Legacy Data Confounds Modern Genomics Studies

Luke Anderson-Trocmé[1,2], **Mathieu Bourgey**[1,2], **Simon Gravel**[1,2]

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

[1] Department of Human Genetics, McGill University, Montreal, QC H3A 0G1, Canada;
[2] McGill 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 using dated sequencing technologies have been known to contain a higher rate of false positives than their more recent counterparts. Is it time to retire legacy data?

Mutation Spectrum

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

Motivations

This project was motivated by this unusual population genetics observation in a recent publication by Harris et. al, that the 1000 Genomes Project (1kGP) 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 mutagen, or a technical bias. In investigating this mutational signature, we were unable to reproduce the results using a larger and higher quality dataset and concluded that this signal can be attributed to sequencing error.

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

Results

A peculiar mutational signature in Japan

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

To further investigate the extent to which low quality individuals were associated to spurious mutations, we performed a genome-wide association (GWA) study 1. GWA studies are commonly used to find genomic regions associated to 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 SNPs associated to individuals with low quality.

Using the linear GWA study function offered by PLINK, we were able to identify $625 \;\; p < 10^{-8}$ and $1048 \;\; p < 10^{-6}$ SNPs that were associated to the average quality of SNPs mapped for an individual 1. This GWA study 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 variants that are associated to the quality of mapped bases have an enrichment in *AC→*CC mutations.

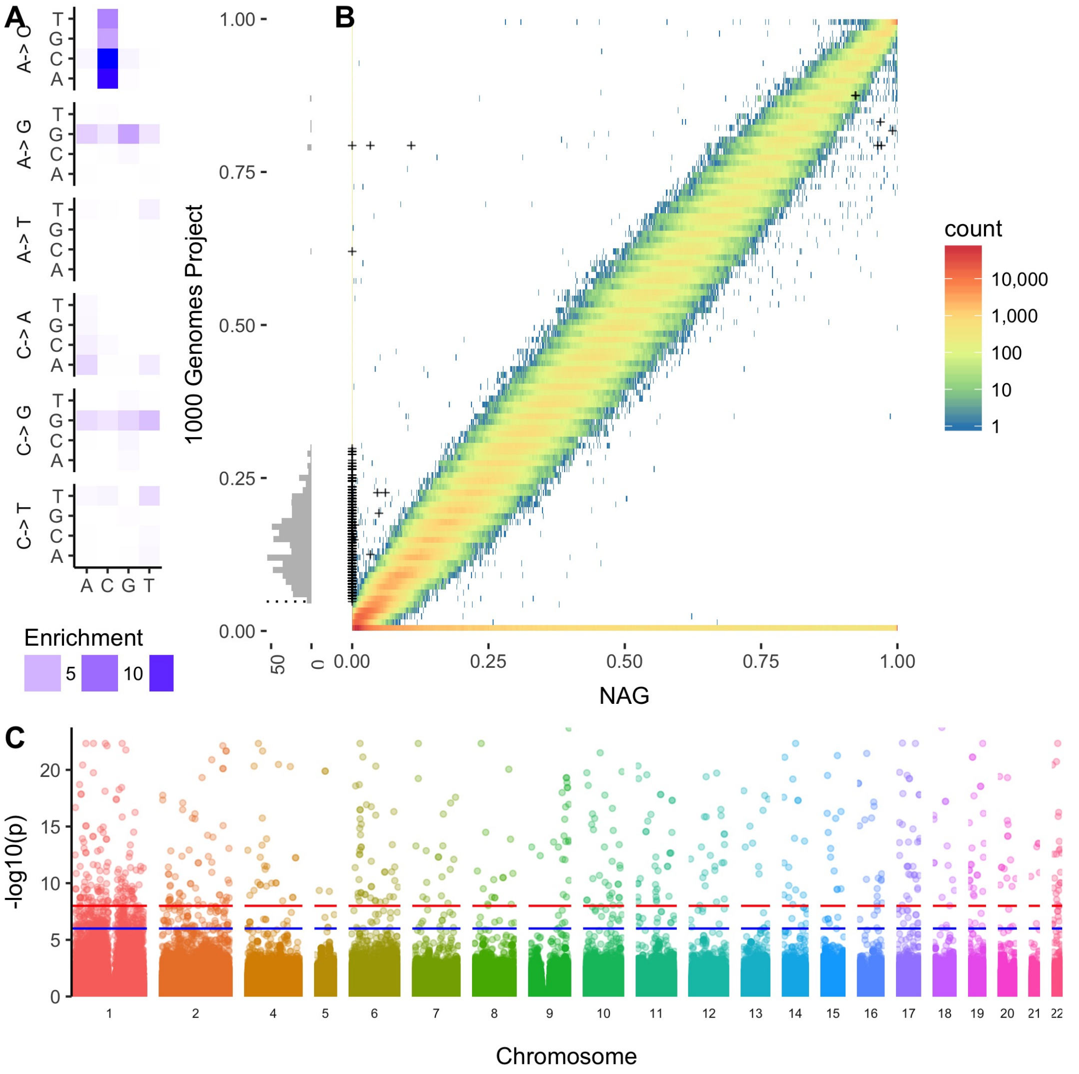

Figure 1. A Joint frequency spectrum plot of the Japanese from the 1000 Genomes Project and a more recent dataset. Crosses (+) are variants that reached genome wide significance in a GWA of sequencing quality. The histogram on the left of the plot is the distribution of significant variants. The lowest 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 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.

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

One way to assess the validity of these unusual variants is to see if they are present in a higher quality dataset. 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.

Sequencing quality over time

We turned our attention to the other populations in the 1kGP, we found that the sequencing done in phase 1 was more variable and overall tended to include lower quality sequencing data 2. The sequencing quality of individuals increases over time 2. By 2011 the sequencing quality seems to level off, this also coincides with the phasing out of older sequencing technologies.

Overlap of significant SNPs

Comparing the results of each independent GWA study, 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 strongly suggests that these variants might not be genuine 3.

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.

Found to be included in other GWA studies

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 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 reached genome wide significance for biological traits.

Discussion

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. Polygenic risk scores take into account all snps reaching significance, without much manual curation. It's likely that the variants we have identified are being included in these multi-locus risk scores.

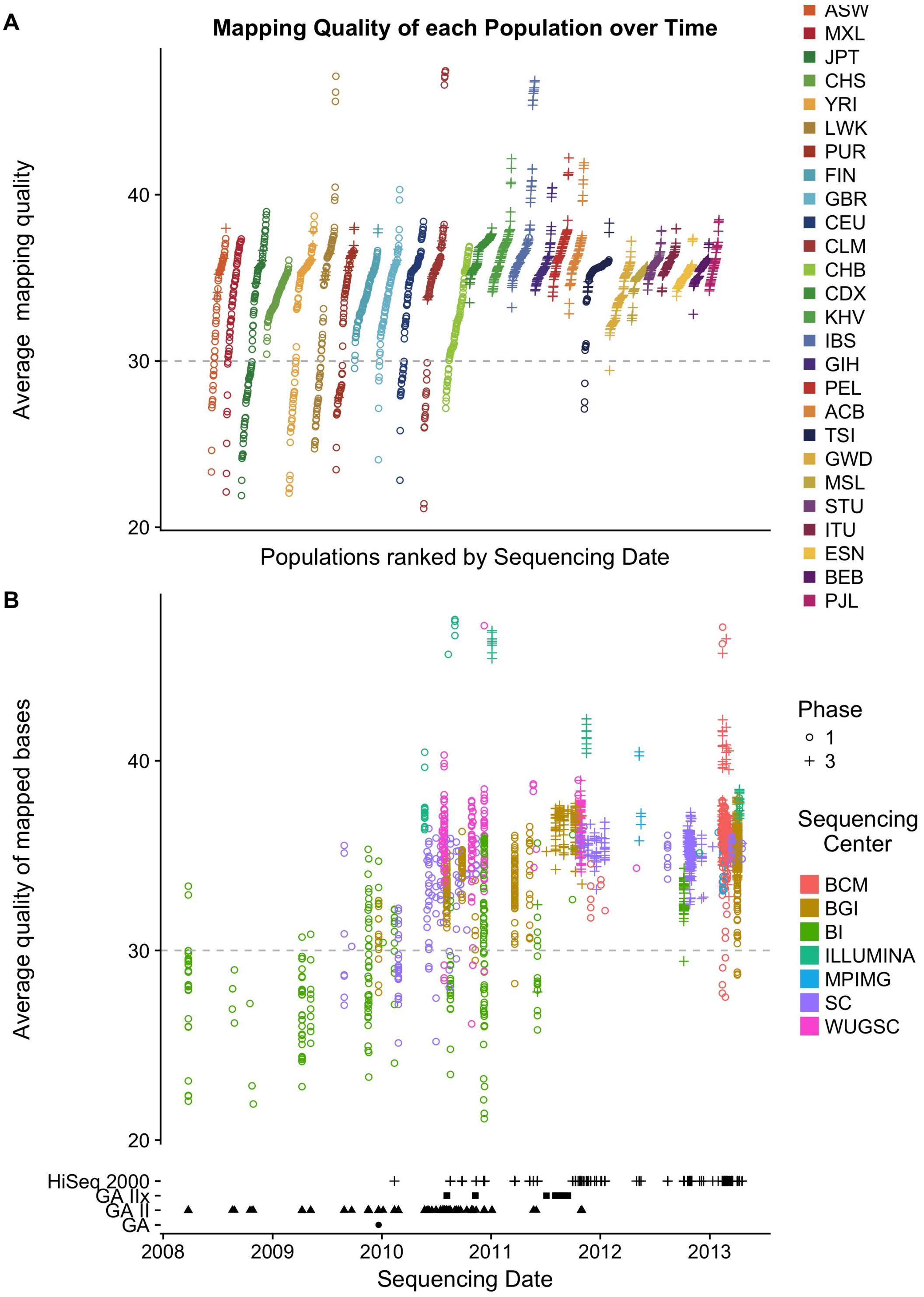

Figure 2. **A** The average mapping quality of each individual per population included in the 1000 Genomes sequencing project. The x-axis is ranked by populations with the lease to the most variance, followed by average mapping quality per individual. **B** Same data as in **A** except the x-axis is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

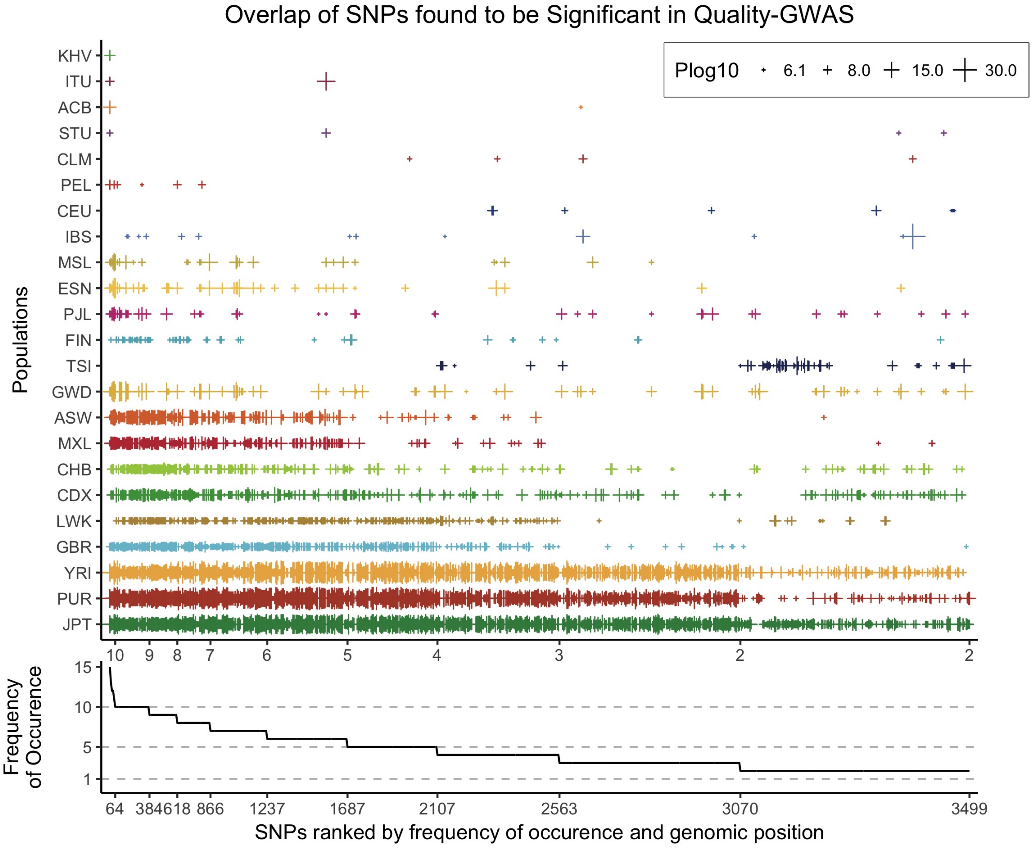

Figure 3. Overlap of SNPs identified independently to be associated with quality. The populations that have the most low-quality individuals also have the most low-quality variants. What is interesting here, is that the same variants identified as being low quality independently in each population are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

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

Recommendations

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

Imputation

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

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

GWA studies 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.

Luke: 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?

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.

Methods

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 submission dates and sequencing centres for each sample in the analysis was available in the sequence index files. 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?*

Data Availability

Index of BAS files available here.

Phase3 analysis sequence index file available here

link to my compiled metadata file here

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.

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*

Mutation Spectrum

We calculated the mutation spectrum for each list of significant SNPs for each population.

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

Imputation

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

Code Availability

here is where I will put my git hub

Acknowledgments

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