

Legacy Data Confounds Modern Genomics Studies

Luke Anderson-Trocmé^{1,2} and Simon Gravel^{1,2}

*For correspondence:
simon.gravel@mcgill.ca (SG)

¹Department of Human Genetics, McGill University, Montreal, QC H3A 0G1, Canada;

²McGill University and Genome Quebec Innovation Centre, Montreal, QC H3A 0G1, Canada

Abstract In investigating a population genetics signal, we noticed some discrepancies between the Japanese samples from the 1000 Genomes Project (1kGP) and a more recent high quality dataset. We found that the variants causing the signal were only present in individuals with low average sequencing quality. We identified 625 $p < 10^{-8}$ and 1048 $p < 10^{-6}$ single nucleotide polymorphisms (SNPs) that were associated to quality in the 1kGP dataset Japanese individuals. These variants are present in nearly all of the populations from the 1kGP. We then turned our attention to the rest of the 1000 Genomes populations and saw that there were similar batch effects in many of these populations. Some of these variants are being imputed onto genotype data and reach genome wide significance in recent publications. [*update values for all populations*](#)

Introduction

Reference Cohorts

The last 5 years has seen an increase in the number of individuals genotyped through private companies or medical research cohorts. Larger sample sizes allow researchers to identify finer resolution statistically significant differences between groups of individuals. Mutations associated to a disease can be identified by comparing the genomes of healthy individuals to those afflicted by a disease. The mutations commonly found in patients and rarely in controls might be associated to the disease in question. However, demonstrating that these mutations are biologically relevant can be difficult. Especially with the increasing size of cohorts, spurious associations are increasingly becoming an issue. For this reason, careful consideration must be taken when including individuals from different ancestral origins in these association studies. Benign mutations at high frequencies in one populations might be exceedingly rare in another population. Therefore, population wide differences in mutations must be included as covariates to avoid spurious associations.

Imputation

Genome wide associations using large cohorts have lead to developments in the identification of rare genetic diseases as well as the risk prediction to certain types of cancer and diseases. Despite drastic reductions in cost of whole genome sequencing, it remains an expensive test for large sample sizes. For this reason, genotype data is often imputed to increase power for association studies in a cost effective way.

[Luke: cut paragraph to discussion](#)

Combining data can be an issue when the quality of the data produced can vary between sequencing technologies and even sequencing centres. The errors in one dataset don't disappear when they are combined with another higher quality one. Many of the legacy datasets produced

41 using dated sequencing technologies have been known to contain a higher rate of false positives
42 than their more recent counterparts. Is it time to retire legacy data?

43 **Mutation Spectrum**

44 A genome wide mutational signature can be measured by taking the sum of all the different
45 types of mutations in an individual. The signatures from individuals in the same population will
46 have a tendency to be more similar due to shared ancestry. In 2015, Harris et al reported an
47 overrepresentation of the TCC to TTC mutation in European populations. Harris and Pritchard
48 published a follow up study comparing how various ethnic groups differ from one another with
49 respect to their mutational spectrum. In this paper, they proposed that not only is there a distinct
50 mutational spectrum difference in Europeans, but that these changes are present in most human
51 populations. One result was of particular interest to our research group : the heterogeneously
52 distributed signal found in a subset of Japanese individuals. A mutational signature that is present
53 unevenly across a population is unexpected because it suggested strong population structure;
54 unlikely for a dense population living on an archipelago.

55 **Motivations**

56 This project was motivated by this unusual population genetics observation in a recent publication
57 by Harris et. al, that the 1000 Genomes Project (1kGP) Japanese population seem to be partitioned
58 in two clusters of diff mutation rates. This heterogeneously distributed mutational signal was
59 unexpected as a signal of this nature could either be due to population structure, a mutagen,
60 or a technical bias. In investigating this mutational signature, we were unable to reproduce the
61 results using a larger and higher quality dataset and concluded that this signal can be attributed to
62 sequencing error.

63 This study is the result of an investigation in the quality of the 1kGP dataset. To begin, we will
64 discuss the discrepancy between the Japanese samples from the 1kGP and a more recent high
65 quality dataset. We then consider methods to discriminate sequencing errors resulting from dated
66 technologies. Next, we explore more broadly how low quality variants remain embedded in other
67 populations of the 1kGP. Finally, we will discuss the impact these variants have on modern analyses.

68 **Results**

69 **A peculiar mutational signature in Japan**

70 If two groups of individuals are sampled from the same population, we expect there to be little
71 deviation in allele frequencies between the two groups. However, when comparing the allele
72 frequencies between the Japanese individuals from the 1kGP and a more recent dataset, we
73 observed an unusually large number of private single nucleotide polymorphisms (SNPs), only found
74 in one of the two groups. Surprisingly, the variants responsible for the signal observed by Harris
75 et. al (2017) were missing in the more recent dataset 1. We also noticed that the individuals
76 carrying these mutations all had lower average quality scores of mapped bases. This suggested
77 that individuals with lower quality were driving this signal.

78 To further investigate the extent to which low quality individuals were associated to spurious
79 mutations, we performed a genome-wide association (GWA) study 1. GWA studies are commonly
80 used to find genomic regions associated to biologically relevant traits. In this case, we are using
81 this analysis in a more unconventional way to identify regions of the genome that are associated to
82 biologically irrelevant traits like SNPs associated to individuals with low quality.

83 Using the linear GWA study function offered by PLINK, we were able to identify 625 $p < 10^{-8}$ and
84 1048 $p < 10^{-6}$ SNPs that were associated to the average quality of SNPs mapped for an individual 1.
85 This GWA study included 104 individuals, *000* of them were sequenced in phase 1 while the rest
86 were sequenced in phase 3 of the 1000 Genomes Project. The variants that are associated to the
87 quality of mapped bases have an enrichment in *AC→*CC mutations.

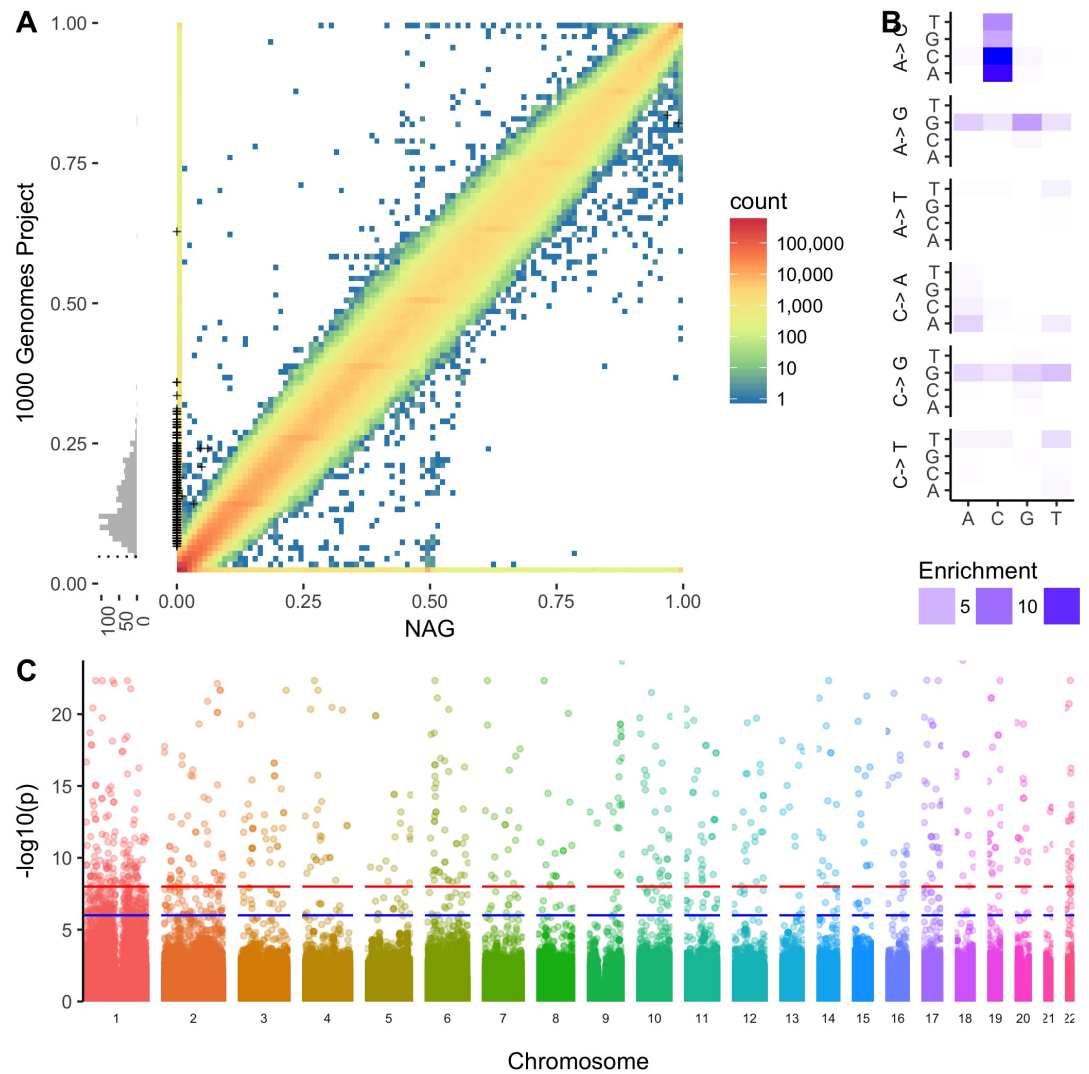


Figure 1. **A** Joint frequency spectrum plot of the Japanese from the 1000 Genomes Project and a more recent dataset. Crosses (+) are variants that reached genome wide significance in a GWA of sequencing quality. The histogram on the left of the plot is the distribution of significant variants. The lowest frequency variant able to reach genome wide significance is 5%. **B** Mutation spectrum of the 1048 variants that reached a genome wide significance with a p value less than in a GWA of sequencing quality. There is a **significant** enrichment in **AC→*CC* mutations. **C** Genome wide association of the average quality of mapped bases for the 104 Japanese individuals included in the 1000 Genomes Project. This GWA study identified 625 $p < 10^{-8}$ and 1048 $p < 10^{-6}$ SNPs that were associated to the average quality of SNPs mapped for an individual. The same analysis was performed independently for each of the populations in the 1000 Genomes Project.

88 Despite removing the variants identified as being low quality, the signal persists. When we
89 remove the variants significantly associated to low quality, the signal identified by Harris et. al
90 (2017) the enrichment in *AC→*CC mutations. When we remove individuals with average quality of
91 mapped SNPs below 30, the signal goes away. We suspect that this has to do with the lack of power
92 we have for more rare alleles. The lowest frequency variants that are significantly associated to
93 quality are at 5%. When we remove all the variants above 1% in 1kGP but missing from the NAG
94 data, the signal disappears. This suggests that while the GWA approach can identify some of the
95 low hanging bad apples, there are likely more of these false positives nested inside these legacy
96 cohorts.

97 [Luke: more here about using higher quality data to validate legacy data](#). Upon comparing the
98 Japanese 1000 Genomes cohort to a higher quality and larger cohort, we identified *000* more
99 variants that are beyond the expected frequency spectrum deviation for individuals from the same
100 population.

101 [*supplementary figures : genome wide, mutspect without sig snps, and without <30*](#)

102 Sequencing quality over time

103 We turned our attention to the other populations in the 1kGP, we found that the sequencing done
104 in phase 1 was more variable and overall tended to include lower quality sequencing data 2. The
105 sequencing quality of individuals increases over time 2. [*What sequencers were used in each
106 phase?*](#) By 2011 the sequencing quality seems to level off.

107 Overlap of significant SNPs

108 Comparing the results of each independent GWA study, we were able to identify over *0000* vari-
109 ants that were independently associated to low quality in multiple populations. This confirmation
110 using more than one GWAS is a strongly suggests that these variants might not be genuine 3.

111 Imputation

112 30% of the SNPs we identified as being associated with low quality were found to be imputed using
113 the Michigan Imputation Server. These should be removed from reference database.

114 Found to be included in other GWA studies

115 Once we identified SNPs that were clearly associated with low quality, we searched the literature for
116 any GWA studies that might have called these erroneous variants as being significantly correlated
117 with some biological trait. Using the NHGRI-EBI Catalog of published genome-wide association
118 studies we queried the rsIDs of the SNPs we identified as being low quality and found 6 recent
119 publications that had found at least one of the variants to have reached genome wide significance
120 in their study.

121 Five of these studies used the 1000 Genomes Project as the reference database for imputation
122 and one used the 1000 Genomes Project cell cultures and sequence data. They used strict quality
123 thresholds, including population genetic statistical tests such as the Hardy-Weinberg equilibrium
124 test, allele frequency differences using reference populations. They also removed rare alleles and
125 alleles with high degrees of missingness. Despite using the state of the art quality controls, these
126 erroneous variants managed not only to be imputed onto real genotype data, but they also reached
127 genome wide significance for biological traits.

128 Discussion

129 Why do we care?

130 These SNPs matter because they reached genome wide significance for medically relevant traits.
131 Could lead to false diagnosis at worst, or spurious correlations at least. Polygenic risk scores
132 take into account all snps reaching significance, without much manual curation. It's likely that the
133 variants we have identified are being included in these multi-locus risk scores.



Figure 2. A The average mapping quality of each individual per population included in the 1000 Genomes sequencing project. The x-axis is ranked by populations with the least to the most variance, followed by average mapping quality per individual. **B** Same data as in **A** except the x-axis is sorted by sequencing date. The colors indicate the sequencing centers that produced the data for each individual.

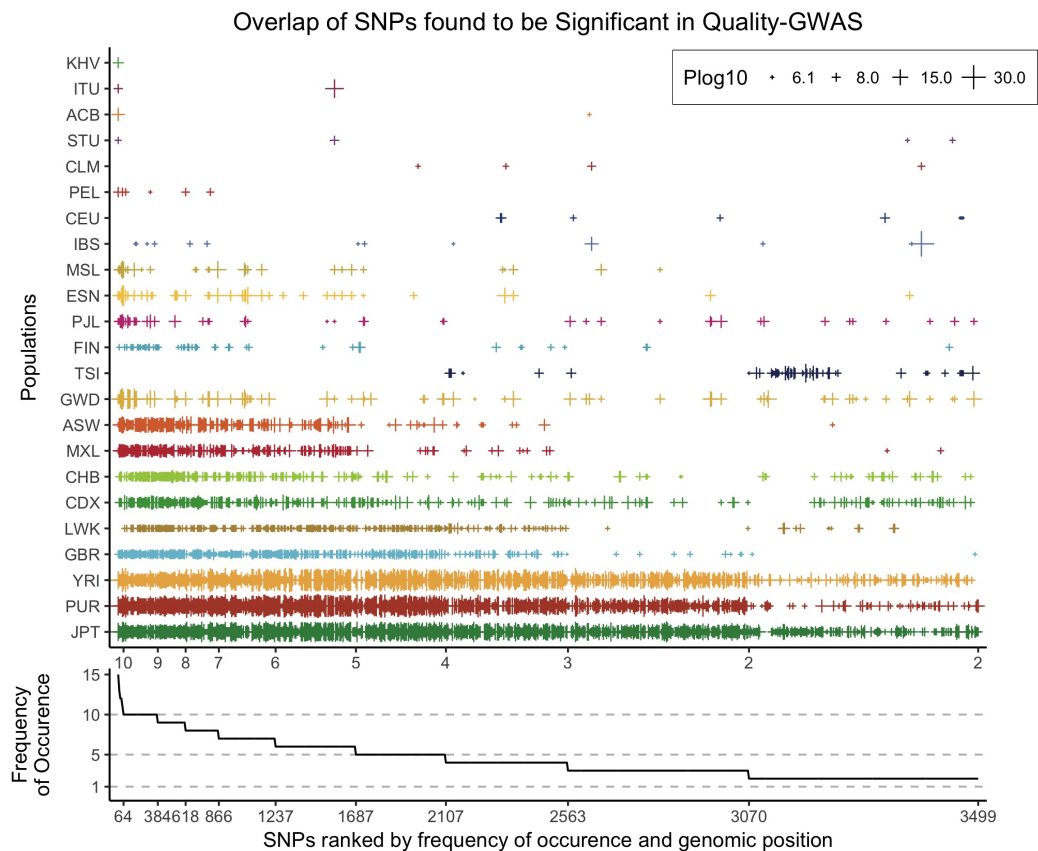


Figure 3. Overlap of SNPs identified independently to be associated with quality. The populations that have the most low-quality individuals also have the most low-quality variants. What is interesting here, is that the same variants identified as being low quality independently in each population are found in other populations.

134 Despite these variants reaching genome wide significance, the majority of researchers did not
135 pursue these variants for further analysis. This is likely because these variants have no peak in LD
136 that is characteristic of a biologically significant variant.

137 Recommendations

138 The most conservative approach would be to remove all individuals that don't meet the quality
139 threshold as well as all the variants associated to low quality. In this case, we used a cut off of an
140 average quality of mapped bases over 30. This threshold has been **previously used by studies**. It
141 is the minimum requirements for **GATK variant calling** for them to have a minimum quality of 30.

142 Imputation

143 Imputation of genotype data is a probabilistic method that infers the bases of a given genome
144 based on its similarity to a set of reference genomes. While on average, two humans differ in about
145 1/10,000 bases, this number is **lower** in closely related individuals, and **higher** in individuals
146 from different continental origins. Modern chip sequencing will provide the genotype information
147 for over 1 million bases of the genome. The unique combination of genotyped bases can be enough
148 to identify haplotype blocks that are identical by descent in individuals from the reference database.

149 The accuracy of imputation depends on the size of the reference database; this varies signifi-
150 cantly from one population to another. This means that individuals with ancestry that is less well
151 represented in the reference database will have lower accuracy of imputation. To overcome this
152 bias, reference databases are often combined to increase the sample size and in turn, the accuracy
153 of imputation.

154 GWA studies from other papers

155 Since these variants are present in more than one population from the 1000 genomes project, they
156 are more likely to be associated to biological traits as they would appear to be like any other variant
157 that is shared among multiple populations. The only way to distinguish some of these more covert
158 false positives is to use statistical tests associating the quality metrics of each position relative to
159 each individual.

160 Luke: due to the temporal nature of the batch effects, and because entire populations were
161 sequenced in one centre on one day, the false positives are more likely to cluster with population
162 structure or case/control?

163 Conclusion

164 Our method identifies spurious mutations by correlating mutations with data quality metrics. We
165 propose including our quality control methods to identify possible false positives in sequencing
166 data. We have focused on the 1000 Genomes Project dataset as its quality metrics were feely
167 available, however the issues of quality control are not limited to this one consortium. This study
168 only used one dataset and one quality metric, but using this same approach can be used to identify
169 more false positives in many more datasets.

170 As more and more large scale genotyping efforts are being imputed on the same legacy datasets,
171 we must scrutinize the quality of the reference databases to avoid the amplification of false positives.
172 These results bring forth many questions regarding the reliability of legacy datasets. Moreover,
173 since there are so many broad applications of imputation, it frames the question for reference data
174 turnover.

175 Methods

176 Metadata

177 The metadata used in this analysis was compiled from each of the index files from the 1000
178 Genomes file system. Average quality of mapped bases per sample was obtained from the BAS

179 files associated with each alignment file. Each BAS file has metadata regarding each sequencing
180 event for each sample. If a sample was sequenced more than once, we took the average of the
181 each quality score from each sequencing instance. The submission dates and sequencing centres
182 for each sample in the analysis was available in the sequence index files. This file also has multiple
183 entries per sample, however, we were unable to match the individual sequencing runs between the
184 bas files and the index file, which lead us to take the average of the quality scores and only kept
185 the earliest sequencing date per sample. The dates of the sequencing are only used to plot Figure.
186 **Average Quality of mapped bases: How was it calculated?**

187 **Data Availability**

188 Index of BAS files [available here](#).
189 Phase3 analysis sequence index file [available here](#)
190 **link to my compiled metadata file here**

191 **Quality Controls**

192 We reproduced the quality control pipelines used by Harris et. al as they applied the current
193 state of the art quality thresholds to remove questionable sequences especially for the high
194 standards for detecting population level differences. Several mask files were applied to remove
195 regions of the genome that might be lower quality, or might have very different mutation rates
196 or basepair complexity compared to the rest of the genome. The 1000 Genomes [strict mask](#)
197 was used to remove low quality regions of the genome , highly conserved regions were removed
198 using the [phastCons100way](#) mask file and highly repetitive regions were also removed using the
199 [NestedRepeats](#) mask file from RepeatMasker. Furthermore, only diallelic autosomal SNPs were
200 considered, with missingness below 0.01, MAF less than 0.1, and MAF greater than 0.9.

201 **Genome Wide Association**

202 Using PLINK v1.90b4.4 we ran a linear genome wide association study independently for each
203 population of the 1000 Genomes Project. We used the average quality of mapped bases per
204 individual as the phenotype for the analysis. We also controlled for population structure by
205 including the first 4 principle components of a PCA of each population using genotype data. **NOT
206 DONE YET**

207 **Mutation Spectrum**

208 We calculated the mutation spectrum for each list of significant SNPs for each population.
209 We also compared the mutation spectrum ratio between populations using a modified version
210 of the methods used in Harris et al. 2017.

211 **Imputation**

212 Using the Michigan Imputation Server, we imputed the genotype data from 1000 Genomes Project

213 **Code Availability**

214 **here is where I will put my git hub**

215 **Acknowledgments**

216 We would like to thank Kelly Harris for sharing her mutation spectrum pipelines.