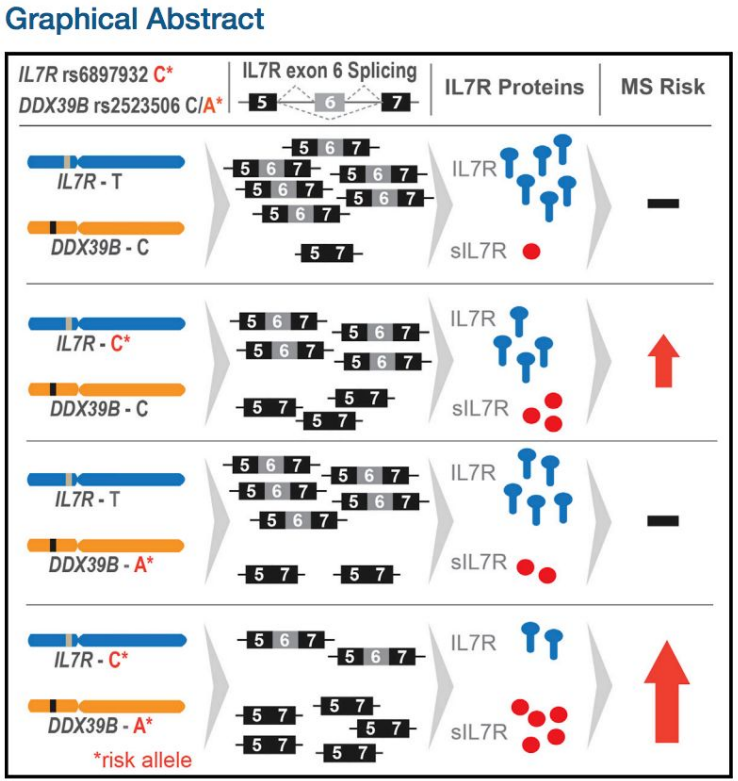
Lectures 6-7: robust statistical approaches for RNAseq

- RNA-seq is biased but quantitative: what are semi-parametric approaches for analyzing expression?
- How to test genetic interactions w/ RNA-expression

Splicing pinpointed as ‘causal factor’ in MS



IL7R splicing changes its interaction with the immune system

The splicing factor has a mutant with epistatic control over this variant

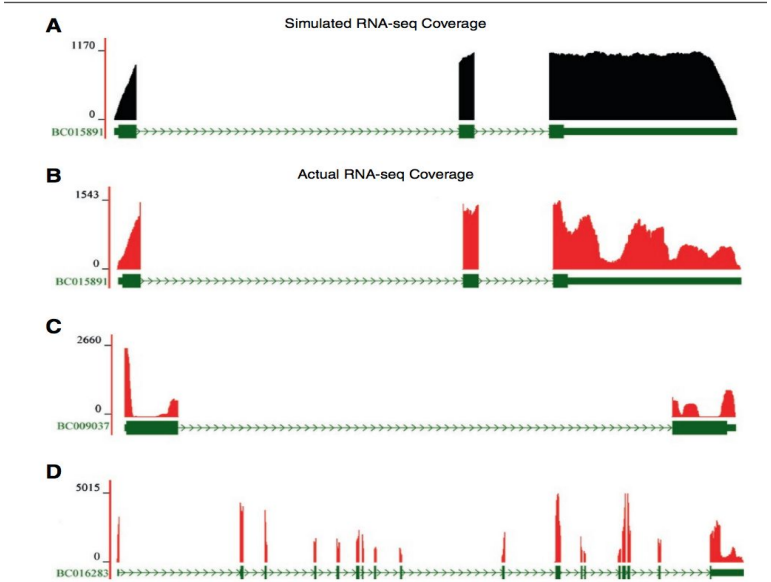
The facebook profile of RNA-seq: perfect statistical models

1. Many models assume each RNA isoform is sampled at Poisson(a \lambda_i) where a is a bias constant proportional to the abundance of the transcript
2. Modified models use the negative binomial
3. These assumptions doesn't hold, as we will see
4. (similar problems with DNA)

Testing for differential expression of RNA requires non-parametric approaches

Why we need robustness: motivation by GTEx and IVT-Seq

Exon level data-- discovering relationships and isoforms?



Lahens *et al. Genome Biology* 2014, **15**:R86
http://genomebiology.com/2014/15/6/R86

Extreme biases in RNA-seq: no theoretical null

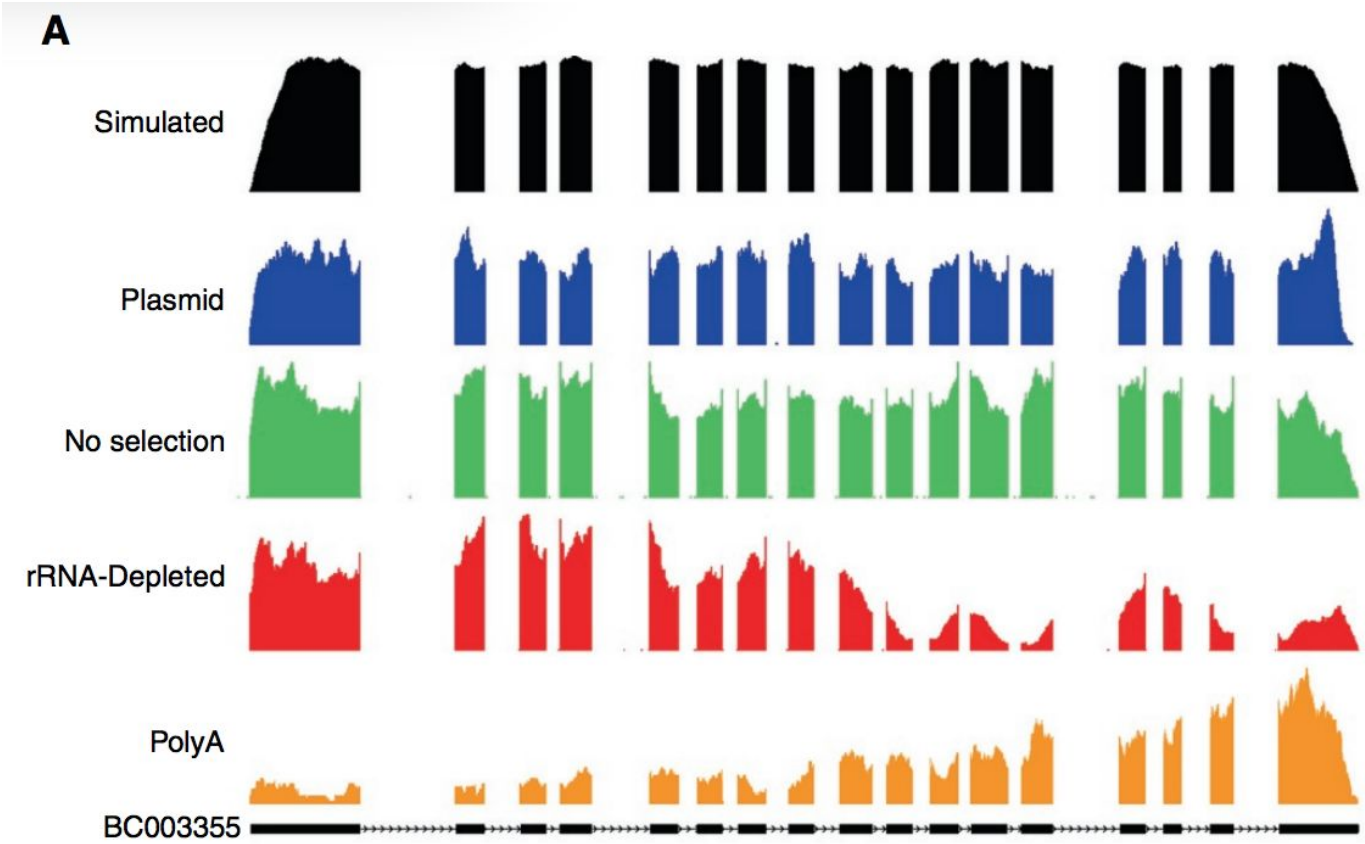
Lahens *et al. Genome Biology* 2014, **15**:R86
http://genomebiology.com/2014/15/6/R86

RESEARCH Open Access

IVT-seq reveals extreme bias in RNA sequencing

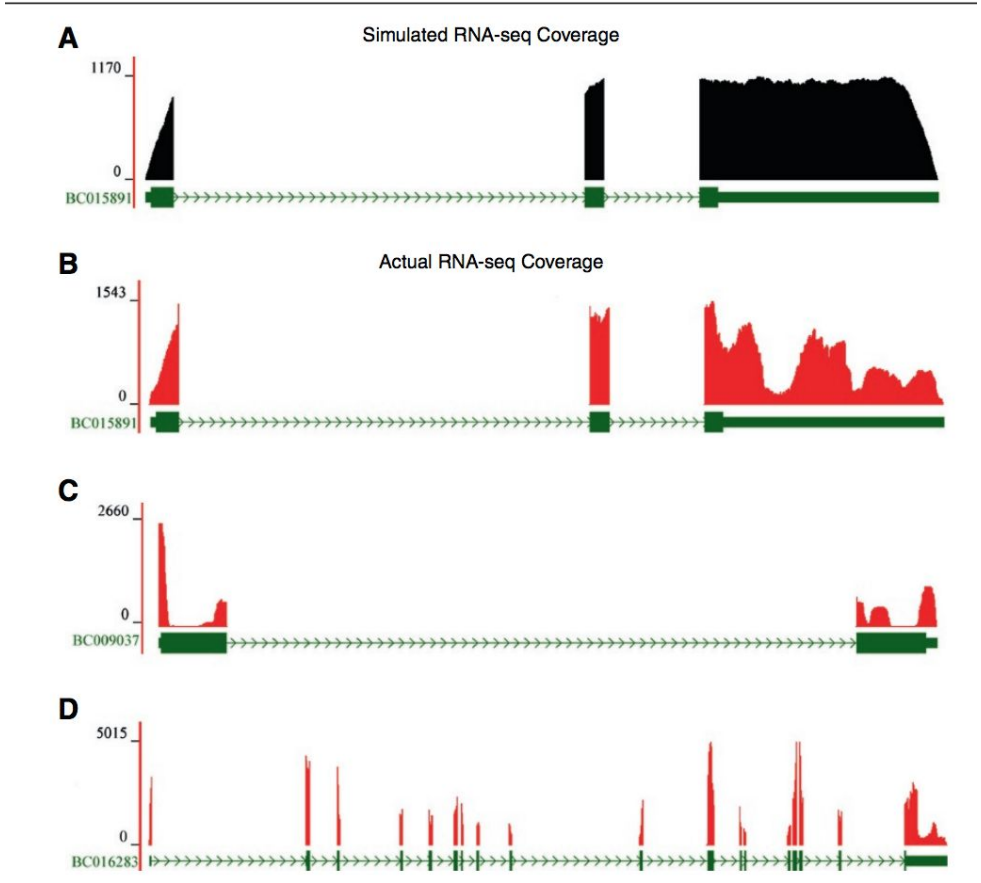
Nicholas F Lahens[1], Ibrahim Halil Kavakli[2,3], Ray Zhang[1], Katharina Hayer[4], Michael B Black[5], Hannah Dueck[6], Angel Pizarro[7], Junhyong Kim[6], Rafael Irizarry[8], Russell S Thomas[5], Gregory R Grant[4,9] and John B Hogenesch[1*]

Extreme bias in RNA-seq



Lahens *et al. Genome Biology* 2014, **15**:R86
http://genomebiology.com/2014/15/6/R86

Simulations and intuition don’t match real data



Lahens *et al. Genome Biology* 2014, **15**:R86
http://genomebiology.com/2014/15/6/R86

Model based approaches

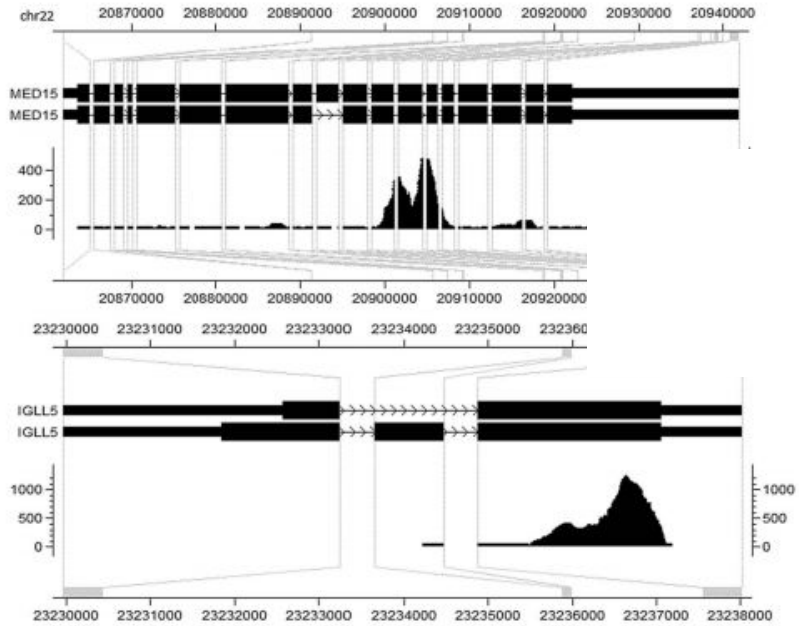


Figure 1.
Visualization of RNA-Seq reads mapped to the genes MED15 and IGLL5 on human chromosome 22 in CisGenome Browser (Jiang et al., 2010). From top to bottom for each

$$f(\theta, b) = l(\theta, b; n, A) - p(b)$$
$$= \sum_{j=1}^{J} \left\{ n_j \ln \left(\sum_{i=1}^{I} \theta_i a_{ij} e^{b_j} \right) - \sum_{i=1}^{I} \theta_i a_{ij} e^{b_j} \right\} - \lambda \sum_{j=1}^{J} |b_j| \quad (2.3)$$

Genes discovered by non-parametric analysis

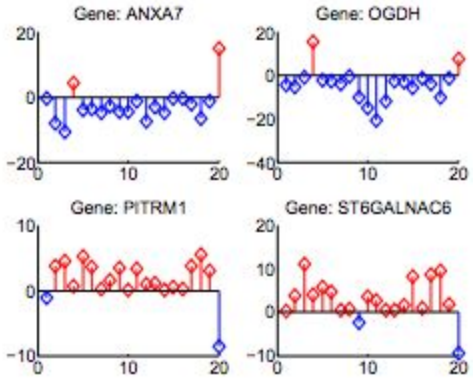


Figure 4: Count difference of the top four genes discovered only by the sign test. X-axis: sample index; Y-axis: gene expression level.

https://arxiv.org/pdf/1801.04005.pdf

Opportunities for discovery using robust statistics and massive data

Motivation by Gtex

Describe data: clinical data https://gtexportal.org/home/datasets

And a great deal of information on genotype/RNA expression

https://gtexportal.org/home/tissueSummaryPage#cause

https://gtexportal.org/home/gene/SMN2

But, statistics are not interpretable

GTEx Analysis V6 (dbGaP Accession phs000424.v6.p1)

Biobank Inventory
GTEx Analysis V6p
GTEx Analysis V6
GTEx Analysis V4
GTEx Analysis Pilot V3

Annotations

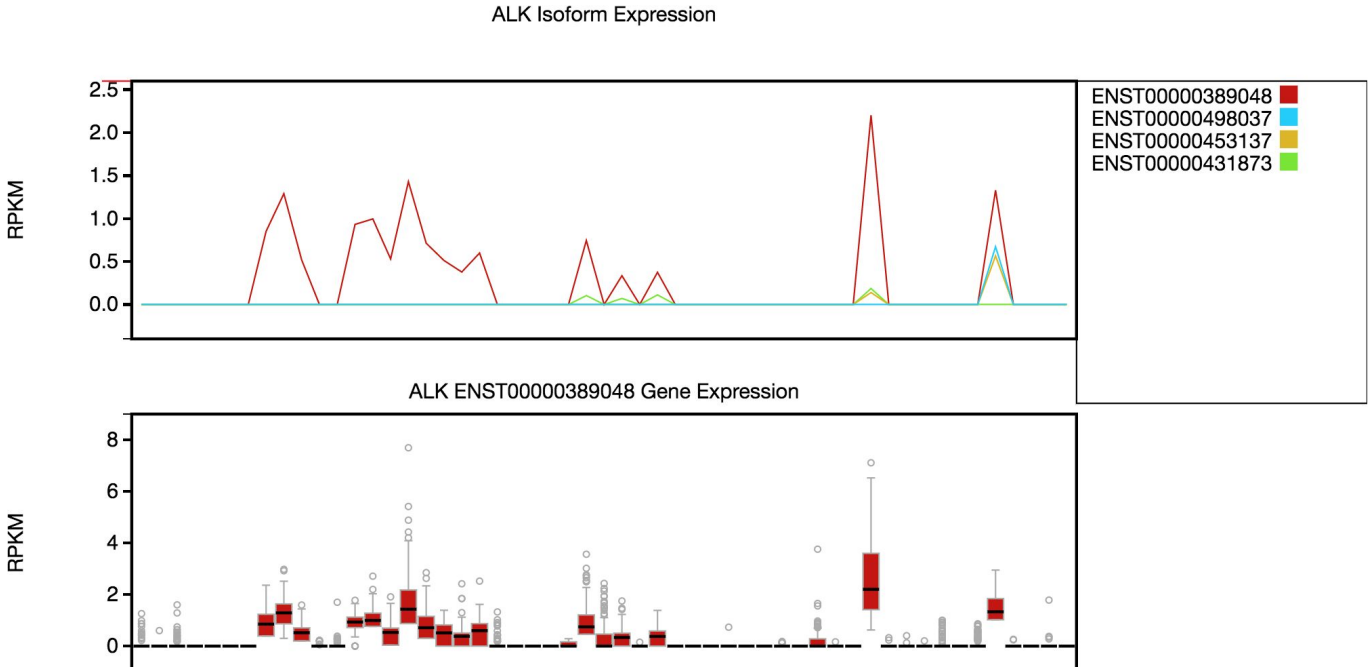
Description	Name	Size
A data dictionary that describes each variable in the GTEx_Data_V6_Annotations_SampleAttributesDS.txt	GTEx_Data_V6_Annotations_SampleAttributesDD.xlsx	32K
A de-identified, open access version of the sample annotations available in dbGaP.	GTEx_Data_V6_Annotations_SampleAttributesDS.txt	5.9M
A de-identified, open access version of the subject phenotypes available in dbGaP.	GTEx_Data_V6_Annotations_SubjectPhenotypesDS.txt	12K
A data dictionary that describes each variable in the GTEx_Data_V6_Annotations_SubjectPhenotypes_DS.txt.	GTEx_Data_V6_Annotations_SubjectPhenotypes_DD.xlsx	22K

RNA-Seq Data

Description	Name	Size
Fraction of intron that is covered by reads.	GTEx_Analysis_v6_RNA-seq_Flux1.6_intron_fraccov.txt.gz	822M
Intron read count.	GTEx_Analysis_v6_RNA-seq_Flux1.6_intron_reads.txt.gz	1.5G
Junction read count.	GTEx_Analysis_v6_RNA-seq_Flux1.6_junction_reads.txt.gz	1.8G
Transcript read count.	GTEx_Analysis_v6_RNA-seq_Flux1.6_transcript_reads.txt.gz	2.8G
Transcript RPKM.	GTEx_Analysis_v6_RNA-seq_Flux1.6_transcript_rpkm.txt.gz	2.8G
Exon read count.	GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_exon_reads.txt.gz	3.7G

Motivation by Gtex

Question: differential isoform expression: example-- real differences, or artifacts?



How do we overcome these problems?

- Learn statistical theory and methods
- Designing our own custom test that captures intuition, then analyze its properties

Does the bootstrap or permutation test break down?

FDR control by knockoffs (Candes')

Bootstrap breakdown (Lehman and Romano)