Notes on Bioinformatics for Population Genetics

Gus Dunn

2014-12-30

Contents

1 Convert VCF files to PLINK format

- http://vcftools.sourceforge.net/documentation.html#plink

From the link above:

> VCFtools can convert VCF files into formats convenient for use in other programs. One such example is the ability to convert into PLINK format. The following function will output the variants in .ped and .map files.

```
vcftools --vcf input_data.vcf --plink --chr 1 --out output_in_plink
```

2 Imputation

- Nature Reviews Genetics 11, 499-511 (July 2010): Box 1 | How genotype imputation works

3 VCF phased vs non-phased

`tags = [VCF, phased, ]`

3.1 Web snippets

- as far as I know, the main reason to use allele phasing information is to increase the correctness of the haplotypes and haplotype blocks inferred from them [source].
- Phased data are ordered along one chromosome and so from these data you know the haplotype. Unphased data are simply the genotypes without regard to which one of the pair of chromosomes holds that allele. [source]
- The ability to distinguish which alleles belong to which chromosome is important when considering how genes are inherited. Generally, a parent passes one of the two copies of each chromosome on to their offspring. While the two chromosomes might both contribute genetic information via a process called recombination, the genes received by a child are typically "linked" and inherited together since they are located on the same chromosome.

  To determine which genes of yours are linked together (and therefore likely to be inherited together by your child), it is first necessary to figure out which alleles (indicated by the variant SNPs) exist together on the same chromosome. This process has been termed "phasing" in the bioinformatics world. [source]
- http://blogs.discovermagazine.com/gnxp/2007/01/basic-concepts-linkage-disequilibrium/#.VKK7BAMAQ

4 Glossary

imputation $\mapsto$ in genetics, imputation refers to the statistical inference of unobserved genotypes. *It is achieved by using known haplotypes in a population*, for instance from the HapMap or the 1000 Genomes Project in humans, thereby allowing to test initially-untyped genetic variants for association with a trait of interest. Genotype imputation hence helps tremendously in narrowing-down the location of probably causal variants in genome-wide association studies. Wikipedia

haplotype $\mapsto$ a collection of specific alleles (particular DNA sequences) in a cluster of tightly-linked genes on a chromosome that are likely to be inherited together. *Put in simple words, haplotype is the group of genes that a progeny inherits from one parent.* Wikipedia

$\mapsto$ A second meaning of the term is a set of single-nucleotide polymorphisms (SNPs) on a single chromatid of a chromosome pair that are associated statistically. It is thought that these associations, and the identification of a few

alleles of a haplotype sequence, can unambiguously identify all other polymorphic sites in its region. Wikipedia

↦ haplotype is a contraction for haploid genotypes. Wikipedia

linkage disequilibrium ↦ the occurrence of some combinations of alleles or genetic markers in a population more often or less often than would be expected from a random formation of haplotypes from alleles based on their frequencies. It is a second order phenomenon derived from linkage, which is the presence of two or more loci on a chromosome with limited recombination between them. The amount of linkage disequilibrium depends on the difference between observed allelic frequencies and those expected from a homogenous, randomly distributed model. Populations where combinations of alleles or genotypes can be found in the expected proportions are said to be in linkage equilibrium. Wikipedia

linkage group ↦ in genetics, all of the genes on a single chromosome. They are inherited as a group; that is, during cell division they act and move as a unit rather than independently. The existence of linkage groups is the reason some traits do not comply with Mendel's law of independent assortment (recombination of genes and the traits they control); *i.e.*, the principle applies only if genes are located on different chromosomes. Variation in the gene composition of a chromosome can occur when a chromosome breaks, and the sections join with the partner chromosome if it has broken in the same places. This exchange of genes between chromosomes, called crossing over, usually occurs during meiosis, when the total number of chromosomes is halved. Encyclopedia Britannica

↦ A pair or set of genes on a chromosome that tend to be transmitted together. American Heritageő Dictionary of the English Language