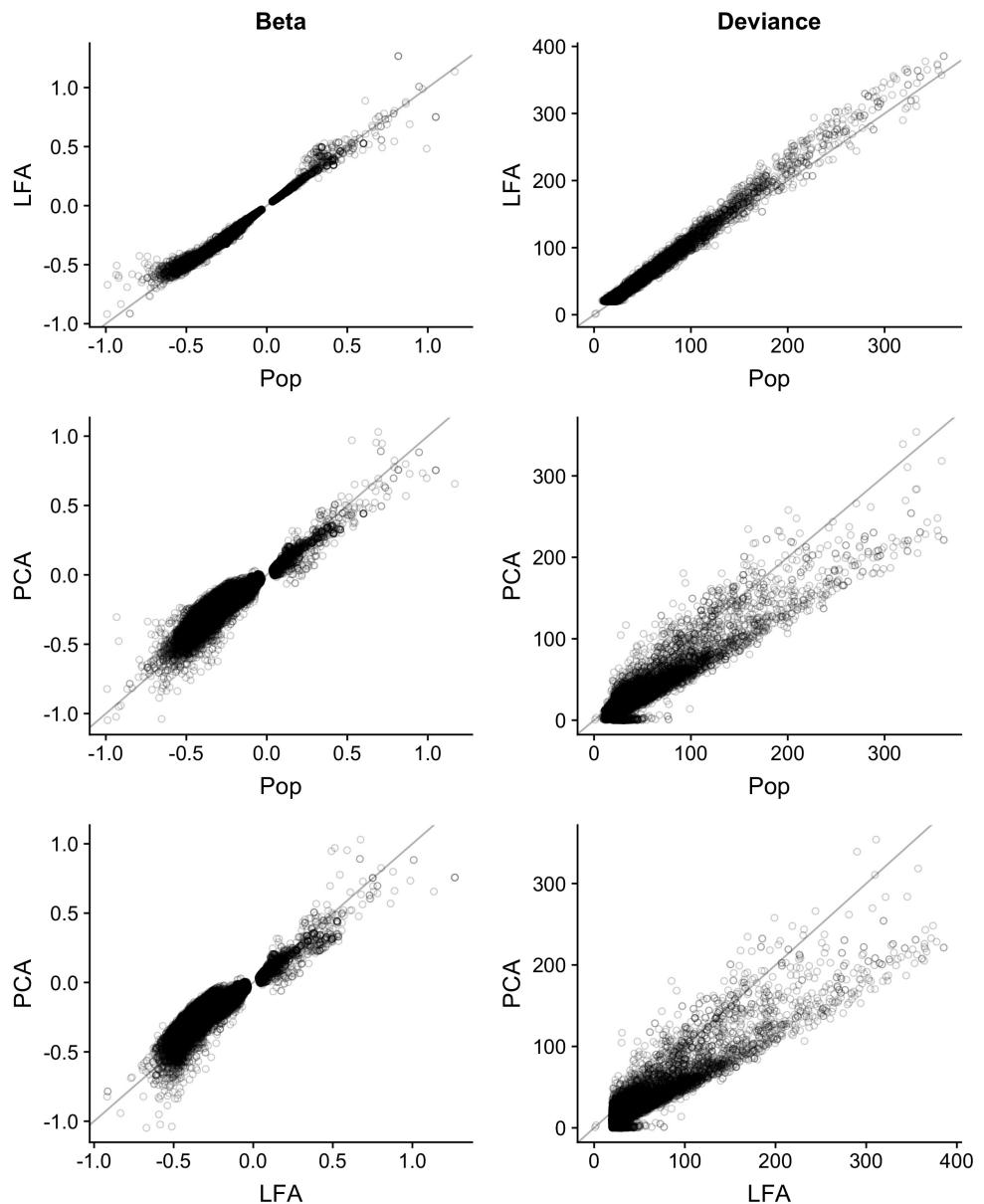
### 334 References

- 335 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73.
- 337 1000 Genomes Project Consortium (2012). An integrated map of genetic variation. *Nature*, 135:0–9.
- 338 Aikens, R. C., Johnson, K. E., and Voight, B. F. (2019). Signals of Variation in Human Mutation Rate at Multiple  
339 Levels of Sequence Context. *Molecular Biology and Evolution*.
- 340 Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N.,  
341 Borg, A., Børresen-Dale, A. L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C.,  
342 Eils, R., Eyjórd, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Illicic, T., Imbeaud, S., Imielinski, M.,  
343 Jäger, N., Jones, D. T., Jonas, D., Knappskog, S., Koo, M., Lakhani, S. R., López-Otín, C., Martin, S., Munshi,  
344 N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., Puente, X. S.,  
345 Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N., Span,  
346 P. N., Teague, J. W., Totoki, Y., Tutt, A. N., Valdés-Mas, R., Van Buuren, M. M., Van 'T Veer, L., Vincent-Salomon,  
347 A., Waddell, N., Yates, L. R., Zucman-Rossi, J., Andrew Futreal, P., McDermott, U., Lichter, P., Meyerson, M.,  
348 Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J., and Stratton, M. R. (2013).  
349 Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421.
- 350 Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., Mead, D., Bouman, H., Riveros-Mckay, F.,  
351 Kostadima, M. A., et al. (2016). The allelic landscape of human blood cell trait variation and links to common  
352 complex disease. *Cell*, 167(5):1415–1429.
- 353 Balding, D. J. and Nichols, R. A. (1995). A method for quantifying differentiation between populations at  
354 multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1-2):3–12.
- 355 Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false  
356 discovery rate. *Biometrika*.
- 357 Consortium, . G. P. et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68.

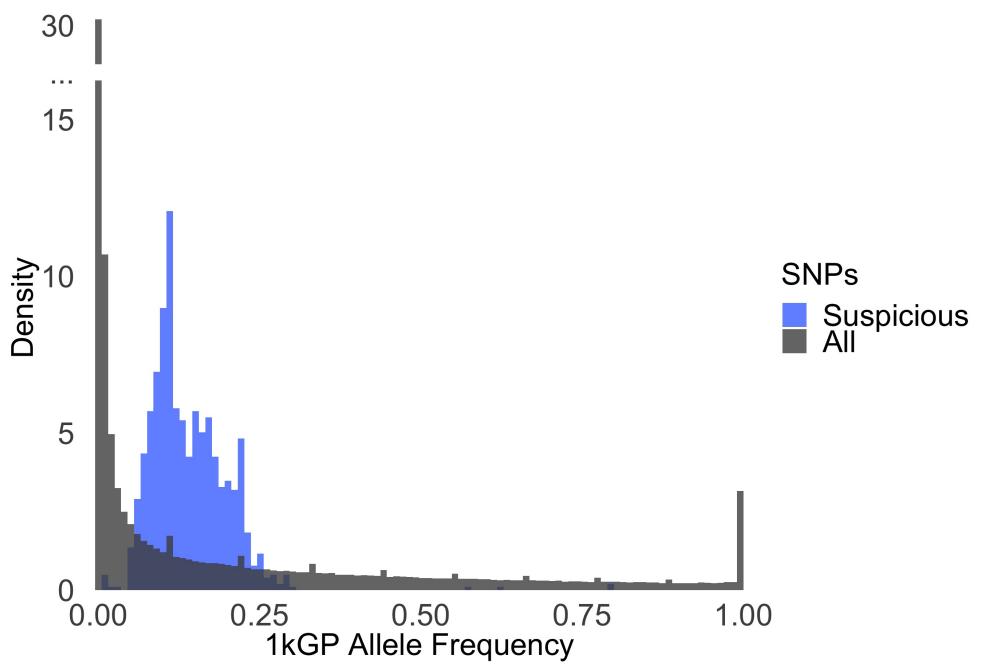
- 358 Ebejer, J. L., Duffy, D. L., Van Der Werf, J., Wright, M. J., Montgomery, G., Gillespie, N. A., Hickie, I. B., Martin,  
359 N. G., and Medland, S. E. (2013). Genome-wide association study of inattention and hyperactivity-impulsivity  
360 measured as quantitative traits. *Twin Research and Human Genetics*.
- 361 Ellinghaus, D., Jostins, L., Spain, S. L., Cortes, A., Bethune, J., Han, B., Park, Y. R., Raychaudhuri, S., Pouget, J. G.,  
362 Hübenthal, M., et al. (2016). Analysis of five chronic inflammatory diseases identifies 27 new associations and  
363 highlights disease-specific patterns at shared loci. *Nature genetics*, 48(5):510.
- 364 Gao, X. R., Huang, H., Nannini, D. R., Fan, F., and Kim, H. (2018). Genome-wide association analyses identify new  
365 loci influencing intraocular pressure. *Human molecular genetics*, 27(12):2205–2213.
- 366 Harris, K. (2015). Evidence for recent, population-specific evolution of the human mutation rate. *Proceedings of  
367 the National Academy of Sciences*, 112(11):3439–3444.
- 368 Harris, K. and Pritchard, J. K. (2017). Rapid evolution of the human mutation spectrum. *eLife*, 6.
- 369 Herold, C., Hooli, B. V., Mullin, K., Liu, T., Roehr, J. T., Mattheisen, M., Parrado, A. R., Bertram, L., Lange, C., and  
370 Tanzi, R. E. (2016). Family-based association analyses of imputed genotypes reveal genome-wide significant  
371 association of alzheimer's disease with osbpl6, ptprg, and pdcl3. *Molecular psychiatry*, 21(11):1608.
- 372 International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–320.
- 373 Kraja, A. T., Vaidya, D., Pankow, J. S., Goodarzi, M. O., Assimes, T. L., Kullo, I. J., Sovio, U., Mathias, R. A., Sun, Y. V.,  
374 Franceschini, N., Absher, D., Li, G., Zhang, Q., Feitosa, M. F., Glazer, N. L., Haritunians, T., Hartikainen, A. L.,  
375 Knowles, J. W., North, K. E., Iribarren, C., Kral, B., Yanek, L., O'Reilly, P. F., McCarthy, M. I., Jaquish, C., Couper,  
376 D. J., Chakravarti, A., Psaty, B. M., Becker, L. C., Province, M. A., Boerwinkle, E., Quertermous, T., Palotie, L.,  
377 Jarvelin, M. R., Becker, D. M., Kardia, S. L., Rotter, J. I., Chen, Y. D. I., and Borecki, I. B. (2011). A bivariate  
378 genome-wide approach to metabolic syndrome: STAMPEED Consortium. *Diabetes*.
- 379 Lan, T., Lin, H., Zhu, W., Laurent, T. C. A. M., Yang, M., Liu, X., Wang, J., Wang, J., Yang, H., Xu, X., and Guo, X. (2017).  
380 Deep whole-genome sequencing of 90 han chinese genomes. *GigaScience*, 6(9):gix067.
- 381 Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S.,  
382 Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F.,  
383 Zou, J., Pierce-Hoffman, E., Bergthout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer,  
384 M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M. I., Moonshine, A. L., Natarajan,  
385 P., Orozco, L., Peloso, G. M., Poplin, R., Rivas, M. A., Ruano-Rubio, V., Rose, S. A., Ruderfer, D. M., Shakir, K.,  
386 Stenson, P. D., Stevens, C., Thomas, B. P., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H. H., Yu, D., Altshuler,  
387 D. M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J. C., Gabriel, S. B., Getz, G., Glatt,  
388 S. J., Hultman, C. M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R.,  
389 Neale, B. M., Palotie, A., Purcell, S. M., Saleheen, D., Scharf, J. M., Sklar, P., Sullivan, P. F., Tuomilehto, J., Tsuang,  
390 M. T., Watkins, H. C., Wilson, J. G., Daly, M. J., and MacArthur, D. G. (2016). Analysis of protein-coding genetic  
391 variation in 60,706 humans. *Nature*, 536(7616):285–291.
- 392 López-Mejías, R., Carmona, F. D., Castañeda, S., Genre, F., Remuzgo-Martínez, S., Sevilla-Perez, B., Ortego-  
393 Centeno, N., Llorca, J., Ubilla, B., Mijares, V., et al. (2017). A genome-wide association study suggests the hla  
394 class ii region as the major susceptibility locus for iga vasculitis. *Scientific reports*, 7(1):5088.
- 395 Lutz, S. M., Cho, M. H., Young, K., Hersh, C. P., Castaldi, P. J., McDonald, M.-L., Regan, E., Mattheisen, M., DeMeo,  
396 D. L., Parker, M., et al. (2015). A genome-wide association study identifies risk loci for spirometric measures  
397 among smokers of european and african ancestry. *BMC genetics*, 16(1):138.
- 398 Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S.,  
399 Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T.,  
400 Gallo, C., Spence, J. P., Song, Y. S., Poletti, G., Balloux, F., Van Driem, G., De Knijff, P., Romero, I. G., Jha, A. R.,  
401 Behar, D. M., Bravi, C. M., Capelli, C., Hervig, T., Moreno-Estrada, A., Posukh, O. L., Balanovska, E., Balanovsky,  
402 O., Karachanak-Yankova, S., Sahakyan, H., Toncheva, D., Yepiskoposyan, L., Tyler-Smith, C., Xue, Y., Abdullah,  
403 M. S., Ruiz-Linares, A., Beall, C. M., Di Rienzo, A., Jeong, C., Starikovskaya, E. B., Metspalu, E., Parik, J., Villemans,  
404 R., Henn, B. M., Hodoglugil, U., Mahley, R., Sajantila, A., Stamatoyannopoulos, G., Wee, J. T., Khusainova, R.,  
405 Khusnutdinova, E., Litvinov, S., Ayodo, G., Comas, D., Hammer, M. F., Kivisild, T., Klitz, W., Winkler, C. A., Labuda,  
406 D., Bamshad, M., Jorde, L. B., Tishkoff, S. A., Watkins, W. S., Metspalu, M., Dryomov, S., Sukernik, R., Singh, L.,  
407 Thangaraj, K., Paäbo, S., Kelso, J., Patterson, N., and Reich, D. (2016). The Simons Genome Diversity Project:  
408 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206.

- 409 Mandage, R., Telford, M., Rodríguez, J. A., Farré, X., Layouni, H., Marigorta, U. M., Cundiff, C., Heredia-Genestar,  
410 J. M., Navarro, A., and Santpere, G. (2017). Genetic factors affecting EBV copy number in lymphoblastoid cell  
411 lines derived from the 1000 Genome Project samples. *PLoS ONE*.
- 412 Mathieson, I. and Reich, D. (2017). Differences in the rare variant spectrum among human populations. *PLoS  
413 Genetics*, 13(2).
- 414 McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C.,  
415 Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature  
416 genetics*, 48(10):1279.
- 417 Minoche, A. E., Dohm, J. C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing  
418 data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*, 12(11).
- 419 Nagy, R., Boutin, T. S., Marten, J., Huffman, J. E., Kerr, S. M., Campbell, A., Evenden, L., Gibson, J., Amador, C.,  
420 Howard, D. M., et al. (2017). Exploration of haplotype research consortium imputation for genome-wide  
421 association studies in 20,032 generation scotland participants. *Genome medicine*, 9(1):23.
- 422 Nishida, N., Sugiyama, M., Sawai, H., Nishina, S., Sakai, A., Ohashi, J., Khor, S.-S., Kakisaka, K., Tsuchiura, T., Hino,  
423 K., et al. (2018). Key hla-drb1-dqb1 haplotypes and role of the btln2 gene for response to a hepatitis b vaccine.  
424 *Hepatology*, 68(3):848–858.
- 425 Park, S. L., Carmella, S. G., Chen, M., Patel, Y., Stram, D. O., Haiman, C. A., Le Marchand, L., and Hecht, S. S. (2015).  
426 Mercapturic acids derived from the toxicants acrolein and crotonaldehyde in the urine of cigarette smokers  
427 from five ethnic groups with differing risks for lung cancer. *PLoS One*, 10(6):e0124841.
- 428 Pfeifer, G. P., Denissenko, M. F., Olivier, M., Tretyakova, N., Hecht, S. S., and Hainaut, P. (2002). Tobacco smoke  
429 carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene*, 21-48(6):7435–7451.
- 430 Pleasance, E. D., Stephens, P. J., O'Meara, S., McBride, D. J., Meynert, A., Jones, D., Lin, M. L., Beare, D., Lau,  
431 K. W., Greenman, C., Varela, I., Nik-Zainal, S., Davies, H. R., Ordóez, G. R., Mudie, L. J., Latimer, C., Edkins, S.,  
432 Stebbings, L., Chen, L., Jia, M., Leroy, C., Marshall, J., Menzies, A., Butler, A., Teague, J. W., Mangion, J., Sun, Y. A.,  
433 McLaughlin, S. F., Peckham, H. E., Tsung, E. F., Costa, G. L., Lee, C. C., Minna, J. D., Gazdar, A., Birney, E., Rhodes,  
434 M. D., McKernan, K. J., Stratton, M. R., Futreal, P. A., and Campbell, P. J. (2010). A small-cell lung cancer genome  
435 with complex signatures of tobacco exposure. *Nature*, 463(7278):184–190.
- 436 Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype  
437 data. *Genetics*, 155(2):945–959.
- 438 Shiraishi, Y., Tremmel, G., Miyano, S., and Stephens, M. (2015). A Simple Model-Based Approach to Inferring and  
439 Visualizing Cancer Mutation Signatures. *PLoS Genetics*, 11(12).
- 440 Song, M., Hao, W., and Storey, J. D. (2015). Testing for genetic associations in arbitrarily structured populations.  
441 *Nature genetics*, 47(5):550.
- 442 Spracklen, C. N., Chen, P., Kim, Y. J., Wang, X., Cai, H., Li, S., Long, J., Wu, Y., Wang, Y. X., Takeuchi, F., et al. (2017).  
443 Association analyses of east asian individuals and trans-ancestry analyses with european individuals reveal  
444 new loci associated with cholesterol and triglyceride levels. *Human molecular genetics*, 26(9):1770–1784.
- 445 Suhre, K., Arnold, M., Bhagwat, A. M., Cotton, R. J., Engelke, R., Raffler, J., Sarwath, H., Thareja, G., Wahl, A.,  
446 DeLisle, R. K., et al. (2017). Connecting genetic risk to disease end points through the human blood plasma  
447 proteome. *Nature communications*, 8:14357.
- 448 Tian, C., Hromatka, B. S., Kiefer, A. K., Eriksson, N., Noble, S. M., Tung, J. Y., and Hinds, D. A. (2017). Genome-wide  
449 association and hla region fine-mapping studies identify susceptibility loci for multiple common infections.  
450 *Nature communications*, 8(1):599.
- 451 van Dijk, E. L., Auger, H., Jaszczyzyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing  
452 technology. *Trends in Genetics*, 30(9):418–426.
- 453 Xu, J., Mo, Z., Ye, D., Wang, M., Liu, F., Jin, G., Xu, C., Wang, X., Shao, Q., Chen, Z., et al. (2012). Genome-wide  
454 association study in chinese men identifies two new prostate cancer risk loci at 9q31. 2 and 19q13. 4. *Nature  
455 genetics*, 44(11):1231.
- 456 Yucesoy, B., Kaufman, K. M., Lummus, Z. L., Weirauch, M. T., Zhang, G., Cartier, A., Boulet, L.-P., Sastre, J.,  
457 Quirce, S., Tarlo, S. M., et al. (2015). Genome-wide association study identifies novel loci associated with  
458 diisocyanate-induced occupational asthma. *Toxicological Sciences*, 146(1):192–201.

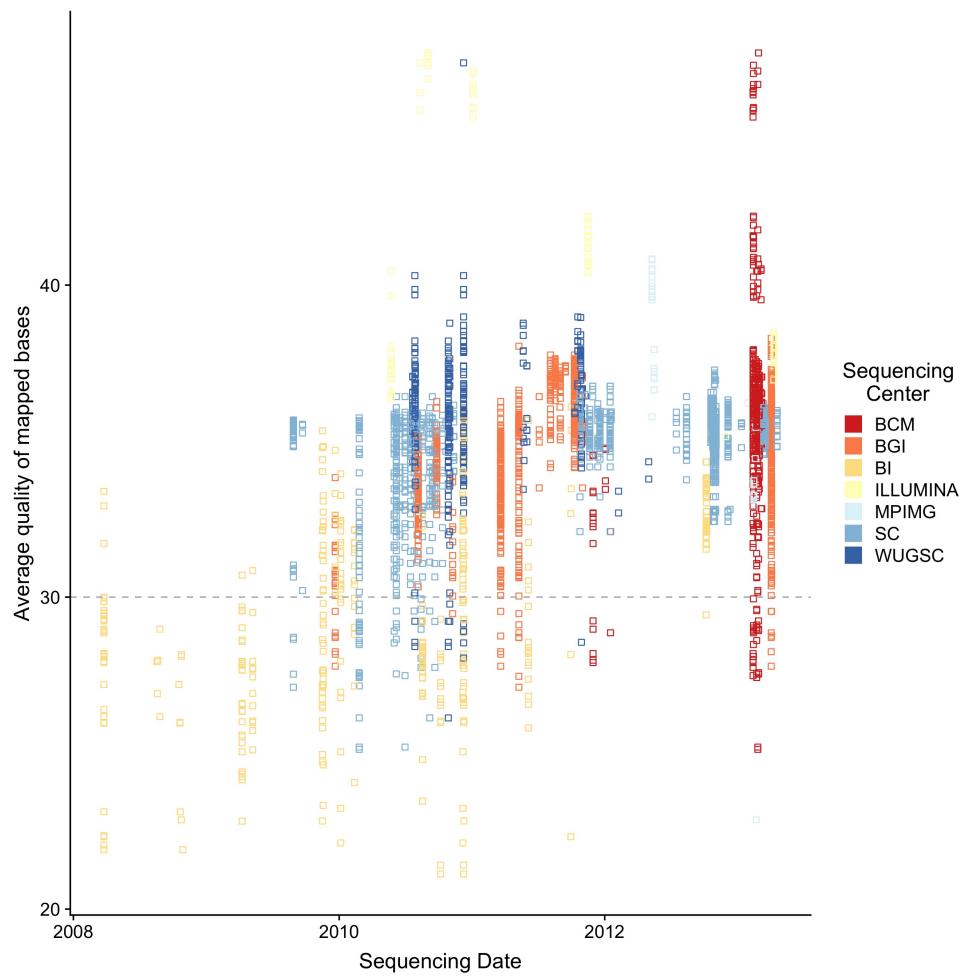
459 **Supplementary Figures**



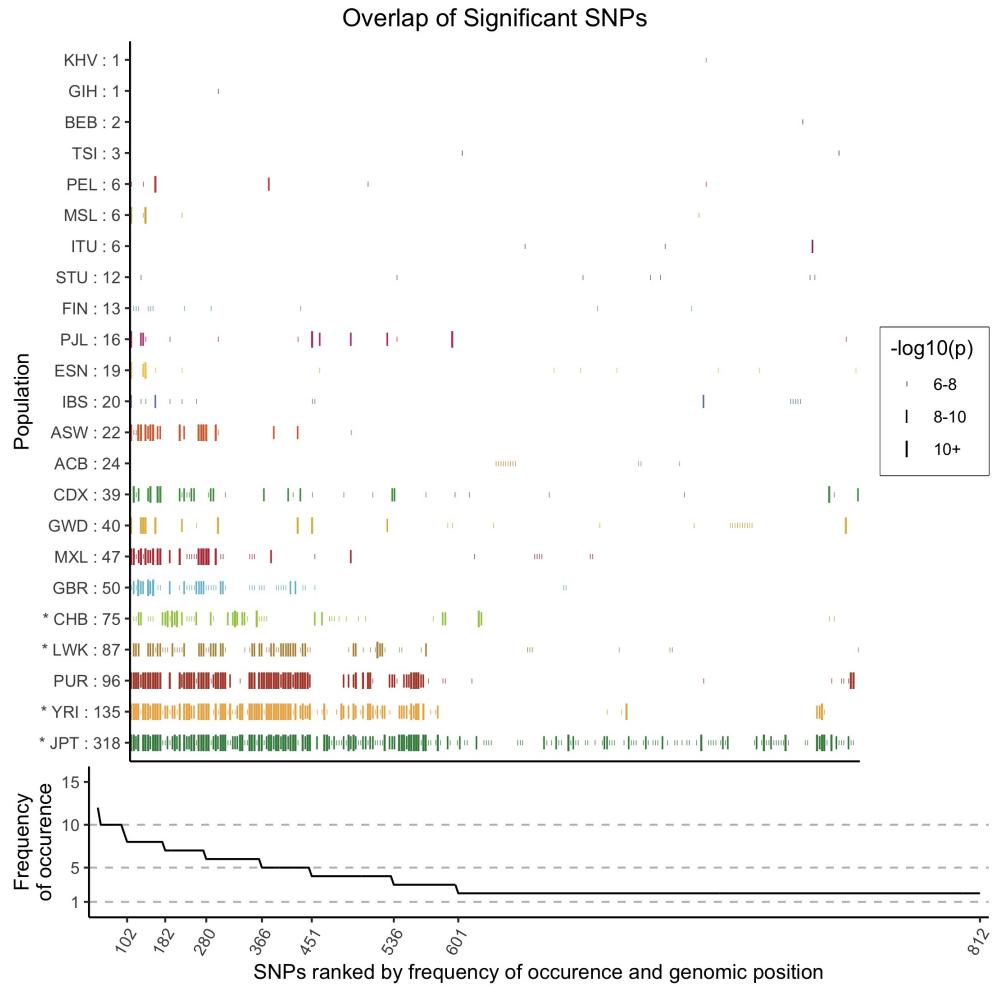
**Figure S1.** Comparison of three logistic regression models for testing association to  $Q$ . These methods model each genotype as a logistic function using principal components (PC), Population membership (Pop) or LFA as an offset. In these plots we are comparing the deviance from the null model in the 15,270 variants identified using the LFA model.



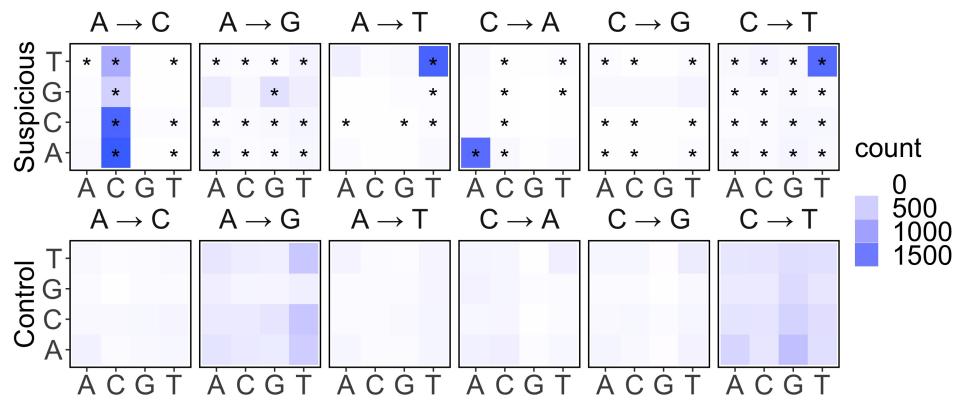
**Figure S2.** Frequency distribution of the 1034  $Q$  associated SNPs from the single population test in the JPT in blue and all SNPs in black. The association test lacks power to identify variants at low frequencies.



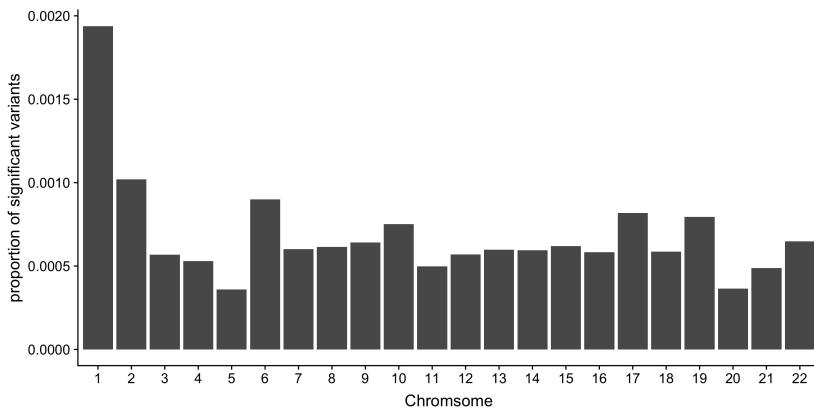
**Figure S3.** Sequencing date of samples included in the 1kGP coloured by sequencing centre. Improvements in sequencing platforms coincides with improvement in sequencing quality.



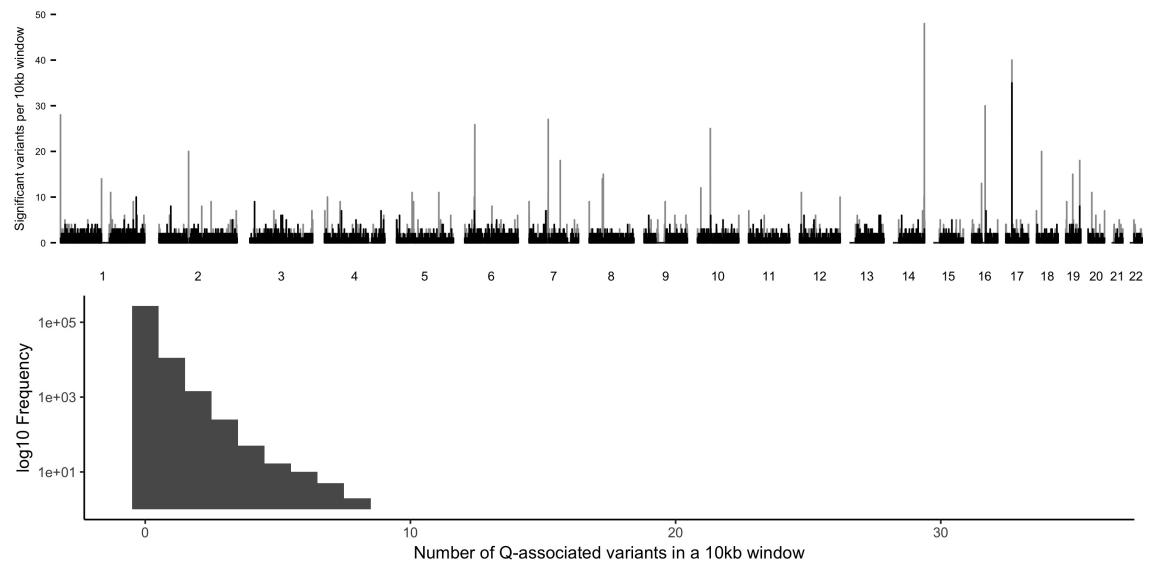
**Figure S4.** Variants associated with average quality of mapped bases  $Q$  in more than one population. The size of the vertical bars (|) are proportional to the  $-\log_{10}(p)$  value of that SNP. The x axis is ranked by the frequency of occurrence of a SNP, then by genomic position. Phase 1 populations are marked by a star (\*). The line plot underneath shows the number of populations for which a variant has reached significance. The populations that tend to have the most individuals with low  $Q$  also tend to have the most variants associated to  $Q$ .



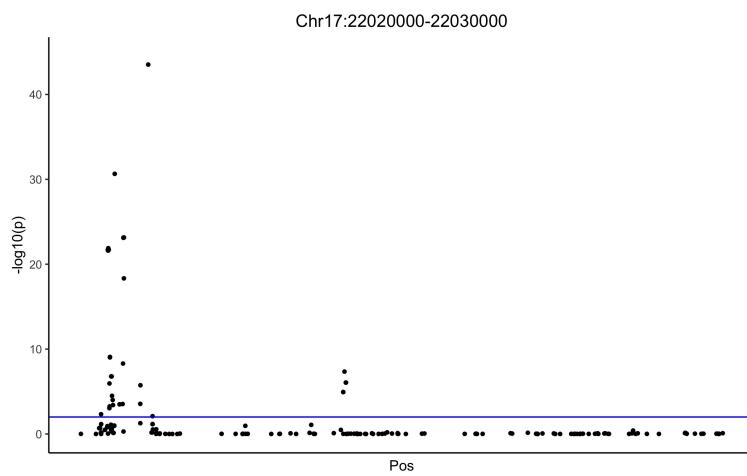
**Figure S5.** Mutation spectrum of the 15,270 *Q*-associated variants that pass the 1kGP strict mask filter, compared to a random set of non-associated variants. A large fraction of the *Q*-associated variants have the \*AC→\*CC mutational pattern. There is also an enrichment in TAT→TTT, TCT→TTT and ACA→AAA (reverse complement : TGT→TTT). These three enrichments can be summarized as T\*T→TTT. Stars (\*) indicate significant difference from the mutational spectrum of a set of control variants of equal size.



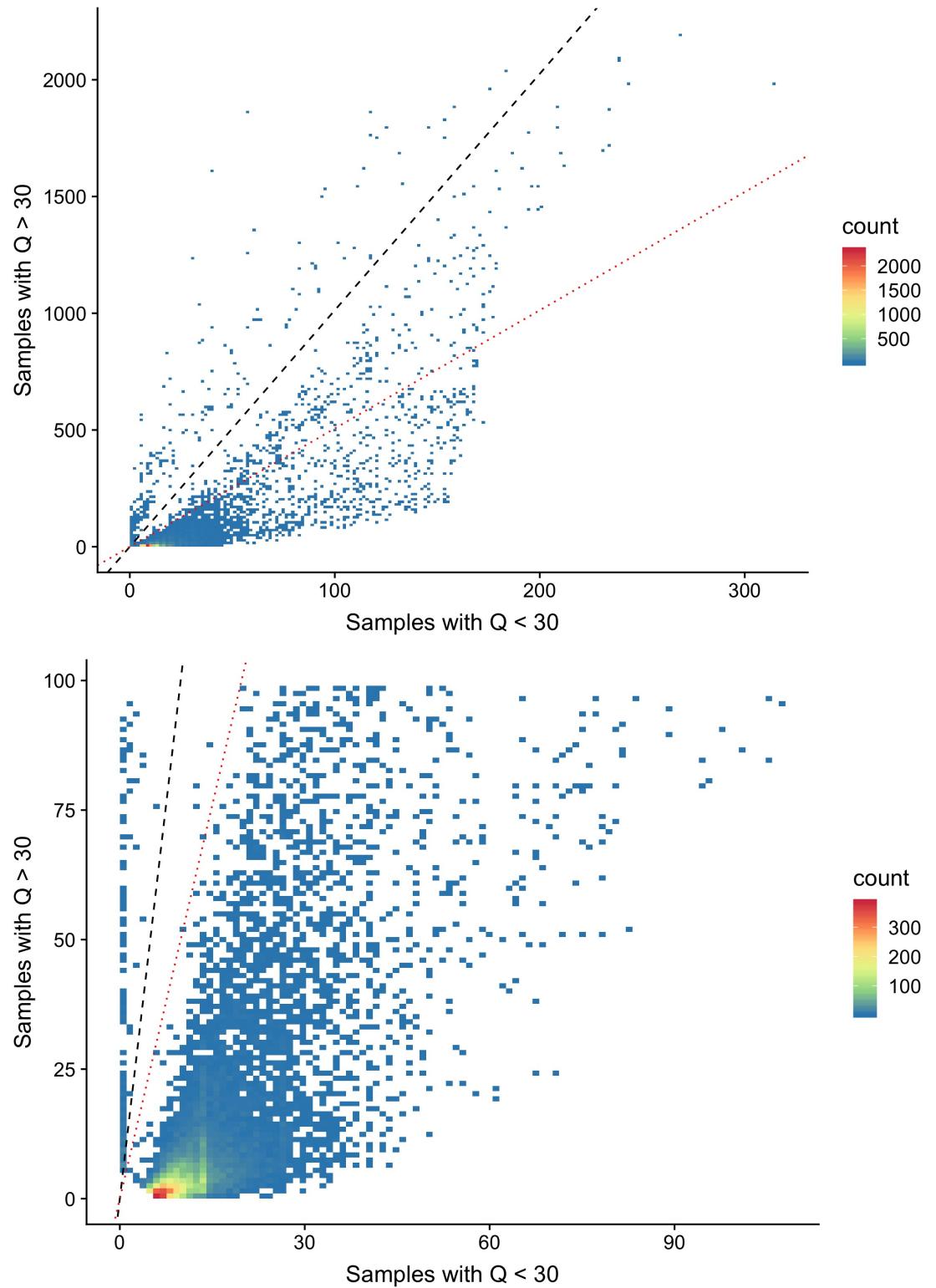
**Figure S6.** The proportion of *Q*-associated variants per chromosome.



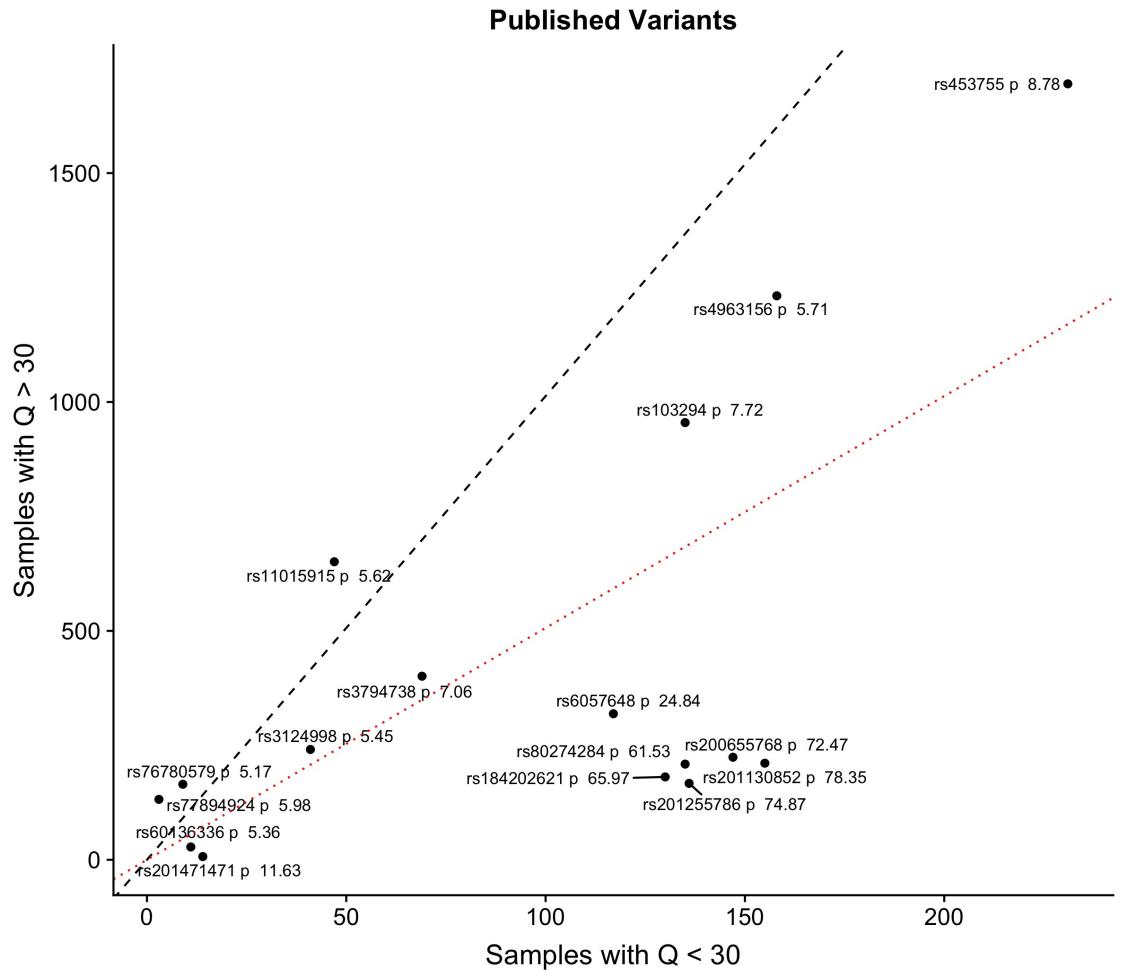
**Figure S7.** The number of *Q*-associated variants per 10kb window across the genome. Grey bars indicate regions within and black bars indicate regions outside the 1000 Genomes Project strict mask. One region not flagged by the 1000 Genomes Project strict mask in chromosome 17 has more than 10 variants per window.



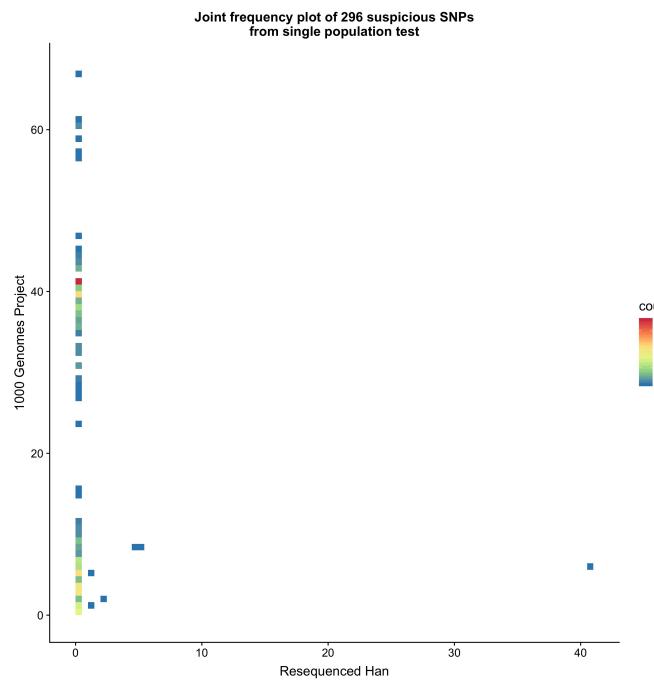
**Figure S8.** Manhattan plot of the  $-\log_{10}(p)$  values for the reverse GWAS logistic regression analysis for the 10kb window with the most *Q*-associated variants per 10kb across the genome.



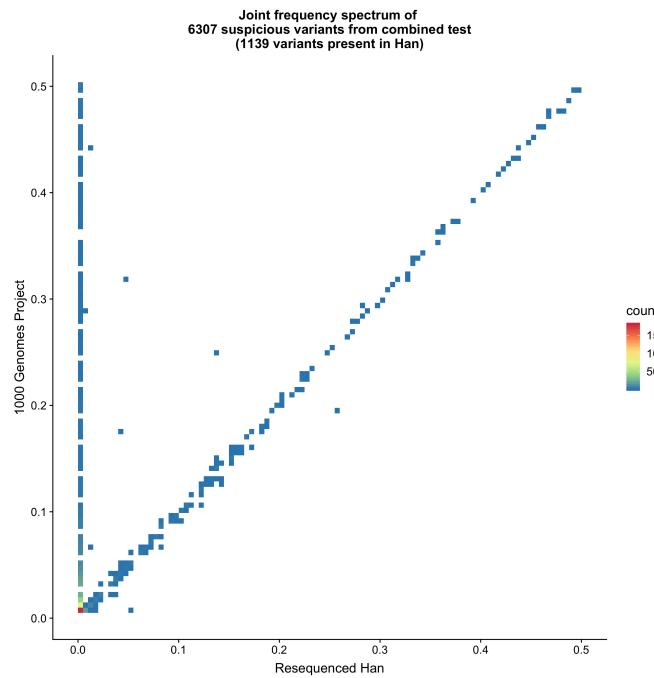
**Figure S9.** Site frequency spectrum plot comparing the allele frequency difference between individuals with low- and high- $Q$ . The black dashed lines indicates equal allele frequencies while the red dotted line for variants twice as frequent in individuals with  $Q$  scores below 30. Two clusters of are visible, where the majority (92.7%) of the  $Q$ -associated variants are more than twice as frequent in individuals with low- $Q$ .



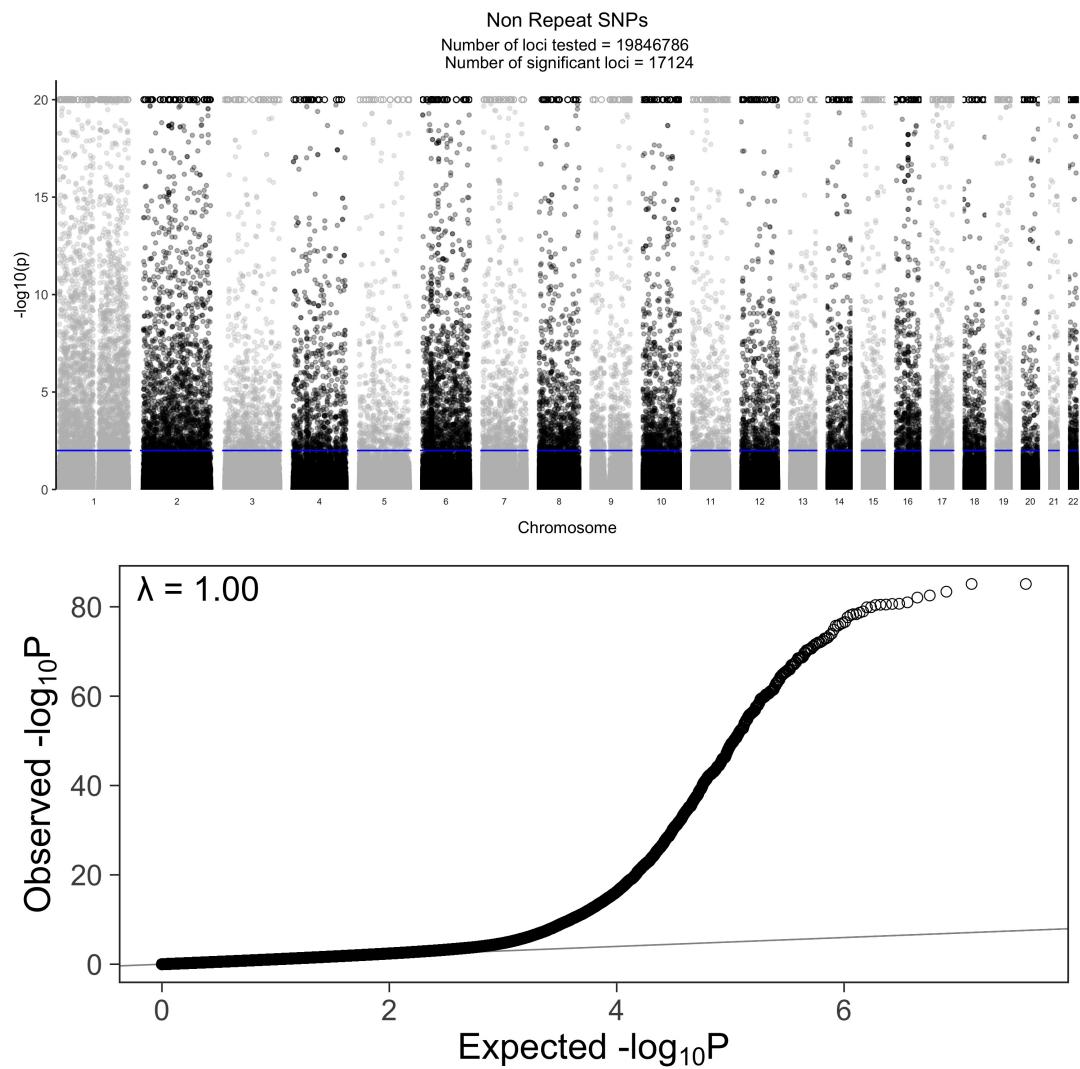
**Figure S10.** Site frequency spectrum plot comparing the frequency of  $Q$ -associated variants identified in publications, for individuals with  $Q$  scores above and below 30. The black dashed lines indicates equal allele frequencies while the red dotted line for variants twice as frequent in individuals with  $Q$  scores below 30. Each of the rsIDs of the variants are labelled for clarity.



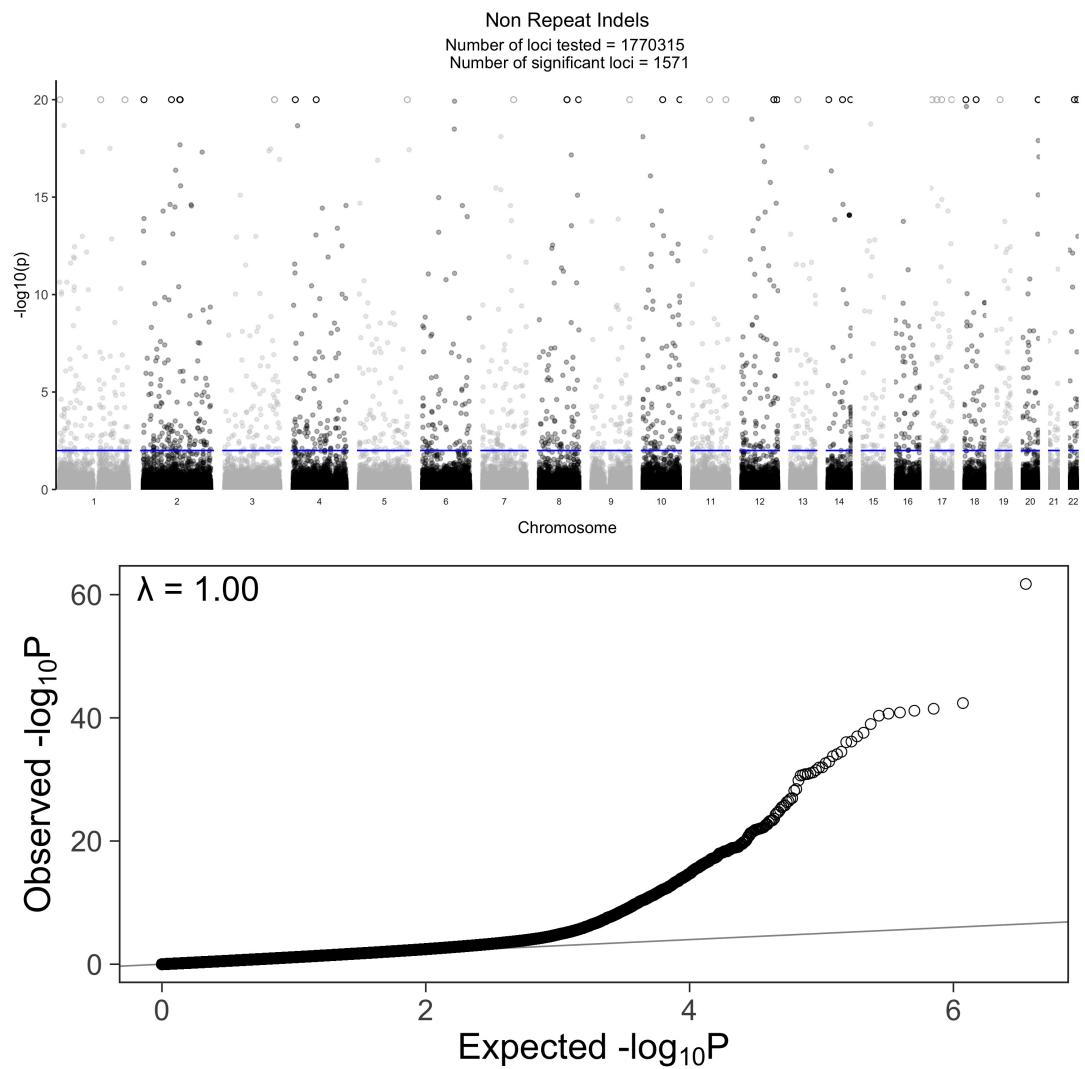
**Figure S11.** Site frequency spectrum plot comparing the original 1000 Genomes Project data to the high depth resequence data for variants that, in the 1000 Genomes Project, are both associated with  $Q$  and polymorphic in the 83 individuals that were resequenced. Among the 296 variants associated with  $Q$  in the single population tests within the 1000 Genomes Project CHB and CHS, 6 are present in the resequenced data ([Lan et al., 2017](#)).



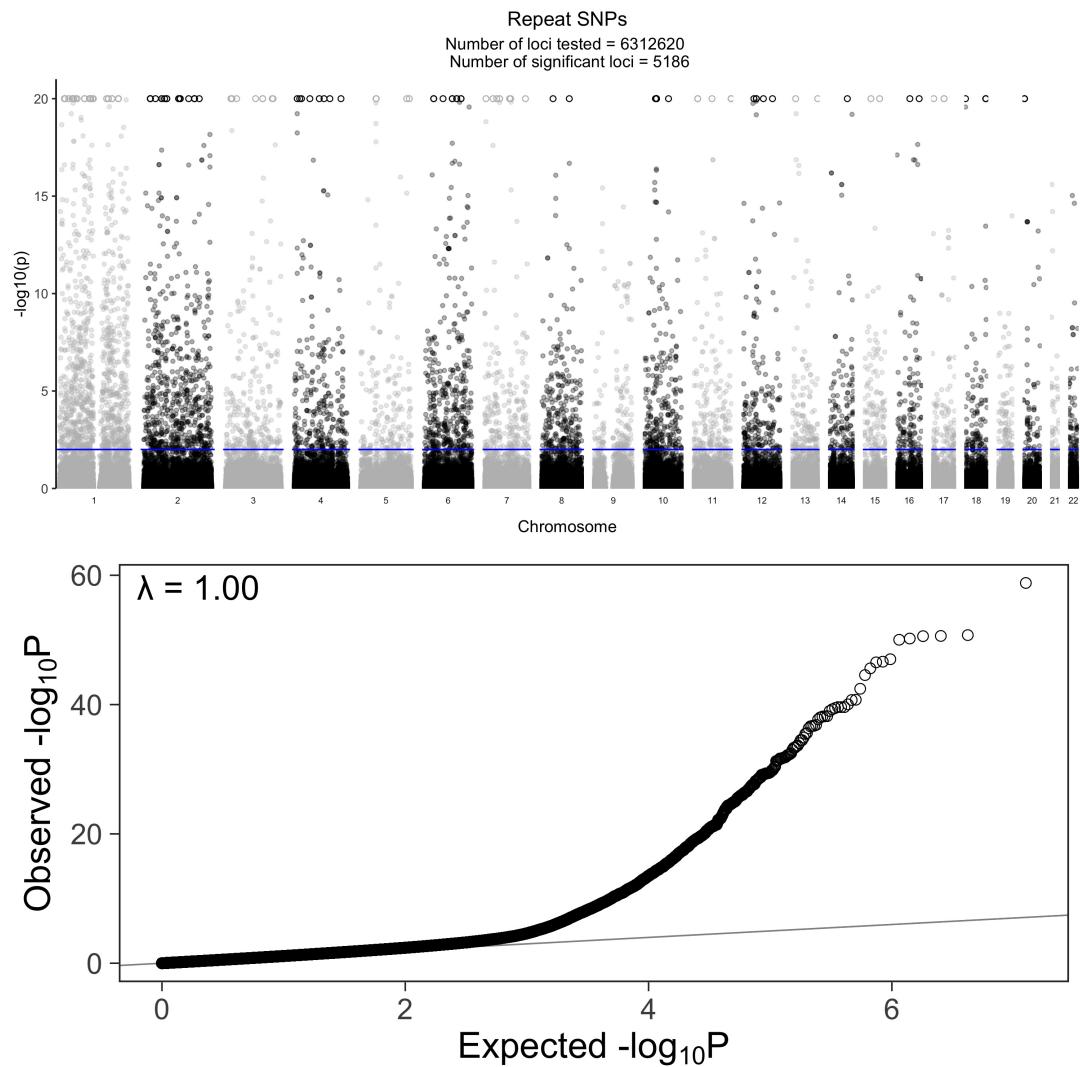
**Figure S12.** Site frequency spectrum plot comparing the original 1000 Genomes Project data to the high depth resequence data for variants that, in the 1000 Genomes Project, are both associated with  $Q$  and polymorphic in the 83 individuals that were resequenced. Among the 6,307 variants associated with  $Q$  in the GCAT model including all populations, 1,139 are present in the high depth resequenced individuals.



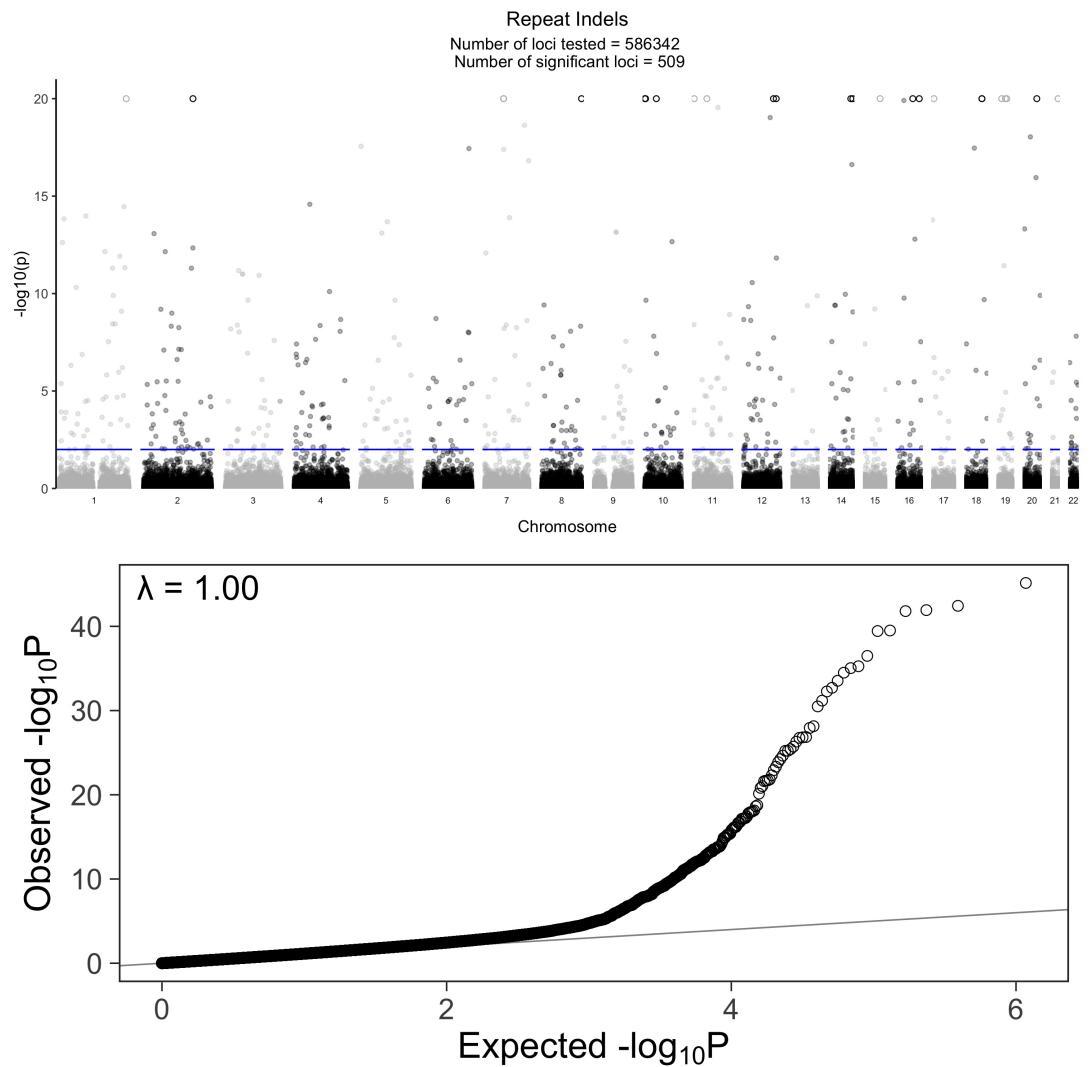
**Figure S13.** Association of SNPs in non-repetitive regions with  $Q$ . **A** Manhattan plot of the  $-\log_{10}(p)$  values for the reverse GWAS logistic regression analysis for SNPs in non repetitive regions. There are 15,018 SNPs that reach  $p$  values greater than  $p < 0.01$  after performing a two-stage Benjamini and Hochberg FDR adjustment. The circles (o) are variants that reached values greater than 20, for clarity we implemented hard ceiling at 20. **B** QQ plot of the unadjusted  $p$  values for the reverse GWAS logistic regression analysis for SNPs in non repetitive regions.



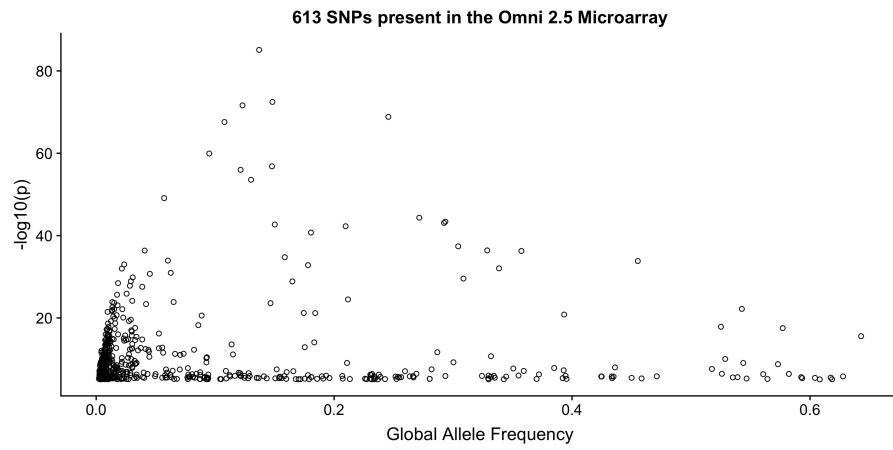
**Figure S14.** Association of indels in non-repetitive regions with  $Q$ . **A** Manhattan plot of the  $-\log_{10}(p)$  values for the reverse GWAS logistic regression analysis for INDELs in non repetitive regions. There are 2,121 INDELs that reach  $p$  values greater than  $p < 0.01$  after performing a two-stage Benjamini and Hochberg FDR adjustment. The circles (o) are variants that reached values greater than 20, for clarity we implemented hard ceiling at 20. **B** QQ plot of the unadjusted  $p$  values for the reverse GWAS logistic regression analysis for INDELs in non repetitive regions.



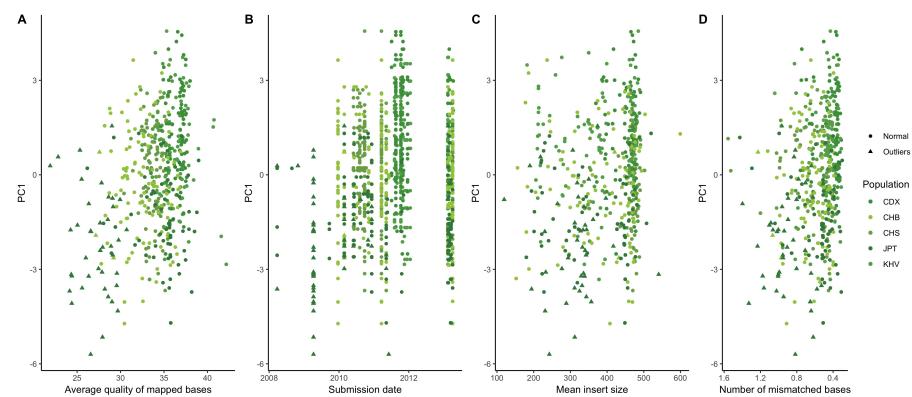
**Figure S15.** Association of SNPs in repetitive regions with  $Q$ . **A** Manhattan plot of the  $-\log_{10}(p)$  values for the reverse GWAS logistic regression analysis for SNPs in repetitive regions. There are 4,405 SNPs that reach  $p$  values greater than  $p < 0.01$  after performing a two-stage Benjamini and Hochberg FDR adjustment. The circles ( $\circ$ ) are variants that reached values greater than 20, for clarity we implemented hard ceiling at 20. **B** QQ plot of the unadjusted  $p$  values for the reverse GWAS logistic regression analysis for SNPs in repetitive regions.



**Figure S16.** Association of indels in repetitive regions with  $Q$ . **A** Manhattan plot of the  $-\log_{10}(p)$  values for the reverse GWAS logistic regression analysis for INDELs in repetitive regions. There are 642 INDELs that reach  $p$  values greater than  $p < 0.01$  after performing a two-stage Benjamini and Hochberg FDR adjustment. The circles ( $\circ$ ) are variants that reached values greater than 20, for clarity we implemented hard ceiling at 20. **B** QQ plot of the unadjusted  $p$  values for the reverse GWAS logistic regression analysis for INDELs in repetitive regions.



**Figure S17.** Estimated frequency and association strength of *Q*- associated variants present on Illumina's Omni 2.5 chip. Variants highly associated to *Q* tend to have low global allele frequencies.



**Figure S18.** Sequencing metrics against the prevalence of the \*AC→\*CC mutational signature in 1000 Genomes Project. The average quality per mapped bases *Q* per individual shows some clustering with individuals with low-quality data showing elevated rates of the signature.