Genotyping and imputation

Kaur Alasoo
19 March 2018
How genotyping works

• https://www.youtube.com/watch?v=Naona1y_l2U&feature=youtu.be&t=59s
Genotype Imputation

Yun Li,¹ Cristen Willer,¹ Serena Sanna,² and Gonçalo Abecasis¹

¹Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109-2029; email: ylwtx@umich.edu; goncalo@umich.edu
²Istituto di Neurogenetica e Neurofarmacologia, Consiglio Nazionale delle Ricerche, Cagliari, Italy
Figure 3

Association of genetic variants near *LDLR* with LDL-cholesterol levels. We use data from the SardiNIA (94) and Diabetes Genetics Initiative (DGI, 90) studies reported by Willer et al. (111). Evidence for association at each SNP, measured as $\log_{10}$ p-value, is represented along the y-axis. The placement of each SNP along the x-axis corresponds to assigned chromosomal location in the current genome build. Results for directly genotyped SNPs are colored in red, imputed SNPs in blue. Note that rs6511720, the SNP showing strongest association in the region, is not well tagged by any of the variants on the Affymetrix genotyping arrays used in the SardiNIA and DGI studies. Evidence for association at the SNP increases to $p < 10^{-25}$ after follow-up in >10,000 individuals in whom the SNP was genotyped directly (111). Association results are superimposed on a gray line that summarizes the local recombination rate map. The bottom panel indicates coding sequences in the region. The putative functional gene, LDLR, is highlighted in blue.
Figure 5

Genome coverage as a function of reference panel size. The accuracy of imputation increases with the number of individuals in the reference panel. To generate the figure, we analyzed genotyped data from the FUSION study (93). For any given $r^2$ threshold, the results illustrate the proportion of markers whose genotypes can be imputed with equal or greater accuracy. The results illustrate how the proportion of markers whose genotypes are recovered accurately (with high $r^2$ between imputed and actual genotypes) increases with larger reference panels.
Homework 3
Get gene ids using the biomart package

```r
genes <- gene_level %>%
  head(3) %>%
  dplyr::select(gene_id) %>%
  unlist

mart <- useDataset("hsapiens_gene_ensembl", useMart("ensembl"))
G_list <- getBM(filters= "ensembl_gene_id", attributes= c("ensembl_gene_id","hgnc_symbol"),values=genes)

G_list
```

```markdown
##
<table>
<thead>
<tr>
<th>ensembl_gene_id</th>
<th>hgnc_symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000111912</td>
<td>NCOA7</td>
</tr>
<tr>
<td>ENSG00000137216</td>
<td>TMEM63B</td>
</tr>
<tr>
<td>ENSG00000146457</td>
<td>WTAP</td>
</tr>
</tbody>
</table>
```
Get gene ids using the g:Profiler package

```r
top3 = head(gene_level, 3)
top3$value = gconvert(top3[["gene_id"]])$name
top3$value = tr
kable(top3[, c("gene_id", "name")])
```

<table>
<thead>
<tr>
<th>gene_id</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000111912</td>
<td>NCOA7</td>
</tr>
<tr>
<td>ENSG00000137216</td>
<td>TMEM63B</td>
</tr>
<tr>
<td>ENSG00000146457</td>
<td>WTAP</td>
</tr>
</tbody>
</table>