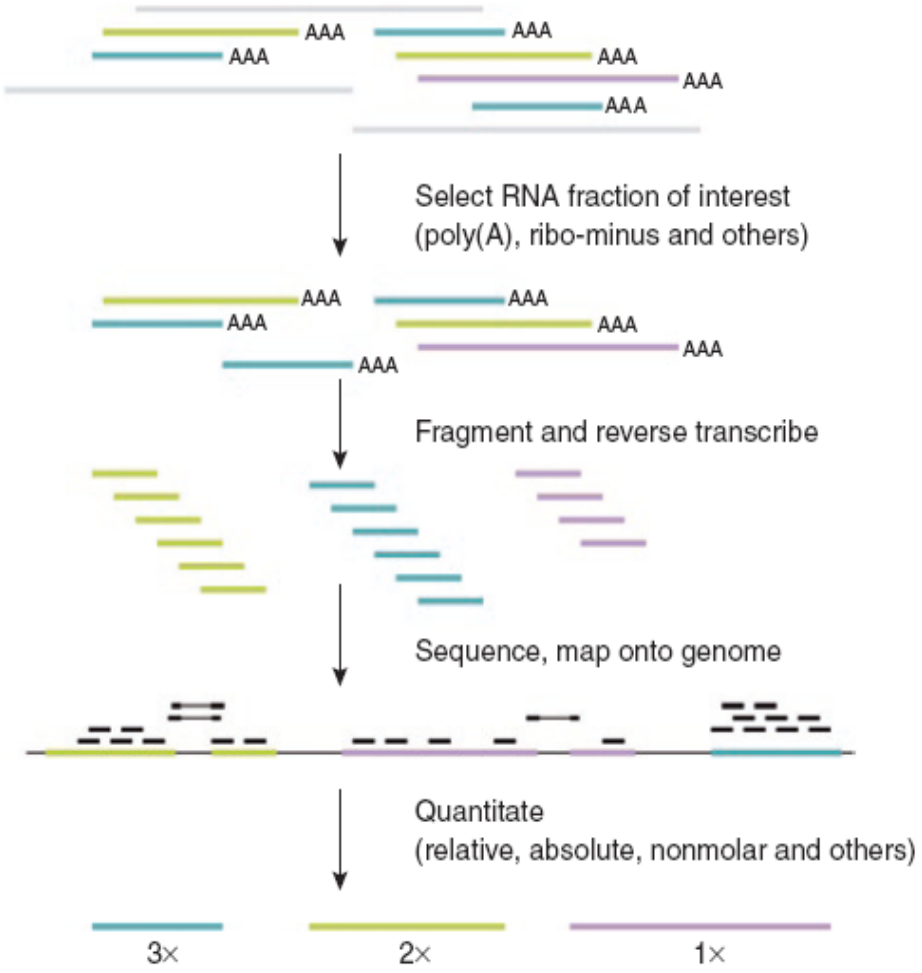
RNA-Seq Data Analysis

Jianlin Cheng, PhD
Computer Science Department
Informatics Institute
University of Missouri, Columbia
Spring, 2012

RNA-Seq – an Emerging, Powerful Approach to Studying Transcriptome

- Study genome-wide gene/RNA expression profiles
- Identify differentially expressed genes
- Recognize alternative splicing, isoforms, SNPs
- Recall very lowly expressed genes
- Identify novel transcripts
- Elucidate genes and pathways perturbed by botanical compounds

RNA-Seq Data Processing Steps



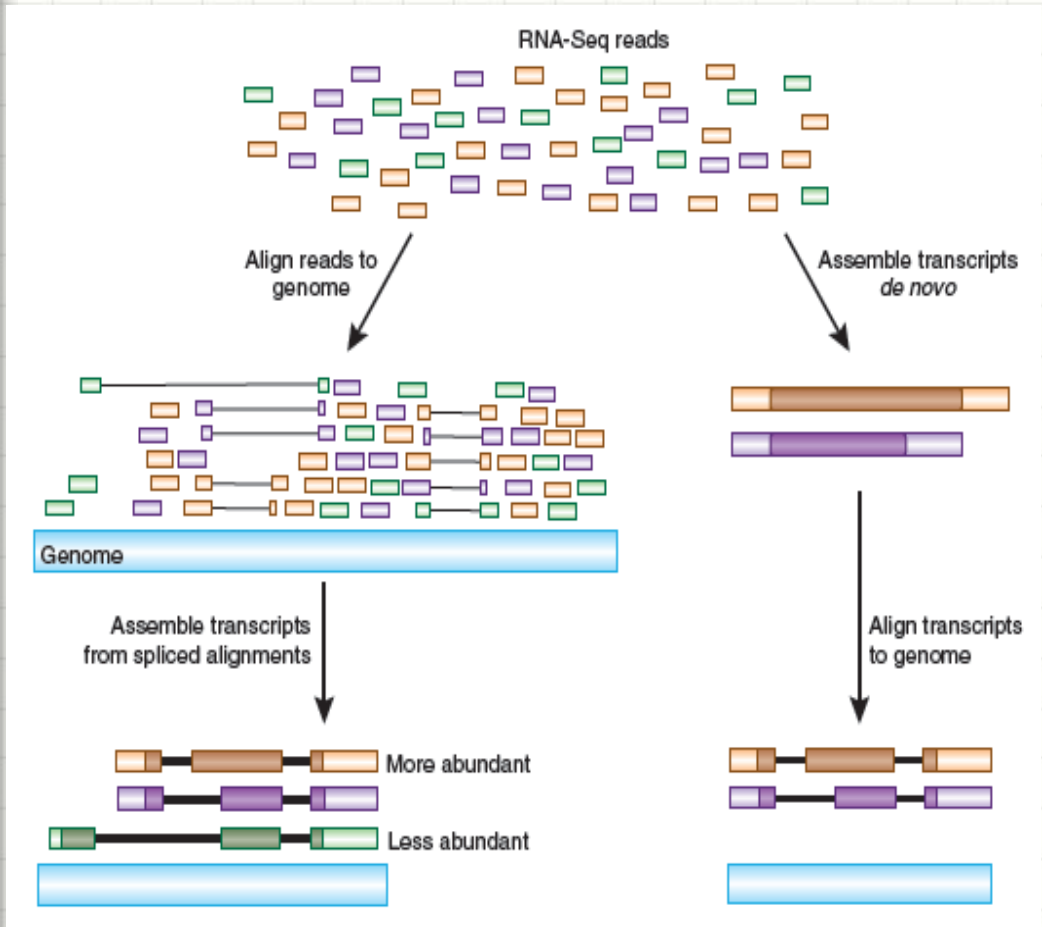
- Isolate RNA
- Prepare a RNA library
- RNA sequencing by NGS
- Reads mapping
- Quantification and analysis

S. Pepke et al., 2009

RNA-Seq Data

- Next Generation Sequencing (e.g. HiSeq2000 from Illumina)
- 8 lanes / samples per flow cell run, ~$120 per lane
- Tens of millions of short sequence single-end / paired-end reads (e.g. 50 nt), a few Gb bases
- Reads format: fastq (sequence + quality scores)

Reads Mapping and Assembly



- Directly map reads to reference genome
- Align map assembled reads to genome
- Unique mapping versus non-unique mapping
- Support splitting reads mapped to spliced exons
- Toleration of sequence variation and noise

Hass & Zody, 2010

Mapping Strategy

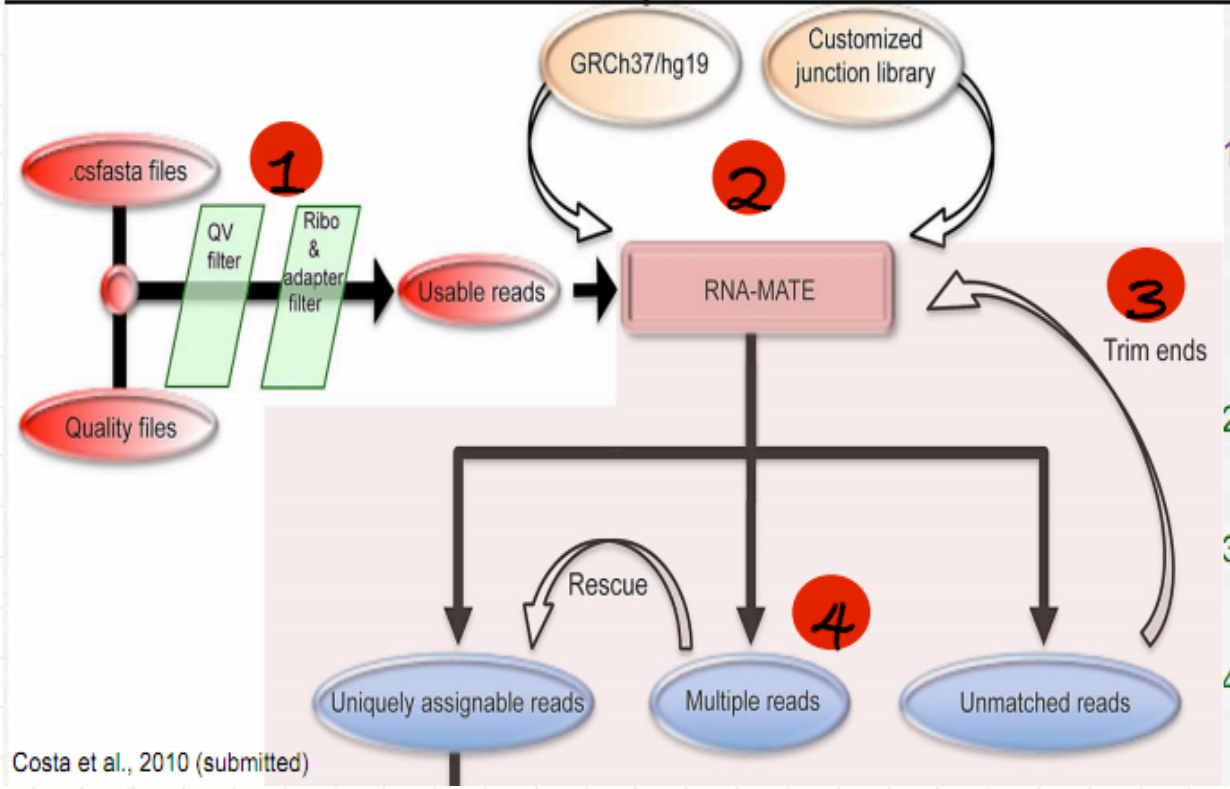
.csfasta

>852_2042_1999_F3T320112011230222013321101020110311320130233210023 03

.qual

>852_2042_1999_F319 20 14 14 8 5 9 16 11 11 6 14 21 14 11 21 -1 20 11 21 12 22 14 18 14 6 11 16 14 16 5 11 23 13 18 4 6 20 13 15 21 17 18 15 11 4 8 7 5 11

A suitable treatment of the multiple matched reads is fundamental to reduce the bias.



Costa et al., 2010 (submitted)

1. Quality assessment and filters (quality plot, remove low quality reads, ribosomal RNA reads, sequencing adapters);
2. Alignment to a reference genome (genome+junction library)
3. "Trim" the rigth-side of the reads and cyclically repeats the step;
4. Handle "multiple" reads;

V. Costa, 2010

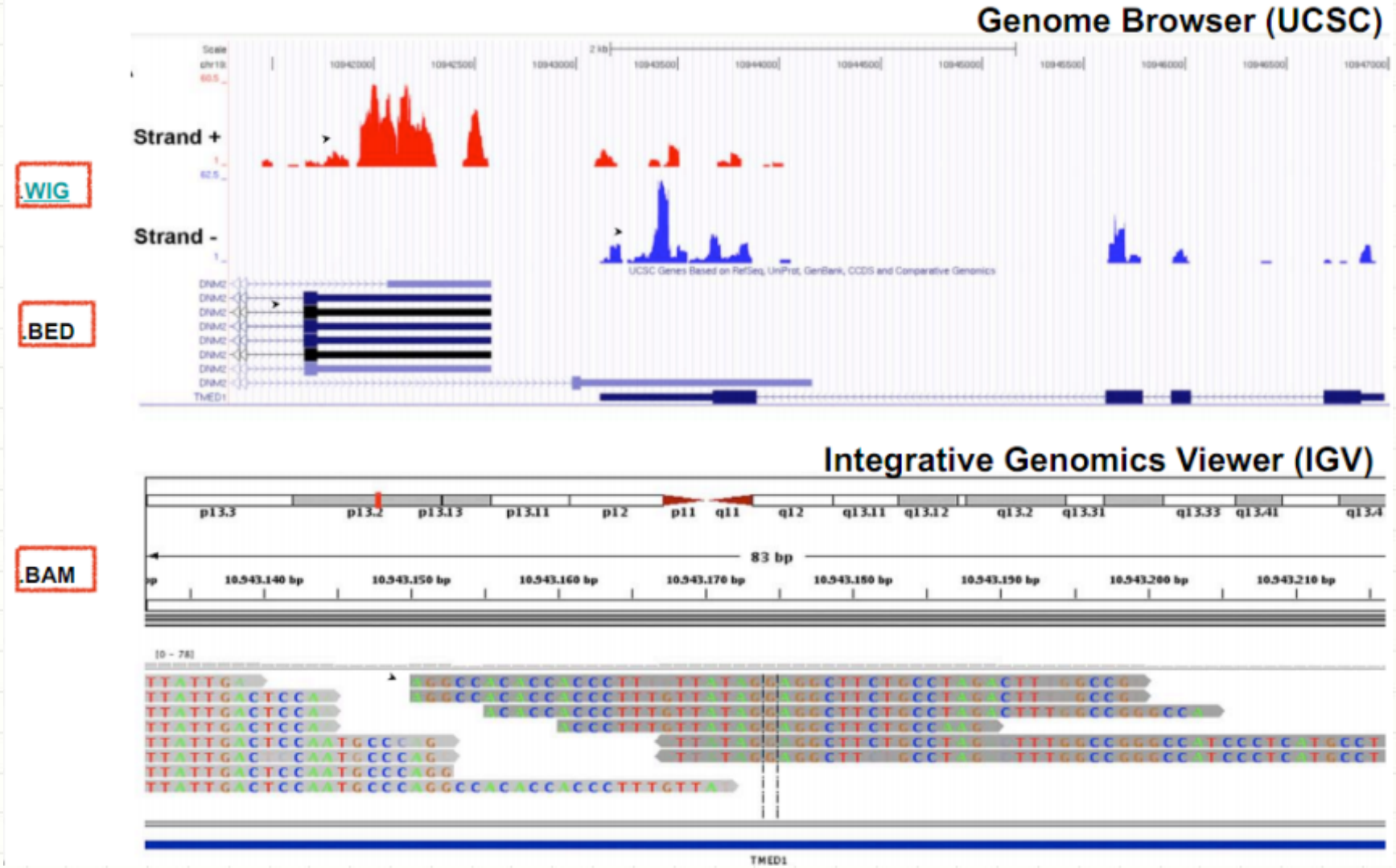
A Mapping Tool - Tophat

- Aligns sequences to the whole genome AND to exon-junctions
- Uses Bowtie, an ultrafast, memory-efficient short read aligner
- Output reported in SAM format
- Independently aligns segments of each read (default 25bp) allowing up to 2 mismatches
- Does not support indels / gapped alignments

http://tophat.cbcb.umd.edu/index.html

L. Ponnala, 2010

Visualization in Genome Browser



V. Costa, 2010

Other RNA-Seq Mapping Tools

Table 3 Bioinformatics tools for short-read sequencing

Program	Categories	Author(s)	Reference	URL
Cross_match	Alignment	Phil Green, Brent Ewing and David Gordon		http://www.phrap.org/phredphrapconsed.html
ELAND	Alignment	Anthony J. Cox		http://www.illumina.com/
Exonerate	Alignment	Guy S. Slater and Ewan Birney	72	http://www.ebi.ac.uk/~guy/exonerate
MAQ	Alignment and variant detection	Heng Li	37	http://maq.sourceforge.net
Mosaik	Alignment	Michael Strömberg and Gabor Marth		http://bioinformatics.bc.edu/marthlab/Mosaik
RMAP	Alignment	Andrew Smith, Zhenyu Xuan and Michael Zhang	73	http://rulai.cshl.edu/rmap
SHRiMP	Alignment	Michael Brudno and Stephen Rumble		http://compbio.cs.toronto.edu/shrimp
SOAP	Alignment	Ruiqiang Li *et al.*	35	http://soap.genomics.org.cn
SSAHA2	Alignment	Zemin Ning *et al.*	36	http://www.sanger.ac.uk/Software/analysis/SSAHA2
SXOligoSearch	Alignment	Synamatix		http://synasite.mgrc.com.my:8080/sxog/NewSXOligoSearch.php
ALLPATHS	Assembly	Jonathan Butler *et al.*	38	
Edena	Assembly	David Hernandez *et al.*	74	http://www.genomic.ch/edena
Euler-SR	Assembly	Mark Chaisson and Pavel Pevzner	75	
SHARCGS	Assembly	Juliane Dohm *et al.*	76	http://sharcgs.molgen.mpg.de
SHRAP	Assembly	Andreas Sundquist *et al.*	39	
SSAKE	Assembly	René Warren *et al.*	40	http://www.bcgsc.ca/platform/bioinfo/software/ssake
VCAKE	Assembly	William Jeck	77	http://sourceforge.net/projects/vcake
Velvet	Assembly	Daniel Zerbino and Ewan Birney	41	http://www.ebi.ac.uk/%7Ezerbino/velvet
PyroBayes	Base caller	Aaron Quinlan *et al.*	34	http://bioinformatics.bc.edu/marthlab/PyroBayes
PbShort	Variant detection	Gabor Marth		http://bioinformatics.bc.edu/marthlab/PbShort
ssahaSNP	Variant detection	Zemin Ning *et al.*		http://www.sanger.ac.uk/Software/analysis/ssahaSNP

Incomplete list compiled from sources, including http://seqanswers.com/forums/showthread.php?t=43 and http://www.sanger.ac.uk/Users/lh3/seq-nt.html.

Shendure & Ji, 2008

Construct Expression Profiles

- Calculate the number of reads mapped to each gene
- Normalize the number into a quantitative expression value – copies of the expressed gene
    - **RPKM**: reads per kilobase per million reads

A Normalization Example

- RPKM : Reads per kilobase per million mapped reads

  1kb transcript with 1000 alignments in a sample of 10 million reads (out of which 8 million reads can be mapped) will have

  $$RPKM = 1000/(1 * 8) = 125$$

- FPKM : for paired-end sequencing
  - A pair of reads constitute one fragment

L. Ponnala, 2010

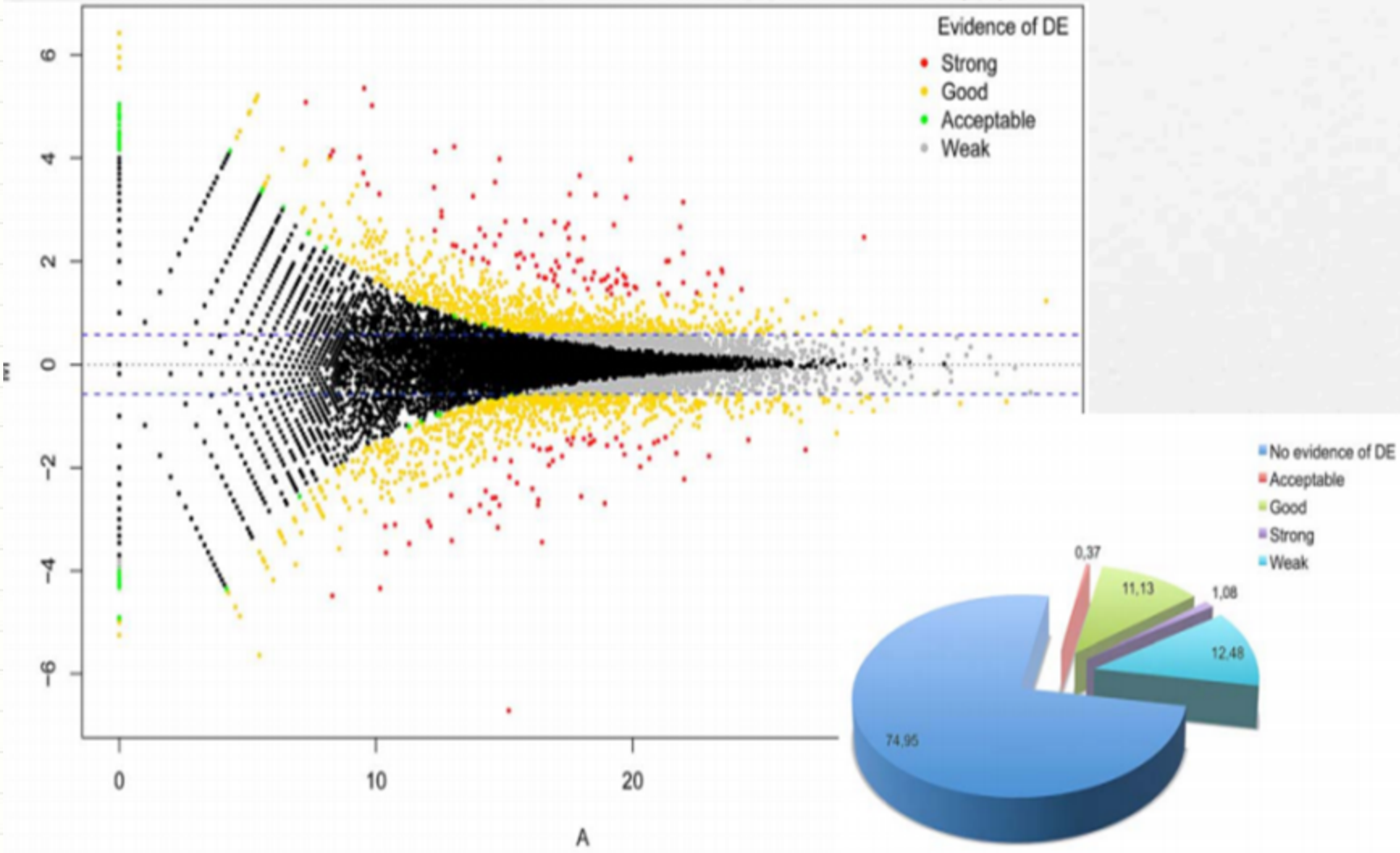
Identify Differentially Expressed Genes

- T-test
- Poisson distribution
- Negative binomial distribution

Tools for Differentially Expression Analysis

- Bio-conductor (t-test)
- edgeR (Poisson distribution; Robinson et al., 2009)
- DEGseq (negative binomial distribution; Wang et al., 2009)

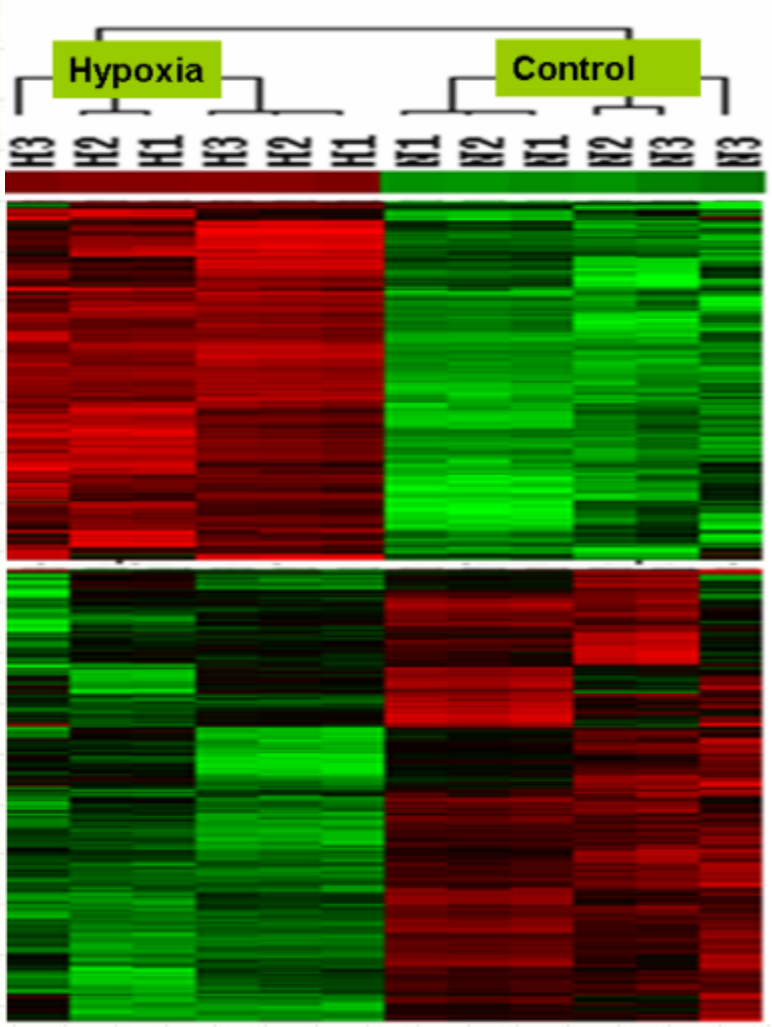
STRONG = detected with all 3 methods
GOOD = detected with 2 methods
ACCEPTABLE = detected with only 1 method
WEAK = below the FC threshold (1,5)



V. Costa, 2010

Generate Modules of Co-Expressed Genes

- K-means clustering
- Expectation-Maximization clustering
- Hierarchical clustering

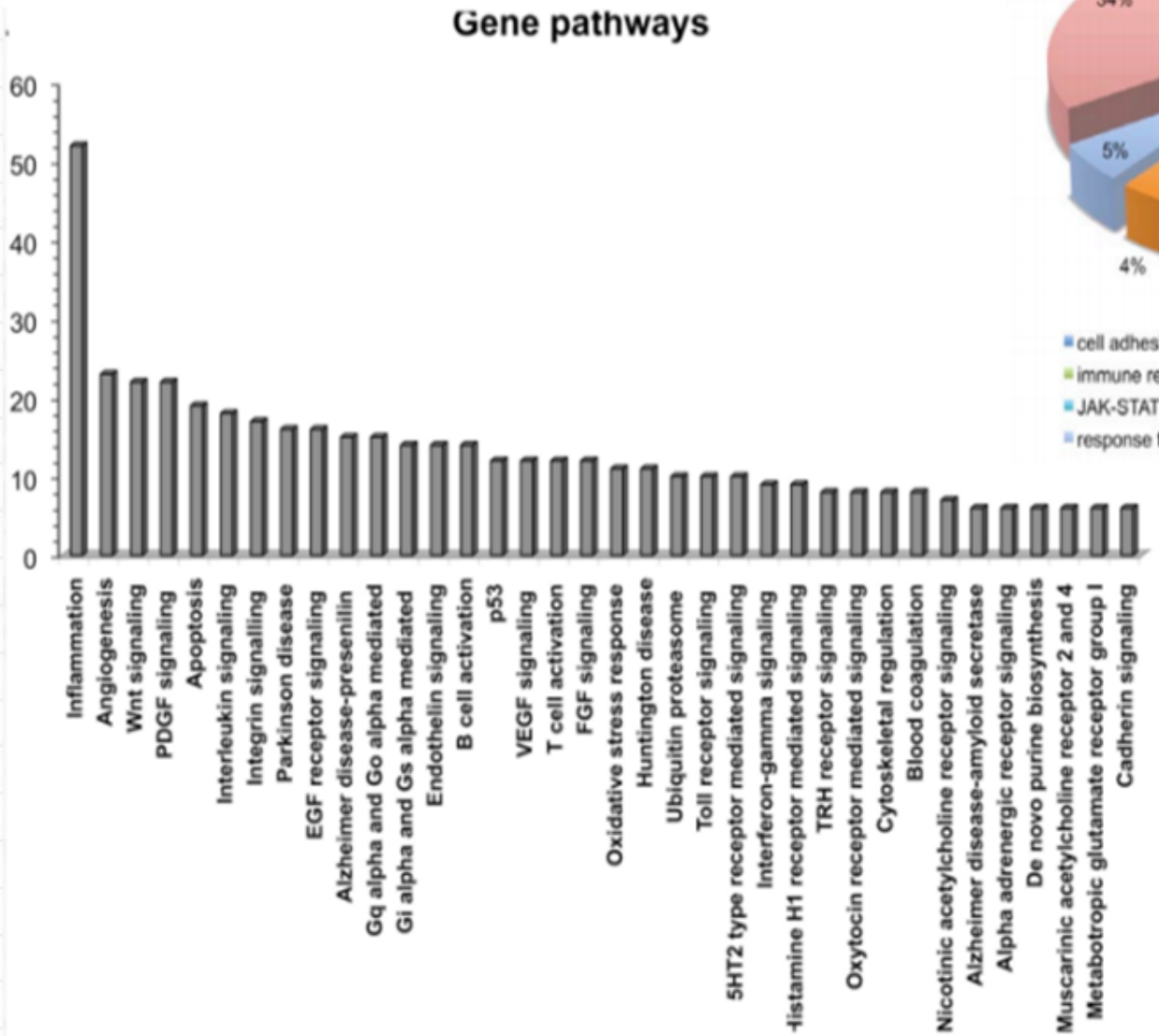


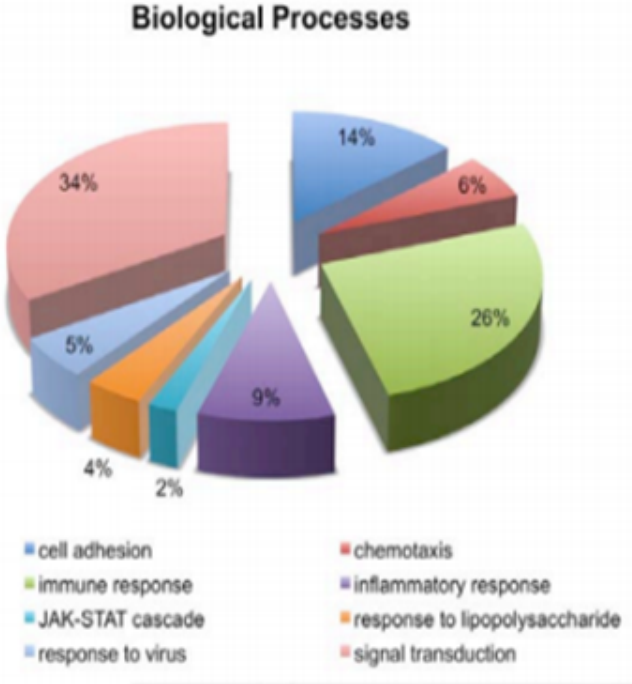
https://intramural.nhlbi.nih.gov/Offices/OCD/CSRP/Pages/ImageGallery.aspx

Gene Function Annotation and Enrichment Test

- MULTICOM protein function prediction pipeline based on UniProt and Gene Ontology (Wang and Cheng, 2011)
- Gene function enrichment test (hypergeometric distribution, chi-square test)

Function Annotation

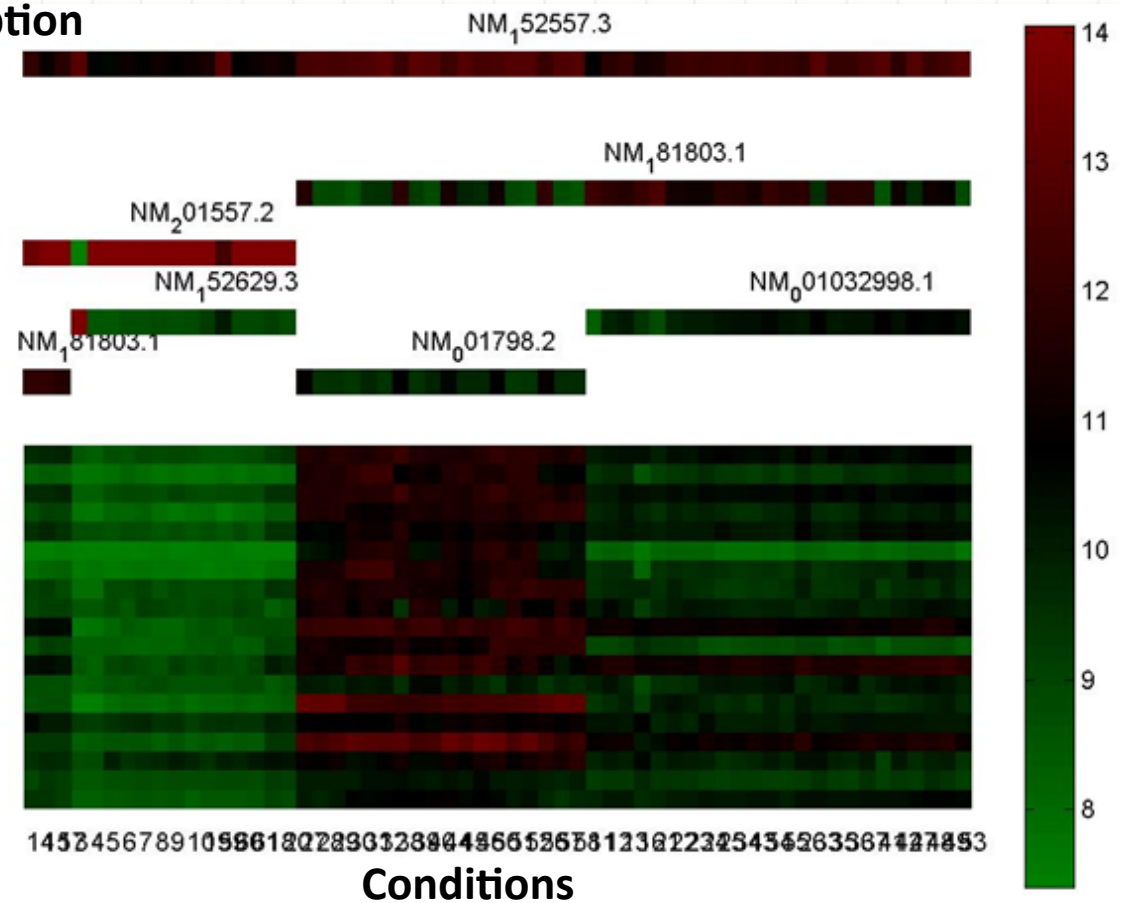




V. Costa, 2010

Construct Gene Regulatory Networks

Transcription factors



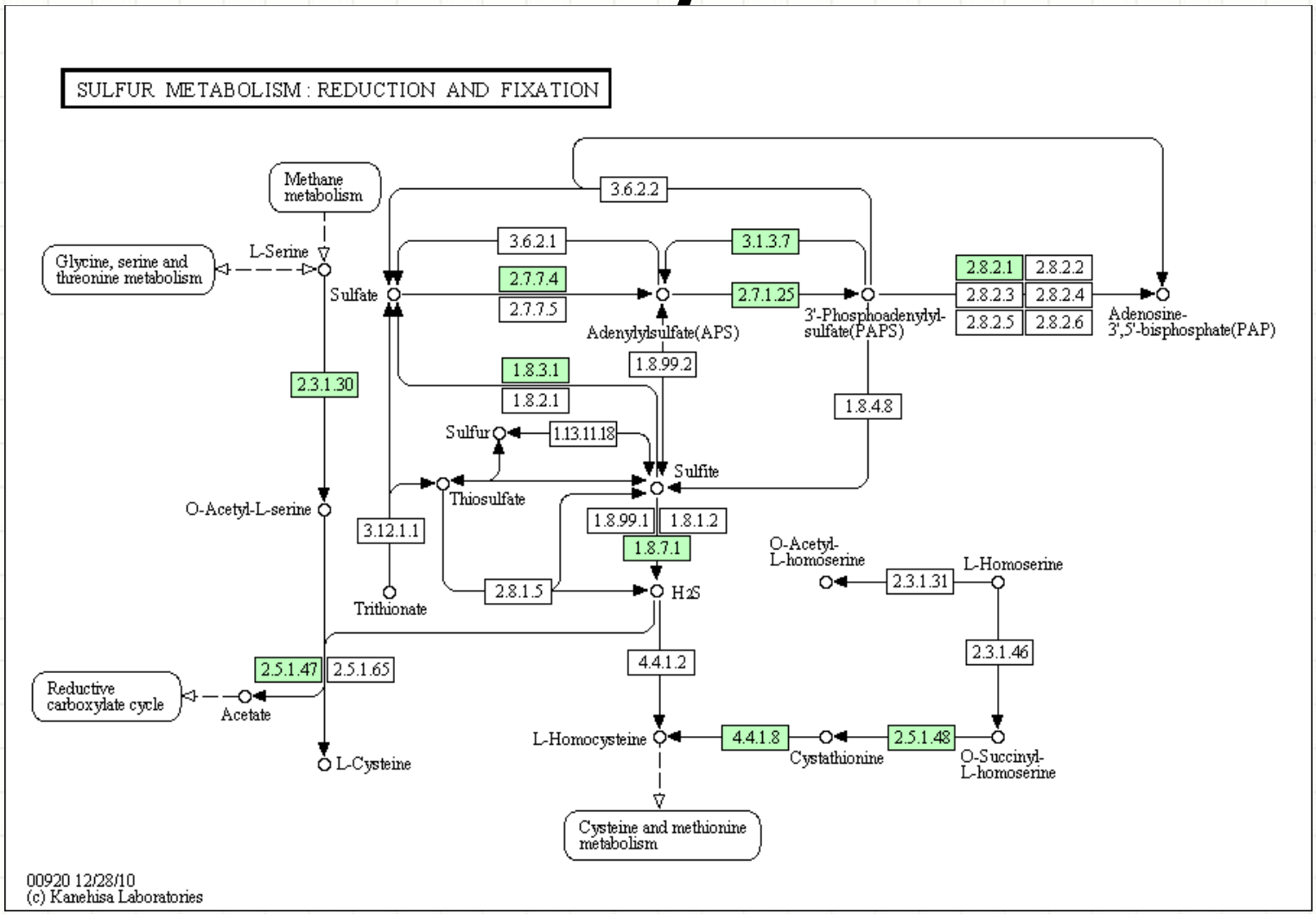
Genes

Conditions

J. Li et al., 2011

GO:0055114	P:oxidation reduction|25|4|15.7017769900593|0.210526315789474
GO:0045454	P:cell redox homeostasis|3|1|10.0659114844212|0.0526315789473684
GO:0006935	P:chemotaxis|3|1|10.0659114844212|0.0526315789473684
GO:0006350	P:transcription|53|2|0.187386682679516|0.105263157894737
GO:0000122	P:negative regulation of transcription from R...|4|1|7.09811589735686|0.0526315789473684
GO:0006954	P:inflammatory response|6|1|4.15813875831599|0.0526315789473684
GO:0006493	P:protein amino acid O-linked glycosylation|2|1|16.0293211065732|0.0526315789473684
GO:0055085	P:transmembrane transport|18|1|0.497808087813482|0.0526315789473684
GO:0045944	P:positive regulation of transcription from R...|4|1|7.09811589735686|0.0526315789473684
GO:0007586	P:digestion|3|1|10.0659114844212|0.0526315789473684

Infer Signal Transduction and Metabolic Pathways



Sulfur Metabolism: Reduction and Fixation

C. Nguyen, 2011

Integrate RNA-Seq Data with other Data

- Protein sequence, function and structure data
- Genomic data
- Chip-Seq data
- Proteomics data
- Protein-protein, protein-ligand interaction data
- Microarray data