Computational Biosciences Institute Workshop 5

Informatics for RNA-sequence analysis

Kelvin Zhang, Ph.D.

CIHR & CBI fellow
Zipursky and Pellegrini Lab

Goals of this workshop

- Introduction to the basic concept of RNA sequencing (RNA-seq) analysis;
  - Rationale, challenges, pipeline, problems, etc.

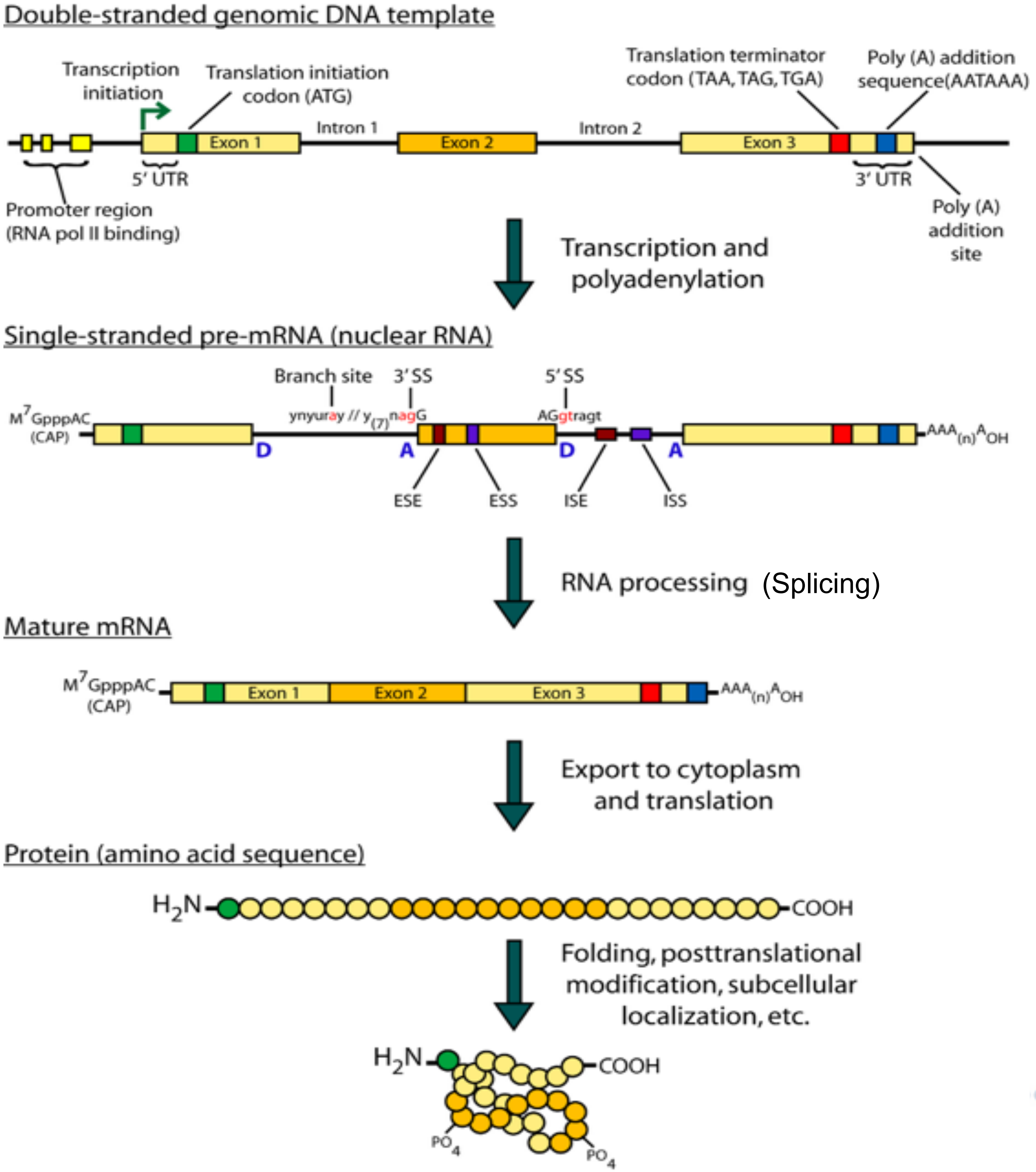
- Provide a practical resource for those new to the topic of RNA-seq analysis;

- Practice a working pipeline of RNA-seq data analysis using galaxy;
  - QC, alignment, gene expression quantification, differential expression analysis, downstream analysis.

Outline

- Day 1: Introduction to the basic concept of RNA sequencing (RNA-seq) analysis;
  - Rationale, challenges, pipeline, problems, etc.
  - Warm up exercises.
- Day 2: Practice a working pipeline of RNA-seq data analysis using galaxy;
  - QC, alignment, gene expression quantification, differential expression analysis.
- Day 3: Practice a working pipeline of RNA-seq data analysis using command lines.

Gene expression



Double-stranded genomic DNA template

Transcription initiation
Translation initiation codon (ATG)
Translation terminator codon (TAA, TAG, TGA)
Poly (A) addition sequence(AATAAA)
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
5′ UTR
3′ UTR
Promoter region (RNA pol II binding)
Poly (A) addition site

Transcription and polyadenylation

Single-stranded pre-mRNA (nuclear RNA)

RNA processing (Splicing)

Mature mRNA

Export to cytoplasm and translation

Protein (amino acid sequence)

Folding, posttranslational modification, subcellular localization, etc.

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

Why RNA-seq?

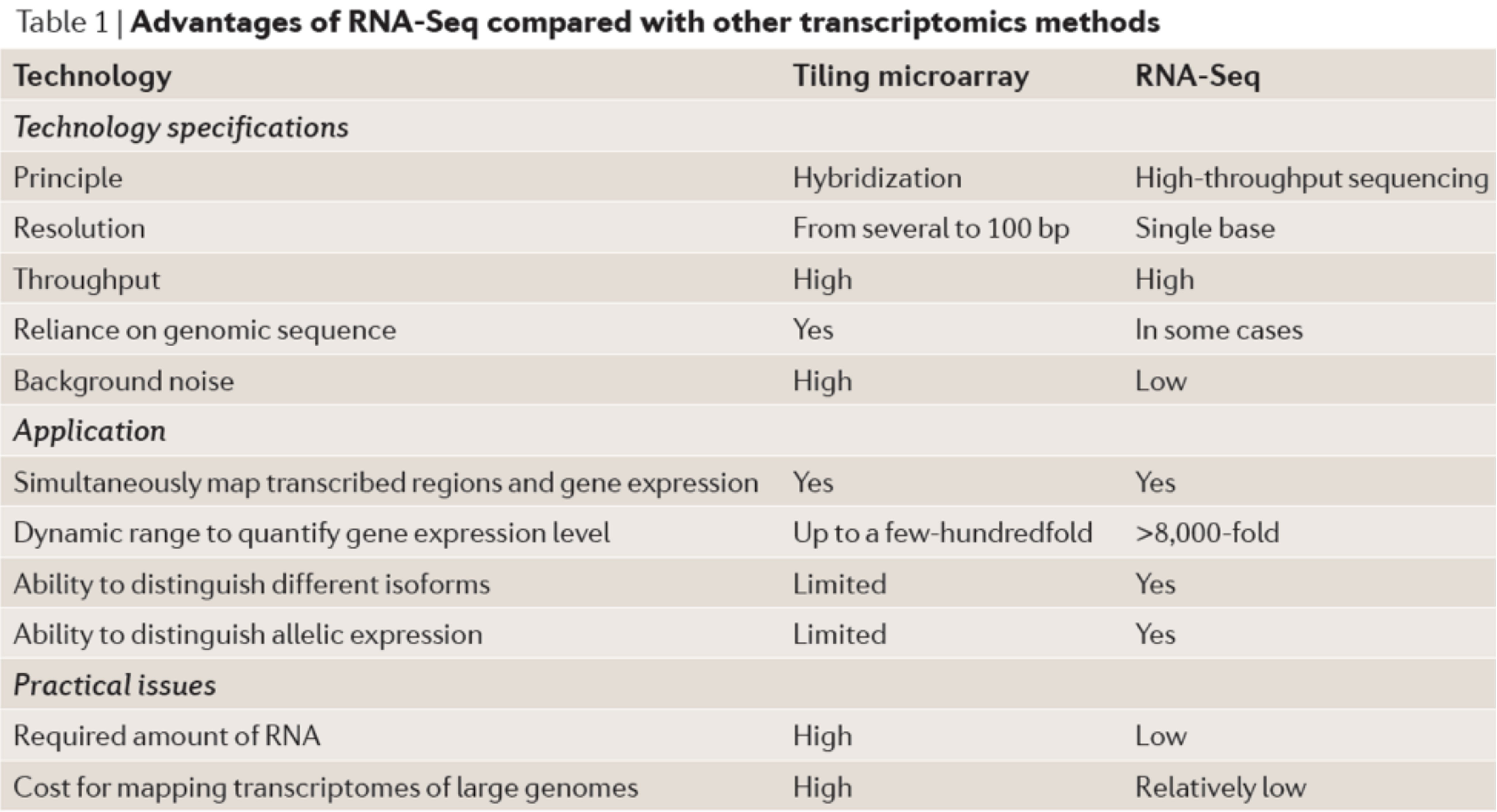


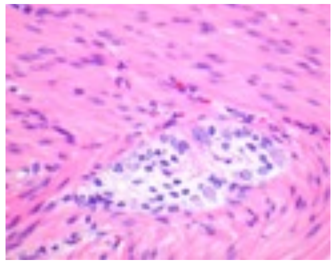
Table 1 | **Advantages of RNA-Seq compared with other transcriptomics methods**

Technology	**Tiling microarray**	**RNA-Seq**
Technology specifications		
Principle	Hybridization	High-throughput sequencing
Resolution	From several to 100 bp	Single base
Throughput	High	High
Reliance on genomic sequence	Yes	In some cases
Background noise	High	Low
Application		
Simultaneously map transcribed regions and gene expression	Yes	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes
Ability to distinguish allelic expression	Limited	Yes
Practical issues		
Required amount of RNA	High	Low
Cost for mapping transcriptomes of large genomes	High	Relatively low

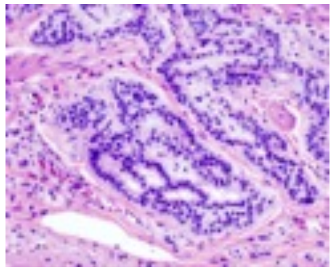
Zhang et al. Nature Reviews Genetics, 2009

RNA sequencing

Samples of interest

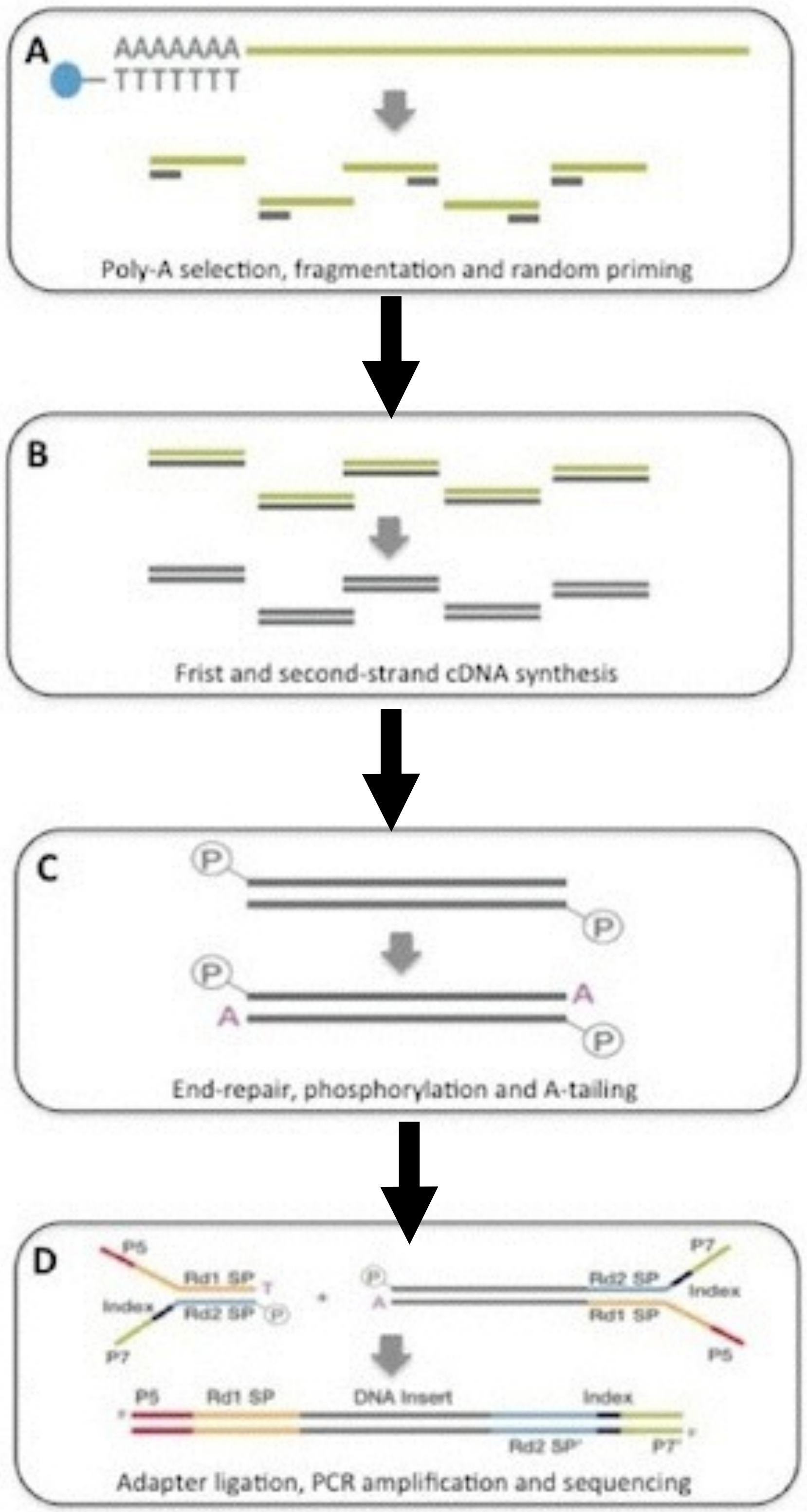


Condition 1
(normal colon)



Condition 2
(colon tumor)

RNA isolation



Poly-A selection, fragmentation and random priming

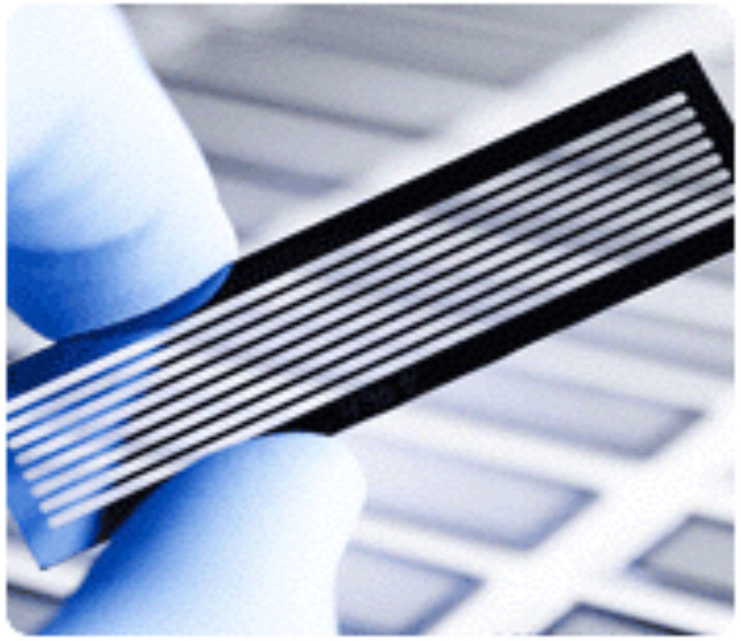
Frist and second-strand cDNA synthesis

End-repair, phosphorylation and A-tailing

Adapter ligation, PCR amplification and sequencing

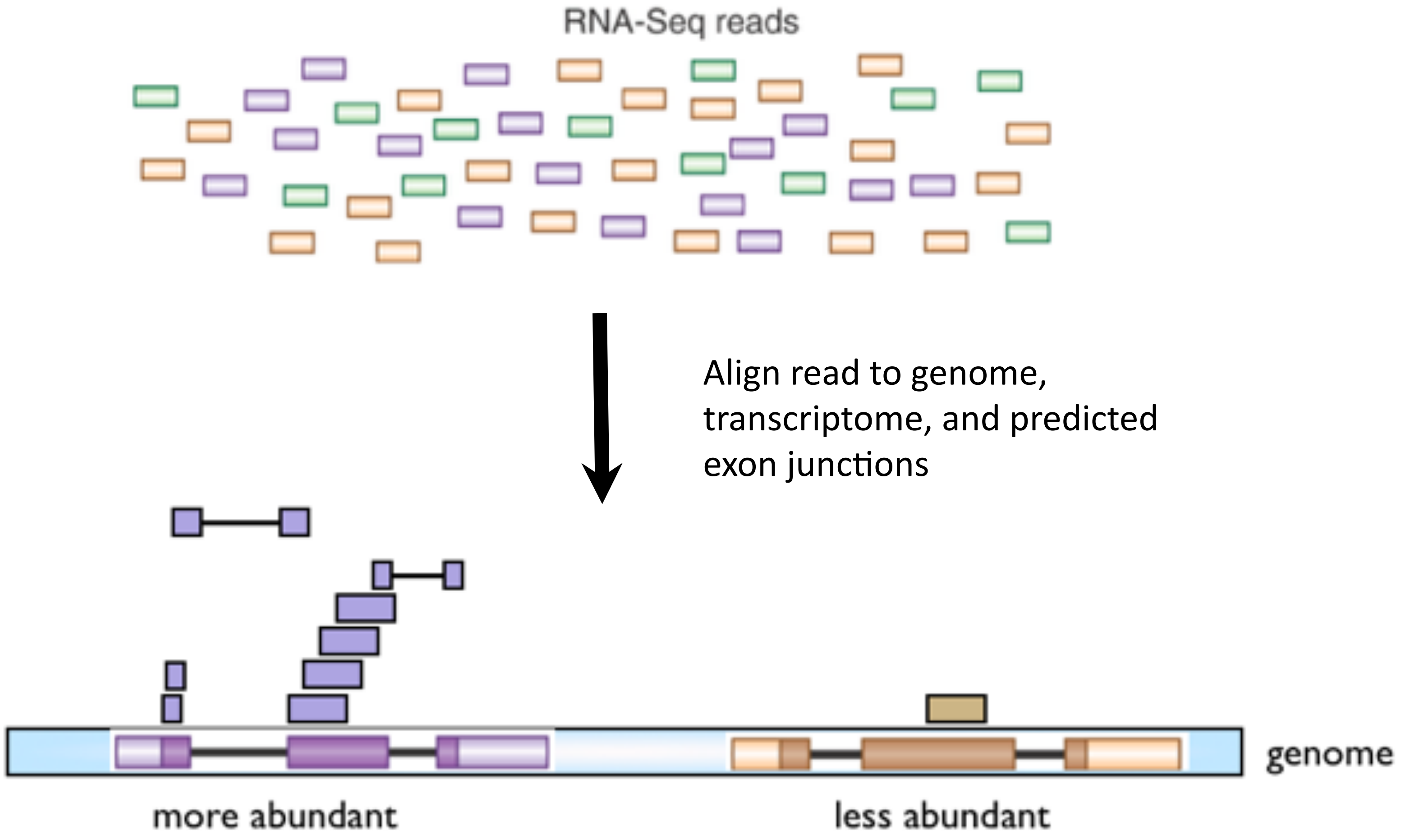
- Ploy-A purification
- Fragmentation
- cDNA synthesis using random primers
- Adapter ligation
- Size selection
- PCR amplification

RNA sequencing



- Flowcell
  - 8 lanes in total
  - about 200 Million reads per lane
  - Multiplex up to 24 samples on one lane using barcodes

RNA sequencing



Align read to genome, transcriptome, and predicted exon junctions

* based on Illumina approach

Downstream analysis

Why we sequence RNA?

- Functional studies
  - Genome may be constant but an experimental condition has a pronounced effect on gene expression
    - e.g. Drug treated vs. untreated cell line
    - e.g. Wild type versus knock out mice
- Some molecular features can only be observed at the RNA level
  - Alternative isoforms, fusion transcripts, RNA editing
- Predicting transcript sequence from genome sequence is difficult
  - Alternative splicing, RNA editing, etc.

Why we sequence RNA?

- Interpreting mutations that do not have an obvious effect on protein sequence
  - “Regulatory” mutations that affect what mRNA isoform is expressed and how much
    - e.g. splice sites, promoters, exonic/intronic splicing motifs, etc.
- Prioritizing protein coding somatic mutations (often heterozygous)
  - If the gene is not expressed, a mutation in that gene would be less interesting
  - If the gene is expressed but only from the wild type allele, this might suggest loss-of-function (haploinsufficiency)
  - If the mutant allele itself is expressed, this might suggest a candidate drug target

Challenges

- RNAs consist of small exons that may be separated by large introns
  - Mapping reads to genome is challenging
- The relative abundance of RNAs vary wildly
  - 10^5 – 10^7 orders of magnitude
  - Since RNA sequencing works by random sampling, a small fraction of highly expressed genes may consume the majority of reads
  - Ribosomal and mitochondrial genes
- RNAs come in a wide range of sizes
  - Small RNAs must be captured separately
  - PolyA selection of large RNAs may result in 3’ end bias
- RNA is fragile compared to DNA (easily degraded)

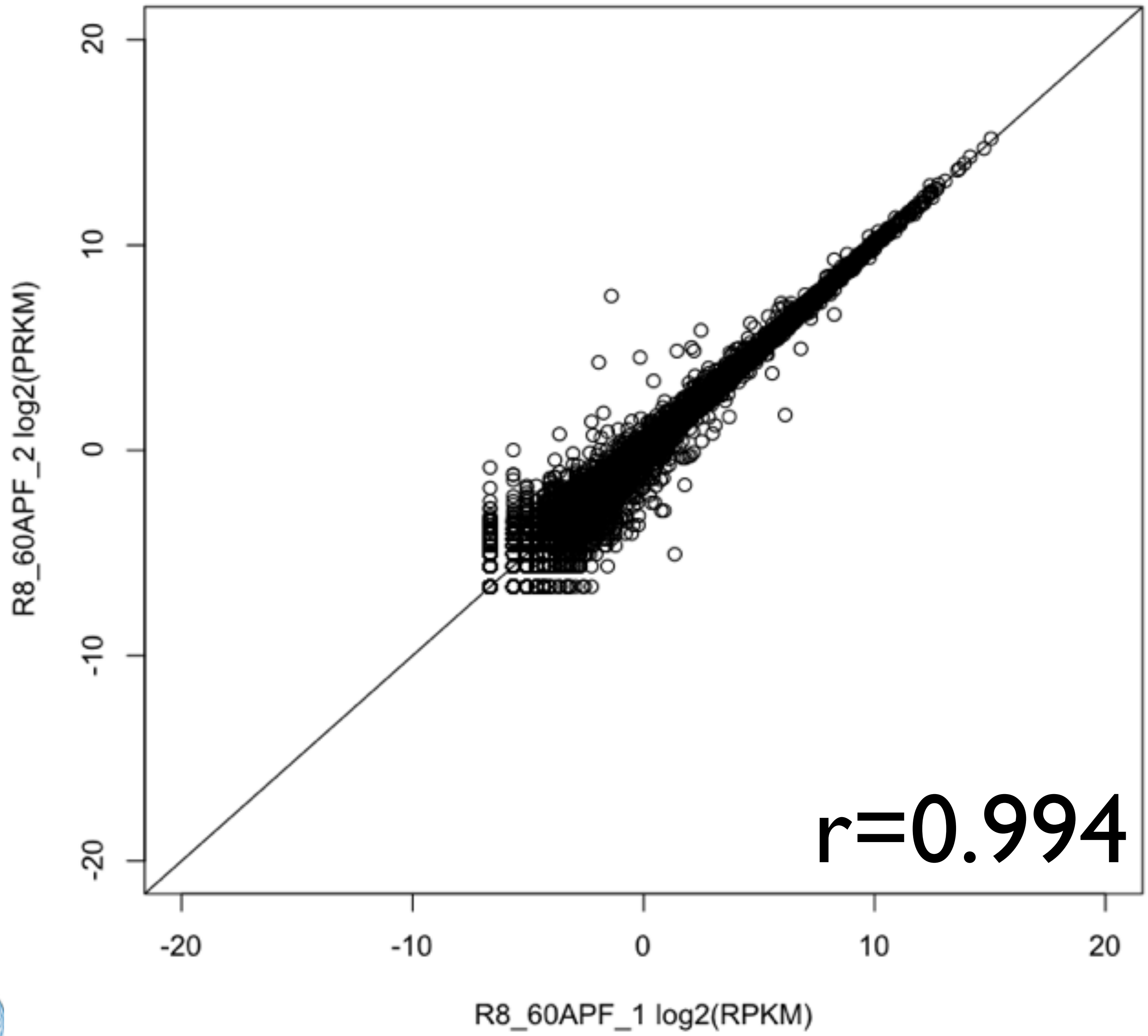
Design considerations

- Standards, Guidelines and Best Practices for RNA-seq
  - The ENCODE Consortium
  - Meta data to supply, replicates, sequencing depth, control experiments, reporting standards, etc
  - http://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf

Replicates

- Technical replicate
  - Multiple instances of sequence generation
    - Flow cells, lanes, indexes
    - not required if they are from the same RNA library
- Biological replicate
  - Multiple isolations of cells showing the same phenotype, stage or other experimental condition
  - Some example concerns/challenges:
    - Environmental factors, growth conditions, time points
  - Correlation Coefficient (> 0.92)

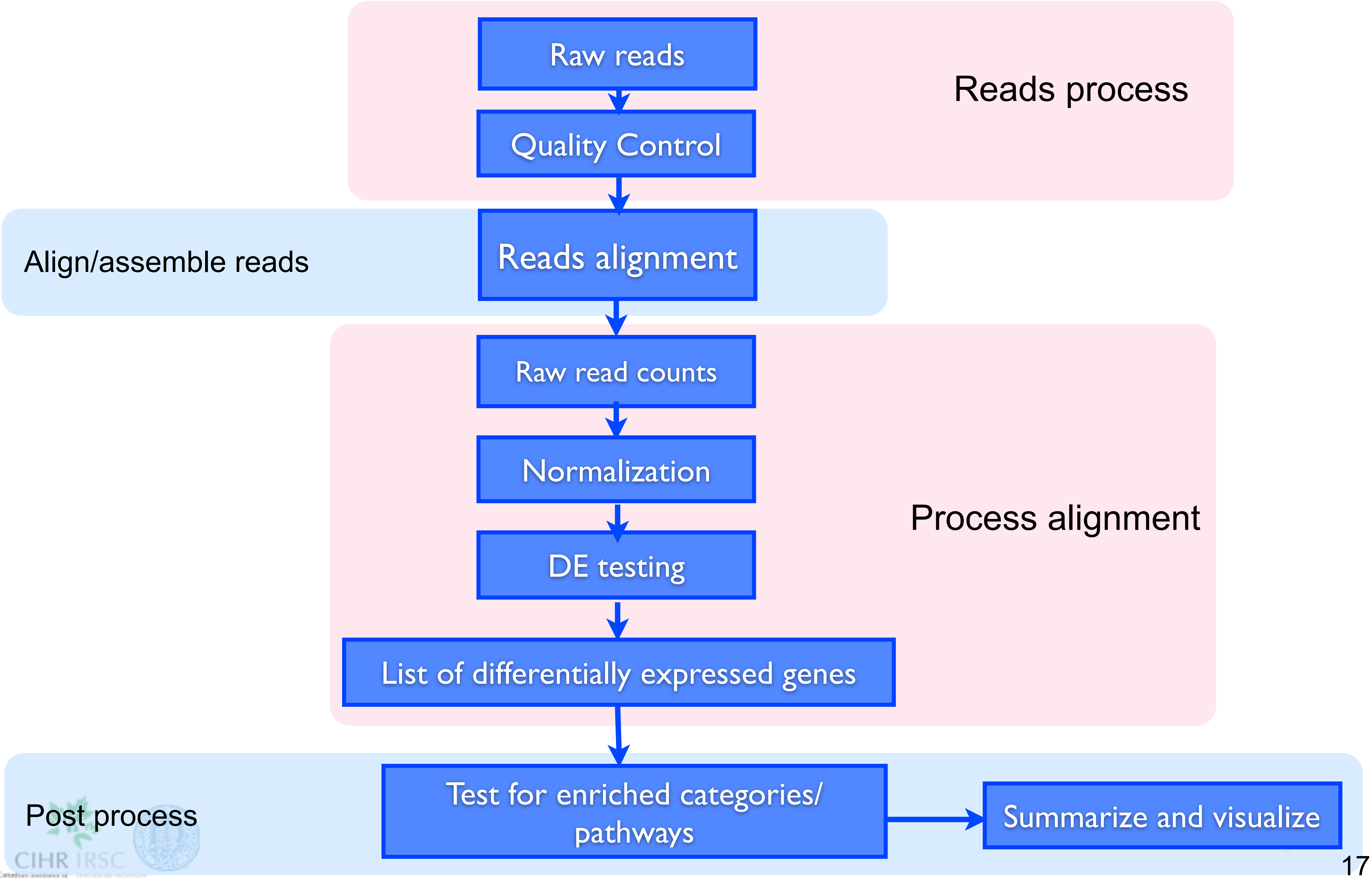
Replicates


20
10
0
-10
-20
R8_60APF_2 log2(PRKM)
-20
-10
0
10
20
R8_60APF_1 log2(RPKM)
r=0.994

What RNA-seq can do?

- Gene expression and differential expression
- Alternative expression analysis
- Transcript discovery and annotation
- Allele specific expression
  - Relating to SNPs or mutations
- Mutation discovery
- Fusion detection
- RNA editing

RNA-seq workflows



Raw reads

Quality Control

Reads process

Align/assemble reads

Reads alignment

Raw read counts

Normalization

DE testing

List of differentially expressed genes

Process alignment

Post process

Test for enriched categories/ pathways

Summarize and visualize

Question 1: Should I remove duplicates for RNA-seq?

- Maybe… more complicated question than for DNA
- Concern.
  - Duplicates may correspond to biased PCR amplification of particular fragments
  - For highly expressed, short genes, duplicates are expected even if there is no amplification bias
  - Removing them may reduce the dynamic range of expression estimates
- Assess library complexity and decide…
- If you do remove them, assess duplicates at the level of paired-end reads (fragments) not single end reads

Question 2: How much library depth is needed for RNA-seq?

- **Library sequence depth** is the average read coverage of target sequences.
  - sequence depth = total number of reads × read length / estimated target sequence length
- For example, for the Drosophila transcriptome (about 30Mbp), if 30 million reads with the length of 50bp are generated,

The depth is: 30 m × 50 bp / 30mbp = 50×

CIHR IRSC

Question 2: How much library depth is needed for RNA-seq?

- My advice. Don't ask this question if you want a simple answer…
- Depends on a number of factors:
  - Question being asked of the data. Gene expression? Alternative expression? Mutation calling?
  - Tissue type, RNA preparation, quality of input RNA, library construction method, etc.
  - Sequencing type: read length, paired vs. unpaired, etc.
  - Computational approach and resources

Question 2: How much library depth is needed for RNA-seq?

Suggestion:

- Identify publications with similar goals
- Pilot experiment
- Good news: 1-2 lanes of recent Illumina HiSeq data should be enough for most purposes

Guidelines:

Project Goals	Differential Gene Expression	De novo Assembly of transcriptome	Refine gene model	Identification of structural variants
Library Types	PE	PE	PE, SE	PE
Sequencing Depth	Moderate (< 50×)	Extensive (> 50×)	Extensive (> 50×)	Extensive (> 50×)

Question 3: What mapping strategy should I use for RNA-seq?

- Depends on read length
- < 50 bp reads
  - Use aligner like BWA and a genome + junction database
  - Junction database needs to be tailored to read length
    - Or you can use a standard junction database for all read lengths and an aligner that allows substring alignments for the junctions only (e.g. BLAST … slow).
  - Assembly strategy may also work (e.g. Trans-ABySS)
- > 50 bp reads
  - Spliced aligner such as Bowtie/TopHat

Question 4: How reliable are expression predictions from RNA-seq?

- Are novel exon-exon junctions real?
  - What proportion validate by RT-PCR and Sanger sequencing?
- Are differential/alternative expression changes observed between tissues accurate?
  - How well do DE values correlate with qPCR?
- Spike-in control

Tool recommendations

- Alignment
  - BWA
    - Align to genome + junction database
  - Tophat (PMID: 19289445)
    - Spliced alignment genome
  - hmmSplicer (PMID: 21079731)
    - Spliced alignment to genome – focus on splice sites specifically
- Expression, differential expression alternative expression
  - Cufflinks/Cuffdiff
  - DESeq, EdgeR
- Fusion detection
  - Defuse
  - Comrad
- Transcript assembly
  - Trans-ABySS (also useful for isoform and fusion discovery).
  - MISO
- Mutation calling
  - SNVMix
- Visit forums for more recommendations and discussion
  - http://seqanswers.com/
  - http://www.biostars.org/

Downstream data analysis

- I have identified a list of differentially expressed genes. What I can do with them?

- How to use known information about gene functions and gene relationships to help understand the biology behind a list of differentially expressed genes?

- Determine pathways containing (many of) the genes concerned and gain biological insight.

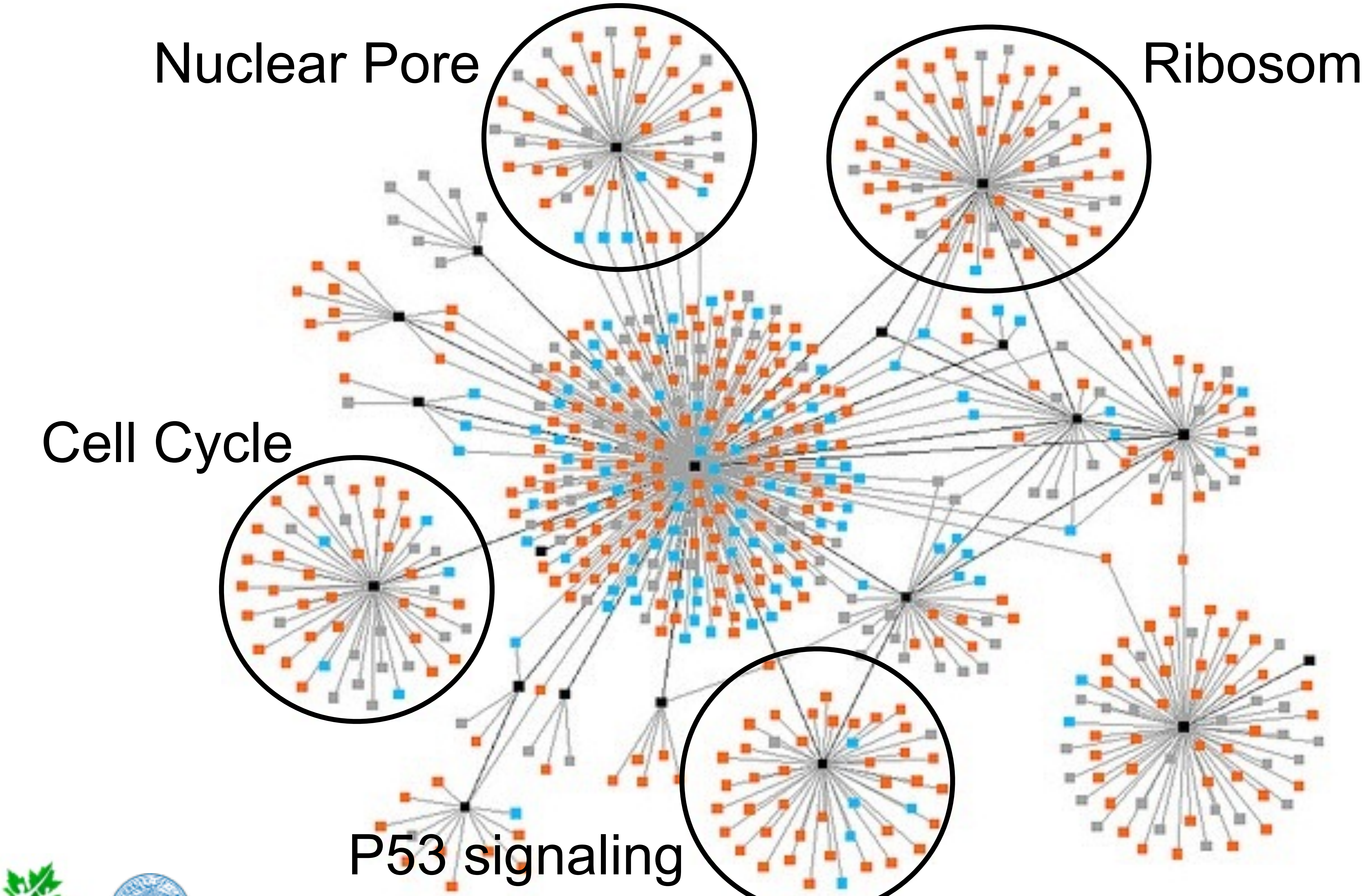
- Gene Set Enrichment Analysis

Gene set enrichment analysis

- We can break down cellular functions into different gene sets.

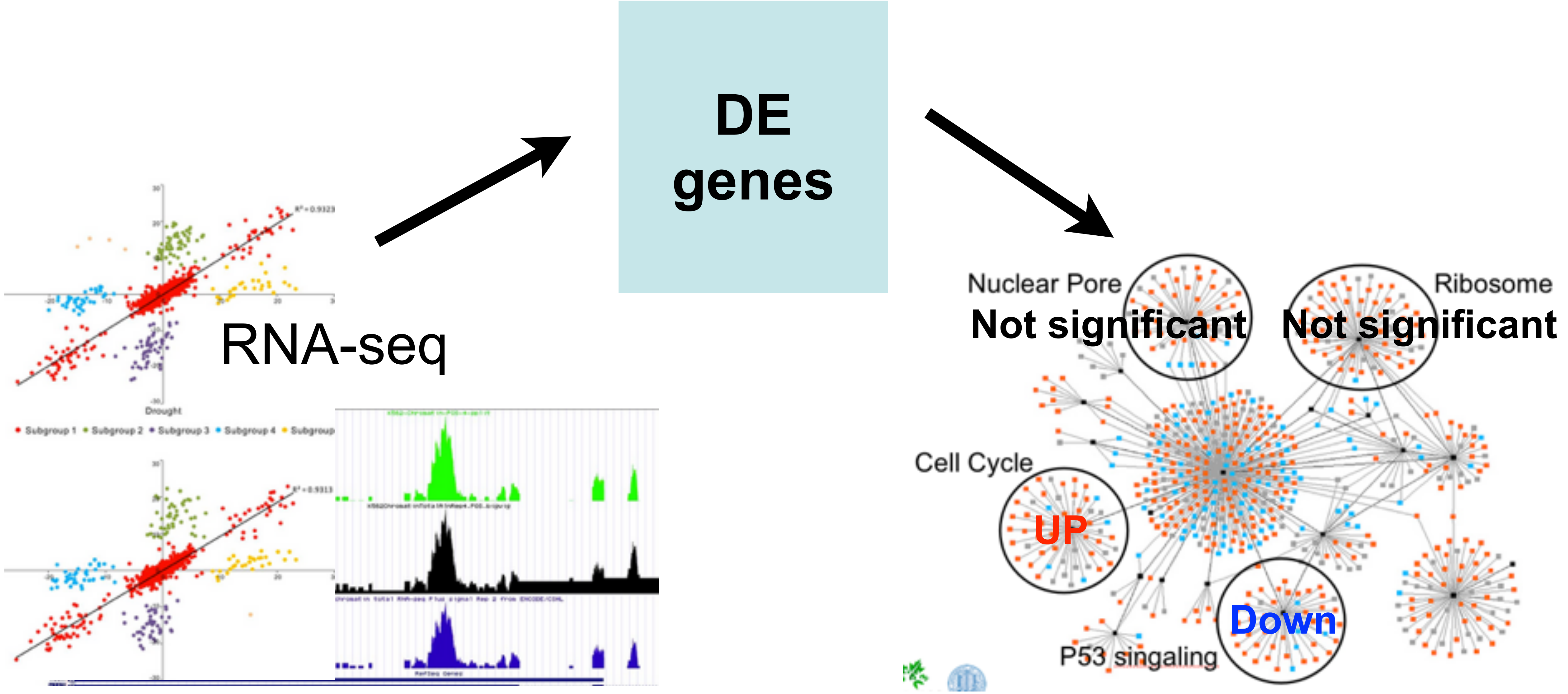
- Each gene set is associated to a specific cellular function, process, component or pathway.

Gene set enrichment analysis



Gene set enrichment analysis

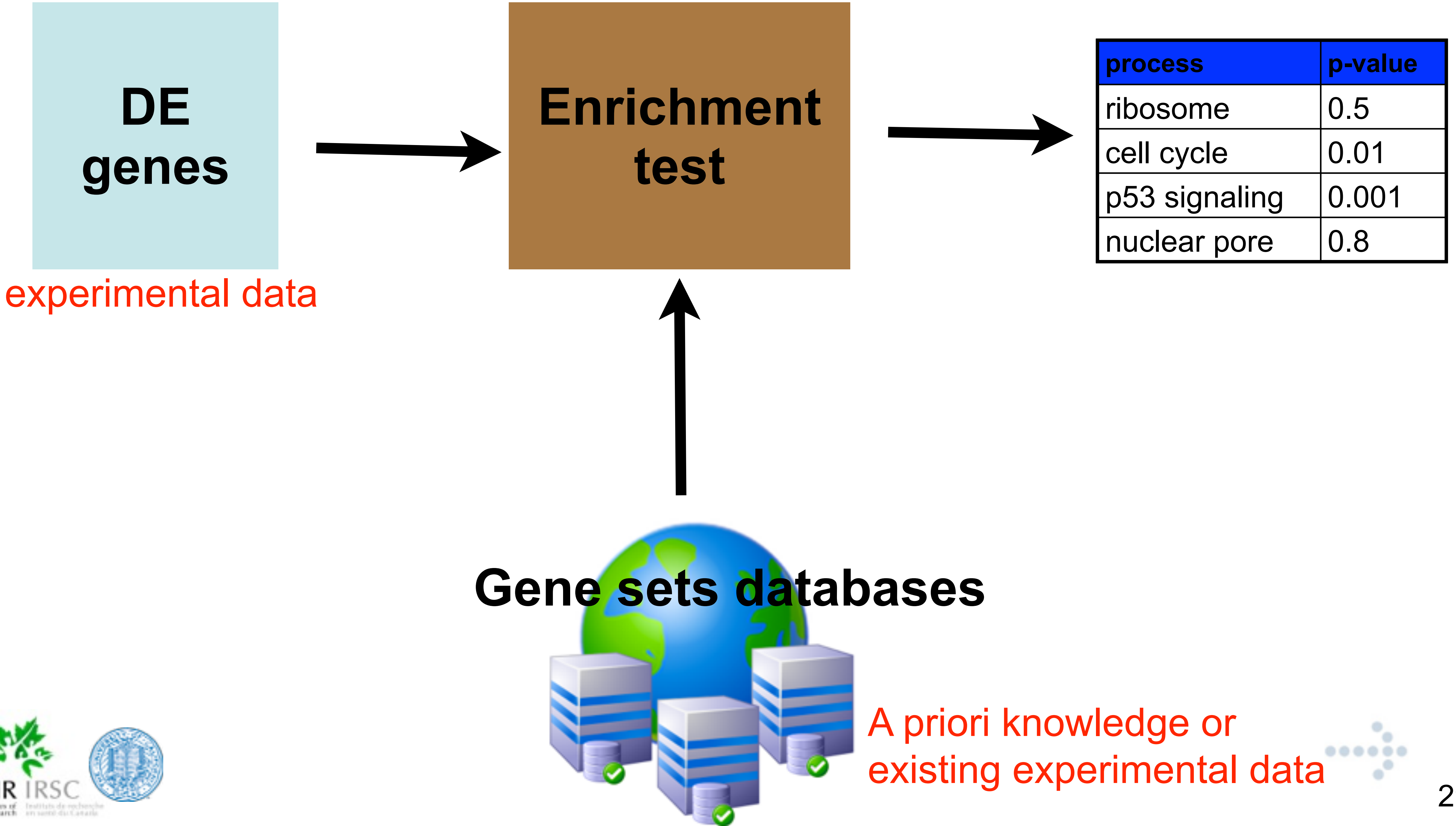
- Find known gene sets (e.g. pathways) enriched in a gene list (e.g. from RNA-seq).



RNA-seq

DE genes

Gene set enrichment analysis



DE genes

experimental data

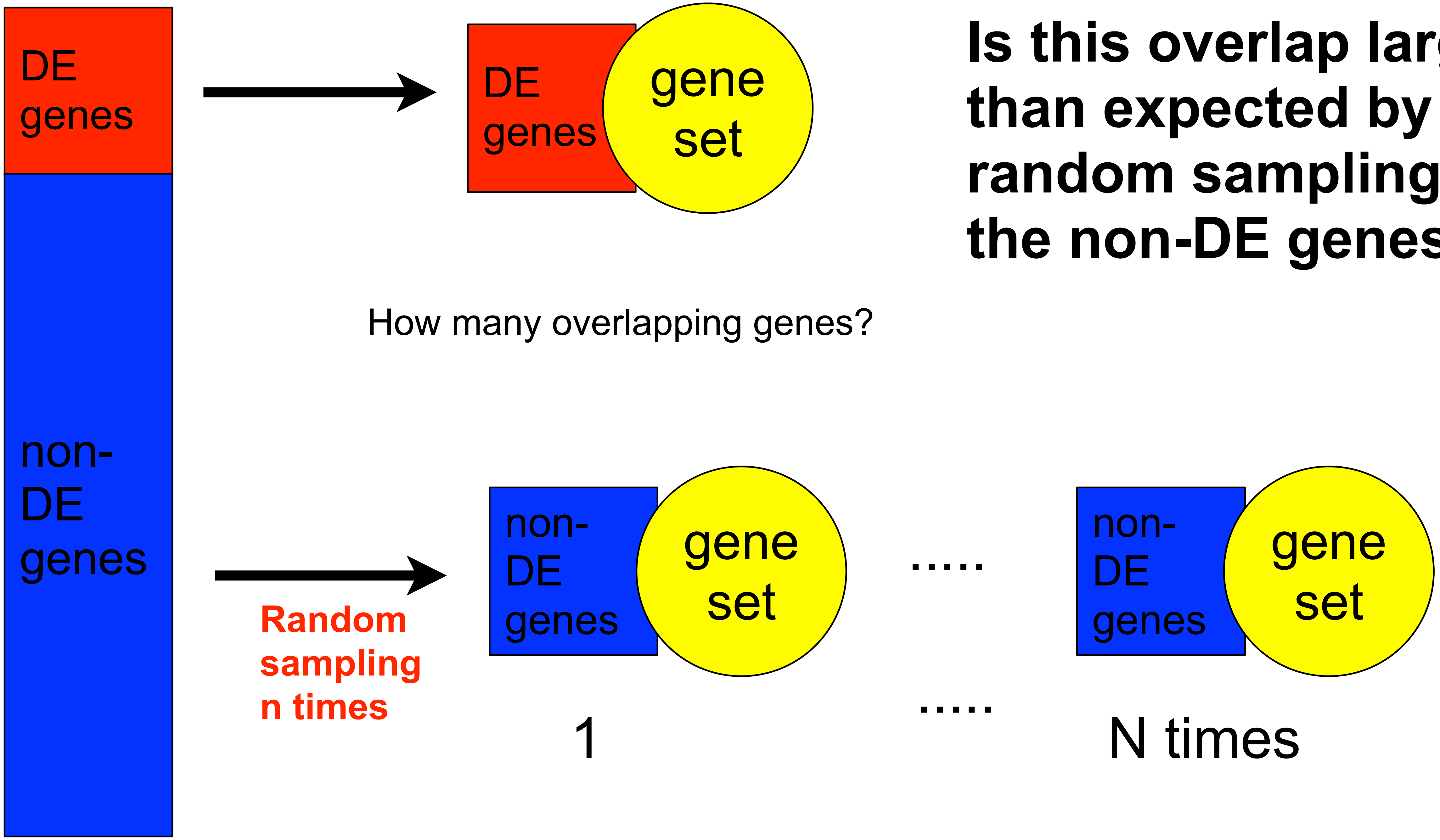
Enrichment test

process	p-value
ribosome	0.5
cell cycle	0.01
p53 signaling	0.001
nuclear pore	0.8

Gene sets databases

A priori knowledge or existing experimental data

Enrichment test

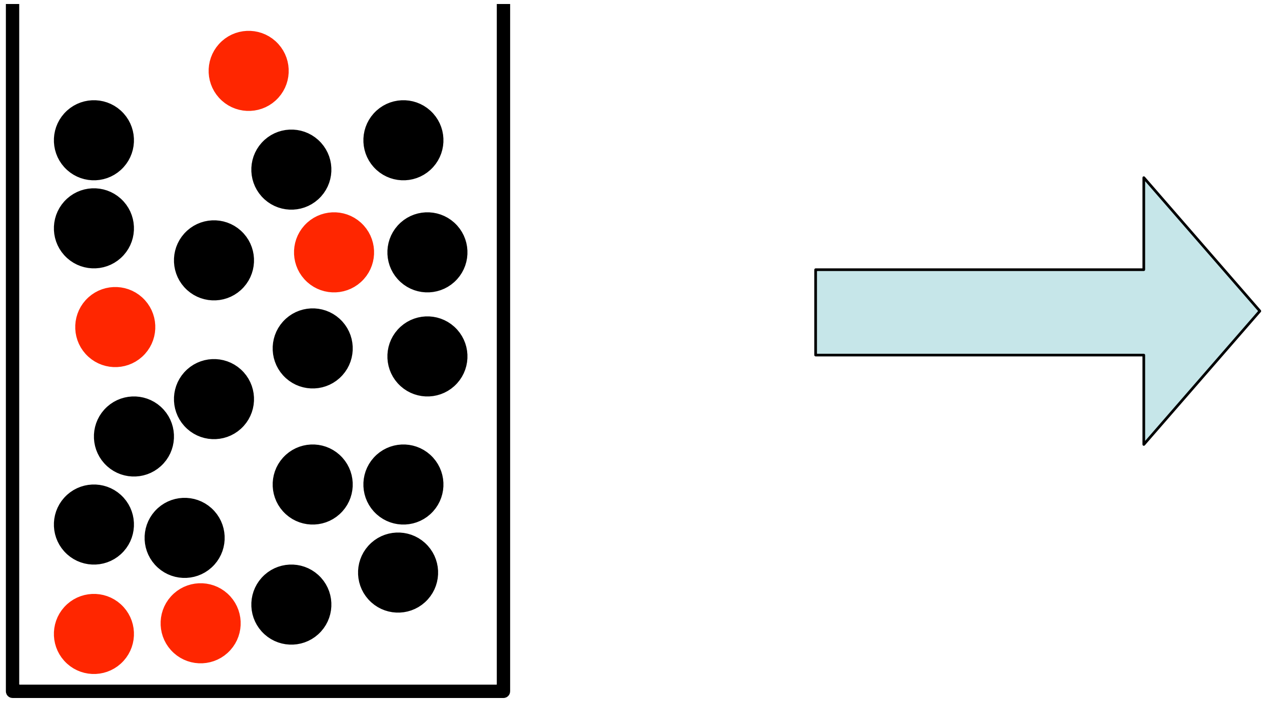


How many overlapping genes?

Is this overlap larger than expected by random sampling of the non-DE genes?

Fisher’s exact test

Null hypothesis: List is a random sample from the whole population.
Alternative hypothesis: More red genes (DE genes) than expected.



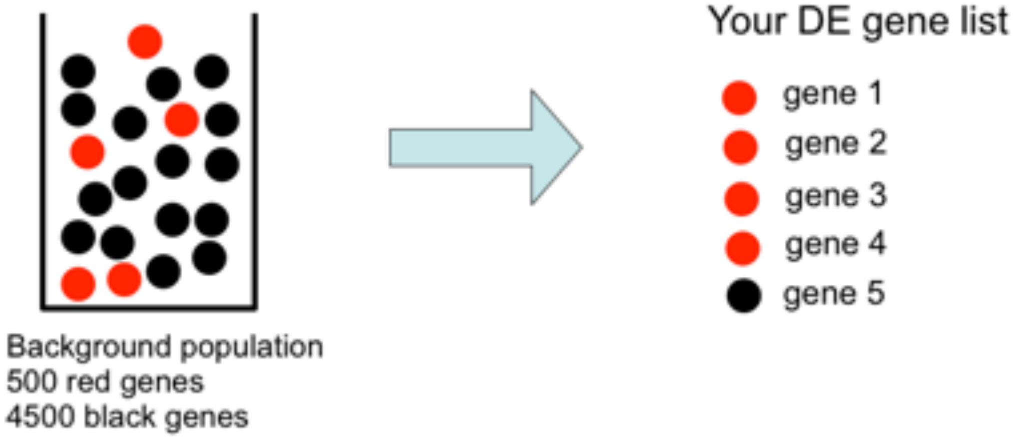
Your DE gene list

🔴 gene 1
🔴 gene 2
🔴 gene 3
🔴 gene 4
● gene 5

Background population
500 red genes
4500 black genes

Fisher’s exact test

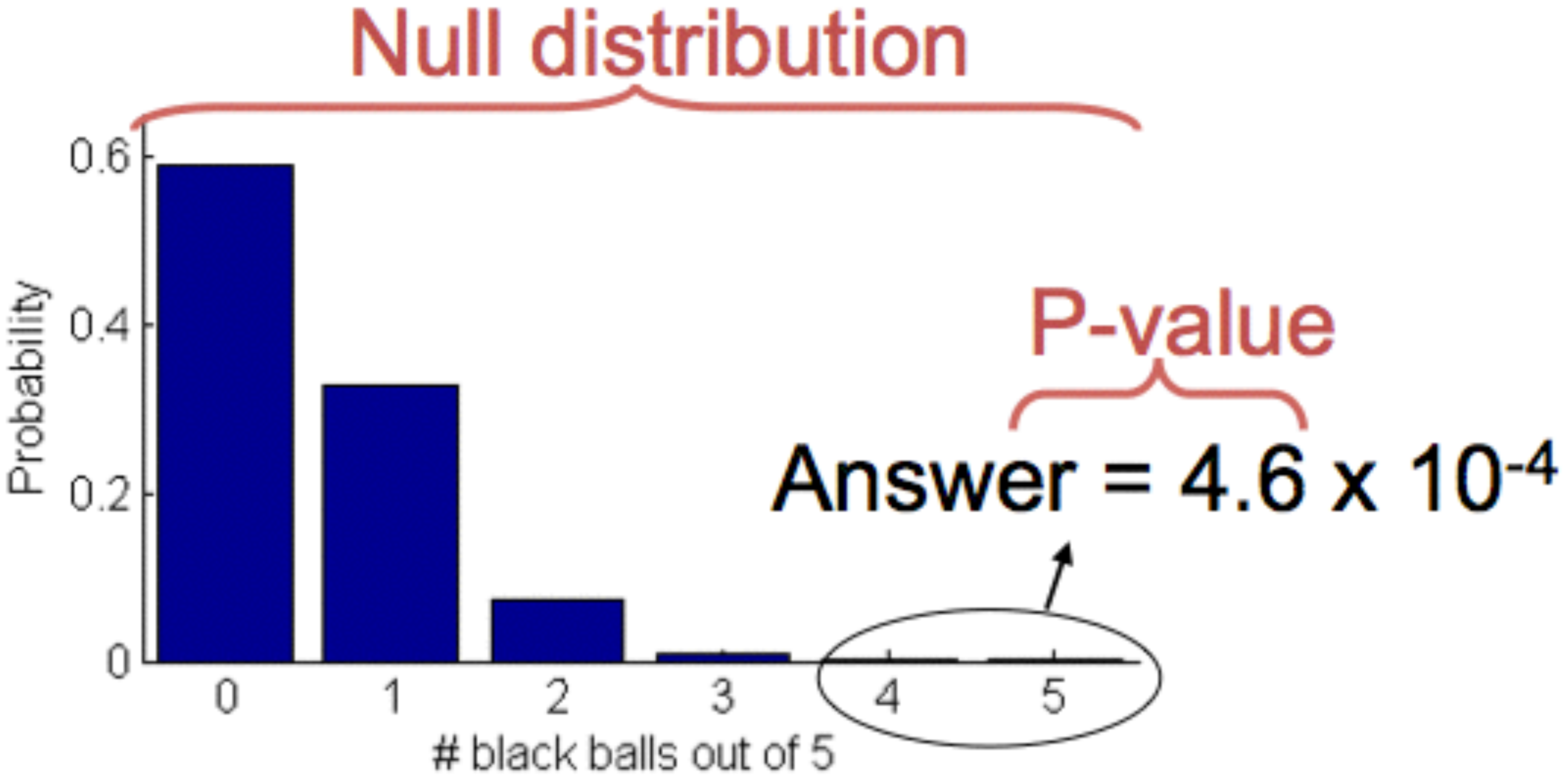
Null hypothesis: List is a random sample from the whole population.
Alternative hypothesis: More red genes (DE genes) than expected.



Background population
500 red genes
4500 black genes

Your DE gene list

- ● gene 1
- ● gene 2
- ● gene 3
- ● gene 4
- ● gene 5



Null distribution

P-value

Answer = 4.6×10^{-4}

Fisher’s exact test - tips

- Could test either over-enrichment or under-enrichment.
- Need to choose “background population”.
- Need to correct P-value for multiple testing problem. Multiple testing corrections adjust p-values derived from multiple statistical tests to correct for occurrence of false positives.

P-values vs. Q-values

- Corrected P-value is greater than or equal to the probability that any single one of the observed enrichments could be due to random draws.
    - Bonferroni: Corrected P-value = number of test * original P-value
- Bonferroni correction is very stringent and can wash away real enrichments.
- Often users are willing to accept a less stringent condition, the "false discovery rate" (FDR), which leads to a gentler correction when there are real enrichments.
- Typically FDR corrections are calculated using the Benjamini-Hochberg procedure.
- FDR threshold is often called the "q-value".

Benjamini-Hochberg

1) The p-values of each gene are ranked from the smallest to the largest.

2) The largest p-value remains as it is.

3) The second largest p-value is multiplied by the total number of genes in gene list divided by its rank. If less than 0.05, it is significant.

   - Corrected p-value = p-value*(n/n-1) < 0.05, if so, gene is significant.

4) The third p-value is multiplied as in step 3:

   - Corrected p-value = p-value*(n/n-2) < 0.05, if so, gene is significant.

Enrichment test tools

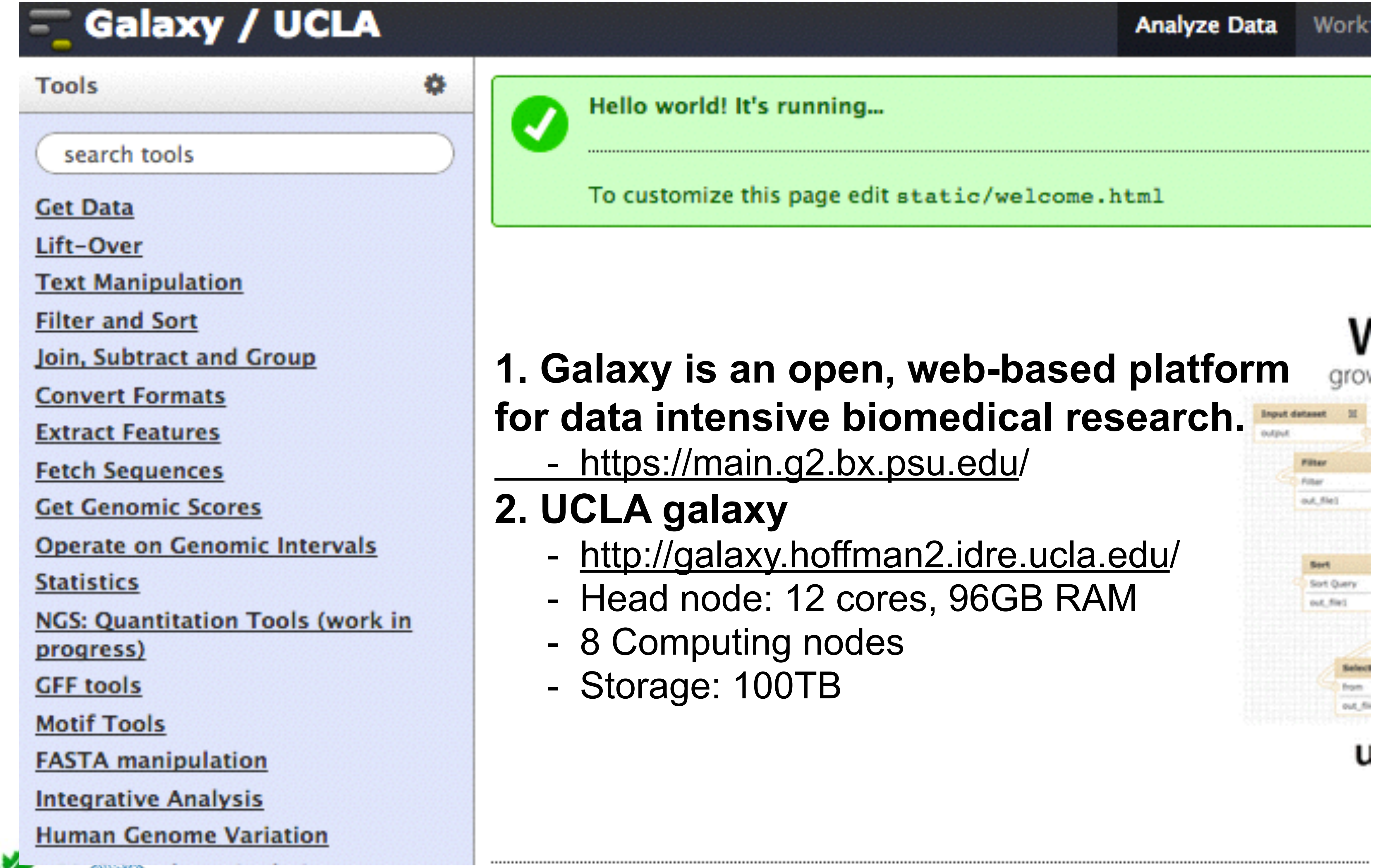
Ranked list (semi- quantitative)

Commercialized tool

Warm up exercise

1. Please download datasets:
   - http://tinyurl.com/kg3kxkt
2. Unzip it
3. Open UCLA Galaxy and log in:
   - http://galaxy.hoffman2.idre.ucla.edu/

Warm up exercise



Hello world! It's running...

To customize this page edit `static/welcome.html`

1. Galaxy is an open, web-based platform for data intensive biomedical research.

- https://main.g2.bx.psu.edu/

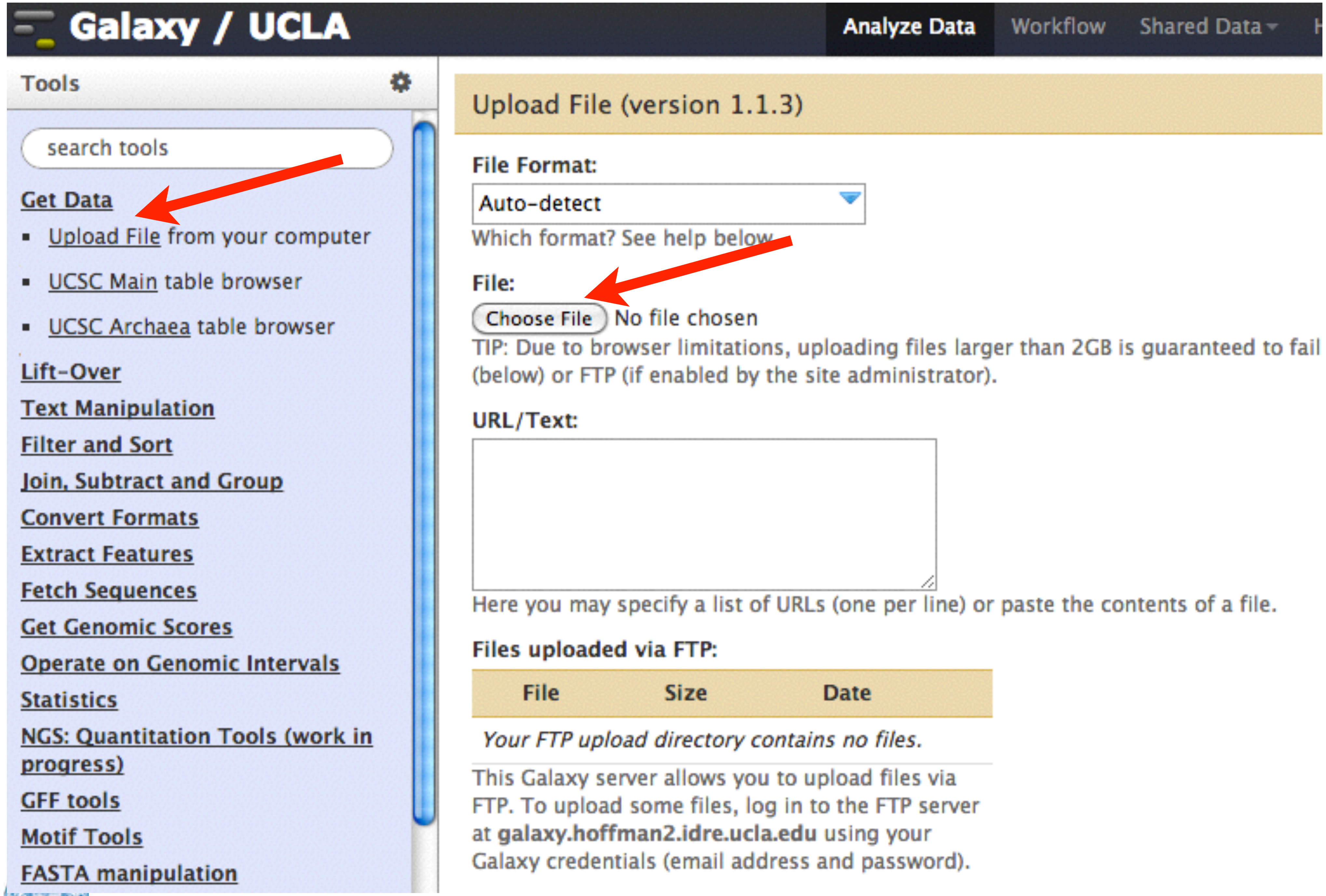
2. UCLA galaxy

- http://galaxy.hoffman2.idre.ucla.edu/
- Head node: 12 cores, 96GB RAM
- 8 Computing nodes
- Storage: 100TB

CIHR IRSC

Warm up exercise

4. Load the data onto Galaxy



Upload File (version 1.1.3)

File Format:

Auto-detect

Which format? See help below

File:

No file chosen

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail (below) or FTP (if enabled by the site administrator).

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

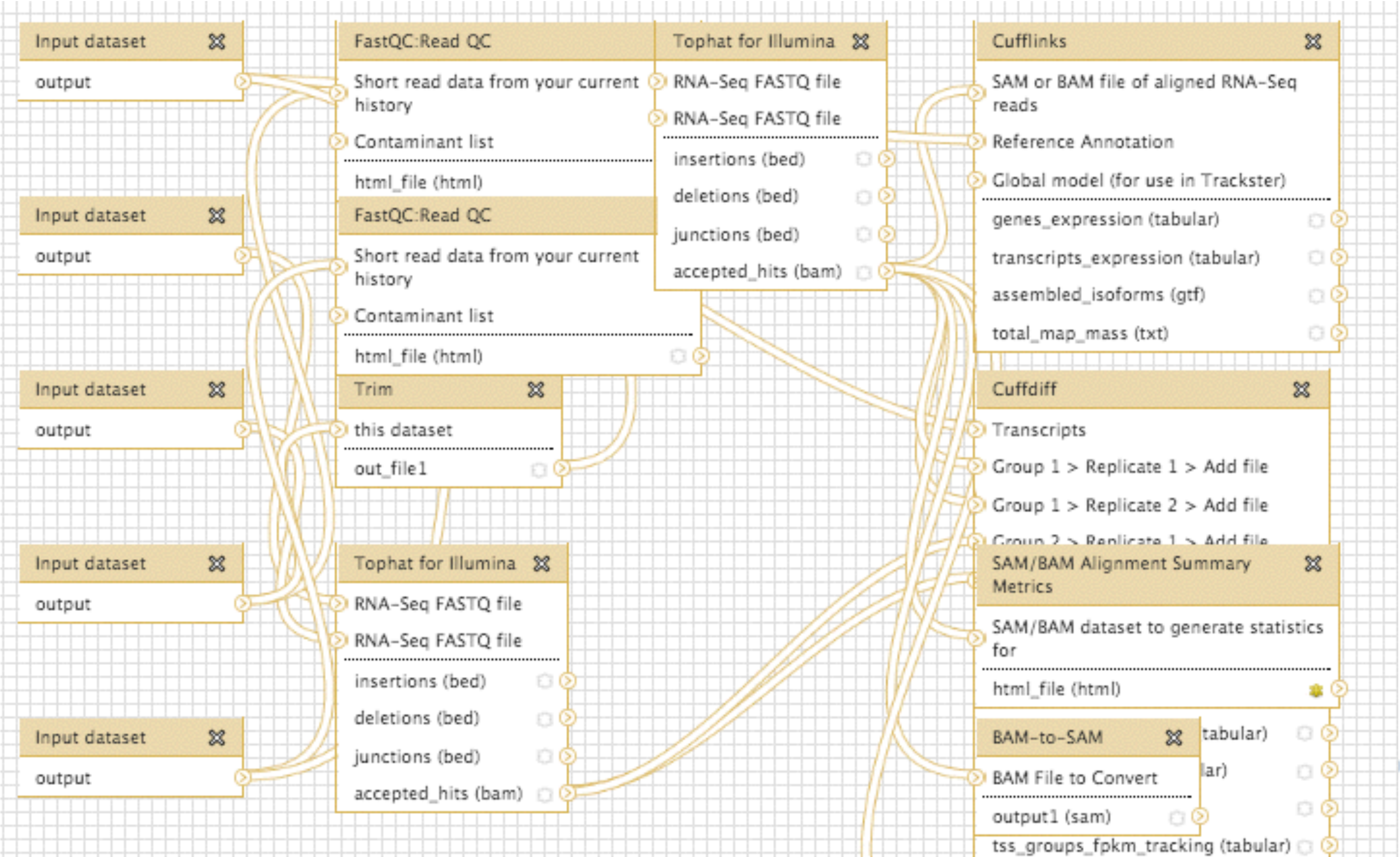
File	Size	Date
Your FTP upload directory contains no files.		

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at **galaxy.hoffman2.idre.ucla.edu** using your Galaxy credentials (email address and password).

Warm up exercise

5. Download the “Workflow”

A workflow is a **sequential** collection of Galaxy operations to complete an analysis.



Input dataset
output

Input dataset
output

Input dataset
output

Input dataset
output

Input dataset
output

FastQC:Read QC
Short read data from your current history
Contaminant list
html_file (html)

FastQC:Read QC
Short read data from your current history
Contaminant list
html_file (html)

Trim
this dataset
out_file1

Tophat for Illumina
RNA-Seq FASTQ file
RNA-Seq FASTQ file
insertions (bed)
deletions (bed)
junctions (bed)
accepted_hits (bam)

Tophat for Illumina
RNA-Seq FASTQ file
RNA-Seq FASTQ file
insertions (bed)
deletions (bed)
junctions (bed)
accepted_hits (bam)

Cufflinks
SAM or BAM file of aligned RNA-Seq reads
Reference Annotation
Global model (for use in Trackster)
genes_expression (tabular)
transcripts_expression (tabular)
assembled_isoforms (gtf)
total_map_mass (txt)

Cuffdiff
Transcripts
Group 1 > Replicate 1 > Add file
Group 1 > Replicate 2 > Add file
Group 2 > Replicate 1 > Add file

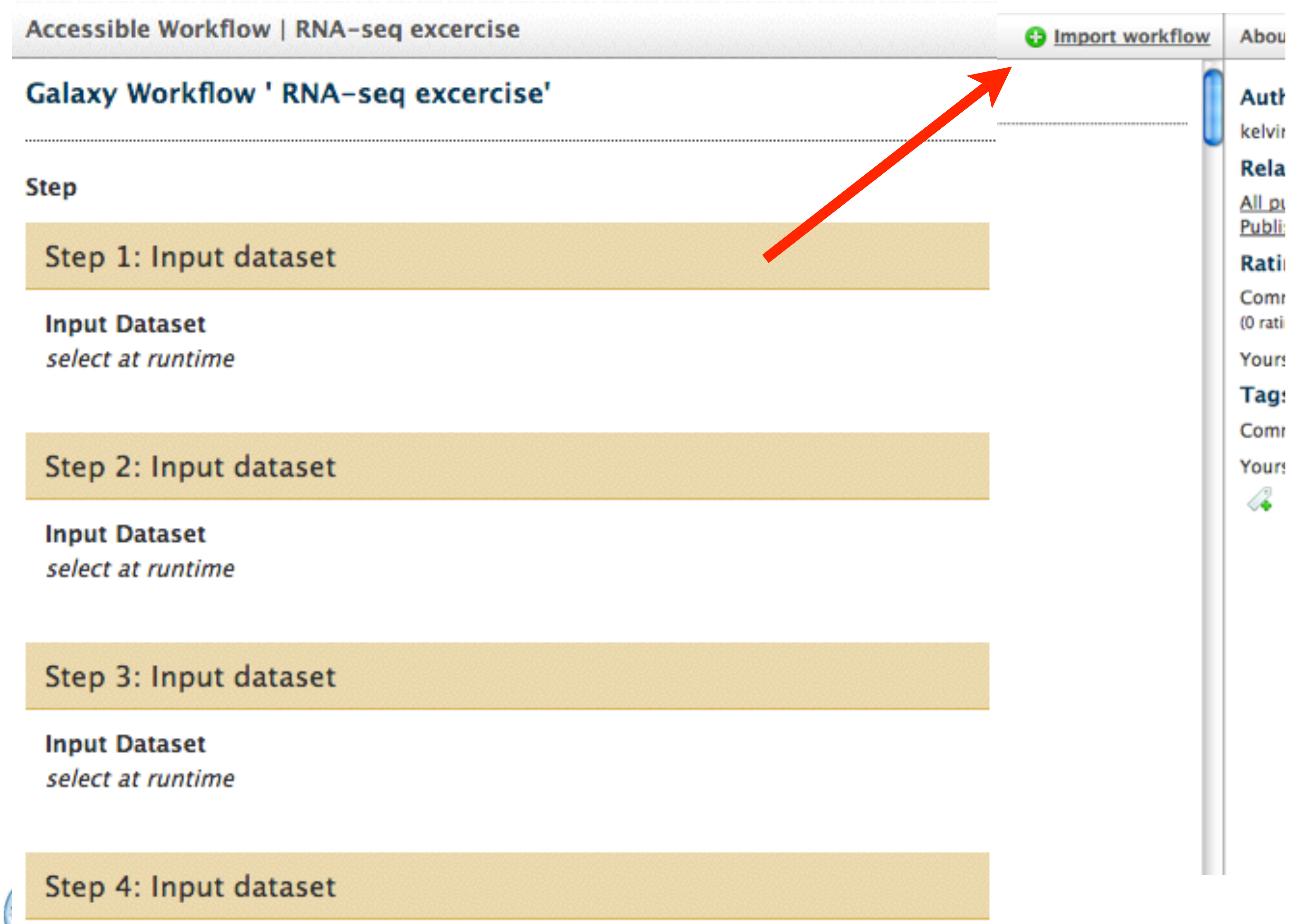
SAM/BAM Alignment Summary Metrics
SAM/BAM dataset to generate statistics for
html_file (html)

BAM-to-SAM
BAM File to Convert
output1 (sam)

tabular)
lar)
tss_groups_fpkm_tracking (tabular)

Warm up exercise

http://galaxy.hoffman2.idre.ucla.edu/u/kelvin-zhang/w/rna-seq-excercise



Accessible Workflow | RNA-seq excercise

Galaxy Workflow ' RNA-seq excercise'

Step

Step 1: Input dataset

Input Dataset
select at runtime

Step 2: Input dataset

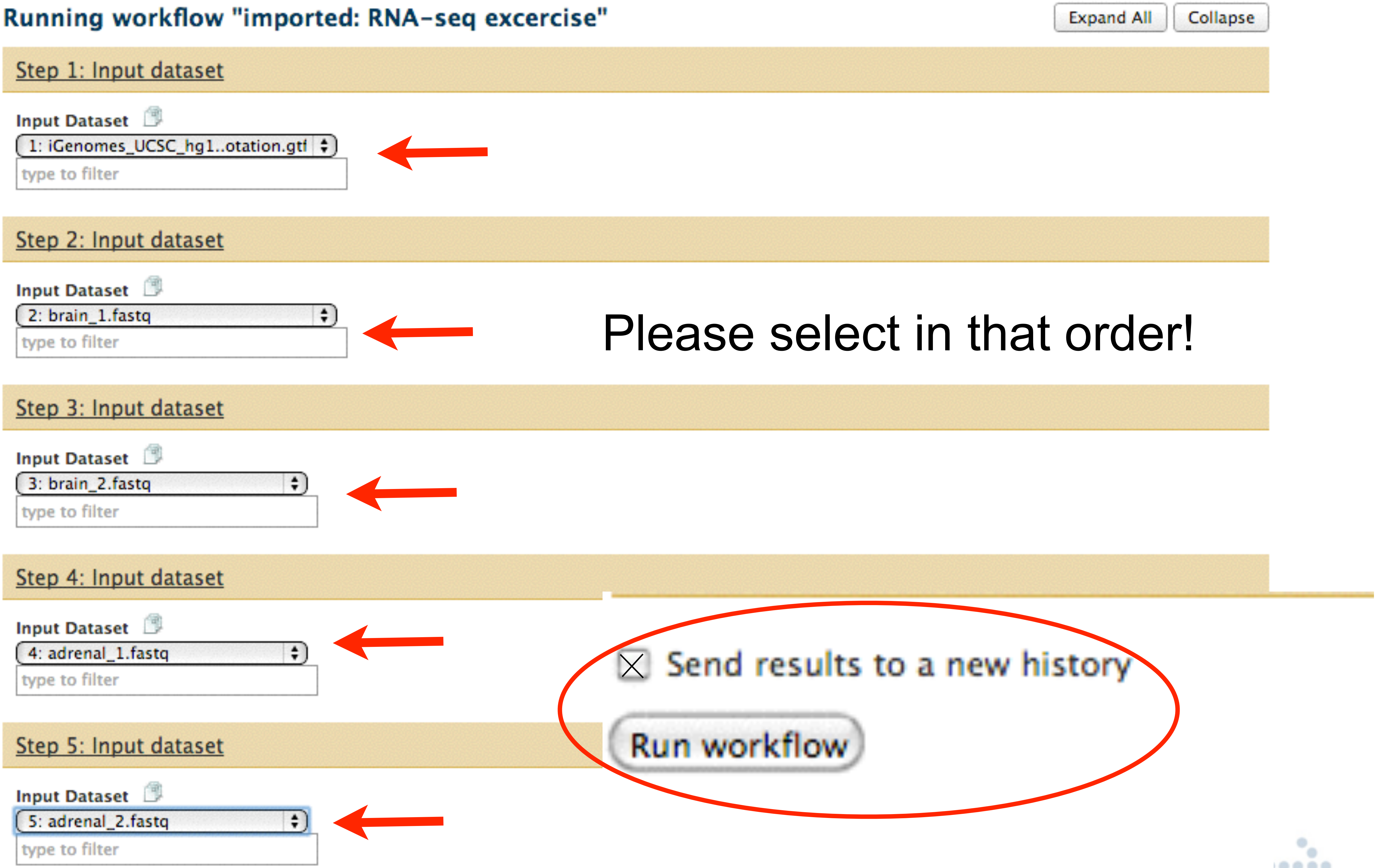
Input Dataset
select at runtime

Step 3: Input dataset

Input Dataset
select at runtime

Step 4: Input dataset

Warm up exercise



Running workflow "imported: RNA-seq excercise"

Expand All | Collapse

Step 1: Input dataset

Input Dataset
1: iGenomes_UCSC_hg1..otation.gtf
type to filter

Step 2: Input dataset

Input Dataset
2: brain_1.fastq
type to filter

Step 3: Input dataset

Input Dataset
3: brain_2.fastq
type to filter

Step 4: Input dataset

Input Dataset
4: adrenal_1.fastq
type to filter

Step 5: Input dataset

Input Dataset
5: adrenal_2.fastq
type to filter

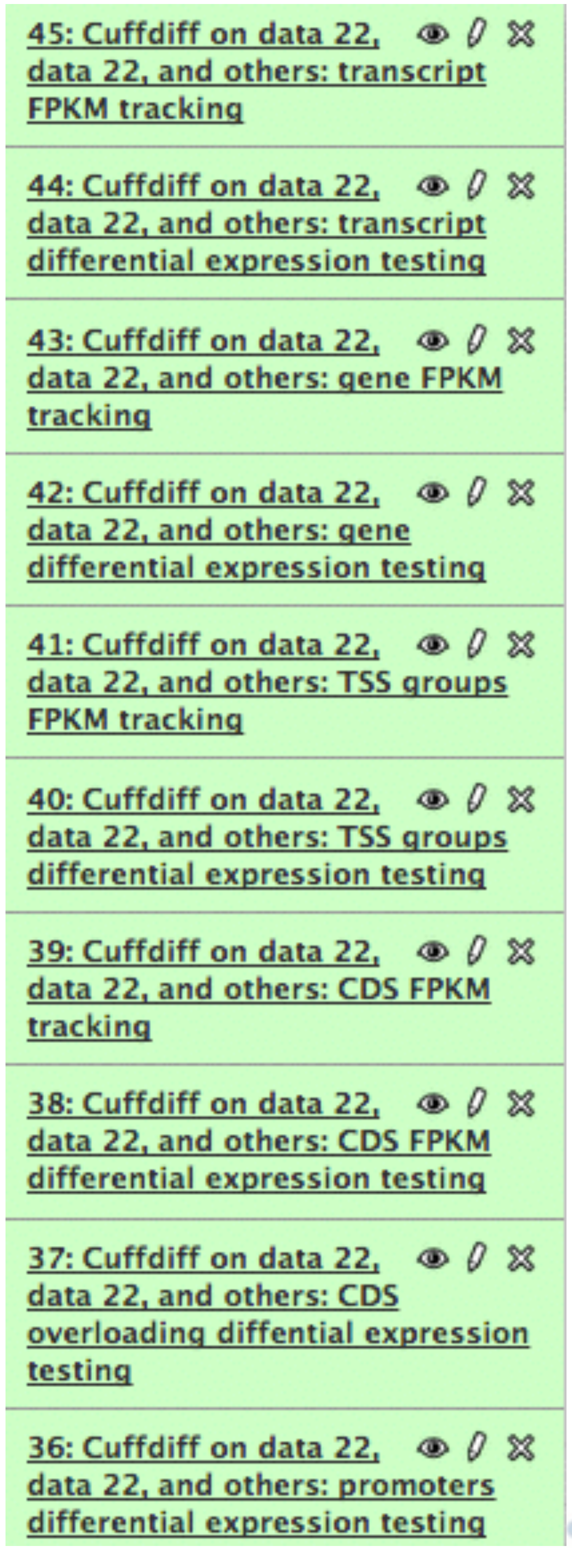
☒ Send results to a new history

Run workflow

Please select in that order!

Warm up exercise

1. Check the scheduled jobs.
2. It usually takes several hours to finish this workflow.



45: Cuffdiff on data 22, data 22, and others: transcript FPKM tracking

44: Cuffdiff on data 22, data 22, and others: transcript differential expression testing

43: Cuffdiff on data 22, data 22, and others: gene FPKM tracking

42: Cuffdiff on data 22, data 22, and others: gene differential expression testing

41: Cuffdiff on data 22, data 22, and others: TSS groups FPKM tracking

40: Cuffdiff on data 22, data 22, and others: TSS groups differential expression testing

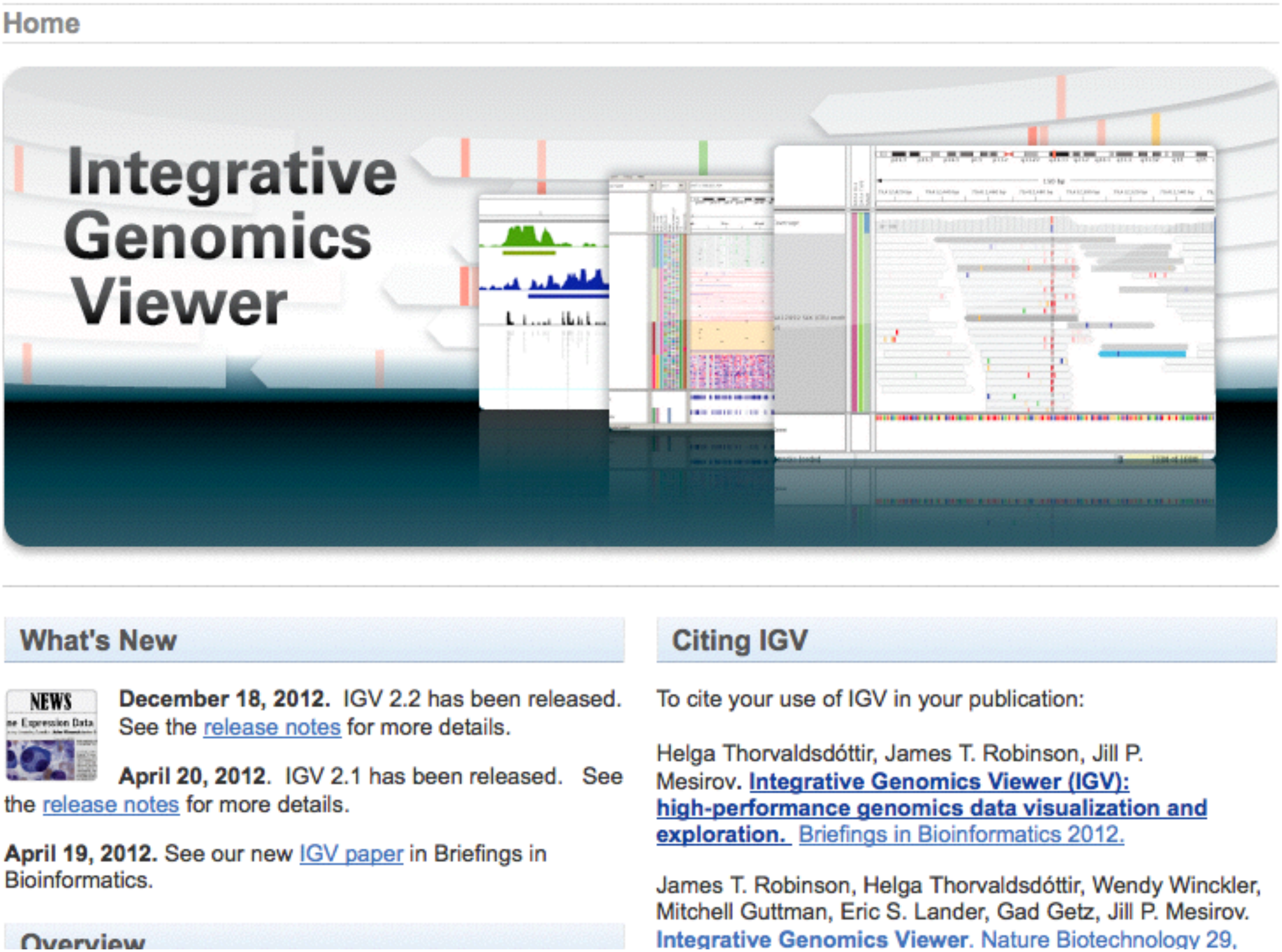
39: Cuffdiff on data 22, data 22, and others: CDS FPKM tracking

38: Cuffdiff on data 22, data 22, and others: CDS FPKM differential expression testing

37: Cuffdiff on data 22, data 22, and others: CDS overloading diffential expression testing

36: Cuffdiff on data 22, data 22, and others: promoters differential expression testing

http://www.broadinstitute.org/igv/



Home

What's New

December 18, 2012. IGV 2.2 has been released. See the release notes for more details.

April 20, 2012. IGV 2.1 has been released. See the release notes for more details.

April 19, 2012. See our new IGV paper in Briefings in Bioinformatics.

Overview

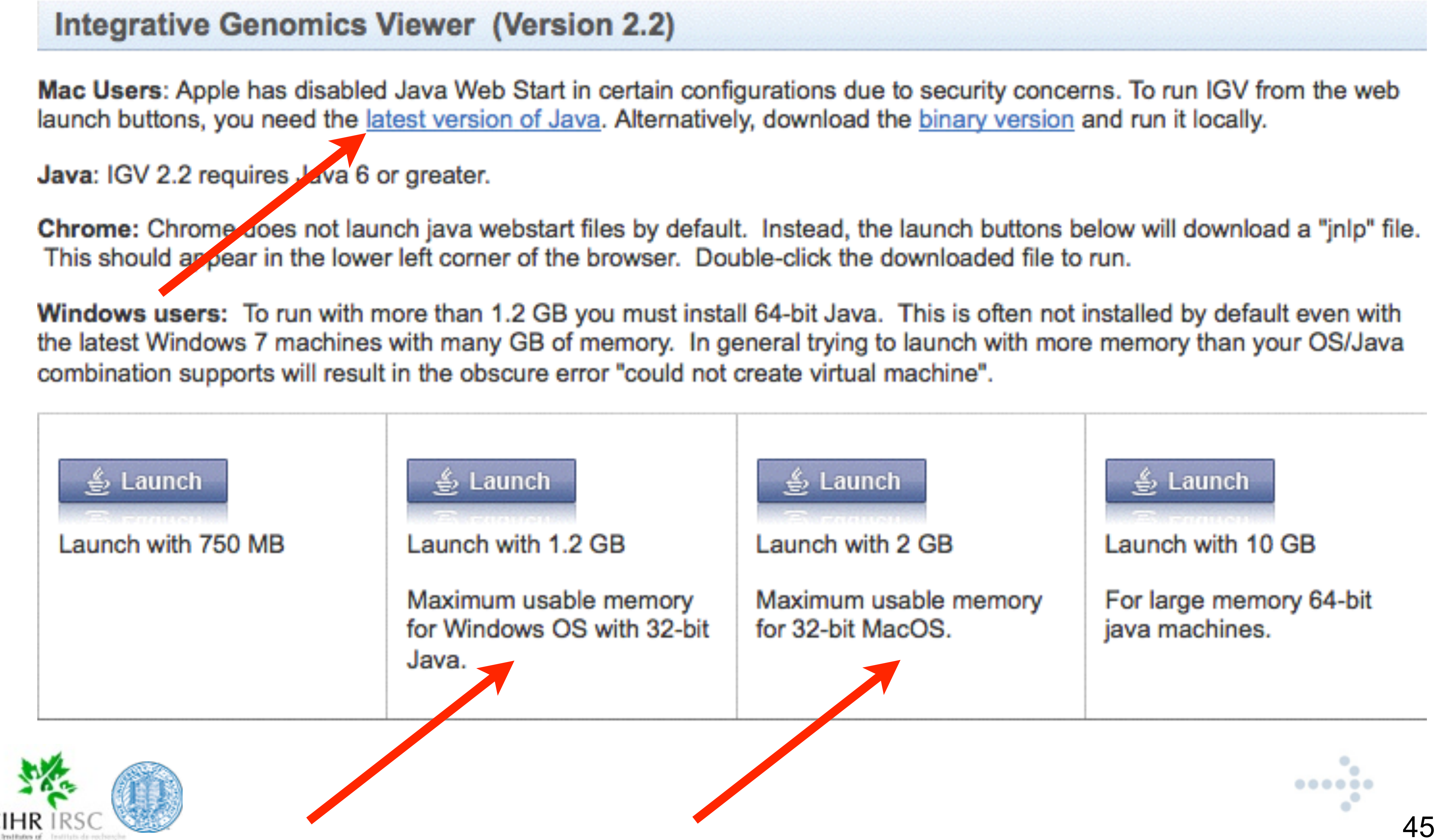
Citing IGV

To cite your use of IGV in your publication:

Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.** Briefings in Bioinformatics 2012.

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. **Integrative Genomics Viewer**. Nature Biotechnology 29,

http://www.broadinstitute.org/igv/



Integrative Genomics Viewer (Version 2.2)

Mac Users: Apple has disabled Java Web Start in certain configurations due to security concerns. To run IGV from the web launch buttons, you need the latest version of Java. Alternatively, download the binary version and run it locally.

Java: IGV 2.2 requires Java 6 or greater.

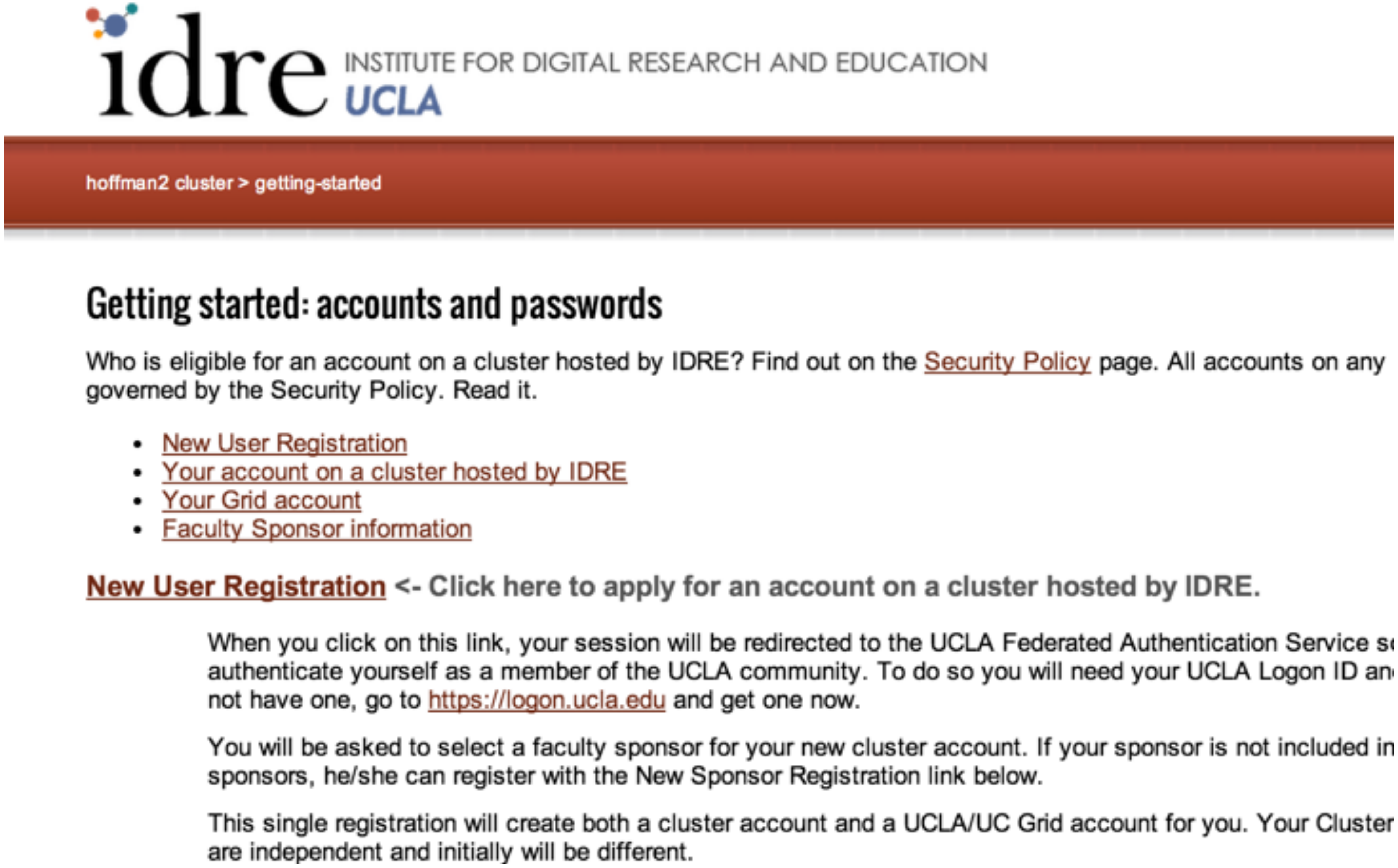
Chrome: Chrome does not launch java webstart files by default. Instead, the launch buttons below will download a "jnlp" file. This should appear in the lower left corner of the browser. Double-click the downloaded file to run.

Windows users: To run with more than 1.2 GB you must install 64-bit Java. This is often not installed by default even with the latest Windows 7 machines with many GB of memory. In general trying to launch with more memory than your OS/Java combination supports will result in the obscure error "could not create virtual machine".

Launch Launch with 750 MB	Launch Launch with 1.2 GB Maximum usable memory for Windows OS with 32-bit Java.	Launch Launch with 2 GB Maximum usable memory for 32-bit MacOS.	Launch Launch with 10 GB For large memory 64-bit java machines.

How to get a hoffman account?

http://hpc.ucla.edu/hoffman2/getting-started/getting-started.php



idre INSTITUTE FOR DIGITAL RESEARCH AND EDUCATION UCLA

hoffman2 cluster > getting-started

Getting started: accounts and passwords

Who is eligible for an account on a cluster hosted by IDRE? Find out on the Security Policy page. All accounts on any governed by the Security Policy. Read it.

- New User Registration
- Your account on a cluster hosted by IDRE
- Your Grid account
- Faculty Sponsor information

New User Registration <- Click here to apply for an account on a cluster hosted by IDRE.

When you click on this link, your session will be redirected to the UCLA Federated Authentication Service s authenticate yourself as a member of the UCLA community. To do so you will need your UCLA Logon ID an not have one, go to https://logon.ucla.edu and get one now.

You will be asked to select a faculty sponsor for your new cluster account. If your sponsor is not included in sponsors, he/she can register with the New Sponsor Registration link below.

This single registration will create both a cluster account and a UCLA/UC Grid account for you. Your Cluster are independent and initially will be different.

Questions/Discussion.