Using regulatory genomics data to interpret the function of disease variants and prioritise genes from expression studies

Enrico Ferrero

Abstract The identification of therapeutic targets is a critical step in the research and developement of new drugs, with several drug discovery programmes failing because of a weak linkage between target and disease. Genome-wide association studies and large-scale gene expression experiments are providing insights into the biology of several common and complex diseases, but the complexity of transcriptional regulation mechanisms often limit our understanding of how genetic variation can influence changes in gene expression. Several initiatives in the field of regulatory genomics are aiming to close this gap by systematically identifying and cataloguing regulatory elements such as promoters and enhacers across different tissues and cell types. In this Bioconductor workflow, we will explore how different types of regulatory genomic data can be used for the functional interpretation of disease-associated variants and for the prioritisation of gene lists from gene expression experiments.

Keywords

bioconductor; r; rstats; regulatory genomics; functional genomics; genetics; gwas; transcriptomics; integration; multiomics

Introduction

Discovering and bringing new drugs to the market is a long, expensive and inefficient process [1;2]. The majority of drug discovery programmes fail for efficacy reasons [3], with up to 40% of these failures due to lack of a clear link between the target and the disease under investigation [4].

Target selection, the first step in drug discovery programmes, is thus a critical decision point. It has previously been shown that therapeutic targets with a genetic link to the disease under investigation are more likely to progress through the drug discovery pipeline, suggesting that genetics can be used as a tool to prioritise and validate drug targets in early discovery [5;6].

One of the biggest challenges in translating findings from genome-wide association studies (GWASs) to therapies is that the great majority of single nucleotide polymorphisms (SNPs) associated with disease are found in non-coding regions of the genome, and therefore cannot be easily linked to a target gene [7]. Many of these SNPs could be regulatory variants, affecting the expression of nearby or distal genes by interfering with the transcriptional process [8].

The most established way to map disease-associated regulatory variants to target genes is to use expression quantitative trait loci (eQTLs) [9], variants that affect the expression of specific genes. The GTEx consortium profiled eQTLs across 44 human tissues by performing a large-scale mapping of genome-wide correlations between genetic variants and gene expression [10].

However, depending on the power of the study, it might not be possible to detect all existing regulatory variants as eQTLs. An alternative is to use information on the location of promoters and distal enhancers across the genome and link these regulatory elements to their target genes. Large, multi-centre initiatives such as ENCODE [11], Roadmap Epigenomics [12] and BLUEPRINT [13;14] mapped regulatory elements in the genome by profiling a number of chromatin features, including DNase hypersensitive sites (DHSs), several types of histone marks and binding of chromatin-associated proteins in a large number of cells and tissues. Similarly, the FANTOM consortium used cap analysis of gene expression (CAGE) to identify promoters and enhancers across hundreds of cells and tissues [15].

Knowing that a certain stretch of DNA is an enhancer is however not informative of the target gene(s). One way to infer links between enhancers and promoters *in silico* is to identify significant correlations across a large panel of cell types, an approach that was used for distal and promoter DHSs [16] as well as for CAGE-defined promoters and enhancers [17]. Experimental methods to assay interactions between regulatory elements also exist. Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) [18;19] couples chromatin immunoprecipitation with DNA ligation to identify DNA regions interacting thanks to the binding of a specific protein. Promoter capture Hi-C [20;21] extends chromatin conformation capture by using “baits” to enrich for promoter interactions and increase resolution.

Overall, linking genetic variants to their candidate target genes is not straightforward, not only because of the complexity of the human genome and transcriptional regulation, but also because of the variety of data types and approaches that can be used. To address this problem, we developed STOPGAP, a database of disease variants mapped to their most likely target gene(s) using several different types of regulatory genomic data [22]. The database is currently undergoing a major overhaul and will eventually be superseded by POSTGAP. A valid and recent alternative is INFERNO [23], though it does only rely on eQTL data for target gene assignment. These resources implement some or all of the approaches that will be reviewed in the workflow and constitute good entry points for identifying the most likely target gene(s) of regulatory SNPs. However, as they tend to hide much of the complexity involved in the process, we will not use them and rely on the original datasets instead.

In this workflow, we will explore how regulatory genomic data can be used to connect the genetic and transcriptional layers by providing a framework for the discovery of novel therapeutic targets. We will use eQTL data from GTEx [10], FANTOM5 correlations between promoters and enhancers [17] and promoter capture Hi-C data [21] to annotate significant GWAS variants to putative target genes and to prioritise genes obtained from a differential expression analysis (Figure 1).

Diagram showing a schematic representation of the workflow and the steps involved.

Workflow

Install required packages

R version 3.4.2 and Bioconductor version 3.6 were used for the analysis. The code below will install all required packages and dependencies from Bioconductor and CRAN:

```
source("https://bioconductor.org/biocLite.R")
# uncomment the following line to install packages
#biocLite(c("clusterProfiler", "DESeq2", "GenomicFeatures", "GenomicInteractions", "GenomicRanges",
```

Gene expression data and differential gene expression analysis

We start with a common scenario: we ran a RNA-seq experiment comparing patients with a disease and healthy individuals, and would like to discover key disease genes and potential therapeutic targets by integrating genetic information in our analysis.

The RNA-seq data we will be using comes from blood of patients with systemic lupus erythematosus (SLE) and healthy controls[24]. SLE is a chronic autoimmune disorder that can affect several organs with a significant unmet medical need[25]. It is a complex and remarkably heterogeneous disease, in terms of both genetics and clinical manifestations[26]. Early diagnosis and classification of SLE remain extremely challenging[27].

In the original study[24], the authors explore transcripts bound by Ro60, an RNA-binding protein against which some SLE patients produce autoantibodies. They identify Alu retroelements among these transcripts and use RNA-seq data to check their expression levels, observing that Alu elements are significantly more expressed in SLE patients, and particularly in those patients with anti-Ro antibodies and with a higher interferon signature metric (ISM).

We are going to use `recount`[28] to obtain gene-level counts:

```
library(recount)
# uncomment the following line to download dataset
#download_study("SRP062966")
load(file.path("SRP062966", "rse_gene.Rdata"))
rse <- scale_counts(rse_gene)
```

Other Bioconductor packages that can be used to access data from gene expression experiments directly in R are `GEOquery`[29] and `ArrayExpress`[30].

We have 117 samples overall. This is what the matrix of counts looks like:

```
assay(rse)[1:3, 1:3]
```

```
##                     SRR2443263 SRR2443262 SRR2443261
## ENSG00000000003.14         19          6         10
## ENSG00000000005.5           0          0          0
## ENSG00000000419.12        489        238        224
```

Each gene is a row and each sample is a column. We note that genes are annotated using the GENCODE[31] v25 annotation, which will be useful later on.

To check how we can split samples between cases and controls, we can have a look at the metadata contained in the `characteristics` column, which is a `CharacterList` object:

```
head(rse$characteristics, 3)
```

```
## CharacterList of length 3
## [[1]] disease status: healthy tissue: whole blood anti-ro: control ism: control
## [[2]] disease status: healthy tissue: whole blood anti-ro: control ism: control
## [[3]] disease status: healthy tissue: whole blood anti-ro: control ism: control
```

We have information about the disease status of the sample, the tissue of origin, the presence and level of anti-ro autoantibodies and the value of the ISM. However, we note that basic information such as age or gender is missing.

We can create some new columns with the available information so that they can be used for downstream analyses. We will also make sure that they are encoded as factors and that the correct reference layer is used:

```
# disease status
rse$disease_status <- sapply(rse$characteristics, "[", 1)
rse$disease_status <- sub("disease status: ", "", rse$disease_status)
rse$disease_status <- sub("systemic lupus erythematosus \\(SLE\\)", "SLE", rse$disease_status)
rse$disease_status <- factor(rse$disease_status, levels = c("healthy", "SLE"))
# tissue
rse$tissue <- sapply(rse$characteristics, "[", 2)
rse$tissue <- sub("tissue: ", "", rse$tissue)
rse$tissue <- factor(rse$tissue)
```

```
# anti-ro
rse$anti_ro <- sapply(rse$characteristics, "[", 3)
rse$anti_ro <- sub("anti-ro: ", "", rse$anti_ro)
rse$anti_ro <- factor(rse$anti_ro)
# ism
rse$ism <- sapply(rse$characteristics, "[", 4)
rse$ism <-sub("ism: ", "", rse$ism)
rse$ism <- factor(rse$ism)
```

We can check how many samples we have in each group (note that we ignore `tissue` as it's always `whole blood`):

```
metadata <- data.frame(disease_status = rse$disease_status, anti_ro.ism = paste(rse$anti_ro, rse$ism
table(metadata)
```

```
##               anti_ro.ism
## disease_status control.control high.ISM_high high.ISM_low med.ISM_high
##        healthy              18             0            0            0
##        SLE                   0            23            1           21
##               anti_ro.ism
## disease_status med.ISM_low none.ISM_high none.ISM_low
##        healthy           0             0           0
##        SLE               2            31          21
```

Now we are ready to perform a simple differential gene expression analysis with `DESeq2`[32]. Note that we remove genes with a low number of counts (less than 50 across all 117 samples) to speed up execution and reduce the memory footprint:

```
library(DESeq2)
dds <- DESeqDataSet(rse, ~ disease_status)
dds <- DESeq(dds)
dds <- dds[rowSums(counts(dds)) >= 50, ]
```

We used an extremely simple model; in the real world we should be accounting for co-variables, potential confounders and interactions between them. For example, age and gender are usually included in this type of analysis, but we don't have access to this information for this dataset. Similarly, the value of the ISM and the presence of anti-Ro autoantibodies can't be included in the analysis due to the fact that these variables are collinear with the disease status variable (i.e.: the value of both `anti_ro` and `ism` is `control` for all samples with `disease_status` equal to `healthy`.) Like `DESeq2`, `edgeR`[33] and `limma`[34] can also deal with multiple cofactors and different experimental designs, and constitute good alternatives for performing differential expression analyses.

We can now look at the data in more detail to assess if we can observe a separation between the SLE and healthy samples and whether any batch effect is visible.
We use the variance stabilising transformation (VST)[35] for visualisation purposes:

```
vsd <- vst(dds, blind = FALSE)
```

We will use the `pheatmap` and `RColorBrewer` packages to perform hierarchical clustering of the samples (Figure 2):

```
library(pheatmap)
library(RColorBrewer)
sampleDists <- dist(t(assay(vsd)))
sampleDistMatrix <- as.matrix(sampleDists)
annotation = data.frame(colData(vsd)[c("anti_ro", "ism", "disease_status")], row.names = rownames(sa
colors <- colorRampPalette(rev(brewer.pal(9, "Blues")))(255)
pheatmap(sampleDistMatrix, clustering_distance_rows = sampleDists, clustering_distance_cols = sample
```

While there isn't an unambiguous split between healthy and disease samples, the most distinct clusters (bottom right and top left) are entirely composed of SLE samples, with the central cluster containing all healthy samples and a number of SLE ones. The clusters don't appear to be due to the ISM or the presence of anti-Ro autoantibodies.

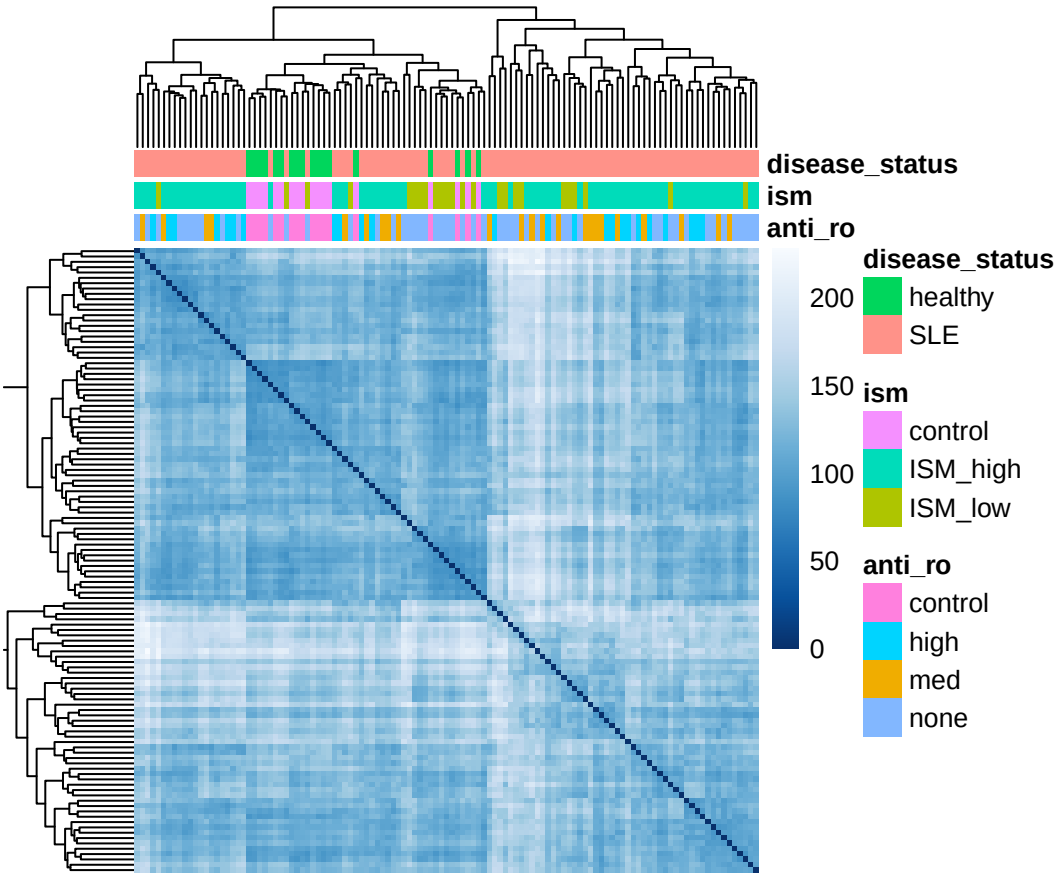

Figure 1. Heatmap showing Euclidean distances between samples clustered using complete linkage. Disease status and other experimental factors are visualised as column annotations.

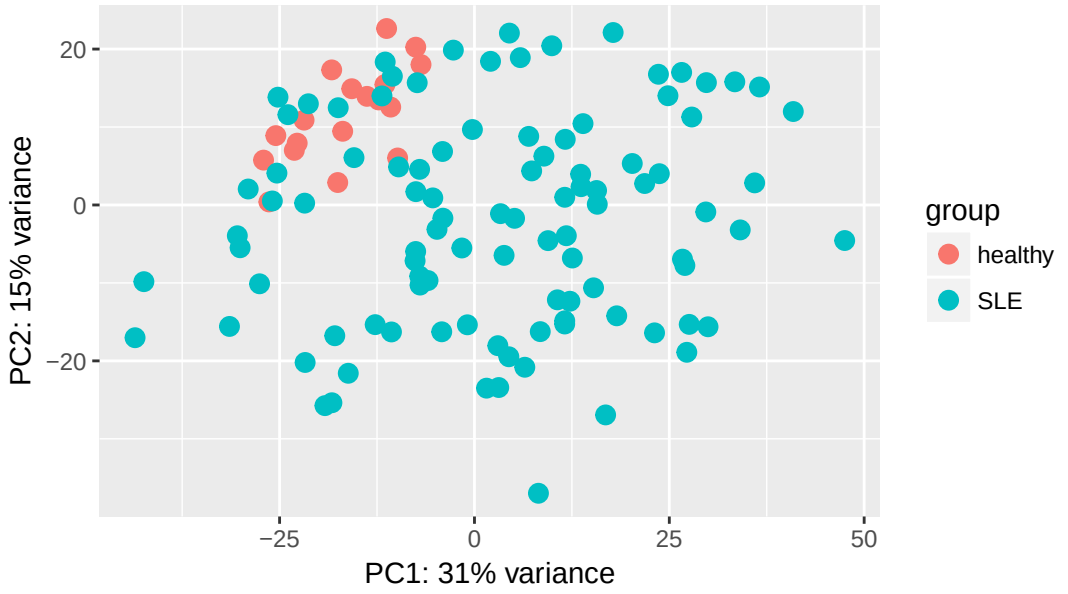

Figure 2. Scatter plot showing results of a PCA with samples coloured according to their disease status.

Similarly, we can perform a principal component analysis (PCA) on the most variable 500 genes (Figure 3). Note that we load `ggplot2`[36] to modify the look of the plot:

```
library(ggplot2)
plotPCA(vsd, intgroup = "disease_status") +
  coord_fixed()
```

We can see some separation of healthy and SLE samples along both PC1 and PC2, though some SLE samples appear very similar to the healthy ones. No obvious batch effects are visible from this plot.

Next, we select genes that are differentially expressed below a 0.05 adjusted *p*-value threshold:

```
res <- results(dds, alpha = 0.05)
summary(res)
```

```
##
## out of 32820 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)      : 4829, 15%
## LFC < 0 (down)    : 2709, 8.3%
## outliers [1]      : 0, 0%
## low counts [2]    : 2548, 7.8%
## (mean count < 1)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

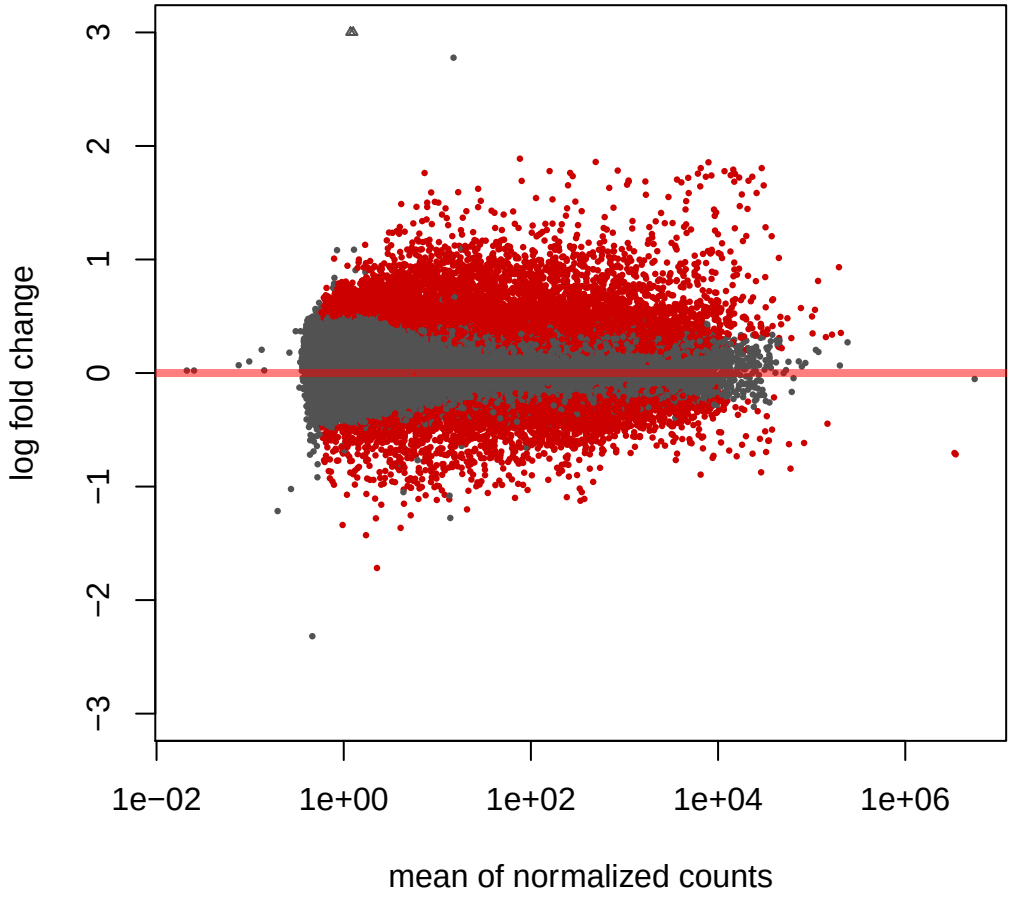

Figure 3. MA plot showing genes differentially expressed (red dots) in SLE patients compared to healthy patients.

We can visualise the shrunken log2 fold changes using an MA plot (Figure 4):

```
res_lfc <- lfcShrink(dds, coef = 2)
plotMA(res_lfc, ylim = c(-3, 3))
```

We observe large numbers of genes differentially expressed in both directions and across a range of fold changes, though the majority of significant genes appear to be upregulated in disease.

For convenience, we will save our differentially expressed genes (DEGs) in another object and map the GENCODE gene IDs to gene symbols using the annotation in the original `RangedSummarizedExperiment` object

```
degs <- subset(res, padj < 0.05)
degs <- merge(rowData(rse), as.data.frame(degs), by.x = "gene_id", by.y = "row.names", all = FALSE)
head(degs, 3)
```

```
## DataFrame with 3 rows and 9 columns
##                          gene_id bp_length symbol   baseMean
##                      <character> <integer> <list>  <numeric>
## ENSG00000000003 ENSG00000000003.14      4535 TSPAN6   8.739822
## ENSG00000000419 ENSG00000000419.12      1207   DPM1 431.485085
## ENSG00000000457 ENSG00000000457.13      6883  SCYL3 686.579323
##                 log2FoldChange      lfcSE      stat       pvalue
##                      <numeric>  <numeric> <numeric>    <numeric>
```

```
## ENSG00000000003     -0.4750382 0.18374822 -2.585267 9.730366e-03
## ENSG00000000419      0.5559772 0.10117967  5.494950 3.908216e-08
## ENSG00000000457      0.1927081 0.05928191  3.250707 1.151185e-03
##                          padj
##                     <numeric>
## ENSG00000000003 4.161281e-02
## ENSG00000000419 2.182977e-06
## ENSG00000000457 7.922475e-03
```

Accessing GWAS data

The differential expression analysis resulted in several thousands of DEGs. Since we know that genes with high levels of differential expression are more likely to harbour disease-associated variants[37] and that therapeutic targets with genetic evidence are more likely to progress through the drug discovery pipeline[6], one way to prioritise them is to check which of these can be genetically linked to SLE. To get hold of relevant GWAS data, we will be using the `gwascat` Bioconductor package[38], which provides an interface to the GWAS catalog[39]. An alternative is to use the GRASP[40] database with the `grasp2db`[41] package.

```
library(gwascat)
# uncomment the following line to download file and build the gwasloc object all in one step
#snps <- makeCurrentGwascat()
# uncomment the following line to download file
#download.file("http://www.ebi.ac.uk/gwas/api/search/downloads/alternative", destfile = "gwas_catalo
snps <- read.delim("gwas_catalog_v1.0.1-associations_e90_r2017-12-04.tsv", check.names = FALSE, stri
snps <- gwascat:::gwdf2GRanges(snps, extractDate = "2017-12-04")
genome(snps) <- "GRCh38"
head(snps, 3)
```

```
## gwasloc instance with 3 records and 37 attributes per record.
## Extracted:  2017-12-04
## Genome:  GRCh38
## Excerpt:
## GRanges object with 3 ranges and 3 metadata columns:
##       seqnames                ranges strand | DISEASE/TRAIT        SNPS
##          <Rle>             <IRanges>  <Rle> |   <character> <character>
##   [1]     chr1 [203186754, 203186754]      * | YKL-40 levels   rs4950928
##   [2]    chr13 [ 39776775,  39776775]      * |     Psoriasis   rs7993214
##   [3]    chr15 [ 78513681,  78513681]      * |   Lung cancer   rs8034191
##         P-VALUE
##       <numeric>
##   [1]     1e-13
##   [2]     2e-06
##   [3]     3e-18
##   -------
##   seqinfo: 23 sequences from GRCh38 genome; no seqlengths
```

`snps` is a `gwasloc` object which is simply a wrapper around a `GRanges` object, the standard way to represent genomic ranges in Bioconductor.

We note here that the GWAS catalog uses GRCh38 coordinates, the same assembly used in the GENCODE v25 annotation. When integrating genomic datasets from different sources it is essential to ensure that the same genome assembly is used, especially because many datasets in the public domain are still using GRCh37 coordinates. As we will see below, it is possible and relatively straightforward to convert genomic coordinates between genome assemblies.

We can select only SNPs that are associated with SLE:

```
snps <- subsetByTraits(snps, tr = "Systemic lupus erythematosus")
```

We can visualise these as a Manhattan plot to look at the distribution of GWAS *p*-values over chromosomes on a negative $\log_{10}$ scale (Figure 5): Note that *p*-values lower than 1e-25 are truncated in the figure:

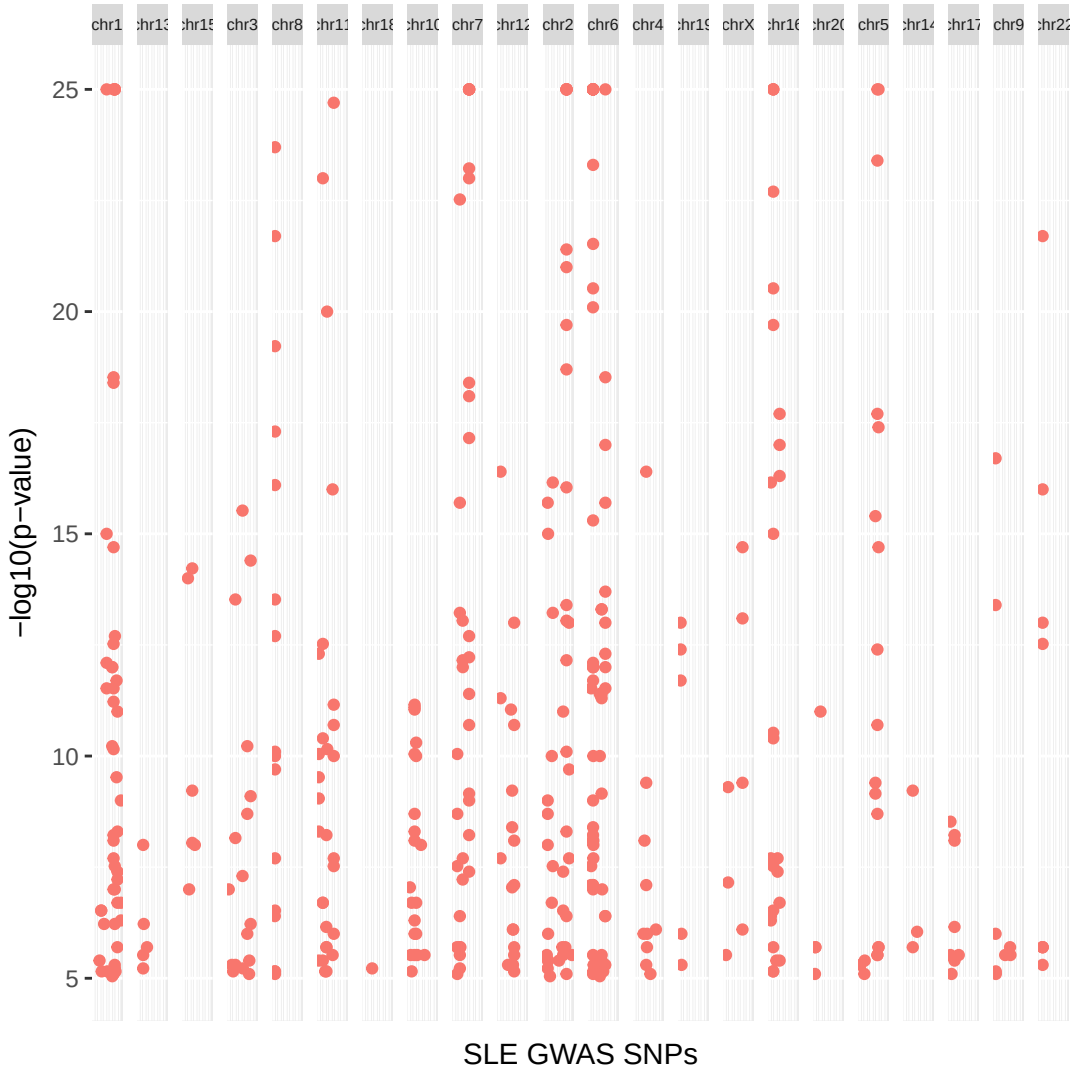

Figure 4. Manhattan plot showing GWAS variants significantly associated with SLE.

```
traitsManh(gwr = snps, sel = snps, traits = "Systemic lupus erythematosus") +
  xlab("SLE GWAS SNPs") +
  ylab("-log10(p-value)") +
  theme(legend.position = "none",
        strip.text.x = element_text(size = 6),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank())
```

We observe several hits across most chromosomes, with many of them below a genome-wide significant threshold (p-value < 1 x 10^{-8}), suggesting that genetics plays an important role in the pathogenesis of SLE.

We note here that genotyping arrays typically include a very small fraction of all possible SNPs in the human genome, and there is no guarantee that the *tag* SNPs on the array are the true casual SNPs[42]. The alleles of other SNPs can be imputed from tag SNPs thanks to the structure of linkage disequilibrium (LD) blocks present in chromosomes. Thus, when linking variants to target genes in a real-world setting, it is important to take into consideration neighbouring SNPs that are in high LD (e.g.: $r^2 > 0.8$) and inherited with the tag SNPs. Unfortunately, at the time of writing there is no straightforward way to perform this LD expansion step using R or Bioconductor packages, possibly because of the large amount of reference data required. The `ldblock` package[43] used to provide this functionality by downloading the HapMap data from the NCBI website, but the dataset was retired in 2016. At present, the best option to do this programmatically is probably to query the Ensembl REST API[44].

Annotation of coding and proximal SNPs to target genes

In order to annotate these variants, we need a a `TxDb` object, a reference of where transcripts are located on the genome. We can build this using the `GenomicFeatures`[45] package and the GENCODE v25 gene annotation:

```
library(GenomicFeatures)
# uncomment the following line to download file
#download.file("ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_25/gencode.v25.annotation.g
txdb <- makeTxDbFromGFF("gencode.v25.annotation.gff3.gz")
txdb <- keepStandardChromosomes(txdb)
```

We also have to convert the `gwasloc` object into a standard `GRanges` object:

```
snps <- GRanges(snps)
```

Let's check if the `gwasloc` and `TxDb` object use the same notation for chromosomes:

```
seqlevelsStyle(snps)
```

```
## [1] "UCSC"
```

```
seqlevelsStyle(txdb)
```

```
## [1] "UCSC"
```

OK, they do. Now we can annotate our SNPs to genes using the `VariantAnnotation`[46] package:

```
library(VariantAnnotation)
snps_anno <- locateVariants(snps, txdb, AllVariants())
snps_anno <- unique(snps_anno)
```

We use the `QUERYID` column in `snps_anno` to recover metadata such as SNP IDs and GWAS *p*-values from the original `snps` object:

```
snps_metadata <- snps[snps_anno$QUERYID]
mcols(snps_anno) <- cbind(mcols(snps_metadata)[c("SNPS", "P-VALUE")], mcols(snps_anno))
```

We can visualise where these SNPs are located (Figure 6):

```
loc <- data.frame(table(snps_anno$LOCATION))
ggplot(data = loc, aes(x = reorder(Var1, -Freq), y = Freq)) +
  geom_bar(stat = "identity") +
  xlab("Genomic location of SNPs") +
  ylab("Number of SNPs")
```

As expected[7], the great majority of SNPs are located within introns and in intergenic regions. For the moment, we will focus on SNPs that are either coding or in promoter and UTR regions, as these can be assigned to target genes rather unambiguously:

```
snps_easy <- subset(snps_anno, LOCATION == "coding" | LOCATION == "promoter" | LOCATION == "threeUTR
snps_easy <- as.data.frame(snps_easy)
```

Now we can check if any of the genes we found to be differentially expressed in SLE is also genetically associated with the disease:

```
snps_easy_in_degs <- merge(degs, snps_easy, by.x = "gene_id", by.y = "GENEID", all = FALSE)
```

We have 14 genes showing differential expression in SLE that are also genetically associated with the disease. While this is an interesting result, these hits are likely to be already well-known as potential SLE targets given their clear genetic association.

We will store essential information about these hits in a results `data.frame`:

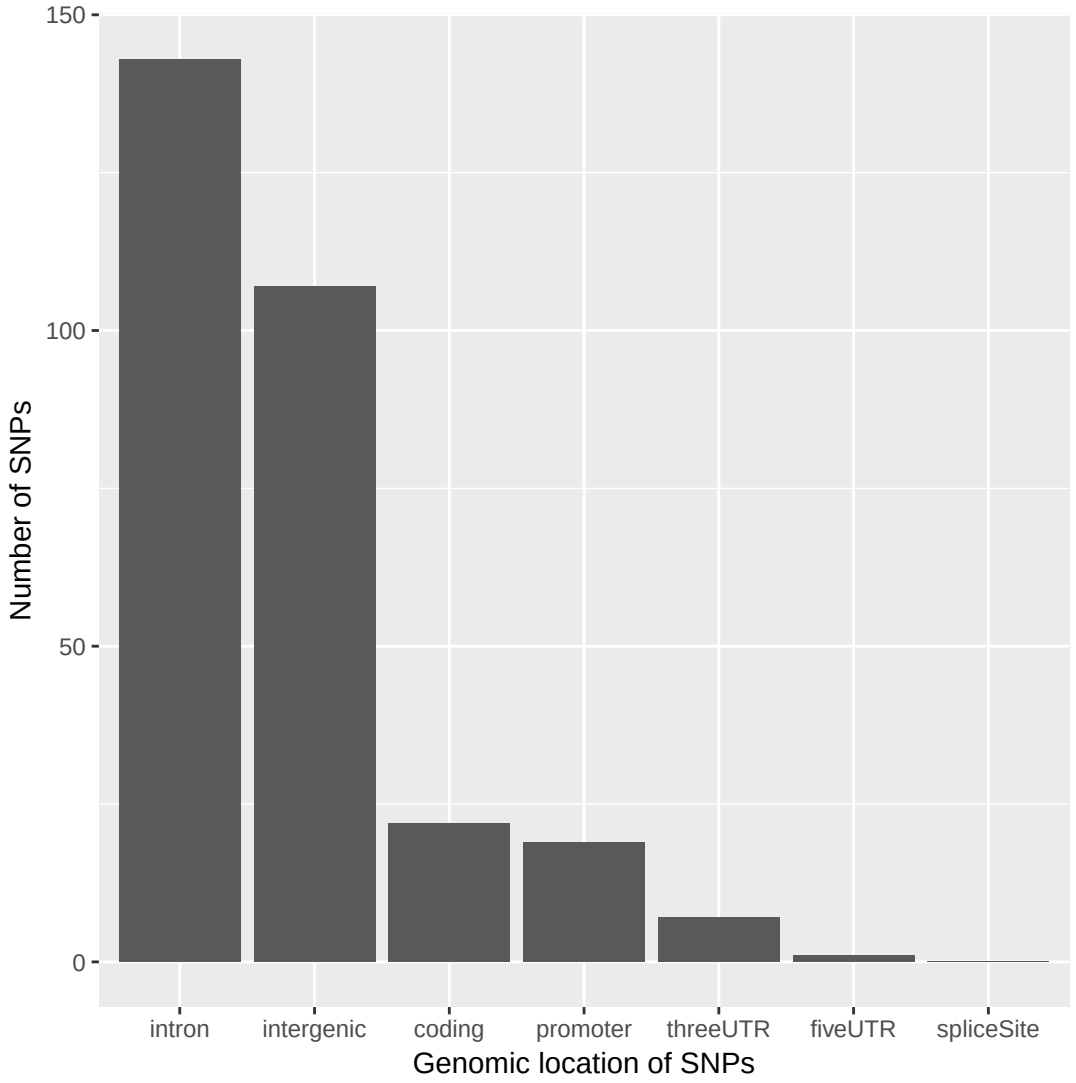

Figure 5. Bar plot showing genomic locations associated with SLE variants.

```
prioritised_hits <- unique(data.frame(
  snp_id = snps_easy_in_degs$SNPS,
  snp_pvalue = snps_easy_in_degs$P.VALUE,
  snp_location = snps_easy_in_degs$LOCATION,
  gene_id = snps_easy_in_degs$gene_id,
  gene_symbol = snps_easy_in_degs$symbol,
  gene_pvalue = snps_easy_in_degs$padj,
  gene_log2foldchange = snps_easy_in_degs$log2FoldChange,
  method = "Direct overlap",
  row.names = NULL))
head(prioritised_hits, 3)
```

```
##       snp_id snp_pvalue snp_location           gene_id gene_symbol
## 1  rs1887428      1e-06      fiveUTR ENSG00000096968.13        JAK2
## 2 rs58688157      5e-13     promoter ENSG00000099834.18       CDHR5
## 3  rs1990760      4e-08       coding  ENSG00000115267.5       IFIH1
##    gene_pvalue gene_log2foldchange         method
## 1 1.951160e-04            0.636590 Direct overlap
## 2 1.455662e-05            1.033372 Direct overlap
## 3 2.719420e-10            1.745324 Direct overlap
```

Use of regulatory genomic data to map intronic and intergenic SNPs to target genes

But what about all the SNPs in introns and intergenic regions? Some of those might be regulatory variants affecting the expression level of their target gene(s) through a distal enhancer. Let's create a dataset of candidate regulatory SNPs that are either intronic or intergenic and remove the annotation obtained with `VariantAnnotation`:

```
snps_hard <- subset(snps_anno, LOCATION == "intron" | LOCATION == "intergenic", select = c("SNPS", "
```

eQTL data

A well-established way to gain insights into target genes of regulatory SNPs is to use eQTL data, where correlations between genetic variants and expression of genes are computed across different tissues or cell types[9]. Here, we will simply match GWAS SNPs and eQTLs according to their genomic locations, which is a rather crude way to integrate these two types of data. More robust alternatives such as PrediXcan[47], TWAS[48] and SMR[49] exist and should be adopted if possible. One downside of these methods is that they require subject-level or complete summary data, making them less practical in some circumstances.

We will use blood eQTL data from the GTEx consortium[10]. To get the data, you will have to register and download the file `GTEx_Analysis_v7_eQTL.tar.gz` from the GTEx portal to the current working directory:

```
# uncomment the following line to extract the gzipped archive file
#untar("GTEx_Analysis_v7_eQTL.tar.gz")
gtex_blood <- read.delim(gzfile("GTEx_Analysis_v7_eQTL/Whole_Blood.v7.signif_variant_gene_pairs.txt.
head(gtex_blood, 3)
```

```
##           variant_id           gene_id tss_distance ma_samples ma_count
## 1 1_231153_CTT_C_b37 ENSG00000223972.4       219284         13       13
## 2    1_61920_G_A_b37 ENSG00000238009.2       -67303         18       20
## 3    1_64649_A_C_b37 ENSG00000238009.2       -64574         16       16
##         maf pval_nominal    slope slope_se pval_nominal_threshold
## 1 0.0191740  3.69025e-08 1.319720 0.233538            1.35366e-04
## 2 0.0281690  7.00836e-07 0.903786 0.178322            8.26088e-05
## 3 0.0220386  5.72066e-07 1.110040 0.217225            8.26088e-05
##   min_pval_nominal   pval_beta
## 1      3.69025e-08 4.67848e-05
## 2      6.50297e-10 1.11312e-06
## 3      6.50297e-10 1.11312e-06
```

We have to extract the genomic locations of the SNPs from the IDs used by GTEx:

```
locs <- strsplit(gtex_blood$variant_id, "_")
gtex_blood$chr <- sapply(locs, "[", 1)
gtex_blood$start <- sapply(locs, "[", 2)
gtex_blood$end <- sapply(locs, "[", 2)
```

We can then convert the `data.frame` into a `GRanges` object:

```
gtex_blood <- makeGRangesFromDataFrame(gtex_blood, keep.extra.columns = TRUE)
```

We also need to ensure that the chromosome notation is consistent with the previous objects:

```
seqlevelsStyle(gtex_blood)
```

```
## [1] "NCBI"    "Ensembl"
```

```
seqlevelsStyle(gtex_blood) <- "UCSC"
```

From the publication[10], we know the genomic coordinates are mapped to genome reference GRCh37, so we will have to uplift them to GRCh38 using `rtracklayer`[50] and a mapping ("chain") file. The `R.utils` package is only required to extract the gzipped file:

```
library(rtracklayer)
library(R.utils)
# uncomment the following line to download file
#download.file("http://hgdownload.cse.ucsc.edu/goldenPath/hg19/liftOver/hg19ToHg38.over.chain.gz", d
# uncomment the following line to extract gzipped file
#gunzip("hg19ToHg38.over.chain.gz")
ch <- import.chain("hg19ToHg38.over.chain")
gtex_blood <- unlist(liftOver(gtex_blood, ch))
```

We will use the `GenomicRanges` package[45] to compute the overlap between GWAS SNPs and blood eQTLs:

```
library(GenomicRanges)
hits <- findOverlaps(snps_hard, gtex_blood)
snps_hard_in_gtex_blood = snps_hard[queryHits(hits)]
gtex_blood_with_snps_hard = gtex_blood[subjectHits(hits)]
mcols(snps_hard_in_gtex_blood) <- cbind(mcols(snps_hard_in_gtex_blood), mcols(gtex_blood_with_snps_h
snps_hard_in_gtex_blood <- as.data.frame(snps_hard_in_gtex_blood)
```

We have 59 blood eQTL variants that are associated with SLE. We can now check whether any of the genes differentially expressed in SLE is an *eGene*, a gene whose expression is influenced by an eQTL. Note that gene IDs in GTEx are mapped to GENCODE v19[10], while we are using the newer v25 for the DEGs. To match the gene IDs in the two objects, we will simply strip the last bit containing the GENCODE gene version, which effectively gives us Ensembl gene IDs:

```
snps_hard_in_gtex_blood$ensembl_id <- sub("(ENSG[0-9]+)\\.[0-9]+", "\\1", snps_hard_in_gtex_blood$ge
degs$ensembl_id <- sub("(ENSG[0-9]+)\\.[0-9]+", "\\1", degs$gene_id)
snps_hard_in_gtex_blood_in_degs <- merge(snps_hard_in_gtex_blood, degs, by = "ensembl_id", all = FAI
```

We can add these 17 genes to our list:

```
prioritised_hits <- unique(rbind(prioritised_hits, data.frame(
  snp_id = snps_hard_in_gtex_blood_in_degs$SNPS,
  snp_pvalue = snps_hard_in_gtex_blood_in_degs$P.VALUE,
  snp_location = snps_hard_in_gtex_blood_in_degs$LOCATION,
  gene_id = snps_hard_in_gtex_blood_in_degs$gene_id.y,
  gene_symbol = snps_hard_in_gtex_blood_in_degs$symbol,
  gene_pvalue = snps_hard_in_gtex_blood_in_degs$padj,
  gene_log2foldchange = snps_hard_in_gtex_blood_in_degs$log2FoldChange,
  method = "GTEx eQTLs",
  row.names = NULL)))
```

FANTOM5 data

The FANTOM consortium profiled gene expression across a large panel of tissues and cell types using CAGE [15;17]. This technology allows mapping of transcription start sites and enhancer RNAs genome-wide. Correlations between these promoter and enhancer elements across a large panel of tissues and cell types can then be calculated to identify significant promoter - enhancer pairs. In turn, we will use these correlations to map distal regulatory SNPs to target genes.

Let's read in the enhancer - promoter correlation data:

```
# uncomment the following line to download the file
#download.file("http://enhancer.binf.ku.dk/presets/enhancer_tss_associations.bed", destfile = "enhan
fantom <- read.delim("enhancer_tss_associations.bed", skip = 1, stringsAsFactors = FALSE)
head(fantom, 3)
```

```
##   X.chrom chromStart chromEnd
## 1    chr1     858252   861621
## 2    chr1     894178   956888
## 3    chr1     901376   956888
##                                                                name
## 1                   chr1:858256-858648;NM_152486;SAMD11;R:0.404;FDR:0
## 2 chr1:956563-956812;NM_015658;NOC2L;R:0.202;FDR:8.01154668254404e-08
## 3     chr1:956563-956812;NM_001160184,NM_032129;PLEKHN1;R:0.422;FDR:0
##   score strand thickStart thickEnd itemRgb blockCount blockSizes
## 1   404      .     858452   858453   0,0,0          2   401,1001
## 2   202      .     956687   956688   0,0,0          2   1001,401
## 3   422      .     956687   956688   0,0,0          2   1001,401
##   chromStarts
## 1      0,2368
## 2     0,62309
## 3     0,55111
```

Everything we need is in the fourth column, `name`: genomic location of the enhancer, gene identifiers, Pearson correlation coefficient and significance. We will use the `splitstackshape` package to parse it:

```
library(splitstackshape)
fantom <- as.data.frame(cSplit(fantom, splitCols = "name", sep = ";", direction = "wide"))
```

Now we can extract the genomic locations of the enhancers and the correlation values:

```
locs <- strsplit(as.character(fantom$name_1), "[:-]")
fantom$chr <- sapply(locs, "[", 1)
fantom$start <- as.numeric(sapply(locs, "[", 2))
fantom$end <- as.numeric(sapply(locs, "[", 3))
fantom$symbol <- fantom$name_3
fantom$corr <- sub("R:", "", fantom$name_4)
fantom$fdr <- sub("FDR:", "", fantom$name_5)
```

We can select only the enhancer - promoter pairs with a decent level of correlation and significance and tidy the data at the same time:

```
fantom <- unique(subset(fantom, corr >= 0.25 & fdr < 1e-5, select = c("chr", "start", "end", "symbol
```

Now we would like to check whether any of our candidate regulatory SNPs are falling in any of these enhancers. To do this, we have to convert the `data.frame` into a `GRanges` object and uplift the GRCh37 coordinates [15] to GRCh38:

```
fantom <- makeGRangesFromDataFrame(fantom, keep.extra.columns = TRUE)
fantom <- unlist(liftOver(fantom, ch))
```

We can now compute the overlap between SNPs and enhancers:

```
hits <- findOverlaps(snps_hard, fantom)
snps_hard_in_fantom = snps_hard[queryHits(hits)]
fantom_with_snps_hard = fantom[subjectHits(hits)]
mcols(snps_hard_in_fantom) <- cbind(mcols(snps_hard_in_fantom), mcols(fantom_with_snps_hard))
snps_hard_in_fantom <- as.data.frame(snps_hard_in_fantom)
```

We can now check if any of these genes is differentially expressed in our RNA-seq data:

```
snps_hard_in_fantom_in_degs <- merge(snps_hard_in_fantom, degs, by = "symbol", all = FALSE)
```

We have identified 7 genes whose putative enhancers contain SLE GWAS SNPs. Let's add these to our list:

```
prioritised_hits <- unique(rbind(prioritised_hits, data.frame(
  snp_id = snps_hard_in_fantom_in_degs$SNPS,
  snp_pvalue = snps_hard_in_fantom_in_degs$P.VALUE,
  snp_location = snps_hard_in_fantom_in_degs$LOCATION,
  gene_id = snps_hard_in_fantom_in_degs$gene_id,
  gene_symbol = snps_hard_in_fantom_in_degs$symbol,
  gene_pvalue = snps_hard_in_fantom_in_degs$padj,
  gene_log2foldchange = snps_hard_in_fantom_in_degs$log2FoldChange,
  method = "FANTOM5 correlations",
  row.names = NULL)))
```

Promoter Capture Hi-C data

More recently, chromatin interaction data was generated across 17 human primary blood cell types using promoter capture Hi-C[21]. More than 30,000 promoter baits were used to capture promoter-interacting regions genome-wide, which were then mapped to enhancers based on annotation present in the Ensembl Regulatory Build[51]. This dataset provides a valuable resource for interpreting complex genomic data, especially in the context of autoimmune diseases (and other conditions where immune cells play a role). Significant interactions between enhancers and promoters can be accessed in the supplementary data of the paper:

```
# uncomment the following line to download file
#download.file("http://www.cell.com/cms/attachment/2086554122/2074217047/mmc4.zip", destfile = "mmc4
# uncomment the following lines to extract zipped files
#unzip("mmc4.zip")
#unzip("DATA_S1.zip")
pchic <- read.delim("ActivePromoterEnhancerLinks.tsv", stringsAsFactors = FALSE)
head(pchic, 3)
```

```
##   baitChr  baitSt baitEnd baitID oeChr    oeSt   oeEnd oeID
## 1    chr1 1206873 1212438    254  chr1  943676  957199  228
## 2    chr1 1206873 1212438    254  chr1 1034268 1040208  235
## 3    chr1 1206873 1212438    254  chr1 1040208 1043143  236
##                       cellType.s.
## 1                            nCD8
## 2 nCD4,nCD8,Mac0,Mac1,Mac2,MK,Mon
## 3     nCD4,nCD8,Mac0,Mac1,Mac2,MK
##
## 1
## 2 S007DDH2,S007G7H4,C0066PH1,S00C2FH1,S00390H1,S001MJH1,S001S7H2,S0022IH2,S00622H1,S00BS4H1,S004
## 3          S007DDH2,S007G7H4,C0066PH1,S00C2FH1,S00390H1,S001MJH1,S001S7H2,S0022IH2,S00622H1,S00BS
```

We will use the `InteractionSet` package[52], which is specifically designed for the representation of chromatin interaction data. We start by creating a `GInteractions` object:

```
library(InteractionSet)
promoters <- GRanges(seqnames = pchic$baitChr, ranges = IRanges(start = pchic$baitSt, end = pchic$ba
enhancers <- GRanges(seqnames = pchic$oeChr, ranges = IRanges(start = pchic$oeSt, end = pchic$oeEnd)
pchic <- GInteractions(promoters, enhancers)
```

As gene identifiers are not provided, we also have to map promoters to the respective genes so that we know which genes are regulated by which enhancers. We can do this by using the `TxDb` object we previously built to extract positions of transcription start sites (TSSs) and then add the GENCODE gene IDs as metadata to the `pchic` object:

```
tsss <- promoters(txdb, upstream = 0, downstream = 1, columns = "gene_id")
hits <- nearest(promoters, tsss)
pchic$gene_id <- unlist(tsss[hits]$gene_id)
```

Next, we calculate the overlaps between SLE GWAS SNPs and enhancers (the *second* region of the `GInteractions` object) :

```
hits <- findOverlaps(snps_hard, pchic, use.region = "second")
snps_hard_in_pchic = snps_hard[queryHits(hits)]
pchic_with_snps_hard = pchic[subjectHits(hits)]
mcols(snps_hard_in_pchic) <- cbind(mcols(snps_hard_in_pchic), mcols(pchic_with_snps_hard))
snps_hard_in_pchic <- as.data.frame(snps_hard_in_pchic)
```

We check if any of these enhancers containing SLE variants are known to putatively regulate genes differentially expressed in SLE:

```
snps_hard_in_pchic_in_degs <- merge(snps_hard_in_pchic, degs, by = "gene_id", all = FALSE)
```

And finally we add these 13 genes to our list:

```
prioritised_hits <- unique(rbind(prioritised_hits, data.frame(
  snp_id = snps_hard_in_pchic_in_degs$SNPS,
  snp_pvalue = snps_hard_in_pchic_in_degs$P.VALUE,
  snp_location = snps_hard_in_pchic_in_degs$LOCATION,
  gene_id = snps_hard_in_pchic_in_degs$gene_id,
  gene_symbol = snps_hard_in_pchic_in_degs$symbol,
  gene_pvalue = snps_hard_in_pchic_in_degs$padj,
  gene_log2foldchange = snps_hard_in_pchic_in_degs$log2FoldChange,
  method = "Promoter capture Hi-C",
  row.names = NULL)))
```

These are the final results of our target identification exercise. We can have a look at the most significant SNPs mapped with each of the methods:

```
top_prioritised_hits <- prioritised_hits[order(prioritised_hits$snp_pvalue),]
top_prioritised_hits <- split(top_prioritised_hits, top_prioritised_hits$method)
do.call(rbind, lapply(top_prioritised_hits, head, 1))
```

```
##                          snp_id snp_pvalue snp_location            gene_id
## Direct overlap        rs3757387      1e-48     promoter ENSG00000128604.18
## GTEx eQTLs            rs1270942     2e-165       intron ENSG00000166278.14
## FANTOM5 correlations  rs1150754      6e-29       intron  ENSG00000204421.2
## Promoter capture Hi-C rs1270942     2e-165       intron  ENSG00000219797.2
##                       gene_symbol  gene_pvalue gene_log2foldchange
## Direct overlap               IRF5 5.006707e-03           0.4041349
## GTEx eQTLs                     C2 1.625111e-03           0.9269526
## FANTOM5 correlations       LY6G6C 3.575357e-05           1.4327915
## Promoter capture Hi-C          NA 1.919459e-04           0.4556364
##                                      method
## Direct overlap                Direct overlap
## GTEx eQTLs                        GTEx eQTLs
## FANTOM5 correlations   FANTOM5 correlations
## Promoter capture Hi-C Promoter capture Hi-C
```

We can also visualise the relative contributions from the different approaches we used (Figure 7):

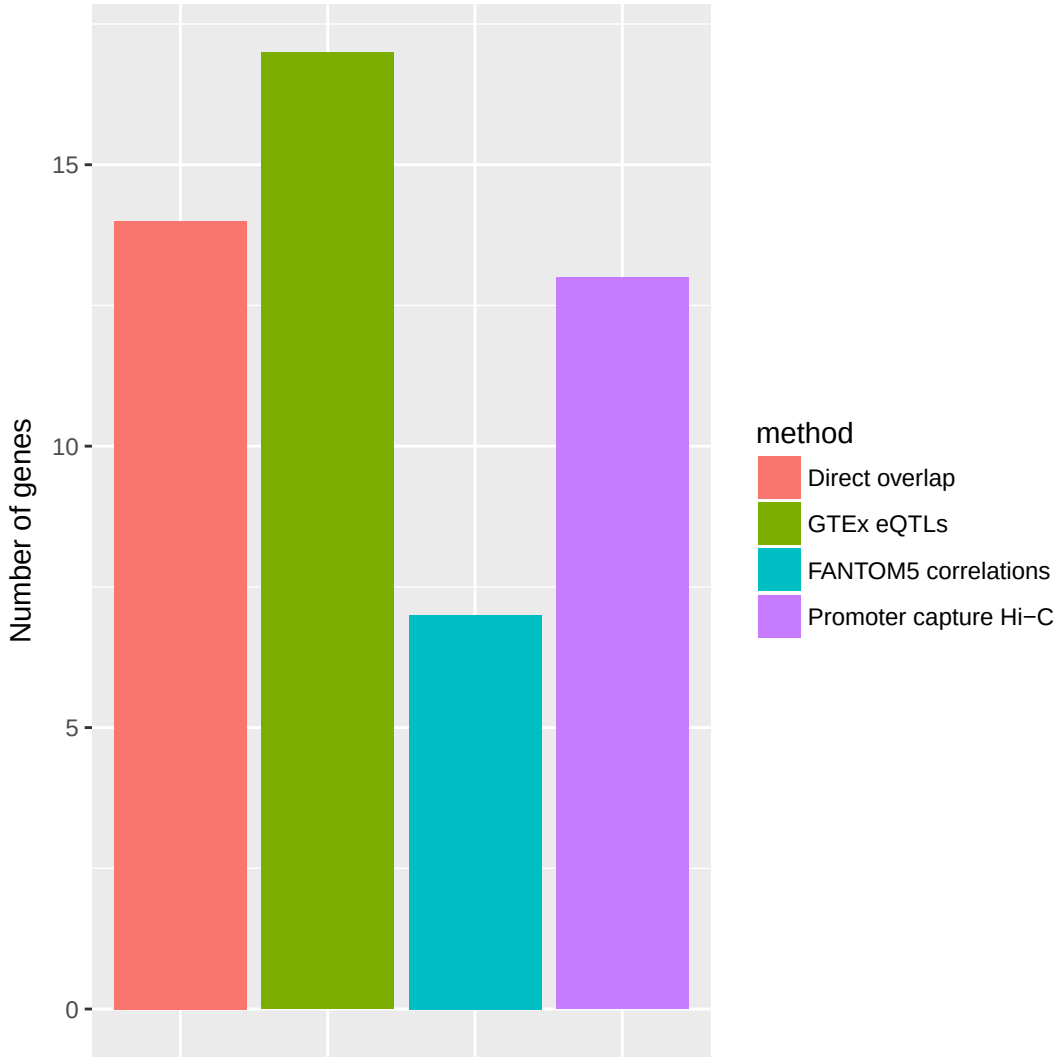

Figure 6. Bar plot showing number of genes identified by each variant mapping method.

```
prioritised_genes <- unique(data.frame(gene_id = prioritised_hits$gene_id, method = prioritised_hits
ggplot(data = prioritised_genes, aes(x = method)) +
  geom_bar(aes(fill = method), stat = "count") +
  ylab("Number of genes") +
  theme(axis.title.x = element_blank(),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank())
```

We observe that all methods significantly contributed to the identification of genes associated with GWAS SNPs. The majority of genes were identified through the integration of the GTEx blood eQTL data, followed by the methods based on direct overlap, promoter capture Hi-C data and FANTOM5 correlations.

Functional analysis of priortised hits

We will use biological processes from the Gene Ontology[53] and the `clusterProfiler` package[54] to functionally characterise our list of genes:

```
library(clusterProfiler)
prioritised_hits_ensembl_ids <- unique(sub("(ENSG[0-9]+)\\.[0-9]+", "\\1", prioritised_hits$gene_id)
all_genes_ensembl_ids <- unique(sub("(ENSG[0-9]+)\\.[0-9]+", "\\1", rownames(rse)))
gobp_enrichment <- enrichGO(prioritised_hits_ensembl_ids,
                            universe = all_genes_ensembl_ids,
                            OrgDb = org.Hs.eg.db,
```

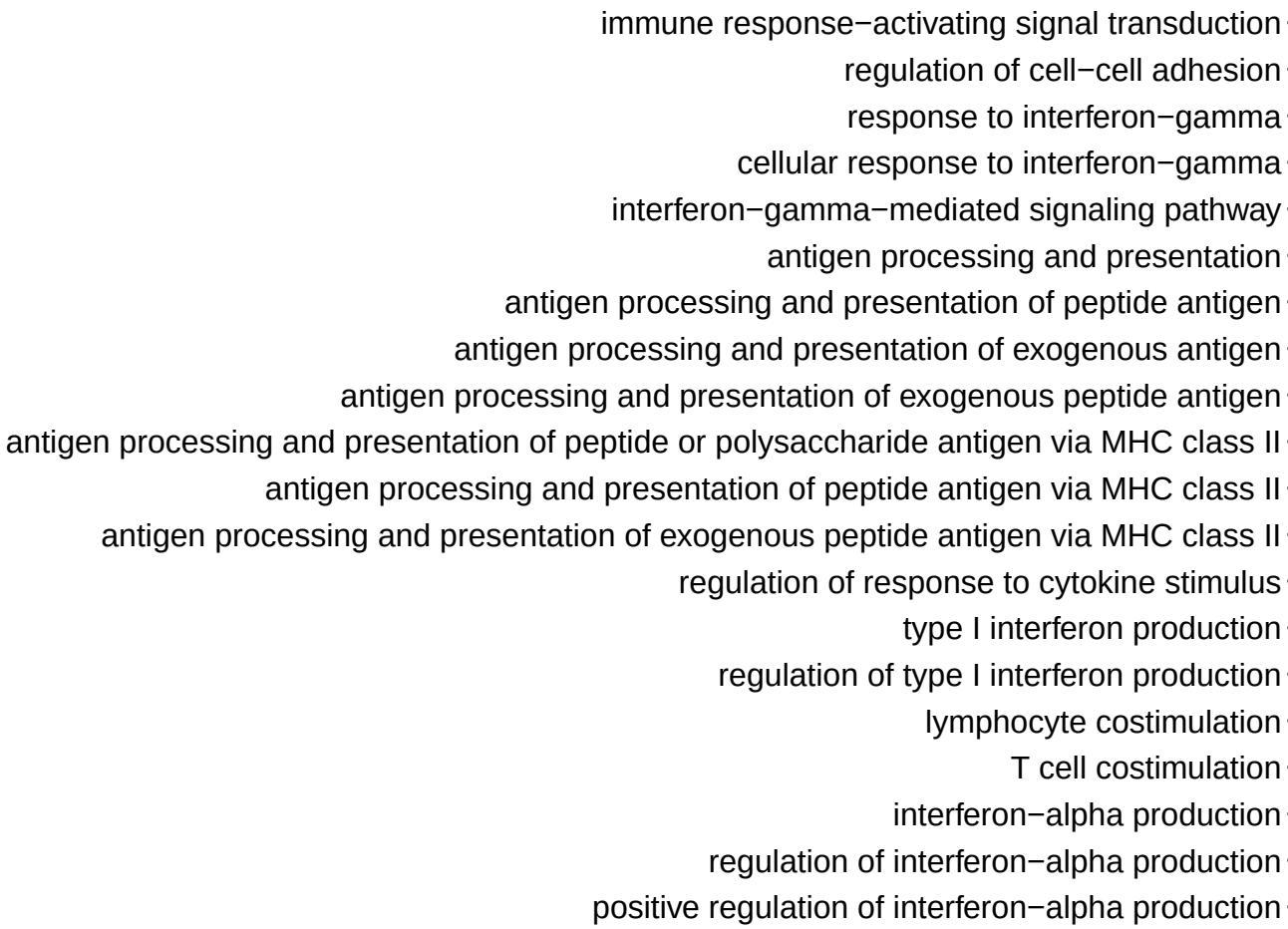

Figure 7. Dot plot showing enrichment of Gene Ontology biological processes for the list of prioritised genes.

```
                                keyType = "ENSEMBL",
                                ont = "BP",
                                pAdjustMethod = "BH",
                                pvalueCutoff = 0.05,
                                qvalueCutoff = 0.05,
                                readable = TRUE)
```

We can visualise the most enriched terms (Figure 8):

```
dotplot(gobp_enrichment, showCategory = 20)
```

We observe a significant enrichment for interferon responses, antigen processing and presentation, and T cell stimulation, all processes which are well-known to play key roles in the pathogenesis of SLE[55;56;57].

From a drug discovery perspective, JAK2 is probably the most attractive target: rs1887428 (p-value = 1 x 10^{-6}) is located in its 5' UTR and the genes is significantly upregulated in disease. Tofacitinib, a pan-JAK inhibitor, showed promising results in mouse[58] and is currently being tested or safety in a phase I clinical trial. We find 7 GWAS SNPs that are blood eQTLs linked to the expression of C2, a protease active in the complement signalling cascade. The most significant variant is rs1270942 (p-value = 2 x 10^{-165}) and is found in an intron of CFB, another component of the complement system. As with other autoimmune diseases, the complement plays a key role SLE in and has been investigated as a therapeutic approach[59]. Another potentially interesting hit is TAX1BP1: rs849142 (p-value = 1 x 9^{-11}) is found within an intron of JAZF1, but can be linked to TAX1BP1

via a chromatin interaction with its promoter. TAX1BP1 inhibits TNF-induced apoptosis[60] and is involved in the IL1 signalling cascade[61], another relevant pathway in SLE that could be therapeutically targeted[62].

Conclusions

In this Bioconductor workflow we have used several packages and datasets to demonstrate how regulatory genomic data can be used to annotate significant hits from GWASs and prioritise gene lists from expression studies, providing an intermediate layer connecting genetics and transcriptomics. Overall, we identified 46 SLE-associated SNPs that we mapped to 49 genes differentially expressed in SLE, using eQTL data[10] and enhancer - promoter relationships from CAGE[15] and promoter capture Hi-C experiments[21]. These genes are involved in key inflammatory signalling pathways and some of them could develop into therapeutic targets for SLE. While options for the visualisations of genomic data and interactions are outside the scope of this workflow, at least three good alternatives exist in Bioconductor: ggbio[63], Sushi[64] and Gviz[65] coupled with the GenomicInteractions package[66]. We refer the reader to these publications and package vignettes for examples.

While simplified, the workflow also demonstrates some real-world challenges encountered when working with genomic data from different sources, such as the use of different genome assemblies and gene annotation systems, the parsing of files with custom formats into Bioconductor objects and the mapping of genomic locations to genes.

As the sample size and power of GWASs and gene expression studies continue to increase, it will become more and more challenging to identify truly significant hits and interpret them. The use of regulatory genomics data as presented here can be an important tool to gain insights into large biomedical datasets and help in the identification of biomarkers and therapeutic targets.

Abbreviations

CAGE: cap analysis of gene expression
DHS: DNase I hypersensitive site
eQTL: expression quantitative trait locus
GWAS: genome-wide association study
ISM: interferon signature metric
SLE: systemic lupus erythematosus
SNP: single nucleotide polymorphism

Data and software availability

Download links for all datasets are part of the workflow.
Software packages required to reproduce the analysis can be installed as part of the workflow.
Source code is available at: `https://github.com/enricoferrero/bioconductor-regulatory-genomics-workflow`.
Archived source code as at the time of publication is available at: `https://doi.org/10.5281/zenodo.1154124`.

Competing interests

EF is a full time employee of GSK.

Grant information

The author declared that no grants were involved in supporting this work.

References

[1] Michael J. Waring, John Arrowsmith, Andrew R. Leach, Paul D. Leeson, Sam Mandrell, Robert M. Owen, Garry Pairaudeau, William D. Pennie, Stephen D. Pickett, Jibo Wang, Owen Wallace, and Alex Weir. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature reviews. Drug discovery*, 14(7):475–86, jul 2015. ISSN 1474-1784.

[2] Joseph A. DiMasi, Henry G. Grabowski, and Ronald W. Hansen. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of health economics*, 47:20–33, may 2016. ISSN 1879-1646.

[3] Richard K Harrison. Phase II and phase III failures: 2013-2015. *Nature reviews. Drug discovery*, 15(12):817–818, dec 2016. ISSN 1474-1784.

[4] David Cook, Dearg Brown, Robert Alexander, Ruth March, Paul Morgan, Gemma Satterthwaite, and Menelas N Pangalos. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nature reviews. Drug discovery*, 13(6):419–31, jun 2014. ISSN 1474-1784.

[5] Robert M. Plenge, Edward M. Scolnick, and David Altshuler. Validating therapeutic targets through human genetics. *Nature reviews. Drug discovery*, 12(8):581–94, aug 2013. ISSN 1474-1784.

[6] Matthew R. Nelson, Hannah Tipney, Jeffery L. Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, Pak Chung Sham, Mulin Jun Li, Junwen Wang, Lon R. Cardon, John C. Whittaker, and Philippe Sanseau. The support of human genetic evidence for approved drug indications. *Nature genetics*, 47(8):856–60, aug 2015. ISSN 1546-1718.

[7] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K Johnson, Theresa K Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R Scott Hansen, Shane Neph, Peter J Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R Sunyaev, Rajinder Kaul, and John A Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science (New York, N.Y.)*, 337(6099):1190–5, sep 2012. ISSN 1095-9203.

[8] Lucas D Ward and Manolis Kellis. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology*, 30(11):1095–106, nov 2012. ISSN 1546-1696.

[9] Frank W. Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature reviews. Genetics*, 16(4):197–212, apr 2015. ISSN 1471-0064.

[10] GTEx Consortium, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group Laboratory, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, University of California Santa Cruz Genome Browser Data Integration &Visualization—UCSC Genomics Institute, Lead analysts:, Data Analysis &Coordinating Center (LDACC): Laboratory, NIH program management:, Biospecimen collection:, Pathology:, EQTL manuscript working group:, Alexis Battle, Christopher D. Brown, Barbara E. Engelhardt, and Stephen B. Montgomery. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, oct 2017. ISSN 1476-4687.

[11] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414): 57–74, sep 2012. ISSN 1476-4687.

[12] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W. Whitaker, Matthew D. Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B. Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K. Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S. Bansal, Annaick Carles, Jesse R. Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C. Sallari, Kyle T. Siebenthall, Nicholas A. Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fisher, David Haussler, Steven J. M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil Sunyaev, James A. Thomson, Thea D. Tlsty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael Q. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30, feb 2015. ISSN 1476-4687.

[13] David Adams, Lucia Altucci, Stylianos E Antonarakis, Juan Ballesteros, Stephan Beck, Adrian Bird, Christoph Bock, Bernhard Boehm, Elias Campo, Andrea Caricasole, Fredrik Dahl, Emmanouil T Dermitzakis, Tariq Enver, Manel Esteller, Xavier Estivill, Anne Ferguson-Smith, Jude Fitzgibbon, Paul Flicek, Claudia Giehl, Thomas Graf, Frank Grosveld, Roderic Guigo, Ivo Gut, Kristian Helin, Jonas Jarvius, Ralf Küppers, Hans Lehrach, Thomas Lengauer, Åke Lernmark, David Leslie, Markus Loeffler, Elizabeth Macintyre, Antonello Mai, Joost H A Martens, Saverio Minucci, Willem H Ouwehand, Pier Giuseppe Pelicci, Hèléne Pendeville, Bo Porse, Vardhman Rakyan, Wolf Reik, Martin Schrappe, Dirk Schübeler, Martin Seifert, Reiner Siebert, David Simmons, Nicole Soranzo, Salvatore Spicuglia, Michael Stratton, Hendrik G Stunnenberg, Amos Tanay, David Torrents, Alfonso Valencia, Edo Vellenga, Martin Vingron, Jörn Walter, and Spike Willcocks. BLUEPRINT to decode the epigenetic signature written in blood. *Nature biotechnology*, 30(3):224–6, mar 2012. ISSN 1546-1696.

[14] Hendrik G. Stunnenberg, International Human Epigenome Consortium, and Martin Hirst. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell*, 167(7):1897, dec 2016. ISSN 1097-4172.

[15] FANTOM Consortium and the RIKEN PMI and CLST (DGT), Alistair R R Forrest, Hideya Kawaji, Michael Rehli, J Kenneth Baillie, Michiel J L de Hoon, Vanja Haberle, Timo Lassmann, Ivan V Kulakovskiy, Marina Lizio, Masayoshi Itoh, Robin Andersson, Christopher J Mungall, Terrence F Meehan, Sebastian Schmeier, Nicolas Bertin, Mette Jørgensen, Emmanuel Dimont, Erik Arner, Christian Schmidl, Ulf Schaefer, Yulia A Medvedeva, Charles Plessy, Morana Vitezic, Jessica Severin, Colin A Semple, Yuri Ishizu, Robert S Young, Margherita Francescatto, Intikhab Alam, Davide Albanese, Gabriel M Altschuler, Takahiro Arakawa, John A C Archer, Peter Arner, Magda Babina, Sarah Rennie, Piotr J Balwierz, Anthony G Beckhouse, Swati Pradhan-Bhatt, Judith A Blake, Antje Blumenthal, Beatrice Bodega, Alessandro Bonetti, James Briggs, Frank Brombacher, A Maxwell Burroughs, Andrea Califano, Carlo V Cannistraci, Daniel Carbajo, Yun Chen, Marco Chierici, Yari Ciani, Hans C Clevers, Emiliano Dalla, Carrie A Davis, Michael Detmar, Alexander D Diehl, Taeko Dohi, Finn Drabløs, Albert S B Edge, Matthias Edinger, Karl Ekwall, Mitsuhiro Endoh, Hideki Enomoto,

Michela Fagiolini, Lynsey Fairbairn, Hai Fang, Mary C Farach-Carson, Geoffrey J Faulkner, Alexander V Favorov, Malcolm E Fisher, Martin C Frith, Rie Fujita, Shiro Fukuda, Cesare Furlanello, Masaaki Furino, Jun-ichi Furusawa, Teunis B Geijtenbeek, Andrew P Gibson, Thomas Gingeras, Daniel Goldowitz, Julian Gough, Sven Guhl, Reto Guler, Stefano Gustincich, Thomas J Ha, Masahide Hamaguchi, Mitsuko Hara, Matthias Harbers, Jayson Harshbarger, Akira Hasegawa, Yuki Hasegawa, Takehiro Hashimoto, Meenhard Herlyn, Kelly J Hitchens, Shannan J Ho Sui, Oliver M Hofmann, Ilka Hoof, Furni Hori, Lukasz Huminiecki, Kei Iida, Tomokatsu Ikawa, Boris R Jankovic, Hui Jia, Anagha Joshi, Giuseppe Jurman, Bogumil Kaczkowski, Chieko Kai, Kaoru Kaida, Ai Kaiho, Kazuhiro Kajiyama, Mutsumi Kanamori-Katayama, Artem S Kasianov, Takeya Kasukawa, Shintaro Katayama, Sachi Kato, Shuji Kawaguchi, Hiroshi Kawamoto, Yuki I Kawamura, Tsugumi Kawashima, Judith S Kempfle, Tony J Kenna, Juha Kere, Levon M Khachigian, Toshio Kitamura, S Peter Klinken, Alan J Knox, Miki Kojima, Soichi Kojima, Naoto Kondo, Haruhiko Koseki, Shigeo Koyasu, Sarah Krampitz, Atsutaka Kubosaki, Andrew T Kwon, Jeroen F J Laros, Weonju Lee, Andreas Lennartsson, Kang Li, Berit Lilje, Leonard Lipovich, Alan Mackay-Sim, Ri-ichiroh Manabe, Jessica C Mar, Benoit Marchand, Anthony Mathelier, Niklas Mejhert, Alison Meynert, Yosuke Mizuno, David A de Lima Morais, Hiromasa Morikawa, Mitsuru Morimoto, Kazuyo Moro, Efthymios Motakis, Hozumi Motohashi, Christine L Mummery, Mitsuyoshi Murata, Sayaka Nagao-Sato, Yutaka Nakachi, Fumio Nakahara, Toshiyuki Nakamura, Yukio Nakamura, Kenichi Nakazato, Erik van Nimwegen, Noriko Ninomiya, Hiromi Nishiyori, Shohei Noma, Shohei Noma, Tadasuke Noazaki, Soichi Ogishima, Naganari Ohkura, Hiroko Ohimiya, Hiroshi Ohno, Mitsuhiro Ohshima, Mariko Okada-Hatakeyama, Yasushi Okazaki, Valerio Orlando, Dmitry A Ovchinnikov, Arnab Pain, Robert Passier, Margaret Patrikakis, Helena Persson, Silvano Piazza, James G D Prendergast, Owen J L Rackham, Jordan A Ramilowski, Mamoon Rashid, Timothy Ravasi, Patrizia Rizzu, Marco Roncador, Sugata Roy, Morten B Rye, Eri Saijyo, Antti Sajantila, Akiko Saka, Shimon Sakaguchi, Mizuho Sakai, Hiroki Sato, Suzana Savvi, Alka Saxena, Claudio Schneider, Erik A Schultes, Gundula G Schulze-Tanzil, Anita Schwegmann, Thierry Sengstag, Guojun Sheng, Hisashi Shimoji, Yishai Shimoni, Jay W Shin, Christophe Simon, Daisuke Sugiyama, Takaai Sugiyama, Masanori Suzuki, Naoko Suzuki, Rolf K Swoboda, Peter A C 't Hoen, Michihira Tagami, Naoko Takahashi, Jun Takai, Hiroshi Tanaka, Hideki Tatsukawa, Zuotian Tatum, Mark Thompson, Hiroo Toyodo, Tetsuro Toyoda, Elvind Valen, Marc van de Wetering, Linda M van den Berg, Roberto Verado, Dipti Vijayan, Ilya E Vorontsov, Wyeth W Wasserman, Shoko Watanabe, Christine A Wells, Louise N Winteringham, Ernst Wolvetang, Emily J Wood, Yoko Yamaguchi, Masayuki Yamamoto, Misako Yoneda, Yohei Yonekura, Shigehiro Yoshida, Susan E Zabierowski, Peter G Zhang, Xiaobei Zhao, Silvia Zucchelli, Kim M Summers, Harukazu Suzuki, Carsten O Daub, Jun Kawai, Peter Heutink, Winston Hide, Tom C Freeman, Boris Lenhard, Vladimir B Bajic, Martin S Taylor, Vsevolod J Makeev, Albin Sandelin, David A Hume, Piero Carninci, and Yoshihide Hayashizaki. A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–70, mar 2014. ISSN 1476-4687.

[16] Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, Kavita Garg, Sam John, Richard Sandstrom, Daniel Bates, Lisa Boatman, Theresa K Canfield, Morgan Diegel, Douglas Dunn, Abigail K Ebersol, Tristan Frum, Erika Giste, Audra K Johnson, Ericka M Johnson, Tanya Kutyavin, Bryan Lajoie, Bum-Kyu Lee, Kristen Lee, Darin London, Dimitra Lotakis, Shane Neph, Fidencio Neri, Eric D Nguyen, Hongzhu Qu, Alex P Reynolds, Vaughn Roach, Alexias Safi, Minerva E Sanchez, Amartya Sanyal, Anthony Shafer, Jeremy M Simon, Lingyun Song, Shinny Vong, Molly Weaver, Yongqi Yan, Zhancheng Zhang, Zhuzhu Zhang, Boris Lenhard, Muneesh Tewari, Michael O Dorschner, R Scott Hansen, Patrick a Navas, George Stamatoyannopoulos, Vishwanath R Iyer, Jason D Lieb, Shamil R Sunyaev, Joshua M Akey, Peter J Sabo, Rajinder Kaul, Terrence S Furey, Job Dekker, Gregory E Crawford, and John a Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, sep 2012. ISSN 1476-4687.

[17] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, Evgenia Ntini, Erik Arner, Eivind Valen, Kang Li, Lucia Schwarzfischer, Dagmar Glatz, Johanna Raithel, Berit Lilje, Nicolas Rapin, Frederik Otzen Bagger, Mette Jørgensen, Peter Refsing Andersen, Nicolas Bertin, Owen Rackham, a Maxwell Burroughs, J Kenneth Baillie, Yuri Ishizu, Yuri Shimizu, Erina Furuhata, Shiori Maeda, Yutaka Negishi, Christopher J Mungall, Terrence F Meehan, Timo Lassmann, Masayoshi Itoh, Hideya Kawaji, Naoto Kondo, Jun Kawai, Andreas Lennartsson, Carsten O Daub, Peter Heutink, David a Hume, Torben Heick Jensen, Harukazu Suzuki, Yoshihide Hayashizaki, Ferenc Müller, Alistair R R Forrest, Piero Carninci, Michael Rehli, and Albin Sandelin. An atlas of active enhancers across human cell types and tissues. *Nature*, 507 (7493):455–461, mar 2014. ISSN 1476-4687.

[18] Melissa J. Fullwood, Chia-Lin Wei, Edison T. Liu, and Yijun Ruan. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome research*, 19(4):521–32, apr 2009. ISSN 1088-9051.

[19] Yubo Zhang, Chee-Hong Wong, Ramon Y. Birnbaum, Guoliang Li, Rebecca Favaro, Chew Yee Ngan, Joanne Lim, Eunice Tai, Huay Mei Poh, Eleanor Wong, Fabianus Hendriyan Mulawadi, Wing-Kin Sung, Silvia Nicolis, Nadav Ahituv, Yijun Ruan, and Chia-Lin Wei. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature*, 504(7479):306–310, dec 2013. ISSN 1476-4687.

[20] Borbala Mifsud, Filipe Tavares-Cadete, Alice N Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W Wingett, Simon Andrews, William Grey, Philip A Ewels, Bram Herman, Scott Happe, Andy Higgs, Emily LeProust, George A Follows, Peter Fraser, Nicholas M Luscombe, and Cameron S Osborne. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature genetics*, 47(6):598–606, jun 2015. ISSN 1546-1718.

[21] Biola M Javierre, Oliver S Burren, Steven P Wilder, Roman Kreuzhuber, Steven M Hill, Sven Sewitz, Jonathan Cairns, Steven W Wingett, Csilla Várnai, Michiel J Thiecke, Frances Burden, Samantha Farrow, Antony J Cutler, Karola Rehnström, Kate Downes, Luigi Grassi, Myrto Kostadima, Paula Freire-Pritchett, Fan Wang, BLUEPRINT Consortium, Hendrik G. Stunnenberg, John A. Todd, Daniel R. Zerbino, Oliver Stegle, Willem H. Ouwehand, Mattia Frontini, Chris Wallace, Mikhail Spivakov, and Peter Fraser. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, 167(5):1369–1384.e19, nov 2016. ISSN 1097-4172.

[22] Judong Shen, Kijoung Song, Andrew J. Slater, Enrico Ferrero, and Matthew R. Nelson. STOPGAP: a database for systematic target opportunity assessment by genetic association predictions. *Bioinformatics (Oxford, England)*, 33 (17):2784–2786, sep 2017. ISSN 1367-4811.

[23] Alexandre Amlie-Wolf, Mitchell Tang, Elisabeth E. Mlynarski, Pavel P. Kuksa, Otto Valladares, Zivadin Katanic, Debby Tsuang, Christopher D. Brown, Gerard D. Schellenberg, and Li-San Wang. INFERNO - INFERring the molecular mechanisms of NOncoding genetic variants. *bioRxiv*, page 211599, oct 2017.

[24] T. Hung, G. A. Pratt, B. Sundararaman, M. J. Townsend, C. Chaivorapol, T. Bhangale, R. R. Graham, W. Ortmann, L. A. Criswell, G. W. Yeo, and T. W. Behrens. The Ro60 autoantigen binds endogenous retroelements and regulates inflammatory gene expression. *Science (New York, N.Y.)*, 350(6259):455–9, oct 2015. ISSN 1095-9203.

[25] Arvind Kaul, Caroline Gordon, Mary K Crow, Zahi Touma, Murray B Urowitz, Ronald van Vollenhoven, Guillermo Ruiz-Irastorza, and Graham Hughes. Systemic lupus erythematosus. *Nature reviews. Disease primers*, 2:16039, jun 2016. ISSN 2056-676X.

[26] Tony N. Marion and Arnold E. Postlethwaite. Chance, genetics, and the heterogeneity of disease and pathogenesis in systemic lupus erythematosus. *Seminars in immunopathology*, 36(5):495–517, sep 2014. ISSN 1863-2300.

[27] Luis M. Amezcua-Guerra, Violeta Higuera-Ortiz, Ulises Arteaga-García, Selma Gallegos-Nava, and Claudia Hübbe-Tena. Performance of the 2012 Systemic Lupus International Collaborating Clinics and the 1997 American College of Rheumatology classification criteria for systemic lupus erythematosus in a real-life scenario. *Arthritis care & research*, 67(3):437–41, mar 2015. ISSN 2151-4658.

[28] Leonardo Collado-Torres, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper D Hansen, Andrew E Jaffe, Ben Langmead, and Jeffrey T Leek. Reproducible RNA-seq analysis using recount2. *Nature biotechnology*, 35(4):319–321, apr 2017. ISSN 1546-1696.

[29] Sean Davis and Paul S. Meltzer. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics (Oxford, England)*, 23(14):1846–7, jul 2007. ISSN 1367-4811.

[30] Audrey Kauffmann, Tim F. Rayner, Helen Parkinson, Misha Kapushesky, Margus Lukk, Alvis Brazma, and Wolfgang Huber. Importing ArrayExpress datasets into R/Bioconductor. *Bioinformatics (Oxford, England)*, 25(16):2092–4, aug 2009. ISSN 1367-4811.

[31] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*, 22(9):1760–74, sep 2012. ISSN 1549-5469.

[32] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550, dec 2014. ISSN 1474-760X.

[33] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–40, jan 2010. ISSN 1367-4811.

[34] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47, apr 2015. ISSN 1362-4962.

[35] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10): R106, 2010. ISSN 1474-760X.

[36] Hadley Wickham. *ggplot2*. Springer New York, New York, NY, 2009. ISBN 978-0-387-98140-6.

[37] Rong Chen, Alex A Morgan, Joel Dudley, Tarangini Deshpande, Li Li, Keiichi Kodama, Annie P Chiang, and Atul J Butte. FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease. *Genome biology*, 9(12):R170, 2008. ISSN 1474-760X.

[38] Vincent J Carey. gwascat, 2017. URL `https://doi.org/doi:10.18129/B9.bioc.gwascat`.

[39] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, Zoe May Pendlington, Danielle Welter, Tony Burdett, Lucia Hindorff, Paul Flicek, Fiona Cunningham, and Helen Parkinson. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research*, 45(D1):D896–D901, jan 2017. ISSN 1362-4962.

[40] John D. Eicher, Christa Landowski, Brian Stackhouse, Arielle Sloan, Wenjie Chen, Nicole Jensen, Ju-Ping Lien, Richard Leslie, and Andrew D. Johnson. GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. *Nucleic acids research*, 43(Database issue):D799–804, jan 2015. ISSN 1362-4962.

[41] Vincent J Carey. grasp2db, 2017. URL `https://doi.org/doi:10.18129/B9.bioc.grasp2db`.

[42] William S. Bush and Jason H. Moore. Chapter 11: Genome-wide association studies. *PLoS computational biology*, 8 (12):e1002822, dec 2012. ISSN 1553-7358.

[43] Vincent J Carey. ldblock, 2017. URL `https://doi.org/doi:10.18129/B9.bioc.ldblock`.

[44] Andrew Yates, Kathryn Beal, Stephen Keenan, William McLaren, Miguel Pignatelli, Graham R. S. Ritchie, Magali Ruffier, Kieron Taylor, Alessandro Vullo, and Paul Flicek. The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics (Oxford, England)*, 31(1):143–5, jan 2015. ISSN 1367-4811.

[45] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. Software for computing and annotating genomic ranges. *PLoS computational biology*, 9(8): e1003118, aug 2013. ISSN 1553-7358.

[46] Valerie Obenchain, Michael Lawrence, Vincent Carey, Stephanie Gogarten, Paul Shannon, and Martin Morgan. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics (Oxford, England)*, 30(14):2076–8, jul 2014. ISSN 1367-4811.

[47] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, Nancy J Cox, and Hae Kyung Im. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091–8, sep 2015. ISSN 1546-1718.

[48] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W J H Penninx, Rick Jansen, Eco J C de Geus, Dorret I Boomsma, Fred A Wright, Patrick F Sullivan, Elina Nikkola, Marcus Alvarez, Mete Civelek, Aldons J Lusis, Terho Lehtimäki, Emma Raitoharju, Mika Kähönen, Ilkka Seppälä, Olli T Raitakari, Johanna Kuusisto, Markku Laakso, Alkes L Price, Päivi Pajukanta, and Bogdan Pasaniuc. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245–52, mar 2016. ISSN 1546-1718.

[49] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, and Jian Yang. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics*, 48(5):481–7, may 2016. ISSN 1546-1718.

[50] Michael Lawrence, Robert Gentleman, and Vincent Carey. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics (Oxford, England)*, 25(14):1841–2, jul 2009. ISSN 1367-4811.

[51] Daniel R Zerbino, Steven P Wilder, Nathan Johnson, Thomas Juettemann, and Paul R Flicek. The ensembl regulatory build. *Genome biology*, 16(1):56, mar 2015. ISSN 1474-760X.

[52] Aaron T. L. Lun, Malcolm Perry, and Elizabeth Ing-Simmons. Infrastructure for genomic interactions: Bioconductor classes for Hi-C, ChIA-PET and related experiments. *F1000Research*, 5:950, jun 2016. ISSN 2046-1402. doi: 10.12688/f1000research.8759.2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4890298/http://f1000research.com/articles/5-950/v2.

[53] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–9, may 2000. ISSN 1061-4036.

[54] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology*, 16(5):284–7, may 2012. ISSN 1557-8100. doi: 10.1089/omi.2011.0118. URL http://www.ncbi.nlm.nih.gov/pubmed/22455463http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3339379http://online.liebertpub.com/doi/abs/10.1089/omi.2011.0118.

[55] Shereen Oon, Nicholas J Wilson, and Ian Wicks. Targeted therapeutics in SLE: emerging strategies to modulate the interferon pathway. *Clinical & translational immunology*, 5(5):e79, may 2016. ISSN 2050-0068. doi: 10.1038/cti.2016.26. URL http://doi.wiley.com/10.1038/cti.2016.26http://www.ncbi.nlm.nih.gov/pubmed/27350879http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4910120.

[56] D L Morris, M M A Fernando, K E Taylor, S A Chung, J Nititham, M E Alarcón-Riquelme, L F Barcellos, T W Behrens, C Cotsapas, P M Gaffney, R R Graham, B A Pons-Estel, P K Gregersen, J B Harley, S L Hauser, G Hom, C D Langefeld, J A Noble, J D Rioux, M F Seldin, Systemic Lupus Erythematosus Genetics Consortium, T J Vyse, and L A Criswell. MHC associations with clinical and autoantibody manifestations in European SLE. *Genes and immunity*, 15(4):210–7, apr 2014. ISSN 1476-5470.

[57] Abel Suárez-Fueyo, Sean J Bradley, and George C Tsokos. T cells in Systemic Lupus Erythematosus. *Current opinion in immunology*, 43:32–38, dec 2016. ISSN 1879-0372.

[58] Yasuko Furumoto, Carolyne K. Smith, Luz Blanco, Wenpu Zhao, Stephen R. Brooks, Seth G. Thacker, Zarzour Abdalrahman, Giuseppe Sciumè, Wanxia L. Tsai, Anna M. Trier, Leti Nunez, Laurel Mast, Victoria Hoffmann, Alan T. Remaley, John J. O'Shea, Mariana J. Kaplan, and Massimo Gadina. Tofacitinib Ameliorates Murine Lupus and Its Associated Vascular Dysfunction. *Arthritis & rheumatology (Hoboken, N.J.)*, 69(1):148–160, jan 2017. ISSN 2326-5205.

[59] Jonatan Leffler, Anders A Bengtsson, and Anna M Blom. The complement system in systemic lupus erythematosus: an update. *Annals of the rheumatic diseases*, 73(9):1601–6, sep 2014. ISSN 1468-2060.

[60] Dirk De Valck, D Y Jin, Karen Heyninck, Marc Van de Craen, Roland Contreras, Walter Fiers, K T Jeang, and Rudi Beyaert. The zinc finger protein A20 interacts with a novel anti-apoptotic protein which is cleaved by specific caspases. *Oncogene*, 18(29):4182–90, jul 1999. ISSN 0950-9232.

[61] L. Ling and D. V. Goeddel. T6BP, a TRAF6-interacting protein involved in IL-1 signaling. *Proceedings of the National Academy of Sciences of the United States of America*, 97(17):9567–72, aug 2000. ISSN 0027-8424. doi: 10.1073/pnas.170279097. URL http://www.ncbi.nlm.nih.gov/pubmed/10920205http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC16905http://www.pnas.org/cgi/doi/10.1073/pnas.170279097.

[62] Lars Rönnblom and Keith B. Elkon. Cytokines as therapeutic targets in SLE. *Nature reviews. Rheumatology*, 6(6):339–47, jun 2010. ISSN 1759-4804.

[63] Tengfei Yin, Dianne Cook, and Michael Lawrence. ggbio: an R package for extending the grammar of graphics for genomic data. *Genome biology*, 13(8):R77, aug 2012. ISSN 1474-760X.

[64] Douglas H. Phanstiel, Alan P. Boyle, Carlos L. Araya, and Michael P. Snyder. *Bioinformatics (Oxford, England)*, 30(19):2808–10, oct 2014. ISSN 1367-4811.

[65] Florian Hahne and Robert Ivanek. Visualizing Genomic Data Using Gviz and Bioconductor. *Methods in molecular biology (Clifton, N.J.)*, 1418:335–51, 2016. ISSN 1940-6029.

[66] Nathan Harmston, Elizabeth Ing-Simmons, Malcolm Perry, Anja Barešić, and Boris Lenhard. GenomicInteractions: An R/Bioconductor package for manipulating and investigating chromatin interaction data. *BMC genomics*, 16:963, nov 2015. ISSN 1471-2164.