Data Wrangling

Using R to interrogate your data

Setting your expectations

This course is...

- … a basic introduction to programming and data manipulation using R
- … intended to create awareness and a foundation to build upon
- … (hopefully) a comfortable place to start if you have never programmed before
- … (hopefully) useful for cleaning, summarizing, organizing your data

This course is **not**...

- … a data analysis course
- … a statistics course
- … a course on Linux/Unix
- … a how-to on becoming a bioinformaticist or “data scientist”
- ...useful if you do not practice/use the techniques consistently

Why learn programming?

Data size and scope is constantly growing.

Saving time- automating analysis, integrating new data

Saving face- reproducible analysis and findings

- (Hopefully!) less human error

Most things have tradeoffs. Questions to ask yourself:

- Is this something I will have to do often?
- Is it tedious?
- Do I already know how to do this?
    - For small tasks that you do not need to repeat, it's often easier/quicker to use what you are comfortable with

Why (and why not) R?

Reasons to use R:

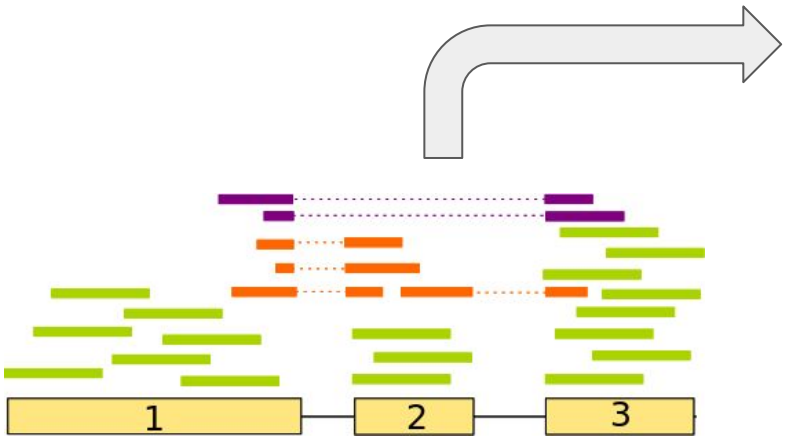
- Very common, especially for statistical tasks
- Open-source (read: free)
- Great online community support
- Bioconductor has a lot of biology-related analyses (http://bioconductor.org/)
    - e.g. ChiP-Seq calling, differential expression

Reasons to *not* use R:

- Origin as domain-specific language for statistics
    - Unfamiliar terms and conventions
- Open-source
    - DIY
- Everyone has preferences
    - coding style/ease
    - programming “flavor”
    - Domain/lab convention

The structure of this course

The data manipulations here will focus on a typical RNA-Seq dataset, where we compare a control and treated sample (or WT vs knockdown). Analysis is based on a “count matrix”, which serves as a proxy measure of gene expression.



gene	sampleD	sampleC	sampleB	sampleA	...
KRAS	757	686	667	680	...
MiR1256	113	107	236	241	...
...	...	...	...	...	...

After sequencing and alignment to the reference genome, count the reads aligning to each gene. Store in a table like above.

A survey of the files we will use:

Download files from the Google Drive for this course.

For ease, save the files in their own folder to keep them separate from your other files (save them somewhere you will remember!)

A survey of the files we will use:

`differential_results.csv`

- Differential expression results (produced by DESeq software)
- Columns separated by comma “,”
- “pval” is raw p-value, “padj” adjusts for “multiple testing correction”

gene	baseMean	C	T	foldChange	log2FoldChange	pval	padj
WASH7P	35.7	36.9	34.5	0.934	-0.097	0.876	1.0
AGRN	25089.9	24060.2	26119.5	1.09	0.118	0.201	0.707
SDF4	9426.4	8805.4	10047.4	1.141	0.19	1.03e-03	1.22e-02
MRT04	4881.5	5346.4	4416.5	0.826	-0.276	1.45e-05	2.99e-04
...							

A survey of the files we will use:

`gene_expression.tsv`

- Normalized gene expression
- Each row corresponds to a gene
- Each column is a sample
- Many rows since many genes (>30,000 for mouse or human)

gene	SW1_Control	SW2_Control	...	SW5_Treated	SW6_Treated
KRAS	757	686	...	667	680
MiR1256	113	107	...	236	241
...	...	...	...	...	...

A survey of the files we will use:

`my_pathway_genes.txt`

- A list of genes in different pathways of interest (here, glycolysis and ras signaling)
- Note that the file does NOT have column headers indicating the name.

GAPDH	glycolysis
EGF	ras_signaling
...	...

A survey of the files we will use:

(Depending on time, we may or may not use this file)

`gene_annotations.tsv`

- Genomic coordinates of the genes
- Modified version of a standard “GTF” (Gene Transfer Format) file

chrom	start	end	strand	name
chr1	2000	5000	+	KRAS
chr1	6000	9000	+	NRAS
...	...	...	...	...

A survey of the files we will use:

(Depending on time, we may or may not use this file)

`mutations.tsv`

- Genomic coordinates for the SNPs
  - Modified version of a “VCF” (Variant Call Format) file
- Note: Other files (gene_annotations.tsv) have chromosomes named like “chr1”. Here, the “chr” prefix for chromosome name is not there and we just have “1”, “2”, …, “X”, “Y”

chrom	pos	ref	alt
1	2000	A	G
1	6000	C	T
...	...	...	...

A biological question we might ask

What is the expression of genes significantly upregulated in our pathway of interest?

Rough steps:

1. Read data files
2. Filter for genes in a pathway of interest
3. Filter for significantly upregulated genes
4. Write results to a file

Following this, plot the expression of the significant genes.

If there is time, see if any of those genes have SNPs