LEGACY DATA CONFOUNDS MODERN STUDIES

LUKE ANDERSON-TROCMÉ , SIMON GRAVEL

Contents

1. Introduction 2
1.1. Context 2
2. Results 3
2.1. Average quality of mapped snps Quality Over Time 3
2.2. Genome Wide Association 3
2.3. Overlap of Significant SNPs 3
2.4. Significant SNPs 3
2.5. Higher quality datasets don't have these SNPs 5
2.6. Imputation 5
2.7. Found to be included in other GWAS 5
3. Discussion 7
3.1. Why do we care? 7
3.2. GWAS from other papers 7
3.3. Why were these SNPs even associated in the first place? 7
4. Conclusion 7
5. Methods 8
5.1. Metadata 8
5.2. Data Availability 8
5.3. Quality Controls 8
5.4. Genome Wide Association 8
5.5. Mutation Spectrum 9
5.6. Imputation 9

1. Introduction

The last 5 years has seen an increase in number of individuals genotyped through private companies as well as medical research cohorts. Genotype data is often imputed to increase power for association studies in a cost effective way. This process is done using a reference database comprised of whole genome sequence data from other individuals in reference populations. We often combine these reference databases to increase the sample size and in turn, the accuracy of imputation. 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 have been known to contain a higher rate of false positives than their more recent counterparts. Is it time to retire legacy datasets?

LAT: here are some words In investigating a mutational signature identified in a Japanese cohort by Harris et. al (2017), we noticed that the individuals responsible for this signal all had lower average quality scores compared to the other Japanese individuals. To further investigate the extent to which low quality individuals were associated to spurious mutations, we then performed a genome-wide association study (GWAS). We discovered TODO: 000 variants that were associated to low mapping quality in TODO: 000 individuals from TODO: 000 of the 26 populations included in the 1000 genomes dataset. One third of these mutations were included in imputation tests. Population specific mutational signatures were also clarified by removing these spurious mutations.

1.1. Context.

This project was motivated by an unusual population genetics observation in a recent publication by Harris et. al, that the 1000 Genomes Project Japanese population seem to be partitioned in two clusters of diff mutation rates. This heterogeneously distributed mutational signal was unexpected as a signal of this nature could either be due to population structure, a recent mutagen, or technical bias.

Our research center had recently obtained access to a large cohort of several thousand high-quality whole genome sequences from Japanese individuals from the Fumi institute. We started off by trying to replicate the results found in Harris et al. 2017, but after several attempts and double checking we had the right quality controls, we were unable to replicate their results with the Fumi dataset. Upon comparing the allele frequencies between the two datasets, we observed an unusually large number of private SNPs despite both cohorts being from the same population. Further investigation in the allele frequency spectra of the two Japanese datasets, indicated that the SNPs responsible for the signal were missing in the more recent dataset. This result raised several red flags. Why would a supposedly genome wide signal be concentrated in private SNPs from a legacy dataset?

Upon investigating the quality control pipelines of the more recent dataset, many of these positions uniquely found in the 1000 Genomes Project had been removed in the standard QC of the sequencing centre. Does this legacy dataset meet the current industry standards of quality? This study is the result of an investigation in the quality of the 1000 Genomes Project dataset.

After compiling all the metadata from each 1000 genomes sequencing run, we noticed a correlation between PC1 of the mutation spectrum signal and the average quality of mapped bases per individual. This suggested that some individuals with lower quality were driving this apparent population genetics signal. [mention fake pop structure] Furthermore, when looking at the days when the individuals were sequenced, we noticed that individuals with this mutational signal were sequenced on the same day in the same sequencing centre. 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.

2. Results

2.1. Average quality of mapped snps Quality Over Time.

As one would expect, the sequencing quality of individuals increases over time 1. Sequencing centers improve on their laboratory techniques and sequencing technologies improve from one generation to the next. The magnitude and impact of the differences in quality within and between populations has not yet been explored.

2.2. Genome Wide Association.

GWAS are commonly used to find genomic regions associated with biologically relevant traits. In this case, we are using this analysis in a more unconventional way to identify regions of the genome that are associated to biologically irrelevant traits like individuals with low average quality of mapped bases.

Using the linear GWAS function offered by PLINK, we were able to identify 000 SNPs that were associated to low quality with p-values below 10^{-8} and 000 snps with p-values below 10^{-6} 2. The results from of the GWAS for other populations are available in supplementary data. A total of 0000 SNPs were associated to low quality in all populations combined.

2.3. Overlap of Significant SNPs.

Comparing the results of each independent GWAS, we were able to identify over 0000 variants that were independently associated to low quality in multiple populations. This confirmation using more than one GWAS is a very strong result suggesting that these variants might not be legitimate 3.

2.4. Significant SNPs.

When we remove individuals below a quality of 30, population structure is clarified. The signal identified in Harris et. al is enriched in sites associated with low quality. [Need to run harris pipeline without 10^{-6}] Therefore, when these low quality sites are removed, the mutational signal is also

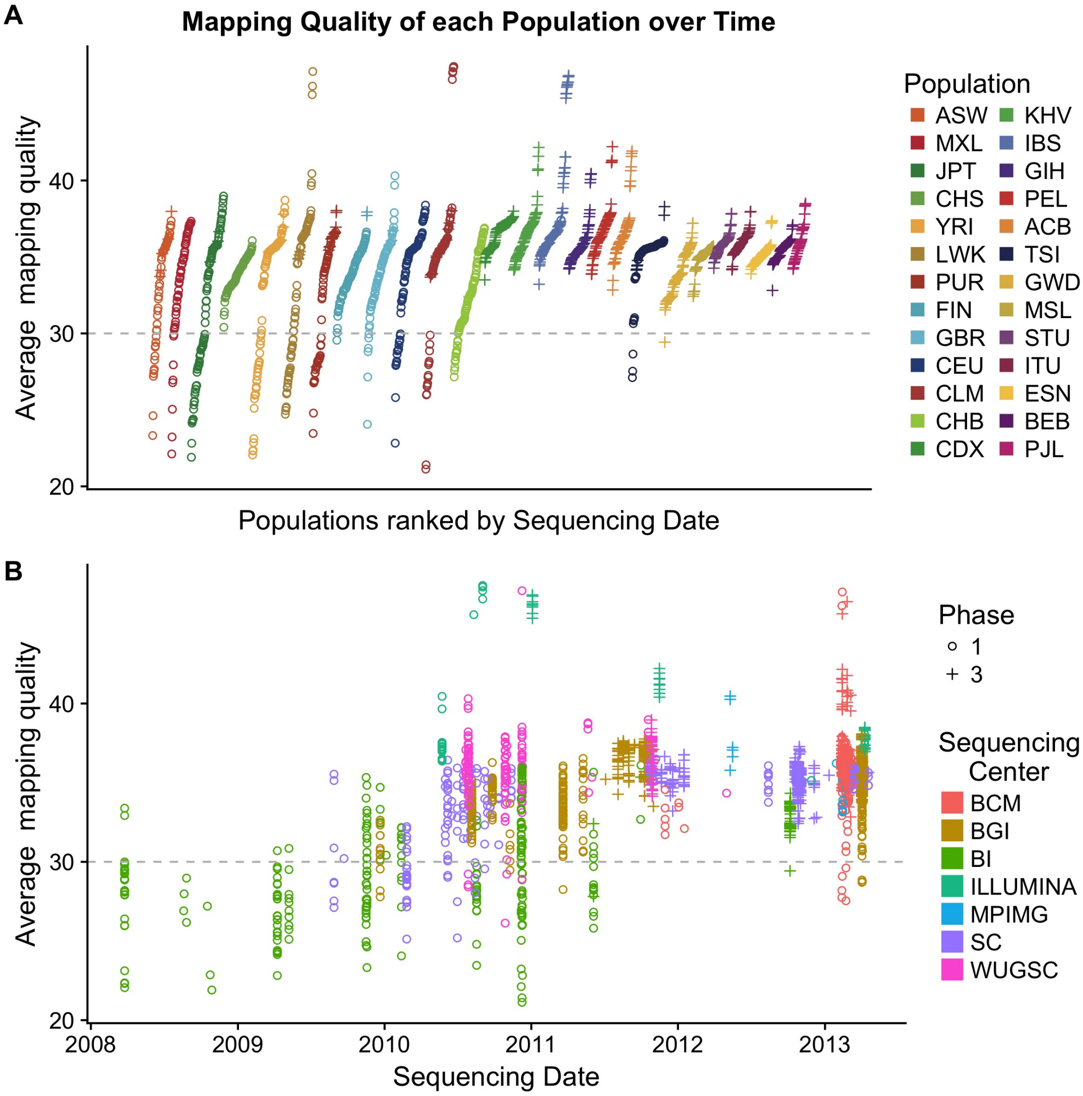

Figure 1. A) The average mapping quality of each individual per population included in the 1000 Genomes sequencing project. The average mapping quality per individual was obtained from the raw alignment index data for each sample. The X axis is ranked by populations with the lease to the most variance, followed by average mapping quality per individual. B) Same data as in A) except the X axis is now sorted by sequencing date, and the colors indicate the sequencing centers that produced the data for each individual.

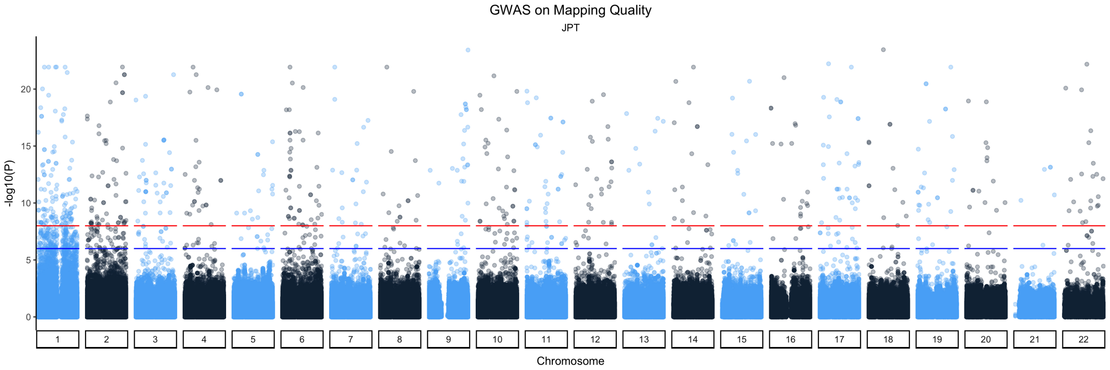

Figure 2. Genome wide association of the average quality of mapped bases for the Japanese individuals included in the 1000 Genomes Project. This GWAS included 104 individuals, 000 of them were sequenced in phase 1 while the rest were sequenced in phase 3 of the 1000 Genomes Project. The same analysis was performed independently for each of the populations in the 1000 Genomes Project.

removed. (Or in this case when we remove the individuals with qualities below 30.)

2.5. **Higher quality datasets don't have these SNPs.** Interestingly, the SNPs we identified had already been removed in higher quality datasets. These HQ datasets also have filtered more SNPs than the GWAS was able to identify. When we compare the allele frequencies between the 1000 Genomes project Japanese population to a more contemporary dataset, there are many more sites that are enriched with the same mutational signature as described above that are not identified using a GWAS. This suggests that we lack the power to identify all the questionable variants from these legacy datasets.

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

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

Five of these studies used the 1000 Genomes Project as the reference database for imputation and one used the 1000 Genomes Project cell cultures and sequence

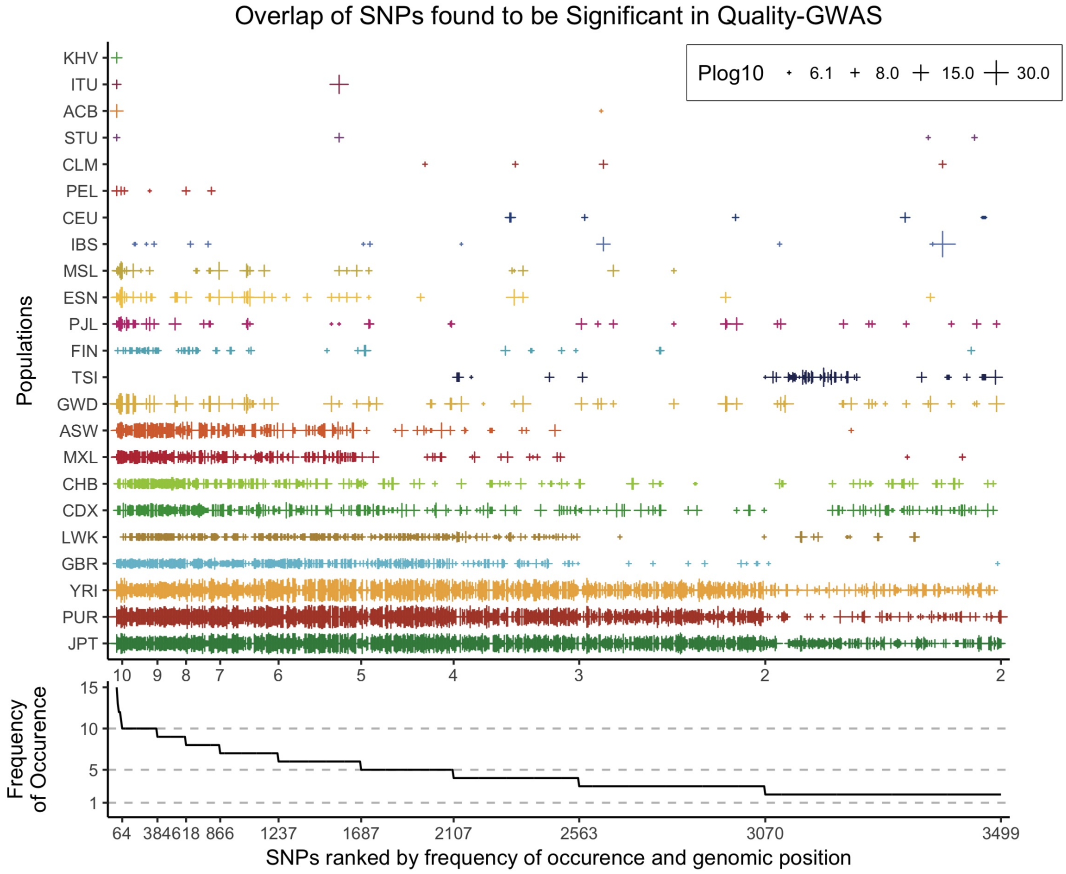

Figure 3. Overlap of SNPs identified independently to be associated with quality. Not surprisingly, the populations that have the most low-quality individuals also have the most low-quality variants. What is interesting here, is that the same variants identified as being low quality independently in each population are found in other populations.

data. They used strict quality thresholds, including population genetic statistical tests such as the Hardy-Weinberg equilibrium test, allele frequency differences using reference populations. They also removed rare alleles and alleles with high degrees of missingness. Despite using the state of the art quality controls, these erroneous variants managed not only to be imputed onto real genotype data, but they also to reach genome wide significance for the traits they were testing for.

3. Discussion

3.1. **Why do we care?** These SNPs matter because they reached genome wide significance for medically relevant traits. Could lead to false diagnosis at worst, or spurious correlations at least. They were not caught sooner because the batch effect of 1000 genomes had never been properly addressed.

Our GWAS method was able to identify 000 variants that are likely to be false positives.

We compared the number of SNPs identified in the GWAS to the number of sites that differed between our higher quality Japanese dataset and found that 100?% of the sites were also missing from the Fumi data. Because we are running the GWAS on a per-population base, for rarer mutations, we are lacking the power to identify these significant SNPs. However, upon comparing the Japanese 1000 Genomes cohort to a higher quality and larger cohort, we identified 000 more variants that are beyond the expected frequency spectrum deviation for individuals from the same population. This suggests that while the GWAS approach can identify some of the low hanging bad apples, there are likely more of these false positives nested inside these legacy cohorts. The most conservative approach would be to remove all individuals that don't meet the quality threshold. In this case, we used a cut off of an average quality of mapped bases over 30. This threshold has been previously used by 000 studies.

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

3.3. **Why were these SNPs even associated in the first place?** [due to the temporal nature of the batch effects, and because entire populations were sequenced in one centre on one day, the false positives are more likely to cluster with population structure or case/control?]

4. Conclusion

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

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

5. Methods

5.1. **Metadata.** The metadata used in this analysis was compiled from each of the index files from the 1000 Genomes file system. Average quality of mapped bases per sample was obtained from the BAS files associated with each alignment file. Each BAS file has metadata regarding each sequencing event for each sample. If a sample was sequenced more than once, we took the average of the each quality score from each sequencing instance. The date of completion of each sequencing occurrence was not available, however we managed to find the submission dates and sequencing centres for each sample in the phase3 analysis sequence index file. This file also has multiple entries per sample, however, we were unable to match the individual sequencing runs between the bas files and the index file, which lead us to take the average of the quality scores and only kept the earliest sequencing date per sample. The dates of the sequencing are only used to plot Figure. [Average Quality of mapped bases: How was it calculated?]

5.2. **Data Availability.** Index of BAS files available here. Phase3 analysis sequence index file available here [link to my compiled metadata file here]

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

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

5.5. **Mutation Spectrum.** We calculated the mutation spectrum for each list of significant SNPs for each population [NOT DONE YET]

We also compared the mutation spectrum ratio between populations using a modified version of the methods used in Harris et al. 2017.

5.6. **Imputation.** Using the Michigan Imputation Server, we imputed the genotype data from 1000 Genomes Project