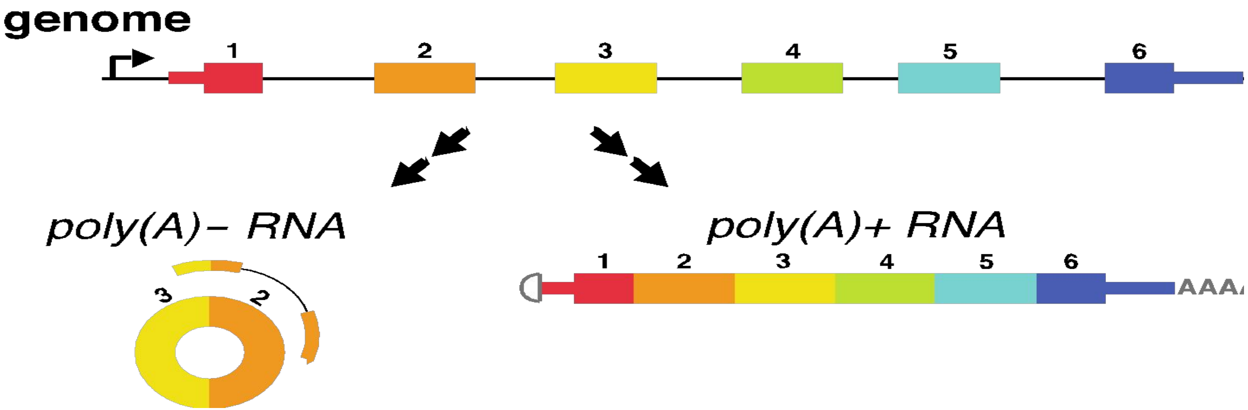
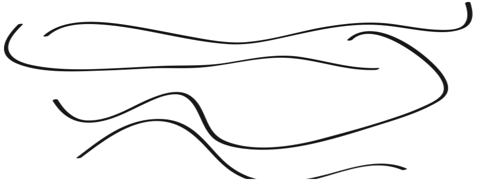
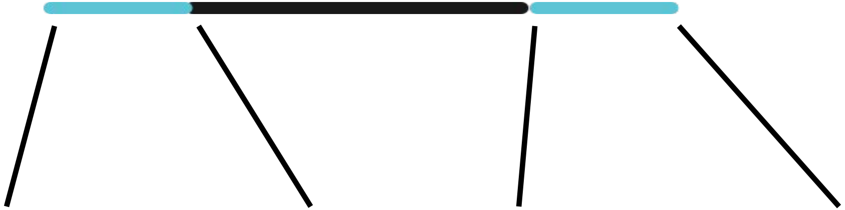
RNA



The data: paired-end RNA-seq



Matched sequences are obtained for each library molecule



CTTC…..G
AAG

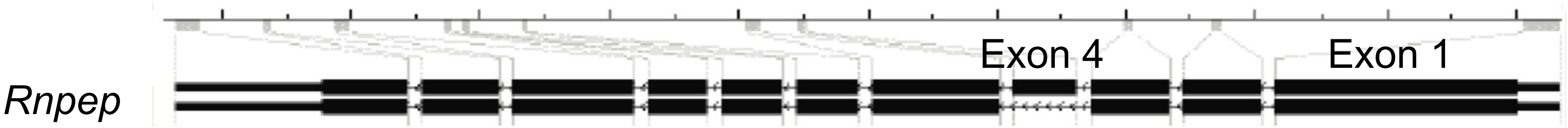
GGAC…..G
CCT

The statistical modeling

- Po(λ), the larger λ, the larger the rate of the rare event
    - Defined as $Po(X=k)=e^{-\lambda} \lambda^{k}/k!$
    - k>0
    - In RNA-Seq, each transcript (compared to all others) will be rare, so each transcript abundance modeled as λ_i
    - A "read" s_j is a sequence matching an RNA at position j
    - simplest model: s_j is generated as Po(λ_i)
- In statistics, we take observed data and use it to estimate parameters, in this case, λ_i

- This is formally accomplished by, for example the MLE
- In RNA seq, "RPKM" is conceptually like λ_i

Intuition for the statistical problem



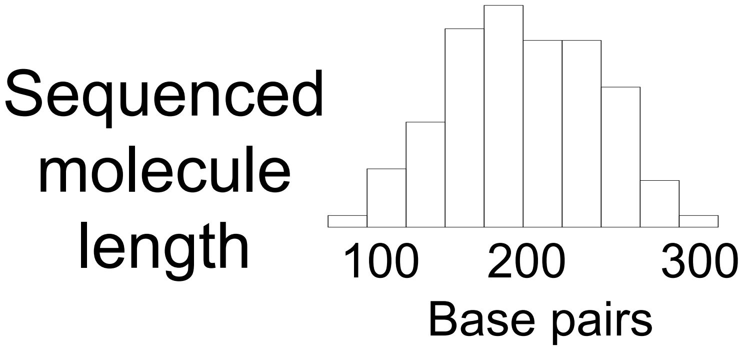
Estimate the expression of each isoform?

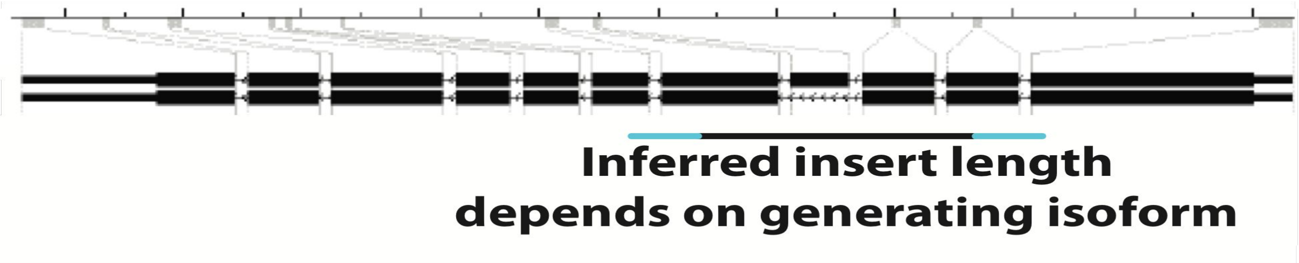
Nontrivial : we only observe fragments of sequences

- Since the size distribution of library molecules is known, inferred insert lengths can be used to increase statistical power and inference

Intuition for the most powerful modeling

- Compute genome-wide insert length distribution





•Mapped to Isoform 1
→ length 150
•Mapped to Isoform 2
→ length 90

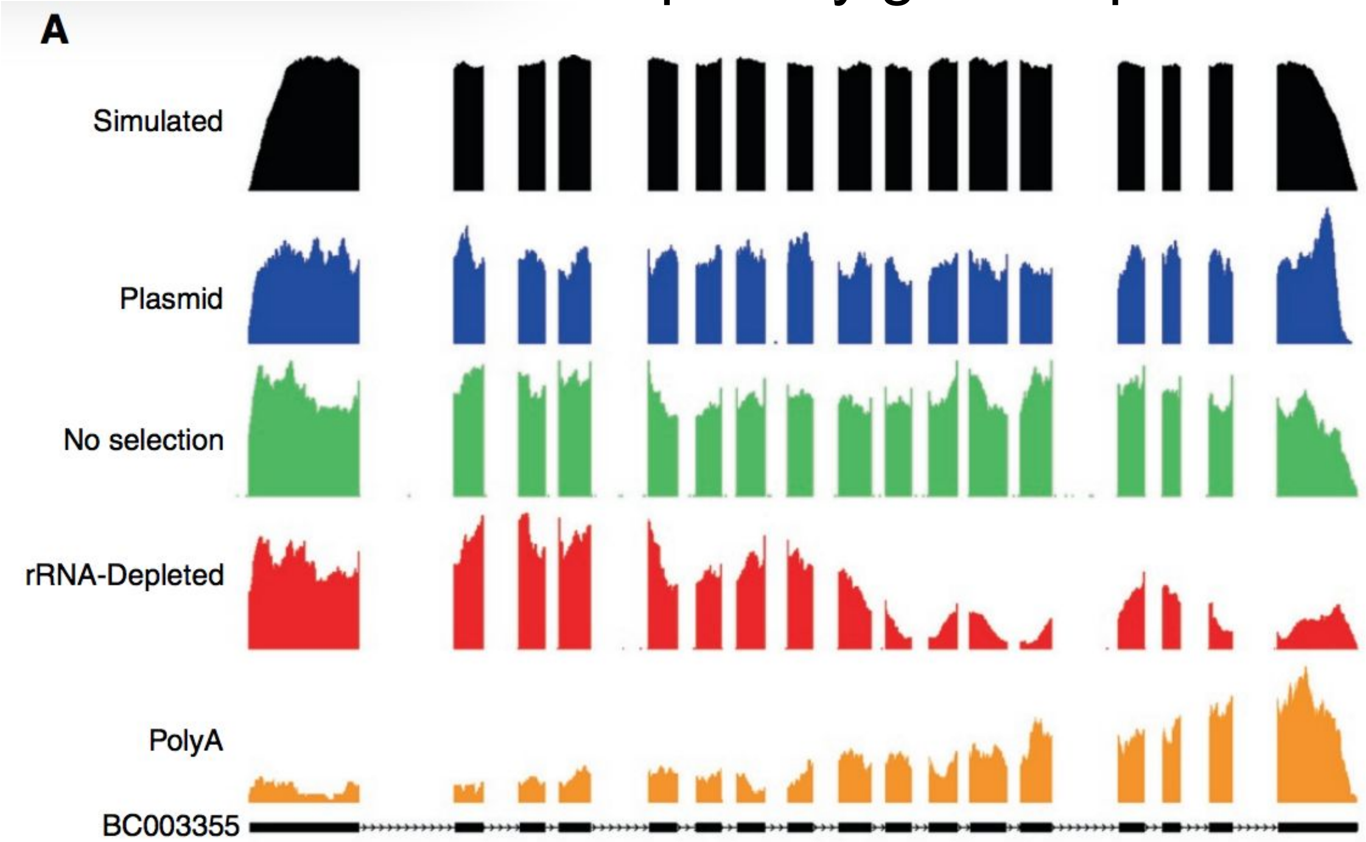
- Statistical improvement over naïve models
- Optimal information reduction
- Quantifies information gain using PE Sequencing

Why do we care: just fun math?

- Not knowing the isoforms means we don’t know the gene level expression
- Off the shelf tools are “mostly right” but many times wrong
- Most labs don’t use their latest published software
- Current tools only provide approximate answers

General problem: alignment as a black box, read densities

Use read densities to quantify gene expression

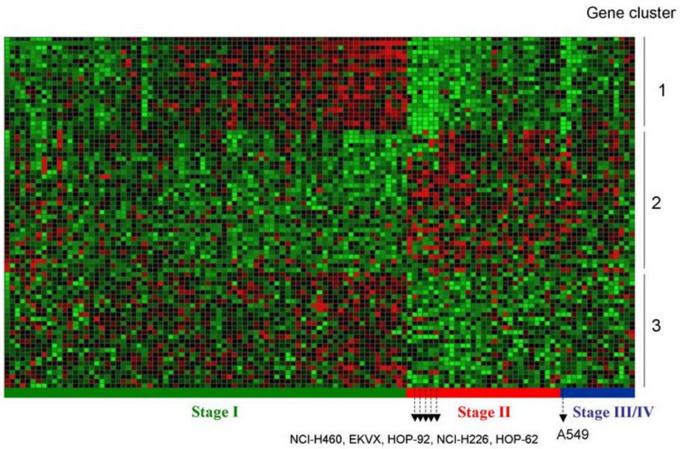


Lahens *et al. Genome Biology* 2014, **15**:R86
http://genomebiology.com/2014/15/6/R86

What are the needed statistical algorithms?

1. Quantifying exon expression, junction expression
2. Deconvolving isoform expression
3. Some are trying to discover new RNA

We want to know the copies of RNA per cell

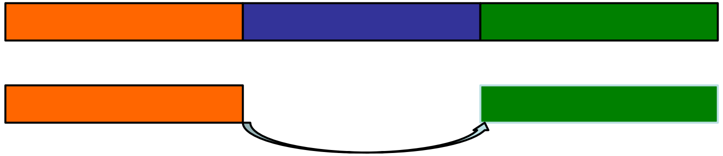


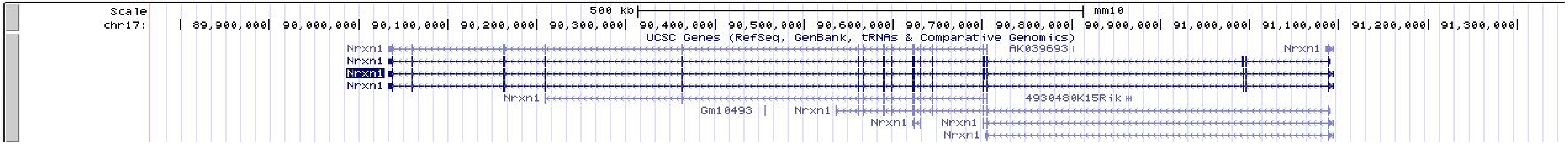
From:
http://media.springernature.com/lw785/springer-stat
mage/art%3A10.1186%2F1471-2164-7-166/MediaC
ects/12864_2006_Article_549_Fig4_HTML.jpg

Intuition for statistically quantifying isoforms

1. Exon-level and junctional reads are observed
2. There is a deconvolution problem
   a. Quantifying exon expression, junction expression
   b. Deconvolving isoform expression

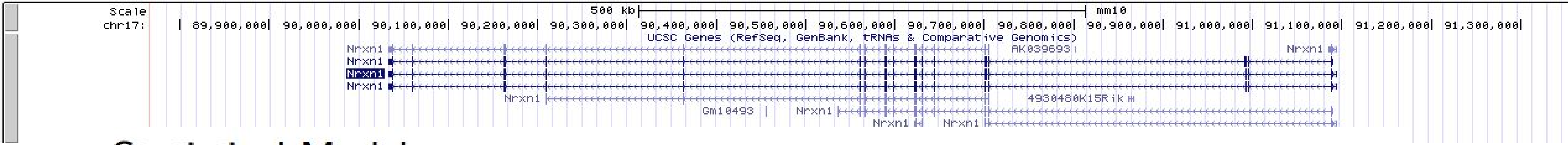
Exon 1 Exon 2 Exon 3





Sufficient statistics, statistical problem, Poisson models

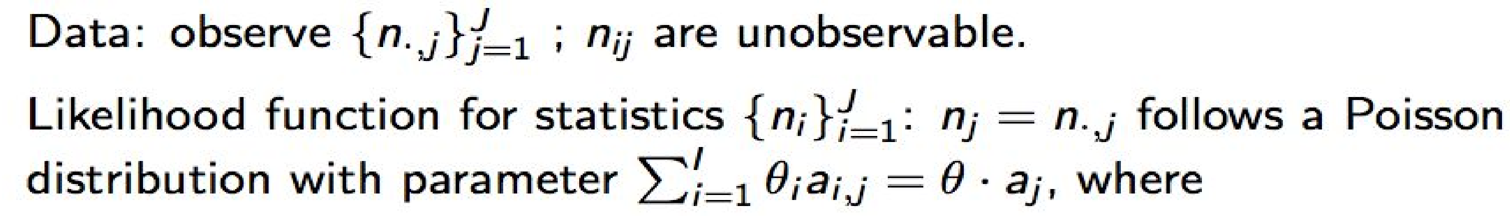
Formalizing the problem and model



Statistical Model

- The relative abundance for the I isoforms are the parameters of interest and denoted $\{\theta_i\}_{i=1}^{I}$.

Solving the problem with statistics



Data: observe $\{n_{\cdot,j}\}_{j=1}^{J}$; n_{ij} are unobservable.

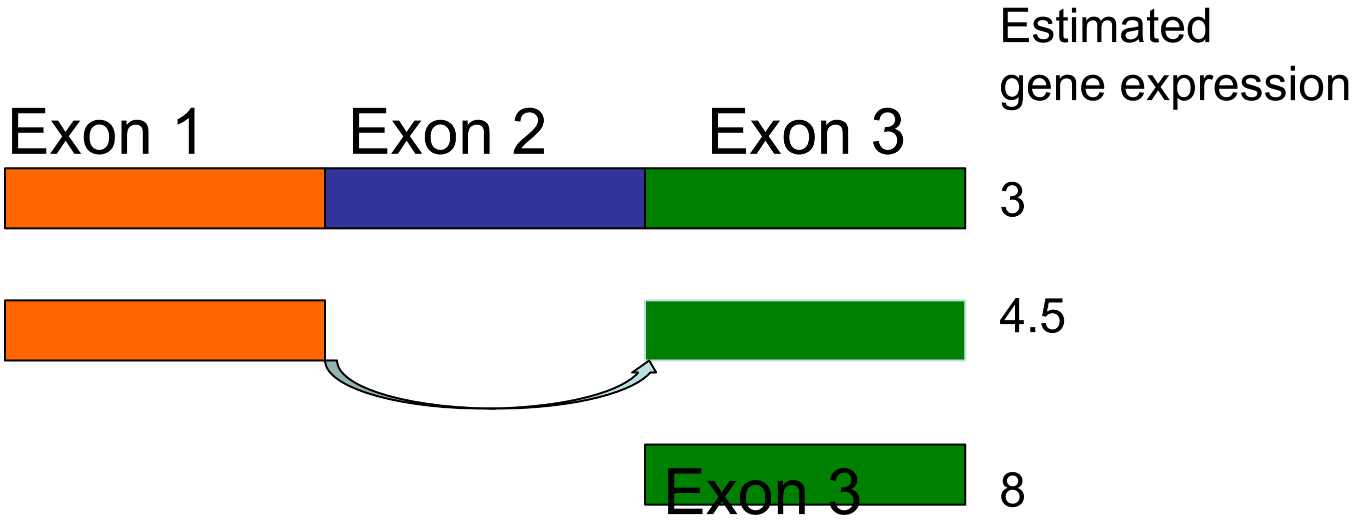
Likelihood function for statistics $\{n_i\}_{i=1}^{J}$: $n_j = n_{\cdot,j}$ follows a Poisson distribution with parameter $\sum_{i=1}^{I} \theta_i a_{i,j} = \theta \cdot a_j$, where

Each isoform expression is independent:

The importance of statistics

Exon	**1**	**2**	**3**
Count	1	0	8

Remember, counts =“expression” in RNA-Seq



Estimated gene expression

Exon 1 Exon 2 Exon 3

3

4.5

Exon 3

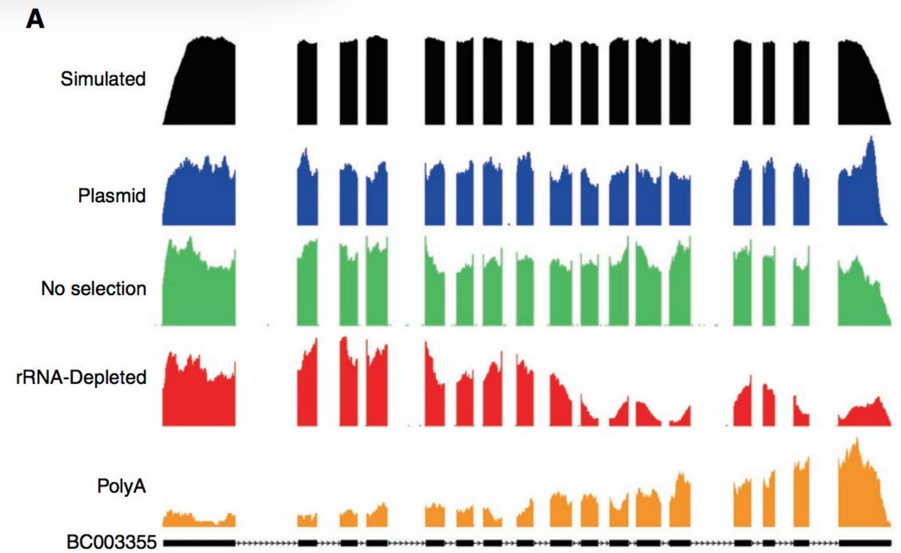
8

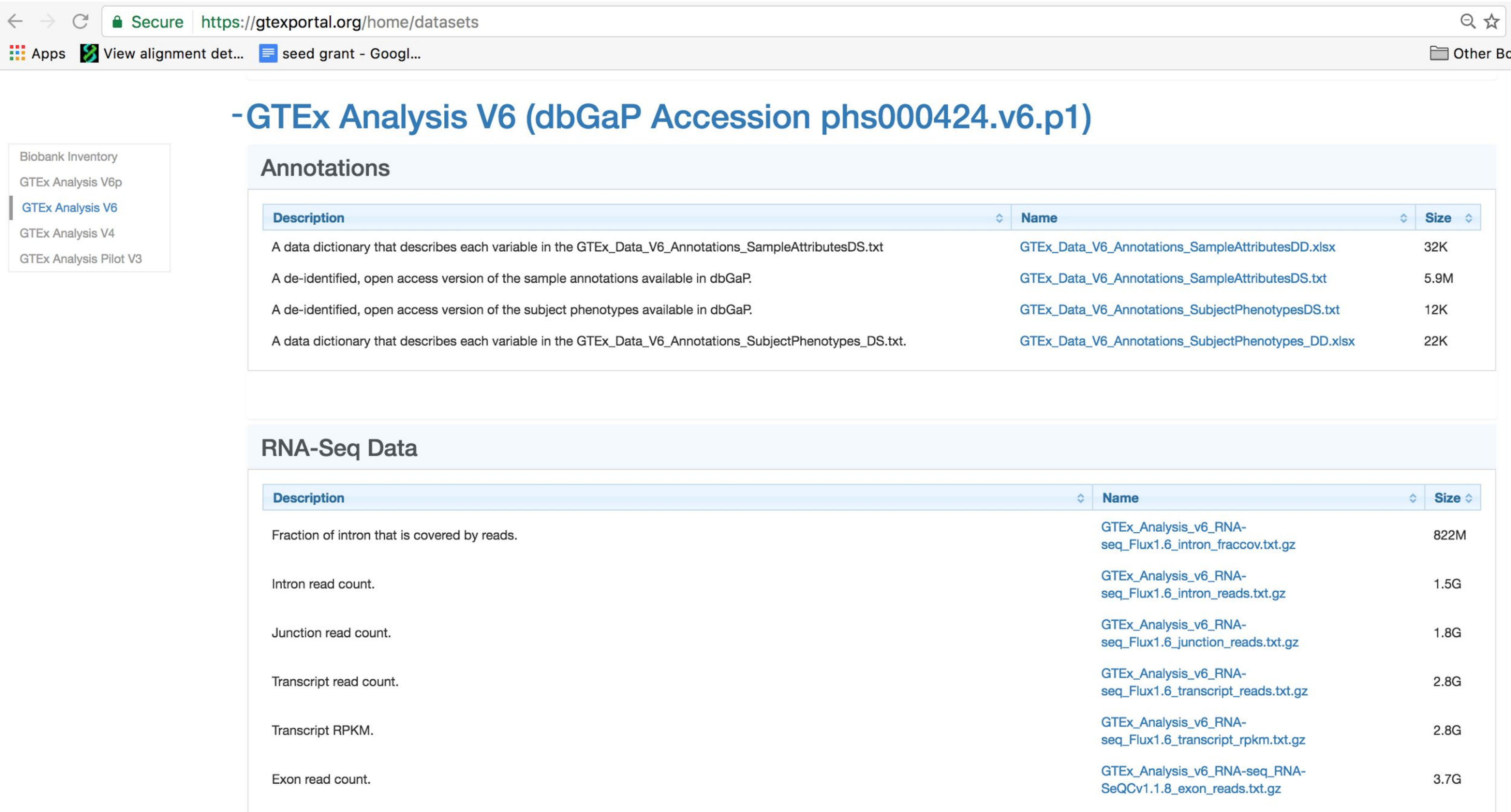
Without taking isoforms into account, gene expression estimates (and differential gene expression will be wrong)!

Even more “problems”: count data is noisy

Example, idea: clean it up w/ robust statistics

Bayesian analysis





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

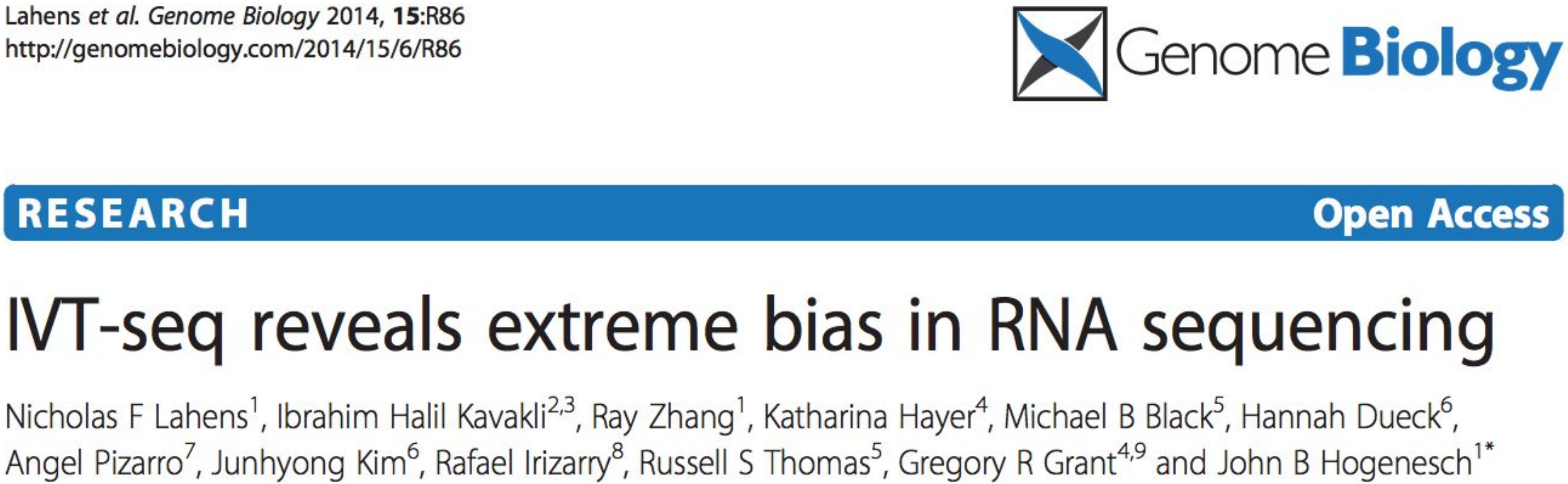
Annotations

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

Extreme biases in RNA-seq: no theoretical null



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

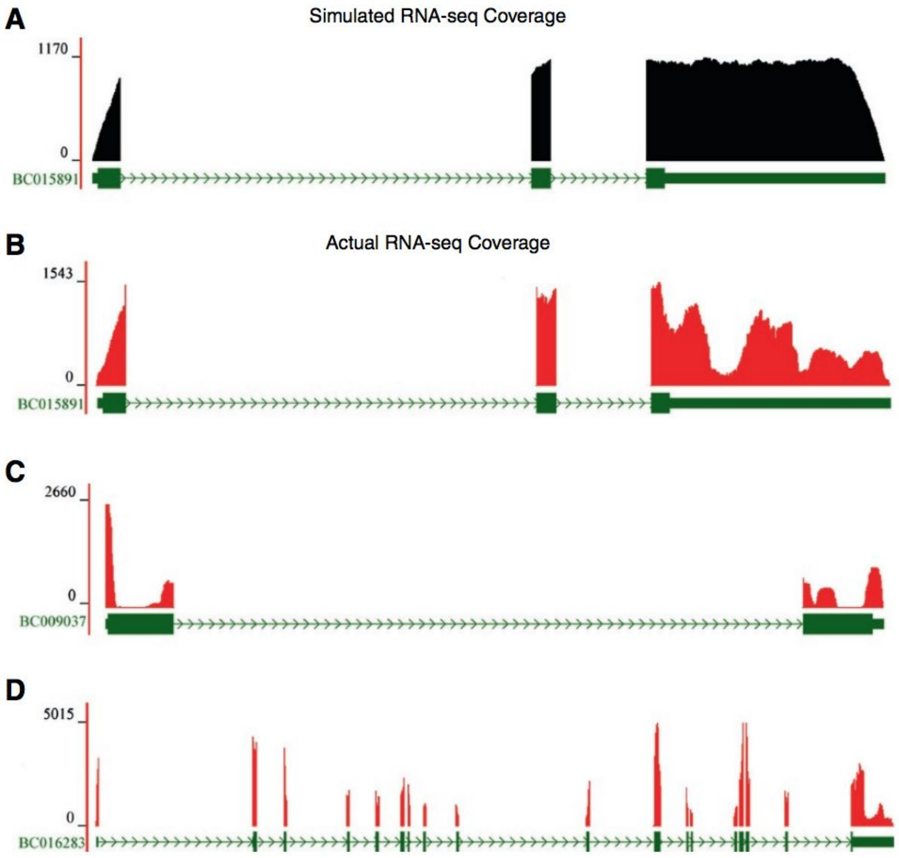
Genome **Biology**

RESEARCH **Open Access**

IVT-seq reveals extreme bias in RNA sequencing

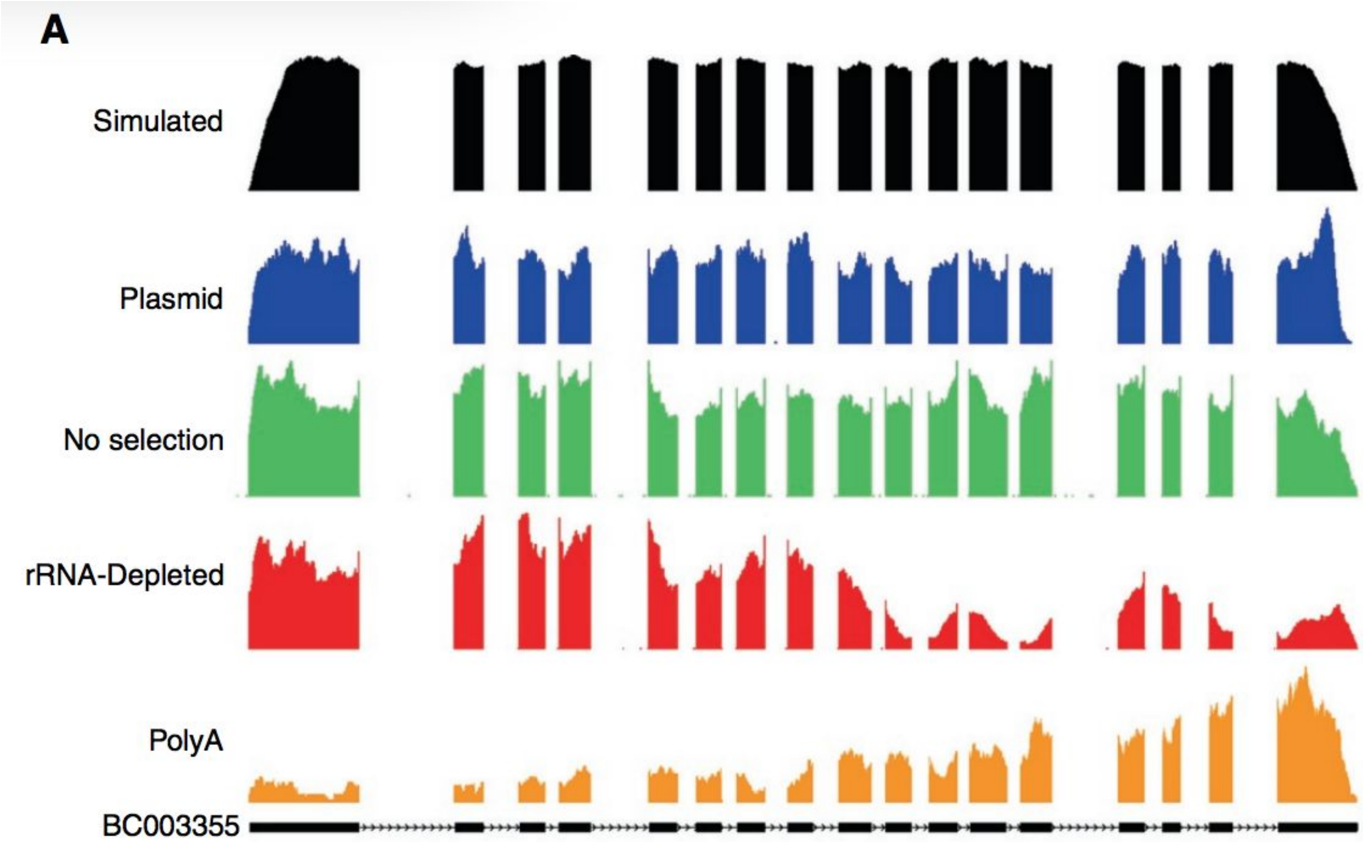
Nicholas F Lahens[1], Ibrahim Halil Kavakli[2,3], Ray Zhang[1], Katharina Hayer[4], Michael B Black[5], Hannah Dueck[6], Angel Pizarro[7], Junhyong Kim[6], Rafael Irizarry[8], Russell S Thomas[5], Gregory R Grant[4,9] and John B Hogenesch[1*]

Simulations and intuition don't match real data



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

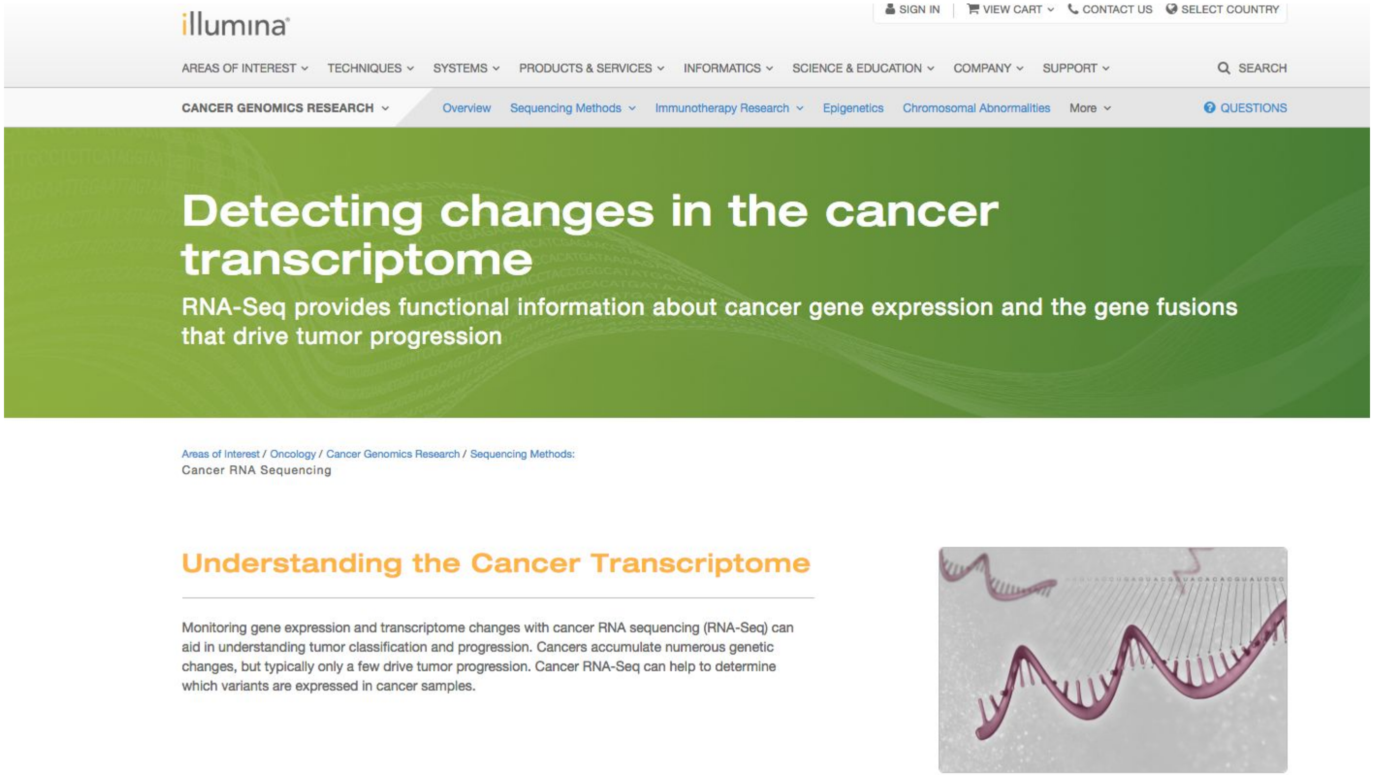
Selection and efficiency confound naive estimation



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

Another motivation: Disease genomics

Targeted therapy based on RNA-seq



Detecting changes in the cancer transcriptome

RNA-Seq provides functional information about cancer gene expression and the gene fusions that drive tumor progression

Areas of Interest / Oncology / Cancer Genomics Research / Sequencing Methods:
Cancer RNA Sequencing

Understanding the Cancer Transcriptome

Monitoring gene expression and transcriptome changes with cancer RNA sequencing (RNA-Seq) can aid in understanding tumor classification and progression. Cancers accumulate numerous genetic changes, but typically only a few drive tumor progression. Cancer RNA-Seq can help to determine which variants are expressed in cancer samples.

Considerations for choice of statistical approach

1. Theoretically best
    a. Under the given null and alternative, it is possible to prove which test is best
    b. Fisher’s efficient estimator
    c. Uniformly Most Powerful test
2. Fast
    a. Inexpensive to store data
        i. Reduction to sufficient or minimal sufficient statistics
    b. Computationally inexpensive
        i. Computing test statistics is simple
3. Mechanistic
    a. Tests and scientific/medical interventions easy to perform
    b. Few predictors, LASSO and NMF move in this direction

Many problems in biomedical science are for mechanistic discovery rather than classification

The first modern, efficient, theoretically tractable tests: Rank tests

1. Theoretically ~~best~~ tractable
2. Fast
   a. Computationally inexpensive
3. Inexpensive to store data

   Downside? Lose power

4. Next lectures will move onto more powerful tests

Rank tests

General idea:

1. Replace data by ranks
2. Perform a test on the ranked data to test if deviation from expectation

Advantage: requires simply sorting the data and a single computation

1. Sort time: O(n log n) (worst case, O(n^2): data storage benefits

Disadvantage: power (brainstorm example)

On board: derivation of Mann-Whitney test and introduction to random permutations

How do we overcome these problems?

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