Homework 7
1. Expression of PTK2B gene in the naive condition and Alzheimer's disease (AD)

Posterior probabilities:

res1 = colocalisation(eQTL_file="PTK2B_eQTL.txt", gwas_file="AD_GWAS_PTK2B_locus.txt", p1 = 1e-4, p2 = 1e-4, p12 = 1e-5, region = 200000)

##)PP.H0.abf)PP.H1.abf)PP.H2.abf)PP.H3.abf)PP.H4.abf)
##))1.02e+06))2.43e+05))6.80e+04))1.51e+02))9.84e-01)
##)
##)[1]) "PP abf for shared variant: 98.4%"
For a given trait (gene expression or disease) are the pD values similar or different?

Does this reflect the colocalisation posterior probabilities (PP4)?

Answer: The pD values are very different for a given trait.

2. Expression of the TRAF1 gene in the IFNγ + Salmonella condition and rheumatoid arthritis (RA).

Posterior probabilities:

\[
\text{res1 = colocalisation(eQTL_file="ICOSLG_eQTL.txt", gwas_file="UC_GWAS_ICOSLG_locus.txt", p1 = 1e-4, p2 = 1e-4, p12 = 1e-5, region = 200000)}
\]

## PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
## 5.19e-11 1.00e-05 5.19e-06 1.00e+00 8.75e-05
## [1] "PP abf for shared variant: 0.00875%"

3. Expression of the ICOSLG gene in the naive condition and ulcerative colitis (UC).

Posterior probabilities:

\[
\text{res2 = colocalisation(eQTL_file="ICOSLG_eQTL.txt", gwas_file="UC_GWAS_ICOSLG_locus.txt", p1 = 1e-4, p2 = 1e-4, p12 = 1e-5, region = 200000)}
\]

## PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
## 8.75e+05 1.00e+00 5.19e+06 1.00e+00 5.19e+11
## [1] "PP abf for shared variant: 0.00875%"
For a given trait (gene expression or disease) are the $p$ values similar or different? Does this reflect the colocalisation posterior probabilities ($PP4$)?

```
res3 = colocalisation(eQTL_file="TRAFF1_eQTL.txt", gwas_file="RA_GWAS_TRAF1_locus.txt", p1 = 1e-4, p2 = 1e-4, p12 = 1e-5, region = 200000)
```

## PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
## 1.38e-06 6.01e-05 1.66e-03 7.12e-02 9.27e-01
## [1] "PP abf for shared variant: 92.7%"
How do the prior probabilities affect colocalisation results?
1. **Expression of PTK2B gene in the naive condition and Alzheimer's disease (AD)**

Posterior probabilities:

```r
res1 = colocalisation(eQTL_file="PTK2B_eQTL.txt",
gwas_file="AD_GWAS_PTK2B_locus.txt", p1 = p1, p2 = p2, p12 = p12, region = 200000)
```

```text
#> PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
#> 8.97e-06 2.13e-04 5.95e-03 1.33e-01 8.61e-01
#> [1] "PP abf for shared variant: 86.1%"
```

2. **Expression of ICOSLG gene in the naive condition and ulcerative colitis (UC).**

Posterior probabilities:

```r
res2 = colocalisation(eQTL_file="ICOSLG_eQTL.txt",
gwas_file="UC_GWAS_ICOSLG_locus.txt", p1 = p1, p2 = p2, p12 = p12, region = 200000)
```

```text
#> PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
#> 5.19e-11 1.00e-05 5.19e-06 1.00e+00 8.75e-06
#> [1] "PP abf for shared variant: 0.000875%"
```

3. **Expression of the TRAF1 gene in the IFNg + Salmonella condition and rheumatoid arthritis (RA).**

```r
res3 = colocalisation(eQTL_file="TRAF1_eQTL.txt",
gwas_file="RA_GWAS_TRAF1_locus.txt", p1 = p1, p2 = p2, p12 = p12, region = 200000)
```

```text
#> PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
#> 8.34e-06 3.63e-04 1.00e-02 4.30e-01 5.60e-01
#> [1] "PP abf for shared variant: 56%"
```
How does the length of the genomic region affect the results?
1. Expression of PTK2B gene in the naive condition and Alzheimer’s disease (AD)

Posterior probabilities:

```r
regions = c(200000, 100000, 50000, 5000)
for(r in regions){
  print(paste0("Region: +/- ", r/1000, " kb"))
  res = colocalisation(eQTL_file="PTK2B_eQTL.txt",
                       gwas_file="AD_GWAS_PTK2B_locus.txt", p1 = p1, p2 = p2, p12 = p12,
                       region = r)
}
```

```
[[1]] "Region: +/- 200 kb"
PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
8.97e-06 2.13e-04 5.95e-03 1.33e-01 8.61e-01
[[1]] "PP abf for shared variant: 86.1%"

[[1]] "Region: +/- 100 kb"
PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
8.97e-06 2.12e-04 5.95e-03 1.32e-01 8.61e-01
[[1]] "PP abf for shared variant: 86.1%"

[[1]] "Region: +/- 50 kb"
PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
8.99e-06 2.10e-04 5.95e-03 1.30e-01 8.64e-01
[[1]] "PP abf for shared variant: 86.4%"

[[1]] "Region: +/- 5 kb"
PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
0.00212 0.01150 0.01770 0.08770 0.88100
[[1]] "PP abf for shared variant: 88.1%"
```

res$plot
2. Expression of ICOSLG gene in the naive condition and ulcerative colitis (UC).

```r
for(r in regions){
    print(paste0("Region: +/- ", r/1000, " kb"))
    res = colocalisation(eQTL_file="ICOSLG_eQTL.txt",
                        gwas_file="UC_GWAS_ICOSLG_locus.txt", p1 = p1, p2 = p2, p12 = p12, region = r)
}
```

```plaintext
## [1] "Region: +/- 200 kb"
## PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
## 5.19e-11 1.00e-05 5.19e-06 1.00e+00 8.75e-06
## [1] "PP abf for shared variant: 0.000875%"
## [1] "Region: +/- 100 kb"
## PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
## 5.19e-11 1.00e-05 5.19e-06 1.00e+00 8.75e-06
## [1] "PP abf for shared variant: 0.000875%"
## [1] "Region: +/- 50 kb"
## PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
## 5.19e-11 1.00e-05 5.19e-06 1.00e+00 8.75e-06
## [1] "PP abf for shared variant: 0.000875%"
## [1] "Region: +/- 5 kb"
## PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
## 8.99e-06 9.18e-01 1.46e-07 1.42e-02 6.79e-02
## [1] "PP abf for shared variant: 6.79%"
res$plot
```
3. Expression of the TRAF1 gene in the IFNg + Salmonella condition and rheumatoid arthritis (RA).

Posterior probabilities:

```r
for(r in regions){
    print(paste0("Region: +/- ", \( r/1000 \), " kb"))
    res = colocalisation(eQTL_file="TRAF1_eQTL.txt",
                        gwas_file="RA_GWAS_TRAF1_locus.txt", p1 = p1, p2 = p2, p12 = p12, region = r)

## [1] "Region: +/- 200 kb"
## PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
##  8.34e-06  3.63e-04  1.00e-02  4.30e-01  5.60e-01
## [1] "PP abf for shared variant: 56%"

## [1] "Region: +/- 100 kb"
## PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
##  8.41e-06  3.59e-04  1.01e-02  4.25e-01  5.64e-01
## [1] "PP abf for shared variant: 56.4%"

## [1] "Region: +/- 50 kb"
## PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
##  8.65e-06  3.59e-04  1.02e-02  4.19e-01  5.70e-01
## [1] "PP abf for shared variant: 57%"

## [1] "Region: +/- 5 kb"
## PP.H0.abf PP.H1.abf PP.H2.abf PP.H3.abf PP.H4.abf
##  7.66e-05  5.85e-04  1.36e-02  9.48e-02  8.91e-01
## [1] "PP abf for shared variant: 89.1%"

res$plot
```
How do the colocalisations results change for the three loci (PTK2B, TRAF1 and ICOSLG)? Can you explain why the changes happen?

Answer: It seems that the smaller the region gets the "stronger" the colocalisation becomes, i.e. the PP4 probability increases. Taking smaller region leaves us less associations and therefore it is easier to get similar patterns in small regions.

Manhattan plot:

P values for lead variants:

<table>
<thead>
<tr>
<th>rsID</th>
<th>Position (bp)</th>
<th>Phenotype</th>
<th>Lead</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs10985070</td>
<td>120873843</td>
<td>RA_GWAS_TRAF1_locus</td>
<td>blue</td>
</tr>
<tr>
<td>rs10985070</td>
<td>120873843</td>
<td>TRAF1_eQTL</td>
<td>blue</td>
</tr>
<tr>
<td>rs2241003</td>
<td>120904499</td>
<td>RA_GWAS_TRAF1_locus</td>
<td>red</td>
</tr>
<tr>
<td>rs2241003</td>
<td>120904499</td>
<td>TRAF1_eQTL</td>
<td>red</td>
</tr>
</tbody>
</table>

For a given trait (gene expression or disease) are the p values similar or different? Does this reflect the colocalisation posterior probabilities (PP4)?
Projects for the Bioinformatics course

Kaur Alasoo
25 April 2018
Reproducing data analysis from published papers
Tasks
1. Start from raw RNA-seq data
2. Reproduce the PCA plot shown above
3. Perform DE analysis
4. Import data into funcExplorer to characterise the differences in gene expression response between Salmonella and Listeria.
Quality control and re-analysis of large published RNA-seq datasets


Tasks
1. Start from raw RNA-seq data
2. Reproduce the PCA plot shown above
3. Perform DE analysis
4. Import data into funcExplorer to characterise the differences in gene expression response between Salmonella and Listeria.
We next searched for genes presenting population differences in their response to treatment, relative to non-stimulated cells (popDRGs). We found 3,841 popDRGs (FDR < 0.05, 70% of popDEGs), the majority of which were treatment specific (2,687 popDRGs; Table S1). popDRGs displaying stronger responses in Africans were enriched in GO functions from metabolic processes to defense responses, while popDRGs responding more strongly in Europeans were essentially restricted to defense functions in the TLR conditions and enriched in translational processes upon IAV infection (Table S1). popDRGs showing the greatest population differences ($\log_2(FC_{pop}) > 1$) were enriched in cytokines and chemokines (Fisher’s exact test, odds ratio [OR] = 36.7, $p < 10^{-5}$), including IL12B and CSF3, responding more strongly to Pam3CSK4 in Africans, and CCL8, CCL13, CCL15, CCL23 and CXCL10, being more responsive to LPS in Europeans (Table 1). These results indicate that while population transcriptional differences of moderate effect are widespread, strong differences predominantly affect antiviral and inflammation-related genes that differ markedly in responsiveness between Africans and Europeans.

Detecting Local Immune-Responsive Regulatory Variation

We next mapped eQTLs by testing for associations between 10,278,745 SNPs (the set of genotyped and imputed SNPs presenting a minor allele frequency [MAF] > 0.05) and gene expression phenotypes. We first mapped local, likely cis-acting eQTLs within 1 Mb of each gene in Africans and Europeans separately. We used an additive linear model (Shabalin, 2012) that included the first two PCs of the genetic data (Figures S5C and S5D) to account for possible population substructure. Considering only eQTLs having an effect size of $|b_{eQTL}| > 0.2$ at a FDR of 5%, we found 2,665 genes with an eQTL in at least one condition (Figure S6A; Table S2A). Of these, 917 genes presented a response eQTL (reQTL), an eQTL with a significantly larger effect size after treatment than at the basal state ($b_{eQTL} > 0$ and $p < 10^{-5}$, Figure 2A). Consistent with data for other cell types or stimuli (Fairfax et al., 2014; Lee et al., 2014), most reQTLs were treatment specific (62%, 570 genes), indicating strong context specificity of the genetic regulation of immune responses.

To investigate the functional features of (r)eQTLs, we used the predicted regulatory elements of CD14+ monocytes (Zerbino...
Figure 4: Estimated cell-type composition of GTEx samples corresponds to image data. Histology images from the GTEx adipose samples with highest (49%) and lowest (0%) macrophage estimates are shown. Both whole biopsy (left) and zoomed in images (right) are presented. Estimated cell type composition of all GTEx samples are provided in Supplementary File 2.

Table 1: TwinsUK macrophage proportion in adipose tissue is associated to obesity-related traits but not age.

Trait \( r^2 \) P-value
BMI 0.22 \( 2.2 \times 10^{-8} \)
Visceral fat 0.29 \( 4.9 \times 10^{-15} \)
Visceral Fat (BMI adjusted) 0.28 \( 1.9 \times 10^{-9} \)
Android/Gynoid ratio 0.36 \( 1.2 \times 10^{-16} \)
Android/Gynoid ratio (BMI adjusted) 0.35 \( 1.8 \times 10^{-12} \)
Age -0.02 0.67

Can you discover the same macrophage signature from the TwinsUK data using only gene expression clustering (funcExplorer)?

Other datasets

- BLUEPRINT
- Fairfax et al monocytes
Benchmarking
Benchmarking RNA-seq aligners for splicing QTL detection

Supervisor: Kaur Alasoo
Problem type: benchmark
Problem description:
Two of the most commonly used aligners for RNA-seq data are STAR and HISAT2. While STAR seems to be more accurate in recent benchmarks (Baruzzo et al), it also uses much more computational resources (e.g. 10x more RAM). Furthermore, it is not clear how differences observed in simulation experiments translate to practical real world applications. In this project, your task is to evaluate STAR against HISAT2 in a specific real world use-case: when you run both aligners on the same dataset and keep all of the other parameters the same then which aligner allows you to detect more genetic variants that are associated with RNA splicing.

Run STAR, STAR 2-pass and HISAT2 (with and without junction database) on the GEUVADIS dataset (Lappalainen et al). Identify genetic variants associated with RNA splicing using Leafcutter + QTLtools and compare the results. Which approach discovers the largest number of genetic associations? What are the main differences?

References
Benchmarking transcript assembly (StringTie, Scallop) against annotated Ensembl transcripts in discovering genetic variants that regulate transcript usage

Supervisor: Kaur Alasoo
Problem type: benchmark
Problem description:
Methods such as StringTie and Scallop that assemble transcripts directly from RNA-seq data are a promising approach to for analysing RNA-seq data. StringTie has also been used to detect genetic variants that regulate which transcript is being used across individuals. However, what is the sensitivity of transcript assembly approaches relative to annotated transcripts from a database is poorly understood.

Your task is to replicate the analysis from the Chun et al paper and apply it to the data generated in the Alasoo et al paper to detect genetic variants that are associated with usage of de novo assembled transcripts from the RNA-seq data. Next, you will compare these associations to the ones detected using Ensembl annotations? Which method detects more associations? Are there specific associations that are missed by one or the other method?

References
Other benchmarks

- txrevise vs Whippet
- Kallisto vs Salmon
Colocalisation
Colocalisation between genetic variants associated with RNA splicing and protein abundance - can RNA splicing change protein abundance without influencing the total amount of RNA produced from the gene?

**Supervisor:** Kaur Alasoo

**Problem description:**
Most human genes have multiple alternative versions known as transcripts. Transcripts can differ from each other in many ways including different start positions, variable middle sections (for example, some transcripts might skip part of the gene sequence via a process called *alternative splicing*) or different end positions. Although genetic differences between individuals that regulate transcript usage have recently been implicated in multiple complex diseases (Li *et al.*, 2016), the immediate functional consequences of most alternative transcription events remain poorly understood.

One possibility is that alternative transcripts alter the amount of protein that is produced from a given gene. However, testing this experimentally for any given transcript and protein is a laborious process. A powerful approach to address this question is to look for genetic variants that are simultaneously associated with transcript expression level as well as protein abundance. Importantly, a recent study by Sun *et al.* identified 1,927 genetic variants that are associated with protein levels in human plasma. You will have access to the full summary statistics from this study (~3.5TB). The aim of this project is to estimate what proportion of these associations could be explained by changes to transcript structure and contrast this to effect of total gene expression level. Identifying robust association between transcript structure and protein abundance can lead to better understanding of the processes involved in protein production.

Using the protein expression summary statistics from Sun *et al.* and transcript usage summary statistics generated by us, you will perform hundreds of colocalisations tests with to coloc package to identify genetic variants associated with both traits.

**References:**
Regulation of transcript usage by DNA binding proteins

**Supervisors:** Kaur Alasoo

**Problem description:**
Most human genes have multiple alternative versions known as transcripts. Transcripts can differ from each other in many ways including different start positions, variable middle sections (for example, some transcripts might skip part of the gene sequence via a process called *alternative splicing*) or different end positions. Genetic differences between individuals that influence which transcript is being used have recently been implicated in multiple complex diseases (Li *et al.*, 2016).

However, the molecular mechanisms how genetic differences influence transcript usage are still poorly understood. While most of these variants are likely to act at the RNA transcript by directly influencing the ‘splicing code’, multiple previous studies have also implicate the role of DNA binding proteins. We have recently developed a computational pipeline to extensively quantify the variation in transcript usage. The first aim of the project is to apply this pipeline to a large RNA-seq dataset from the GEUVADIS project (Lappalainen *et al.*, 2013) to identify a large set of genetic variants that regulate transcript usage. The second aim is to use statistical colocalisation to identify which of these variants regulate protein binding to DNA (‘chromatin accessibility’) in a recently generated dataset of 100 individuals (Kumasaka *et al.*, 2016 and unpublished data). The final aim is to characterise the properties of genetic variants that simultaneously regulate protein binding to DNA and transcript usage.

**References:**
Other projects
Proof-of-concept implementation of g:Profiler for genomic regions to detect enriched transcription factor motifs
Graphical web user interface for visualising summary statistic from large GWAS and QTL studies

- Some implementations exist (e.g. http://www.immunpop.com/), but they are very slow and barely usable. Your task is to explore how modern graphical libraries (eg. higlass.io) could be used to visualise QTL and GWAS summary statistics in a browser.
Leo’s projects

- **Guide-Target mismatch analysis in CRISPR-Cas9 self-targeting data**
  CRISPR-Cas9 is a recently introduced gene editing technique that is currently revolutionizing biological research. This technique is used to cut DNA at a location specified by a 20nt guide RNA. We’ve collected some in-house data to examine the effect of mismatches between the guide RNA sequence and the DNA target. This project would involve analysing some of that data to examine the effect of the type and location of mismatch on the editing capability of each gRNA. The project may also optionally apply machine learning methods to predict the editing outcome of each gRNA from its sequence.

- **Sequence analysis for suppressors of gene essentiality.**
  The deletion of some genes is sometimes tolerated in one individual, but not another. This means the others have additional variants in their genome that compensate the otherwise essential function of the gene. We have sequenced many such examples, and called mutations. This project involves combining these mutation calls (.vcf files) across replicates appropriately, visualising them, and prioritizing candidate causal modifiers.