# A handbook for Computational Genetics

Alfred Pozarickij

2020-07-09

# Contents

# Preface

The scope of this book is to provide an outline of computational methods available for the analysis of genetic data.

First chapter introduces methods to infer population parameters.

Next X chapters focus on population genetics.

In this book, the amount of mathematics and statistics is kept to a minimum. Only methods designed specifically to address issues in genetics are shown. I have produced another book [insert link here], which is intended to familiarise the reader with commonly used approaches and develop some intuition behind them. By no means the list is comprehensive and only serves as a quick guide. The internet provides much more information regarding this topic.

Rather than providing references at the end of each chapter, I decided to combine them into supplementary text [insert link here]. References are arranged according to different topics discussed in this book. I tried to do my best to cause as little confusion as possible.

Finally, don't hesitate to contact me if I have not included your favorite method (apozarickij@gmail.com). I would be more than happy to hear about it.

# Part I

# Quantitative Genetics

# Chapter 1

# Population parameters

## 1.1 Mean

## 1.2 Variance

## 1.3 Covariance

## 1.4 Genetic correlation

## 1.5 Additivity

## 1.6 Dominance/Recesivness

## 1.7 Codominance

## 1.8 Infinitesimal model

### 1.8.1 Omnigenic model

## 1.9 Henetic relationships by Malecot

## 1.10 Genotype simulations

## 1.11 Phenotype simulations

# Chapter 2

# Sequencing technologies

The first step in any genetic analysis is to map sequence reads, callibrate base qualities, and call variants.

Prior to mapping, evaluate base composition along reads. Calculate the proportion of A, C, G, T bases along each read. Flag runs with evidence of unusual patterns of base composition compared to the target genome. Evaluate machine quality scores along reads. Calculate average quality scores per position. Flag runs with evidence of unusual quality score distributions. Calculate the input number of reads and number of bases for each sequenced sample.

## 2.1 While whole genome sequencing and re-sequencing represent ~90% of all DNA based sequencing applications, it's important to not lose sight of the myriad of new protocols available to count or detect epi-genomic features. These include genotyping, measuring DNA-protein interactions and epigenetic markers. Several examples of these protocols are listed below:

### 2.1.1 DNA-protein interactions

DNAse-Seq MNAse-Seq X-ChIP ChIP-Seq FAIRE-Seq ATAC-Seq Chia-PET Hi-C 3-C, 4-C, 5-C Capture-C HiTS-FLIP

### 2.1.2 Epigenetics

Bisulfite-Seq Methyl-Seq RRBS PBAT Me-DIP oxBS-Seq TAB-Seq MBDCap-Seq BisChIP-Seq

### 2.1.3 Genotyping

RAD-Seq ddRAD-Seq nextRAD Capture-Seq ezRAD Low input DNA-Seq MDA DOP-PCR Os-Seq MALBAC Nuc-Seq

## 2.2 Genotype calling algorithms

## 2.3 Sequence alignment

Sequence alignment is a method of arranging sequences of DNA, RNA, or protein to identify regions of similarity. The similarity being identified, may be a result of functional, structural, or evolutionary relationships between the sequences.

If we compare two sequences, it is known as pairwise sequence alignment. If we compare more than two sequences, it is known as multiple sequence alignment.

## 2.4 Sequence assembly

## 2.5 SNP annotation

### 2.5.1 CNV annotation

## 2.6 Gene prediction

# Chapter 3

# Genome-wide association analysis

## 3.1 DNA processing quality control

## 3.2 Batch effects

https://www.bioconductor.org/packages/devel/bioc/vignettes/GWASTools/inst/doc/DataCleaning.pdf

The overall goal of this step is to check the quality of the sample batches. Substantial quality control is done by the genotyping centers prior to releasing the genotype data. However, it is possible that quality control for batches is still lower than desired. If a lower quality batch is detected then it may be necessary to re-run the genotyping for that batch. We can check the batch quality by comparing the missing call rates between batches and looking for significant allele frequency differences between batches.

### 3.2.1 Missing call rate for samples and SNPs

The first step is to calculate the missing call rates for each SNP and for each sample. A high missing call rate for a sample is often indicative of a poorly performing sample. It has been seen that samples from DNA that has undergone whole-genome amplification (WGA) have a relatively higher missing call rate. Similarly a high missing call rate for a SNP is indicative of a problem SNP. Experience from the GENEVA studies has shown that there seem to be a subset of SNPs from which genotype calls are more difficult to make than others. We calculate the missing call rates in a two step process: first the missing call

rates over all samples and SNPs are calculated, then the missing call rates are calculated again, filtering out SNPs and samples that have an initial missing call rate greater than 0.05. The initial SNP missing call rate over all samples is saved in the SNP annotation data file as missing.n1. The analogous idea is applied to the samples: missing.e1 is saved in the sample annotation file and corresponds to the missing call rate per sample over all SNPs, excluding those SNPs with all calls missing. The missing.n2 is calculated as the call rate per SNP over all samples whose missing.e1 is less than 0.05. Again, similarly for the samples, missing.e2 is calculated for each sample over all SNPs with missing.n2 values less than 0.05. It is important to remember that the Y chromosome values should be calculated for males only, since we expect females to have no genotype values for the Y chromosome, although an occasional probe on the Y chromosome is called in a female. If any samples have a high missing rate, we recommend further investigation of what may be causing the missing calls; the samples with a missing call rate greater than 0.05 should be filtered out due to low sample quality.

### 3.2.2 Missing call rates by batch

The missing call rate by batch is calculated to check that there are no batches with comparatively lower call rates. Usually a“batch”is a plate containing samples that were processed together through the genotyping chemistry. In this case all samples were run on different plates (as controls for another dataset).

### 3.2.3 Allele frequency differences across batches

In this step, the chi-square test for differences in allelic frequency is performed between each batch individually and a pool of all the other batches in the study. We then look at the mean  2 statistic over all SNPs for each batch as a function of the ethnic composition of samples in a batch. Next we test for association between batches and population groups, using a  2 contingency test. Then we look at the relationship between the ethnic composition of each batch and the previously calculated  2 test of allelic frequency between each batch and a pool of the other batches. The point is to look for batches that differ from others of similar ethnic composition, which might indicate a batch effect due to genotyping artifact. In this experiment, there are only a few batches and wide variations in race among batches, so it is difficult to interpret the results. In larger GWAS experiments, we generally observe a U-shaped curve of allelic frequency test statistic as a function of ethnic composition. The  2 test is not suitable when the 2×2 tables for each SNP have very small values. For arrays in which many SNPs have very low minor allele frequency, Fisher's exact test is more appropriate.

## 3.3 Sample quality control

### 3.3.1 Cryptic relatedness

### 3.3.2 Population stratification

Sometimes finding an association can be confounded by population stratification. This is because a condition may be more prevalent in one group of people than in a different group, resulting in a spurious association between the condition or trait being tested for and any genetic characteristics which vary between the two different groups of people.

While it is good practice for studies to be based on as homogeneous a group of test subjects as possible, it has been noted in [Price, 2006] that even the mild variation in genetic characteristics among those who classify themselves as belonging to one ethnic group or another can be problematic enough to confound a study done over thousands of genetic markers.

Hidden population stratification may be thought of as a non-zero Fst between unknown groupings of samples.

### 3.3.3 Heterozygosity and missingness outliers

### 3.3.4 Differential missingness

### 3.3.5 Sex chromosome anomalies

## 3.4 Marker quality control

### 3.4.1 Genotyping concordance

In genotyping studies where DNA is directly assayed for positions of variance, concordance is a measure of the percentage of SNPs that are measured as identical. Samples from the same individual or identical twins theoretically have a concordance of 100%, but due to assaying errors and somatic mutations, they are usually found in the range of 99% to 99.95%. Concordance can therefore be used as a method of assessing the accuracy of a genotyping assay platform.

### 3.4.2 Mendelian errors

### 3.4.3 Genotype call rate

### 3.4.4 Minor allele frequency

### 3.4.5 Hardy-Weinberg equilibrium outliers

### 3.4.6 Additional QC for regions like MHC

### 3.4.7 Ambigious nucleotides

If the base and target data were generated using different genotyping chips and the chromosome strand (+/-) for either is unknown, then it is not possible to match ambiguous SNPs (i.e. those with complementary alleles, either C/G or A/T) across the data sets, because it will be unknown whether the base and target data are referring to the same allele or not. While allele frequencies can be used to infer which alleles match, it is recommended to remove all ambiguous SNPs since the allele frequencies provided in base GWAS are often those from resources such as the 1000G project, and so aligning alleles according to their frequency could lead to systematic biases.

### 3.4.8 Non-matching nucleotides

When there is a non-ambiguous mismatch in allele coding between the data sets, such as A/C in the base and G/T in the target data, then this can be resolved by 'flipping' the alleles in the target data to their complementary alleles.

### 3.4.9 Quality control prior to meta-analysis

Allele Frequency Plots (AF Plots): looking for errors in allele frequencies and strand orientations by visually inspecting a plot of the sample allele frequency of filtered SNPs against the frequency in the 1000 Genomes phase 1 version 3 European panel3 for example. P value vs Z-statistic Plots (PZ Plots): looking for the consistency between the reported P values and the P values implied by the coefficient estimates and standard errors in individual cohort. Quantile-Quantile Plots (QQ Plots): looking for the cohort-level QQ plots to look for evidence of unaccounted-for stratification. Predicted vs Reported Standard-Error Plots (PRS Plots): making sure that the standard errors reported in individuals cohorts are approximately consistent with the reported sample size, allele frequency, and phenotype distribution. Use of bivariate LD score regression to verify that the estimated genetic correlations between all large cohorts (defined as $N > 10{,}000$) are large and positive.