RNA-Seq

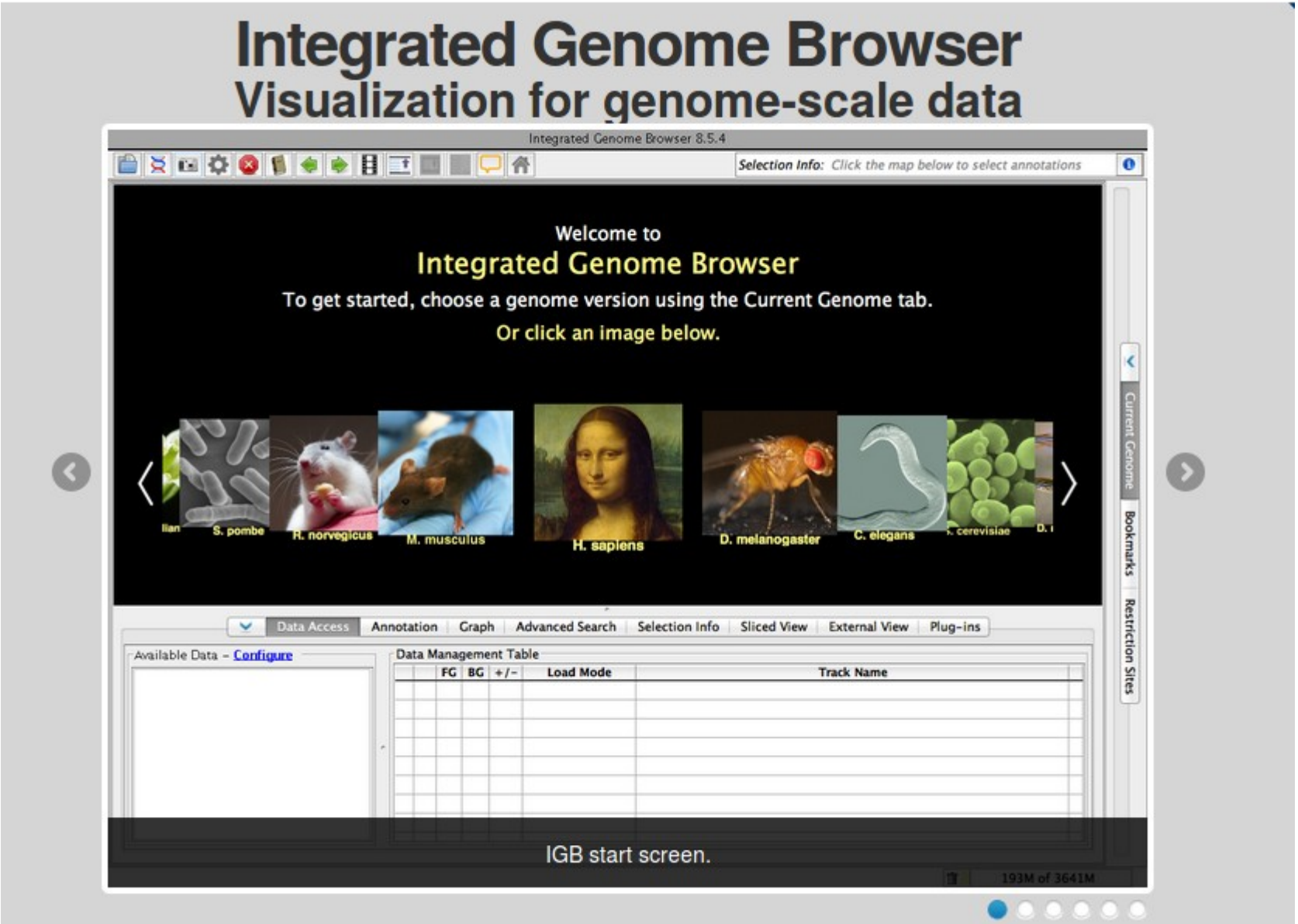
RNA-Seq

- UCSC genome browser
- Gene structure
- BED file
- RNA-Isolation
- Splicing (Intron retention, Exon skipping..)
- Gold Standards (RT-PCR)
- Sampling uncertainty (Urn model)
- Coverage

RNA-Seq

- Biological vs Technical Replication
- High coverage vs More replicates
- Randomize samples over lanes and barcodes
- Getting splice junction reads (Junction Reads)
- Normalization for comparison
- FPKM and RPKM calculations

IGB


Integrated Genome Browser
Visualization for genome-scale data
Integrated Genome Browser 8.5.4
Selection Info: Click the map below to select annotations
Welcome to
Integrated Genome Browser
To get started, choose a genome version using the Current Genome tab.
Or click an image below.
S. pombe
R. norvegicus
M. musculus
H. sapiens
D. melanogaster
C. elegans
Current Genome
Bookmarks
Restriction Sites
Data Access
Annotation
Graph
Advanced Search
Selection Info
Sliced View
External View
Plug-ins
Available Data – Configure
Data Management Table
FG
BG
+/-
Load Mode
Track Name
IGB start screen.
193M of 3641M

Download and Install IGB

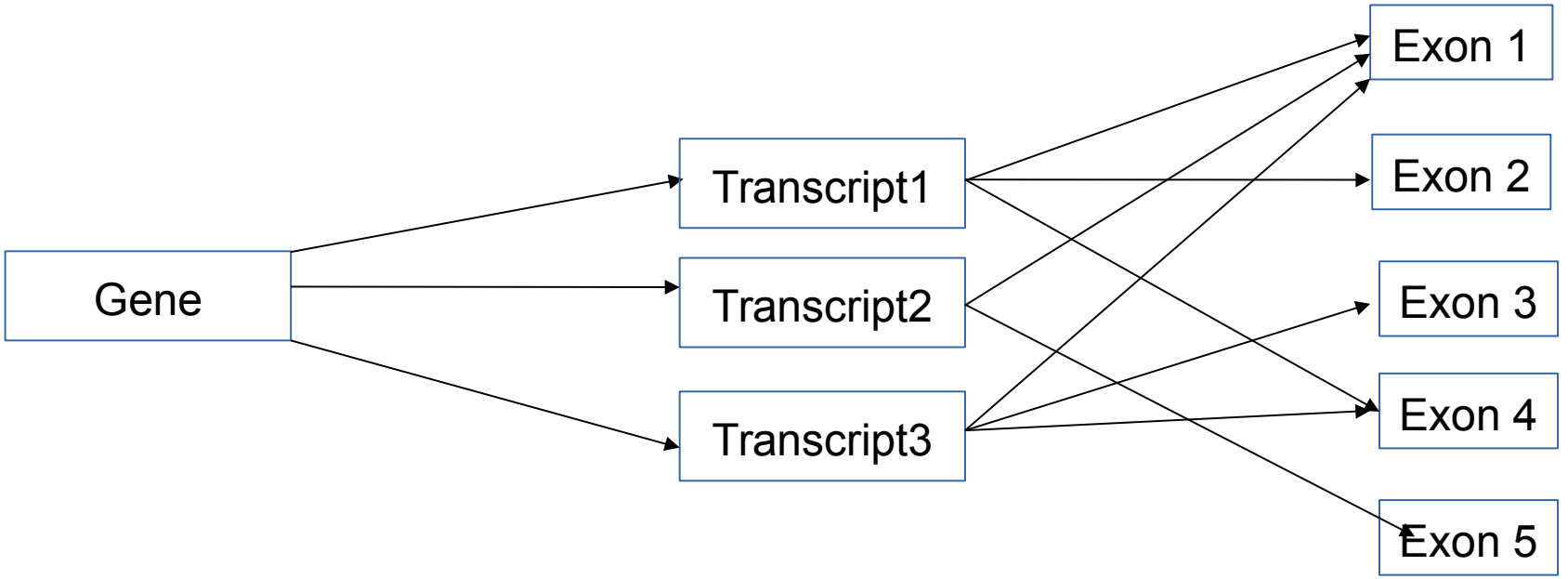
- http://bioviz.org/igb/index.html
- Load Genome arb.fa (~/RNA_Seq/ref/arb.fa)
- Load annotation arb.gtf (~/RNA_Seq/ref/arb.gtf)

Problem

- Browse around IGB with the custom genome arb and find the following for the fist and the last gene-
  - Strand
  - Gene name /transcript name
  - No of Splice variants
  - No of Exons/Introns

GTF file format

- General transfer format
- http://www.ensembl.org/info/website/upload/gff.html
- Inheritance
- Parsing Nightmare



Ding et al. BMC Genomics 2014, 15:431
http://www.biomedcentral.com/1471-2164/15/431

BMC Genomics

RESEARCH ARTICLE Open Access

Genome-wide analysis of alternative splicing of pre-mRNA under salt stress in Arabidopsis

Feng Ding†, Peng Cui†, Zhenyu Wang, Shoudong Zhang, Shahjahan Ali and Liming Xiong*

Abstract

Background: Alternative splicing (AS) of precursor mRNA (pre-mRNA) is an important gene regulation process that potentially regulates many physiological processes in plants, including the response to abiotic stresses such as salt stress.

Results: To analyze global changes in AS under salt stress, we obtained high-coverage (~200 times) RNA sequencing data from *Arabidopsis thaliana* seedlings that were treated with different concentrations of NaCl. We detected that ~49% of all intron-containing genes were alternatively spliced under salt stress, 10% of which experienced significant differential alternative splicing (DAS). Furthermore, AS increased significantly under salt stress compared with under unstressed conditions. We demonstrated that most DAS genes were not differentially regulated by salt stress, suggesting that AS may represent an independent layer of gene regulation in response to stress. Our analysis of functional categories suggested that DAS genes were associated with specific functional pathways, such as the pathways for the responses to stresses and RNA splicing. We revealed that serine/arginine-rich (SR) splicing factors were frequently and specifically regulated in AS under salt stresses, suggesting a complex loop in AS regulation for stress adaptation. We also showed that alternative splicing site selection (SS) occurred most frequently at 4 nucleotides upstream or downstream of the dominant sites and that exon skipping tended to link with alternative SS.

Aim for today

- Get RNA-Seq study from SRA
- Align with splice junction aligner
- Construct transcripts from splice junctions
- Compare with known transcripts
- Get FPKM values for each Gene/Transcript
- Do a differential expression testing
- Look at differentially expressed genes

Genomics

RESEARCH ARTICLE Open Access

Genome-wide analysis of alternative splicing of pre-mRNA under salt stress in Arabidopsis

Feng Ding†, Peng Cui†, Zhenyu Wang, Shoudong Zhang, Shahjahan Ali and Liming Xiong*

Problem : Get the bioproject details. What is the case condition ? Are the library details reported ? How many Replicates for case and control ? Is the study design good ?

Sample Data

- The SRR files are stored in RNA_Seq folder in your home director
- A mock genome (1 Mb) is also kept there (arb.fa)
- Annotation GTF is also kept there (arb.gtf)
- The SRR files contain sampled reads ~ 150000 so that the analysis runs faster
- Locate the files

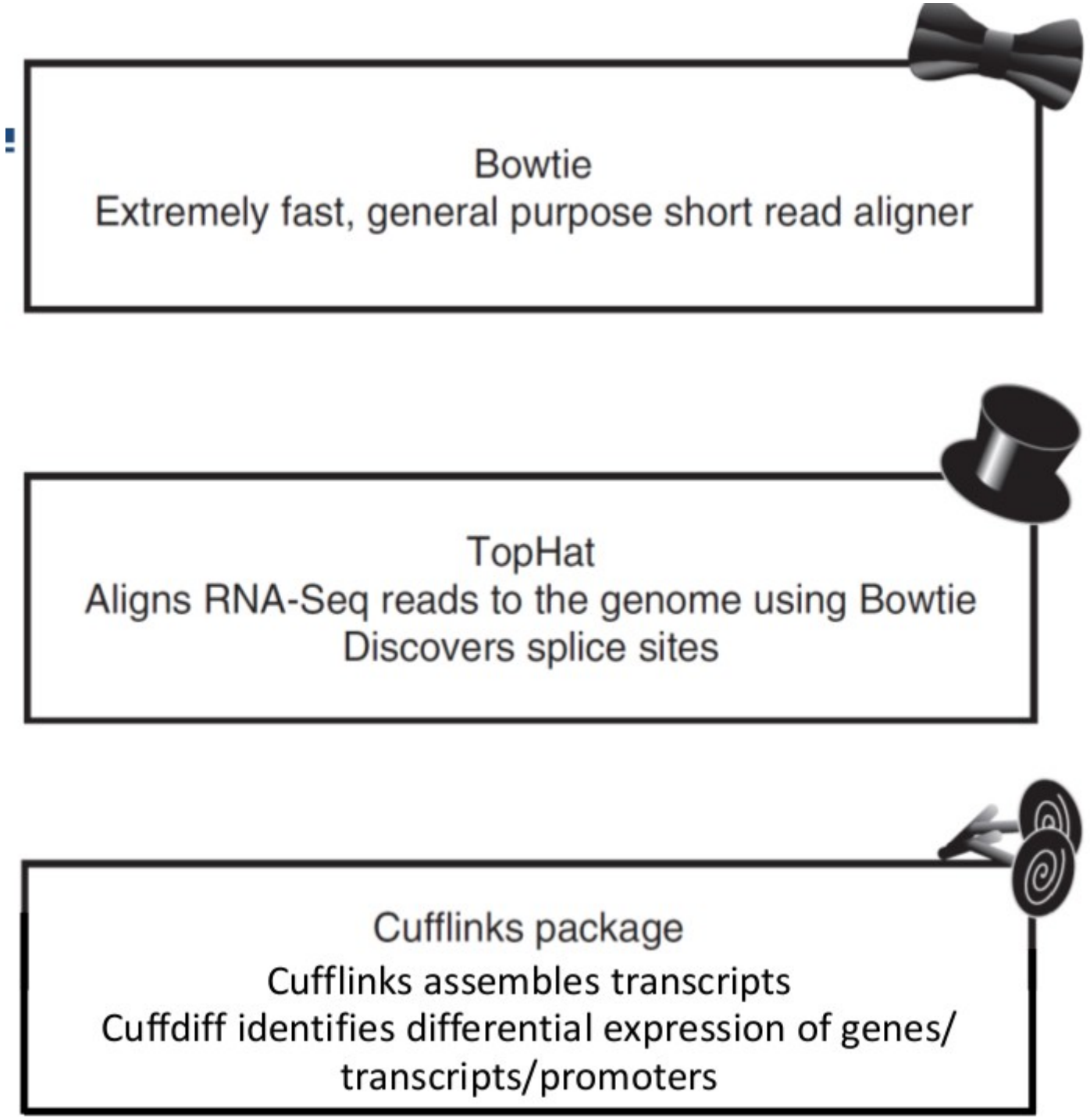
Tuxedo Suite

- Bowtie2 [Genome Alignment]
- Tophat [Splice Junction Alignment]
- Cufflinks [FPKM Estimation, Transcript resconstruction]

Analysis Scenarios

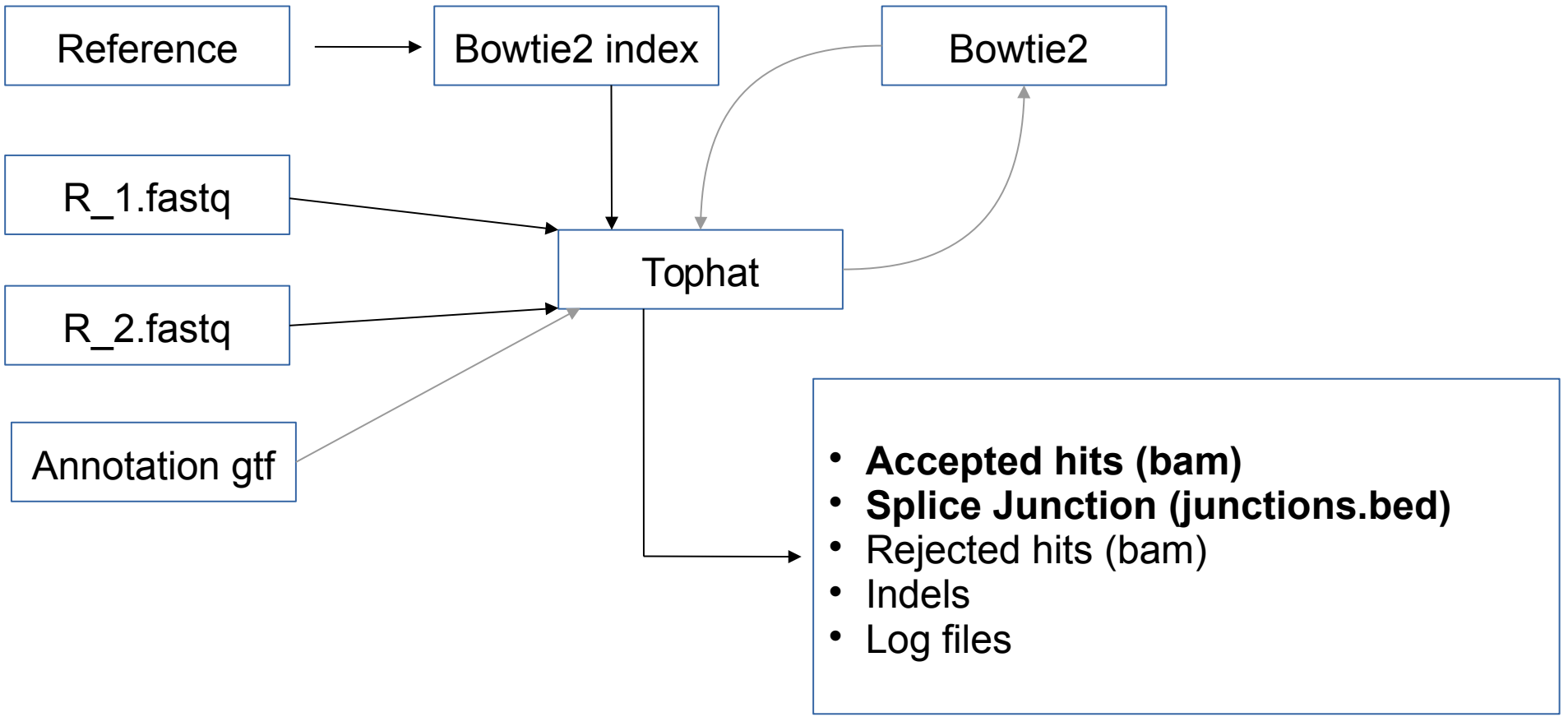
Known Well studied genome
GTF is available

Uknown genome
No GTF


Bowtie
Extremely fast, general purpose short read aligner
TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites
Cufflinks package
Cufflinks assembles transcripts
Cuffdiff identifies differential expression of genes/
transcripts/promoters

TOPHAT Run

- Is a Mapper for mapping reads across splice junctions
- Uses Bowtie2 for normal mapping
- Unmapped reads are then tried to be mapped with large insertions



Task

- Index genome using bowtie2-build
- Map reads using tophat

PROBLEM

- Load the bam file produced by tophat
- Load junction.bed file produced by tophat
- Find a gene with no expression in the sample!
- Find a gene with high expression in the sample!

Transcript Assembly with Cufflinks

- Deconvolutes transcript/splice_variant data
- Create transcripts from junction reads
- Use Coverage information
- Get expression values (FPKM) for each gene
- Uses complicated and sophisticated normalizations and statistical calculations
- Too many options. Might not improve results
- Takes long time to run

Running Cufflinks

- Simplest cufflinks run
  - Cufflinks -o <outdir> <input_bam>
- Output
  - Transcripts.gtf (GTF file of the assembled transcripts)
  - *_fpkm_tracking (FPKM values)

Problem

- Upload accepted_hits.bam in IGB
- Upload transcripts.gtf in IGB
- Upload arb.gtf (known genes) in IGB
- Color the two GTFs differently
- See how good cufflinks can predict splicing and existence of genes ?

Task

- Run Tophat on the remaining 3 samples

Cufflinks with genome info

- Cufflinks -g <known.gtf>
  - Will take help from known genes for assembly
  - Will still discover new genes
- Cufflinks -G <known.gtf>
  - Will only take known genes
  - No new discovery
- Rest is the same
- We use -g

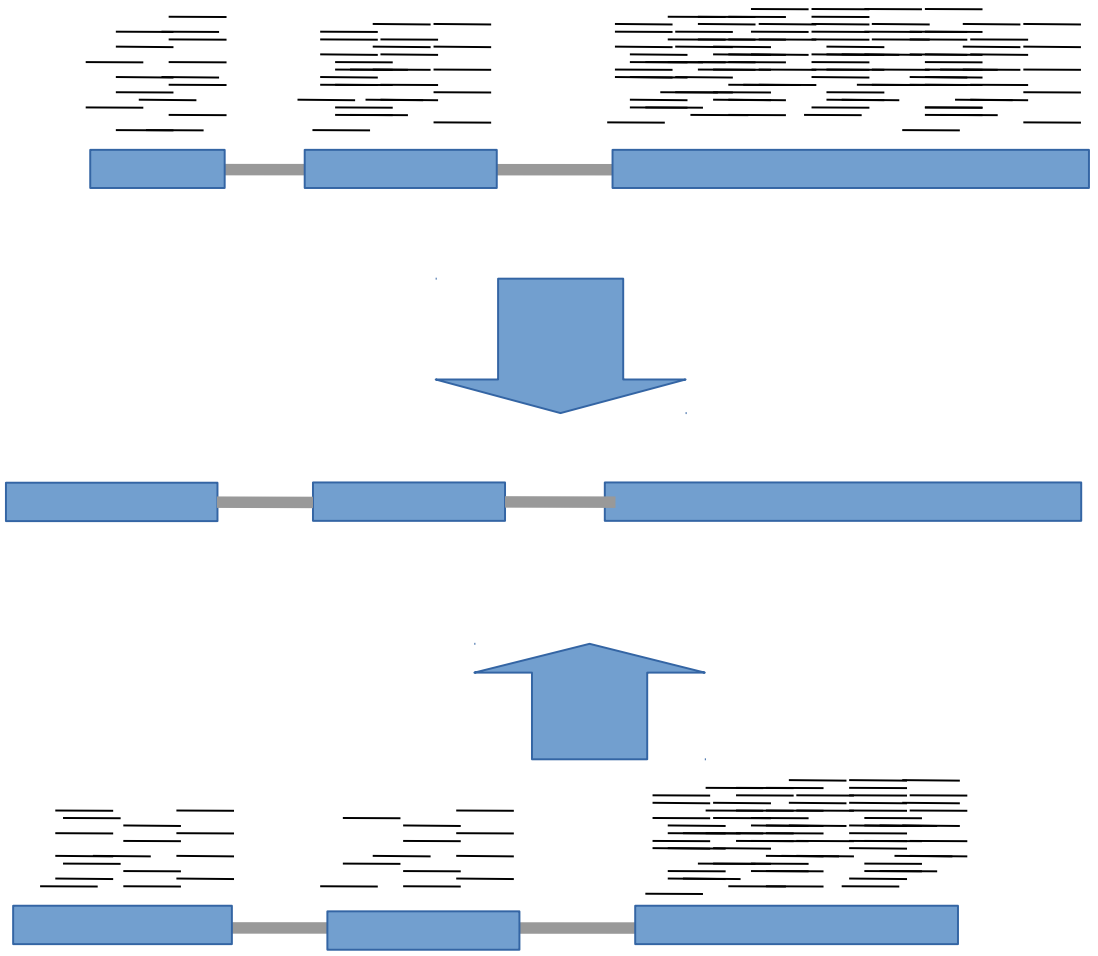
Task

- Run cufflinks in each tophat_out directory
- Use -g <arb.gtf>
- Open transcripts.gtf in a text editor
- Known transcripts will have their old ids (NM_) whereas newly discovered ones will have CUFF ids
- Find a new transcript which was not previously known.

Merging Transcripts

- Is using -g transcripts are called for each condition separately
- So they may differ slightly among conditions
- **For differential expression testing the same transcript must be used in case and control samples**
- Cuffmerge merges transcripts across assemblies

Why Use Cuffmerge



Running Cuffmerge

cuffmerge -g <reference.gtf> -s <ref.fasta> -o <outdir> <file with location of gtf inputs>

copy the path to all gtf files in a file and give as input !!!

Cuffdiff

- Does the actual differential expression testing
- Replicates are given as comma separated bam files
- Several bias correction steps involved
- Calculates expression variability from replicates and uses it for differential expression testing
- Does a FDR correction for multiple testing

Running cuffdiff

Cuffdiff -o <outdir> transcripts.gtf -L case,control
case_rep1.bam,case_rep2.bam...
control_rep1.bam,control_rep2.bam...

Output

- The .diff files will contain differential expression results

Problem

Which Genes are over expressed in response to salt stress ?

Problem

Try running cuffdiff with only one case replicate. How many genes do you find to be differentially expressed ?