Imputation practical

Gibran Hemani

04 June, 2018

All the material for this practical is available here:

https://github.com/explodecomputer/ImputationPractical

In this practical we will

- Look at imputation data formats
- Make some summary plots of imputed data
- Perform associations for imputed SNPs around the FTO region against BMI
- Create a LocusZoom plot of these results

Logging in to the server

Log into bluecrystal using PuTTY. Run the following command to access a compute node:

```
qsub -I -q teaching -l nodes=1:ppn=1,walltime=02:00:00
```

Materials

The scripts for this practical are on GitHub, and the data is located on the bluecrystal server.

Let's download the scripts first

```
module add tools/git-2.18.0
git clone https://github.com/explodecomputer/ImputationPractical.git
```

Navigate to the newly created directory

```
cd ImputationPractical
```

Now let's setup an alias to the data directory:

```
datadir="/panfs/panasas01/sscm/shortcourse/genetic.epidemiology/pract4_Imputation/data"
```

Imputation data format

0. Imputation servers are absolutely the most effective way to perform imputation today. Let's look at them:
    - Michigan server
    - Sanger server
1. As discussed in the lecture, imputation data is **probabilistic**. There are several formats, and we will look at two. First is known as Oxford format (aka gen format). It presents the genotype data in dosages. Specifically, for each individual there are 3 columns, each representing genotype probability.

    The imputed data that we will look at is a chunk of chromosome 16 located here:

`${datadir}/data_chr16.gen.gz`

View the dosage data and accompanying sample information file

```
zless -S ${datadir}/data_chr16.gen.gz
```

and

```
less ${datadir}/data.sample
```

2. We can calculate the minor allele frequencies and info scores using a programme called `qctool`

```
module add apps/qctool-1.4
qctool \
-g ${datadir}/data_chr16.gen.gz \
-snp-stats output/data_chr16.snp-stats

# Compress the output
gzip -f output/data_chr16.snp-stats
```

(takes about 5 minutes)

3. We can create plots of these data in R. Look at

   - the distribution of info scores
   - the distribution of allele frequencies
   - the relationship between info score and allele frequency

```
module add languages/R-3.3.3-ATLAS
Rscript scripts/maf_info_plots.R
```

In order to see these plots we will have to copy them across from the server to our local computers using an SFTP client.

Open up WinSCP (from the Start menu), and connect using the same credentials as you have used in Putty. Once connected you should be able to navigate to the folder `ImputationPractical/`

4. We can also look at another format - VCF (variant call format). This is emerging as a much more popular format, and is currently generated as output by both Sanger and Michigan imputation servers. The software to use for this format is vcftools or bcftools.

```
zless -S ${datadir}/data_chr16.vcf.gz
```

Performing associations with imputed data

It is possible to convert the imputed data to plink's binary format. This will **destroy** information, because it takes the dosages and reduces them to 'best guess' genotypes - discarding the uncertainty inherent in probabilistic dosages. With best guess data one can perform associations as usual with plink or other software.

Alternatively, there is software that can perform associations on the probabilistic dosage data itself, using the uncertainty as part of the association test statistic. For Oxford or VCF format data we can use software called SNPTEST.

1. Here we will perform associations of all the SNPs in our file against BMI

```
module add apps/snptest.2.5.0
snptest \
-data ${datadir}/data_chr16.gen.gz ${datadir}/data.sample \
-pheno bmi \
-cov_all \
```

```
-use_raw_phenotypes \
-frequentist 1 \
-method em \
-o output/bmi.txt
```

This will take a long time, but the results have been pre-computed so you can cancel it (`ctrl+c`). The precomputed results are here:

```
zless results/bmi.txt.gz
```

Note that the `data_chr16.vcf.gz` file can be used here in lieu of `data_chr16.gen.gz`.

2. Remove results with low info scores from the results

```
zgrep -v "#" results/bmi.txt.gz \
| awk '{ if(NR == 1 || $9 > 0.5) { print $0 }}' \
> output/bmi_filtered.txt
```

3. Create a LocusZoom plot. This requires that you use WinSCP to download the `output/bmi_filtered.txt` file from Bluecrystal3, and then upload that file to LocusZoom.

   The "Marker Column Name" is `rsid` and the "P-Value Column Name" is `frequentist_add_pvalue`. The column delimiter is `Space`. For the region specify `rs8050136` as the SNP. How does this plot compare to the results that you obtained from the GWAS session?