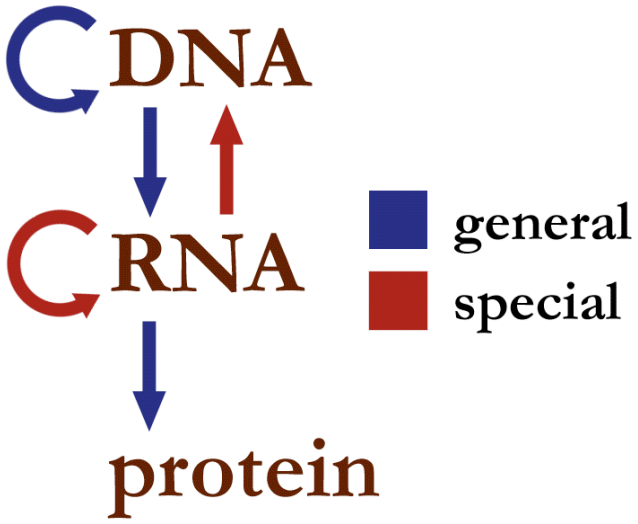
Statistical methods for RNA-seq data

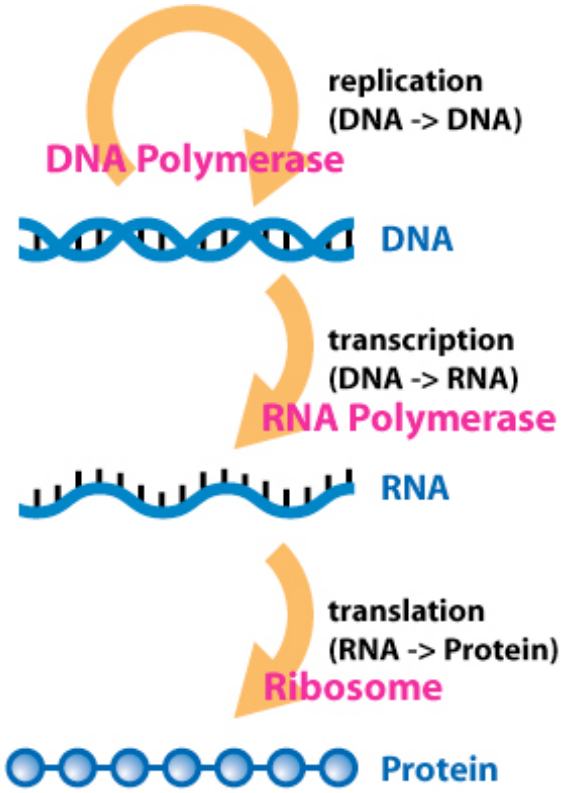
Wei Sun
Department of Biostatistics
Department of Genetics
University of North Carolina, Chapel Hill

Outline

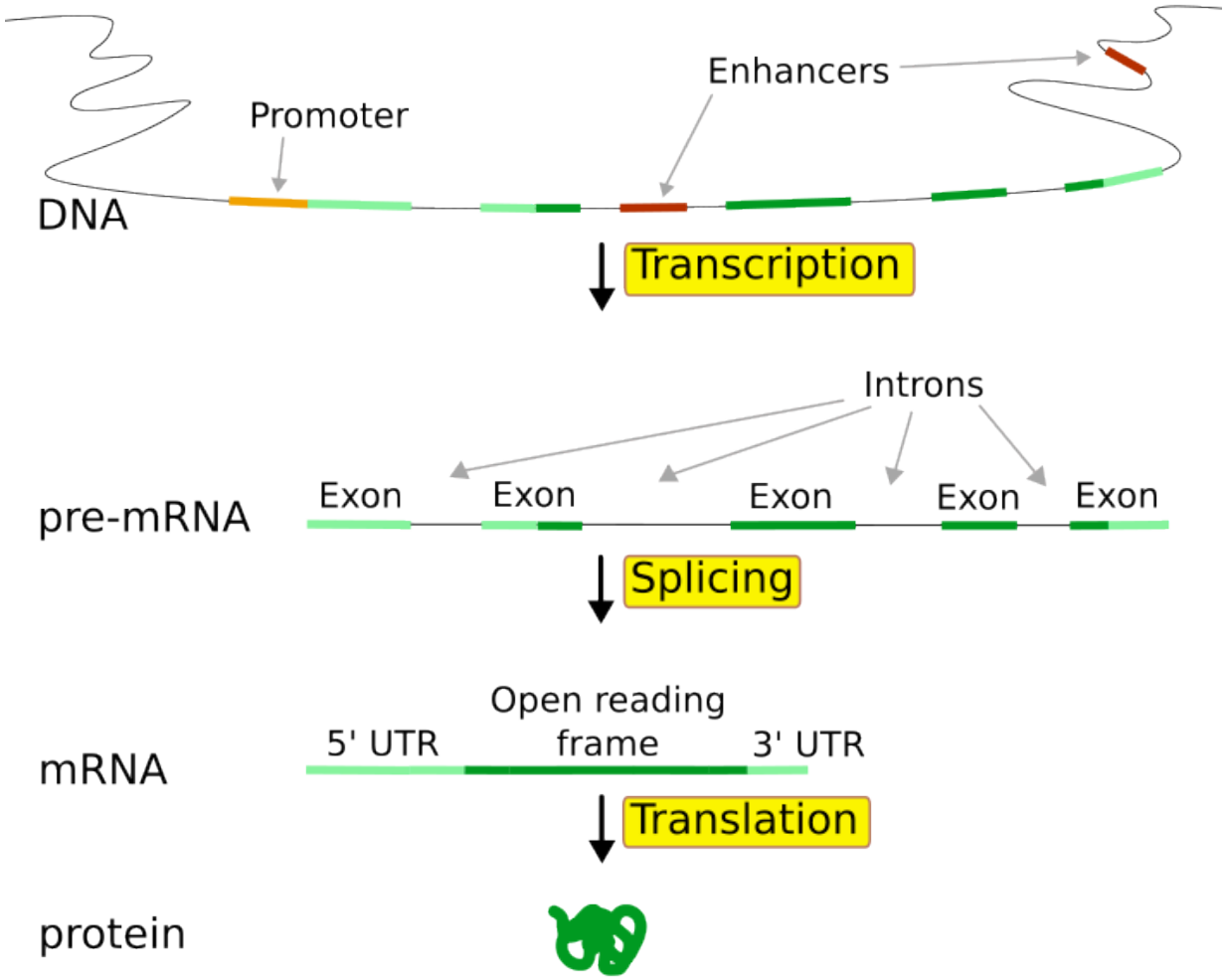
- 1. An brief introduction to RNAseq
- 2. Allele-specific expression
- 3. Isoform-specific expression

Central Dogma

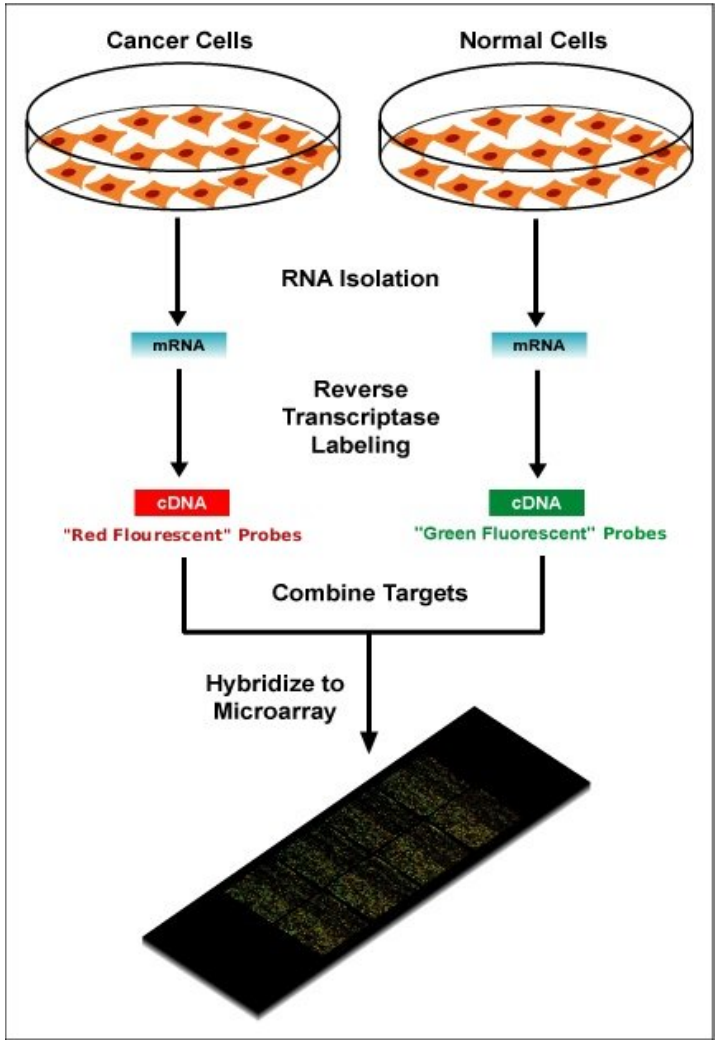




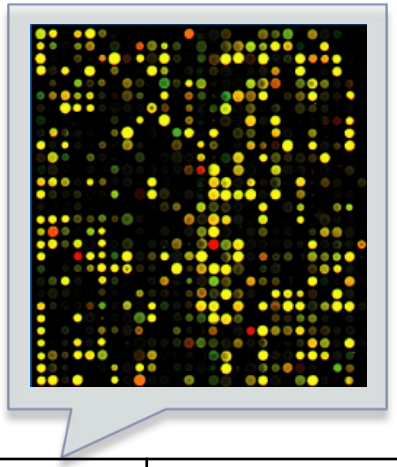
Transcription and Translation



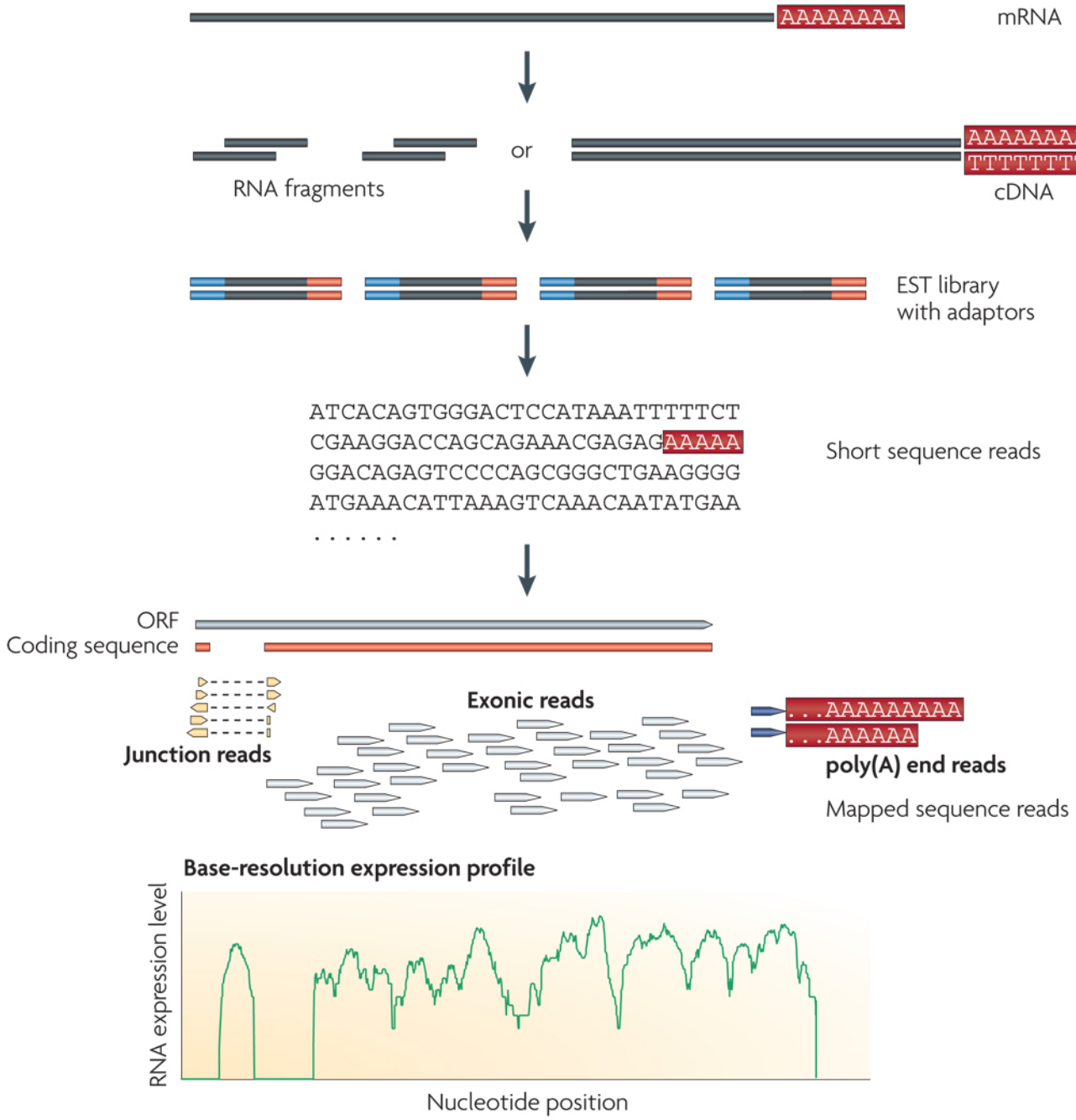
Gene expression microarray



Gene expression microarray



	Sample 1	**Sample 2**	**…**	**Sample n**
Gene 1	10.45	10.52		10.40
Gene 2	4.63	4.76		4.70
Gene 3	8.80	8.96		8.82
…				
…				
Gene p	8.30	8.44		8.27

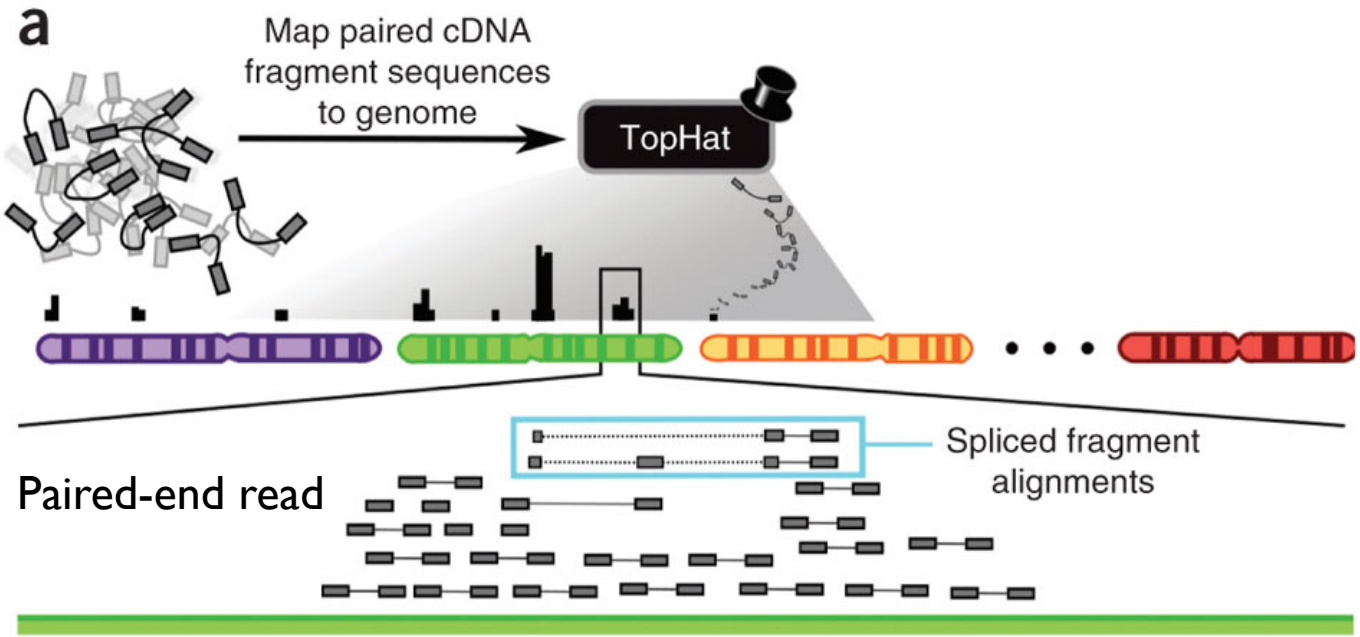


Nat Rev Genet. 2009 Jan;10(1):57-63

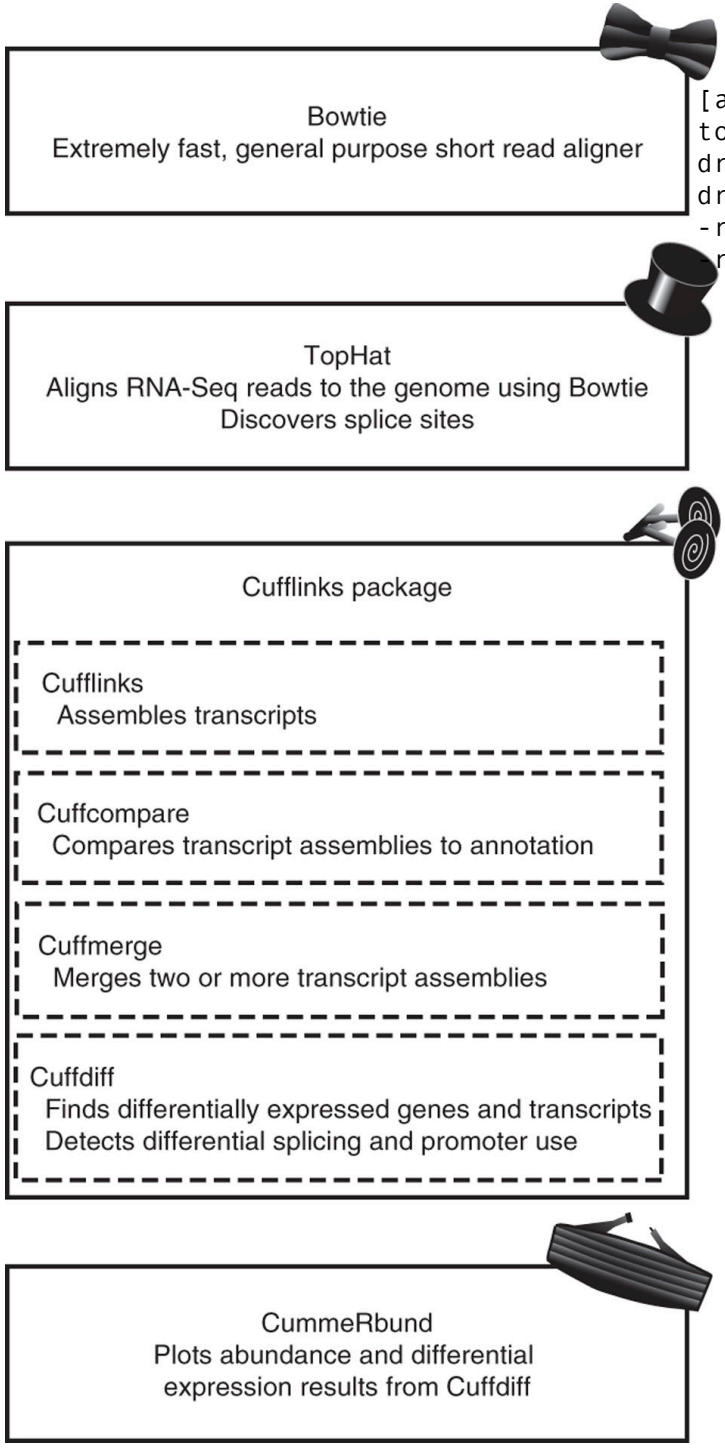
RNA-seq

RNA are cut into small fragments, select those fragments around certain base pairs (e.g., 400bps).

Sequence one end or both ends of the sequence fragments. Each sequenced part of an RNA-seq fragment is called an RNA-seq read. Map the RNA-seq reads to the genome, then count the # of fragments per gene



Bioinformatics. 2009 May 1;25(9):1105-11




```
[ammon:eQTL_seq/CEU/fastq] suninsky% ls -al
total 287346216
drwxr-xr-x  162 suninsky  staff        5508 Jun 16  2012 .
drwxr-xr-x   12 suninsky  staff         408 Sep 14  2012 ..
-rw-rw-rw-@   1 suninsky  staff   483776198 Feb 26  2010 1184_1_1.fastq
-rw-rw-rw-@   1 suninsky  staff   483776198 Feb 26  2010 1184_1_2.fastq
```

bsub -M 16 -q week tophat --segment-length 17
-G ~/research/eQTL_seq/CEU/anno/Homo_sapiens.GRCh37.66.updated.gtf
-o /lustre/scr/w/e/weisun/CEU/tophat/NA12892 -r 8
~/bin/bowtie/indexes/hg19 1184_1_1.fastq 1184_1_2.fastq

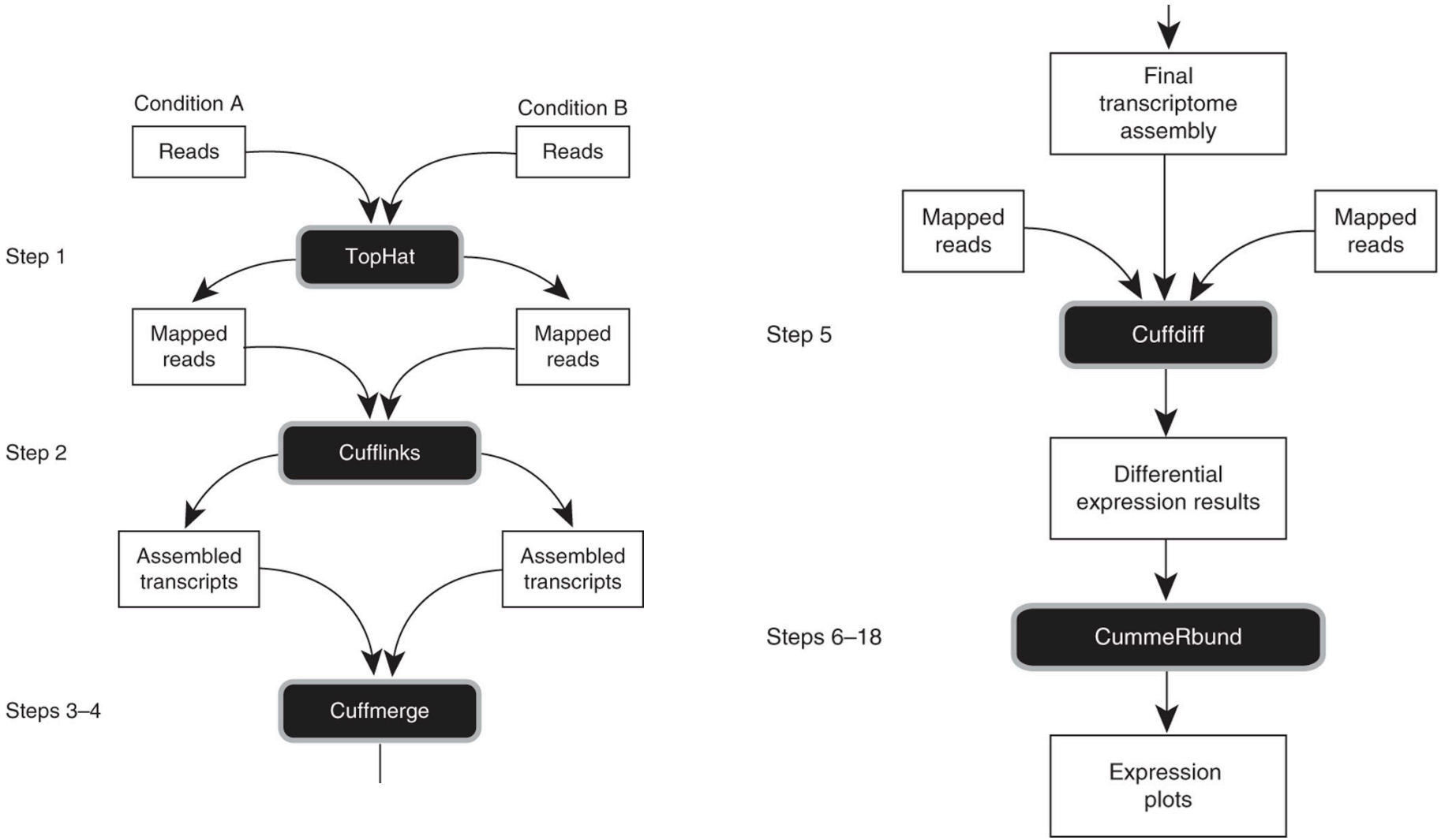
cufflinks -p 8 -o C1_R1_clout C1_R1_thout/accepted_hits.bam

Create a file called assemblies.txt that lists the assembly file for each sample.
The file should contain the following lines:
./C1_R1_clout/transcripts.gtf
./C2_R2_clout/transcripts.gtf

cuffmerge -g genes.gtf -s genome.fa -p 8 assemblies.txt

cuffdiff -o diff_out -b genome.fa -p 8 –L C1,C2 -u merged_asm/merged.gtf

Nat Protoc. 2012 Mar1;7(3):562-78.


Condition A
Condition B
Reads
Reads
Step 1
TopHat
Mapped reads
Mapped reads
Step 2
Cufflinks
Assembled transcripts
Assembled transcripts
Steps 3–4
Cuffmerge
Final transcriptome assembly
Mapped reads
Mapped reads
Step 5
Cuffdiff
Differential expression results
Steps 6–18
CummeRbund
Expression plots

RNA-seq

```
@HWI-EAS134:1:1:0:1446#0/1
GGGACTTAATCAACGCAAGCTTATGACCCGCACTT
+
BB<2A81(66B<2>BB@>?B=BA6=8@BCBB@?3/
```

bed-tools

	Sample 1	**Sample 2**	**…**	**Sample n**
Gene 1	512	339		286
Gene 2	1043	1212		888
Gene 3	20	12		10
…				
…				
Gene p	78	65		42

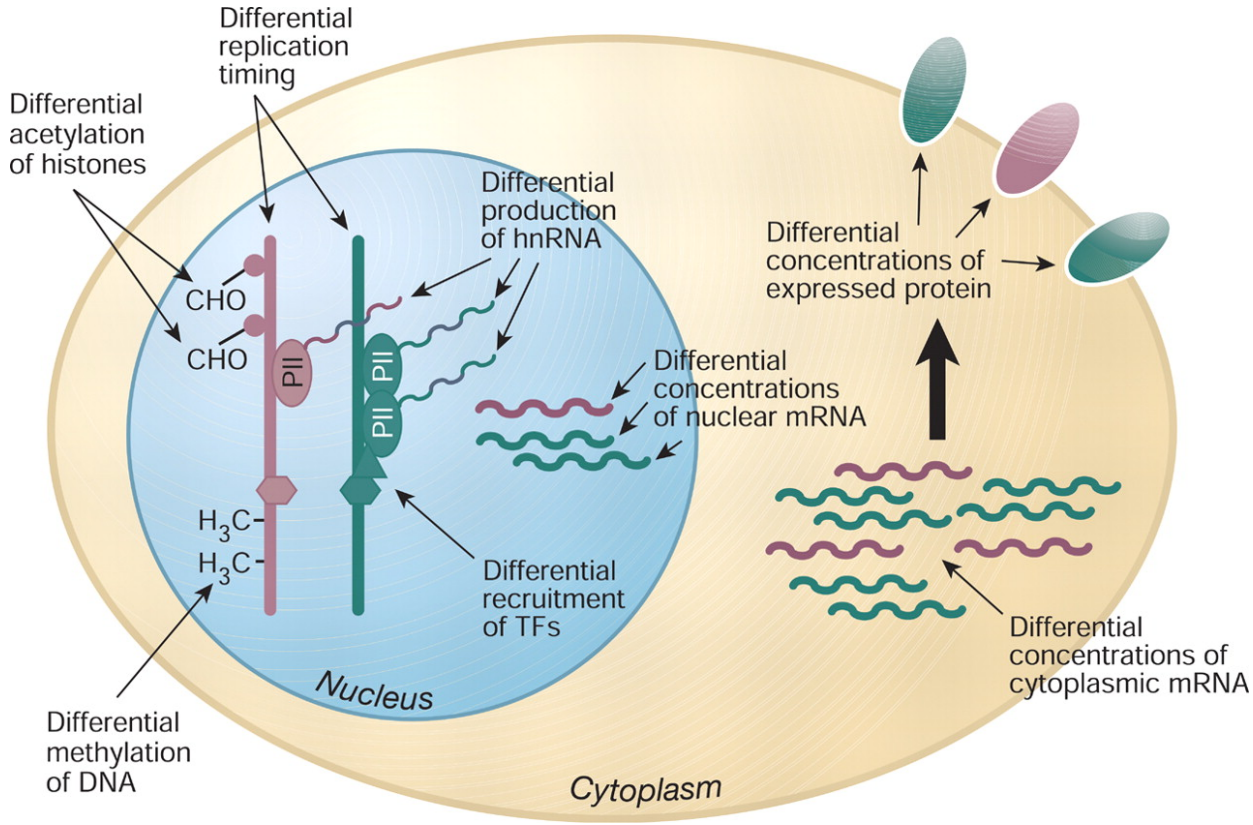
Microarray vs. RNA-seq

- RNA-seq is more accurate
  - Larger dynamic range
- RNA-seq is more expensive
- RNA-seq is less high throughput
- What else?
  - Allele-specific expression
  - RNA isoform specific expression

Outline

- 1. An brief introduction to RNAseq
- 2. Allele-specific expression
- 3. Isoform-specific expression

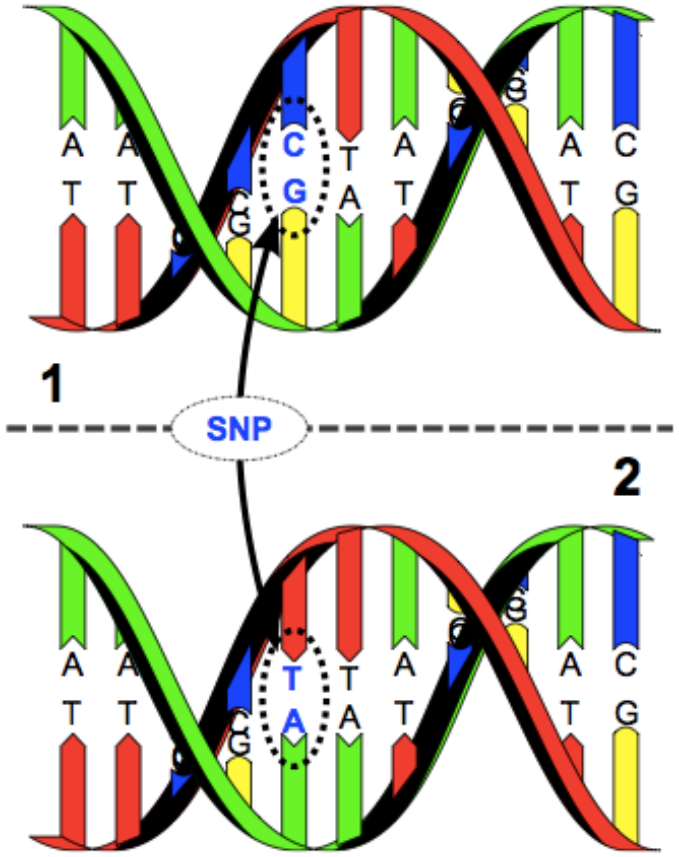
Allele-specific expression (ASE)



Science 22 October 2004: vol. 306 no. 5696 647-650

SNP

A single-nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide — A, T, C or G — in the genome differs between members of a biological species or paired chromosomes in an individual



We have two sets of chromosomes, i.e., two double-helix structure

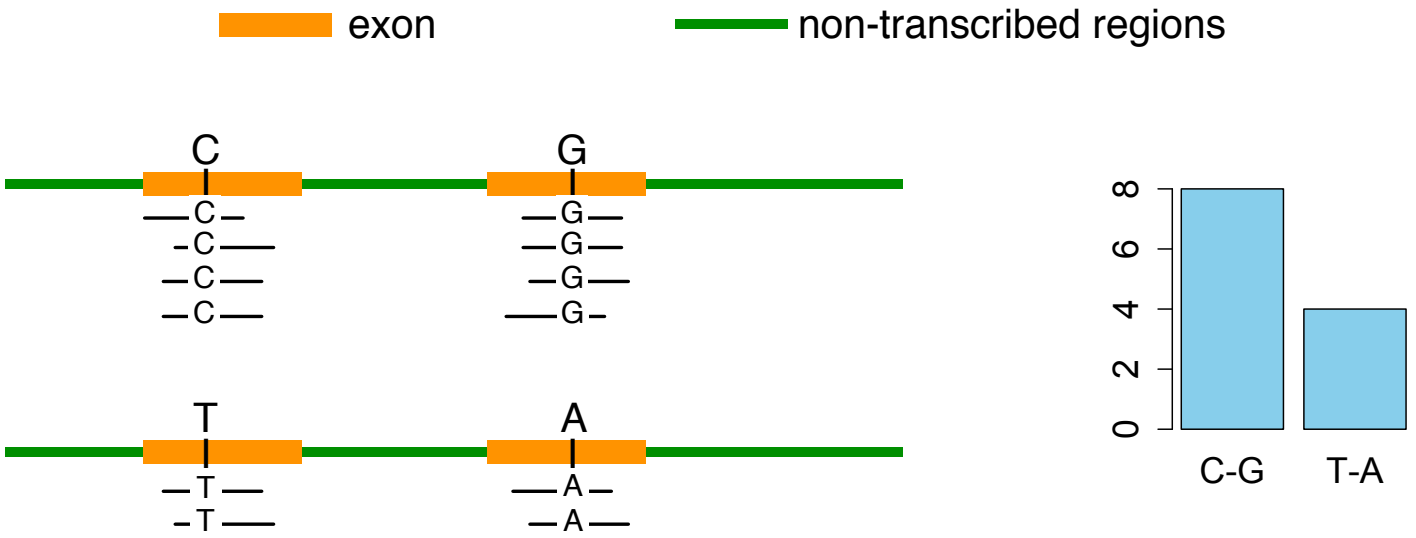
The genotype of this SNP of this individual is CT (heterozygous)

The genotypes at other locations are AA, AA, CC, TT, AA, AA, CC (homozygous)

Allele-specific expression (ASE)

- ASE can not be measured by microarray
  - Microarray use the same probe for both alleles of a gene
- ASE can be measured by RNA-seq
  - Only for those RNA-seq reads that overlap with heterozygous SNPs

Gene-level allele-specific expression (ASE)

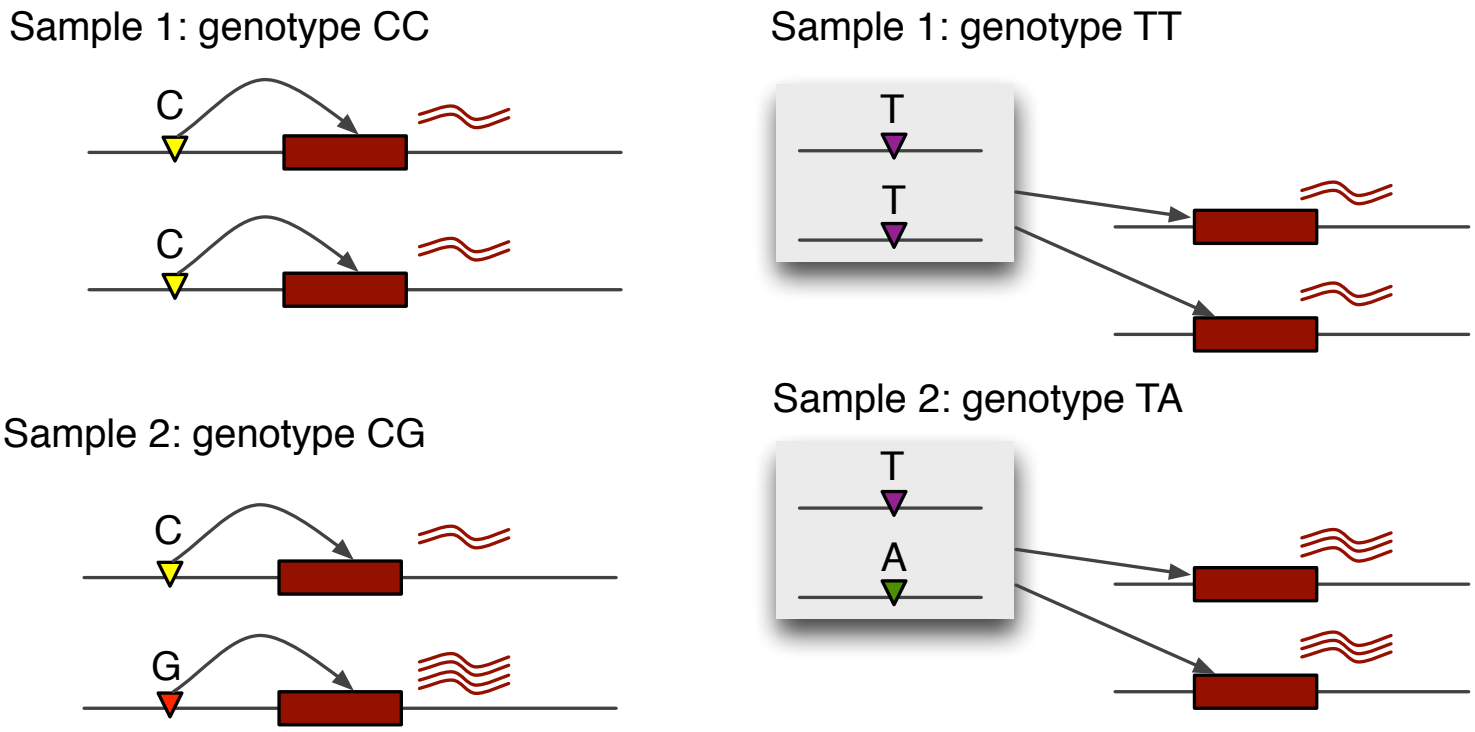


- ✓ A pipeline for appropriate sequence alignment.
- ✓ Need haplotype information to combine the ASE at each SNP/indel to obtain gene level ASE

Sun (2012), A Statistical Framework for eQTL mapping using RNA-seq, *Biometrics, 68(1):1-11*

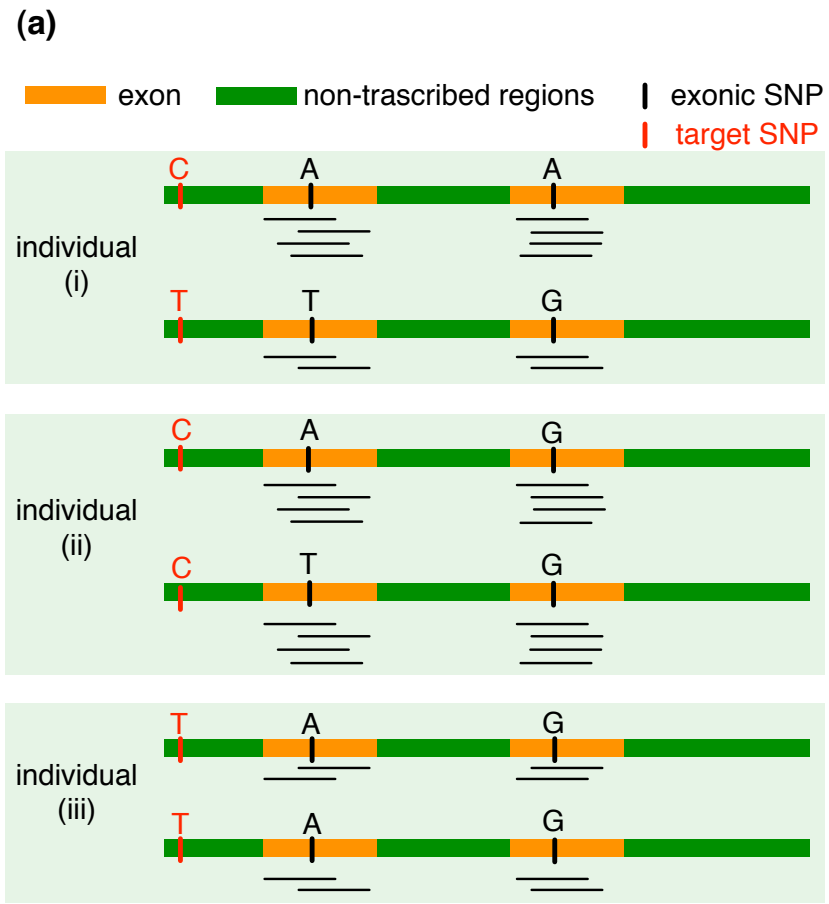
eQTL (gene expression QTL) Mapping Using ASE

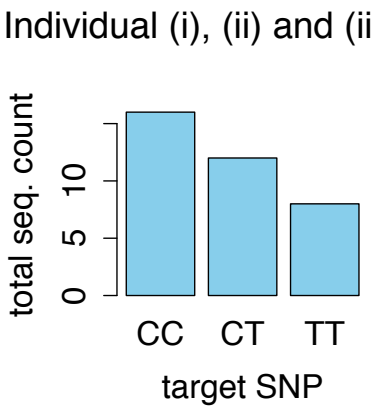
cis-acting eQTL vs. *trans*-acting eQTL

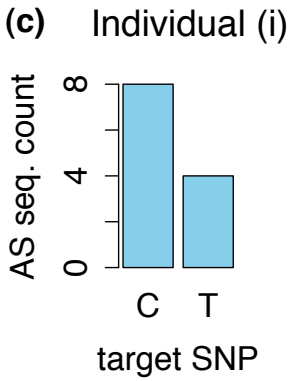


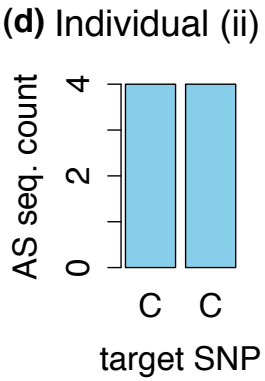
- Sun (2012), A Statistical Framework for eQTL mapping using RNA-seq, *Biometrics, 68(1):1-11*

eQTL Mapping Using ASE









- Sun (2012), A Statistical Framework for eQTL mapping using RNA-seq, *Biometrics, 68(1):1-11*

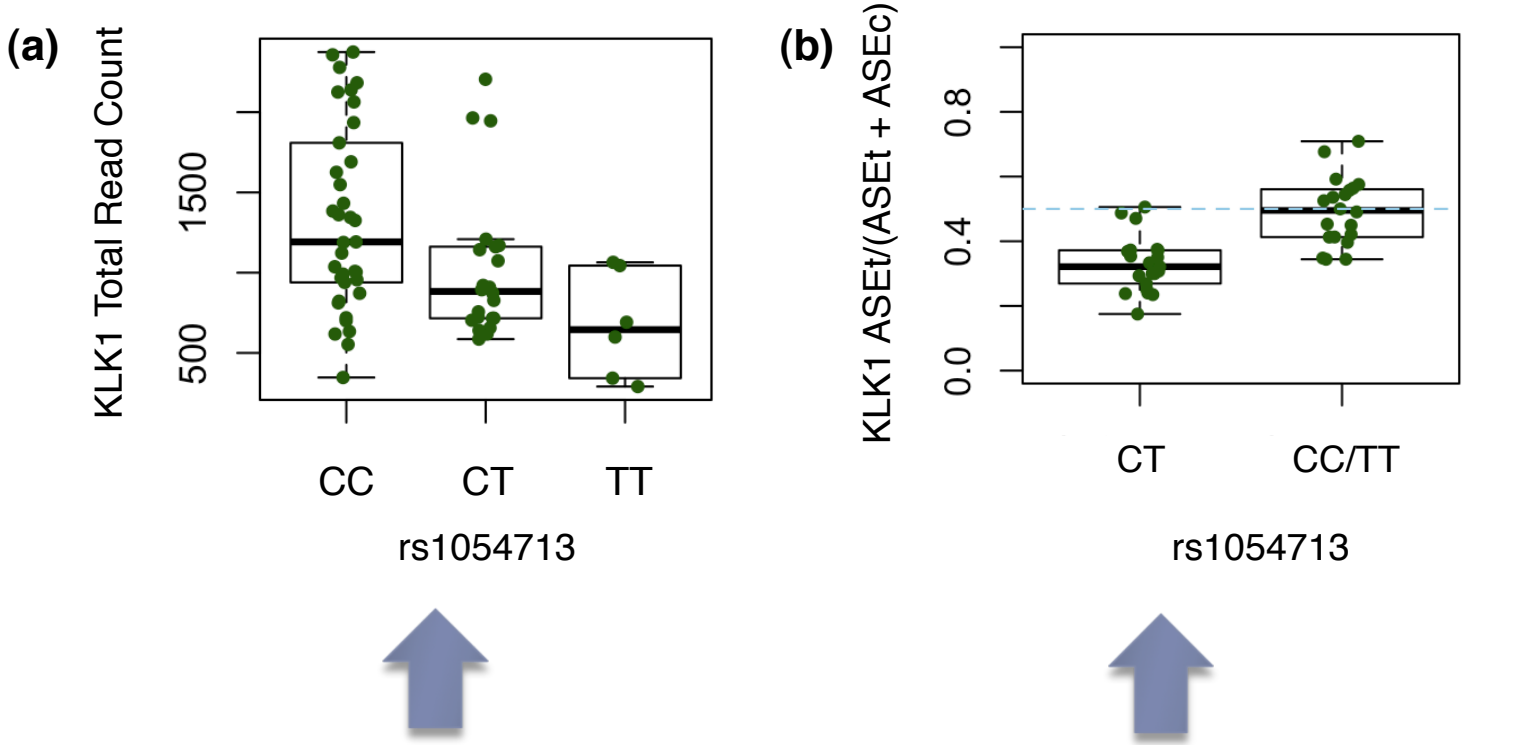
eQTL Mapping Using ASE

- Why?
  - Combine the information of total expression and ASE to improves the power of eQTL mapping
    - Using ASE, we compare the expression of paternal allele vs. maternal allele within an individual. One allele is an internal control of the other allele, and thus many confounding factors are canceled out.
  - Identify *cis*-acting eQTL that directly influences gene expression in an allele-specific manner

- Sun (2012), A Statistical Framework for eQTL mapping using RNA-seq, *Biometrics, 68(1):1-11*

eQTL Mapping Using ASE

Real data analysis of 65 HapMap YRI samples (Pickrell et al. (2010) Nature 464, 768–772)



Negative Binomial distribution for TReC

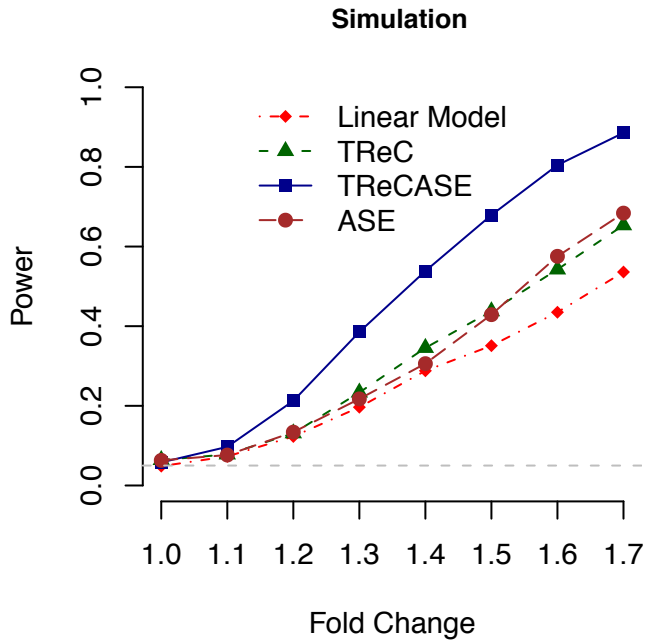
Beta-Binomial distribution for ASE

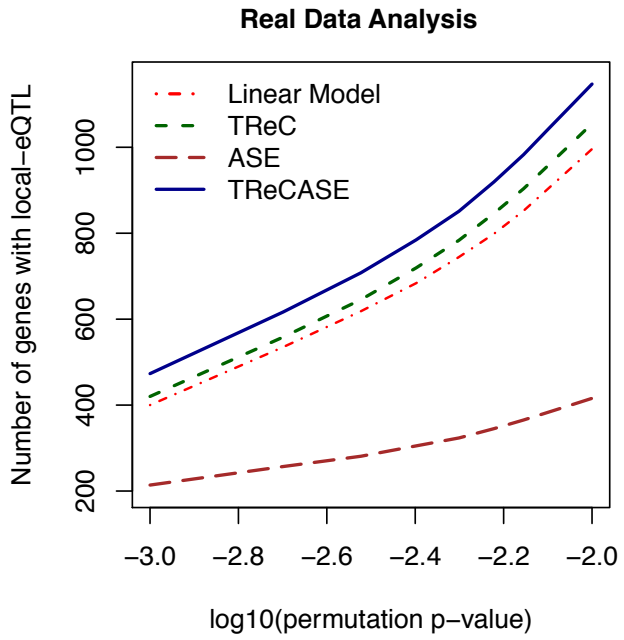
- Sun (2012), A Statistical Framework for eQTL mapping using RNA-seq, *Biometrics, 68(1):1-11*

eQTL Mapping Using ASE

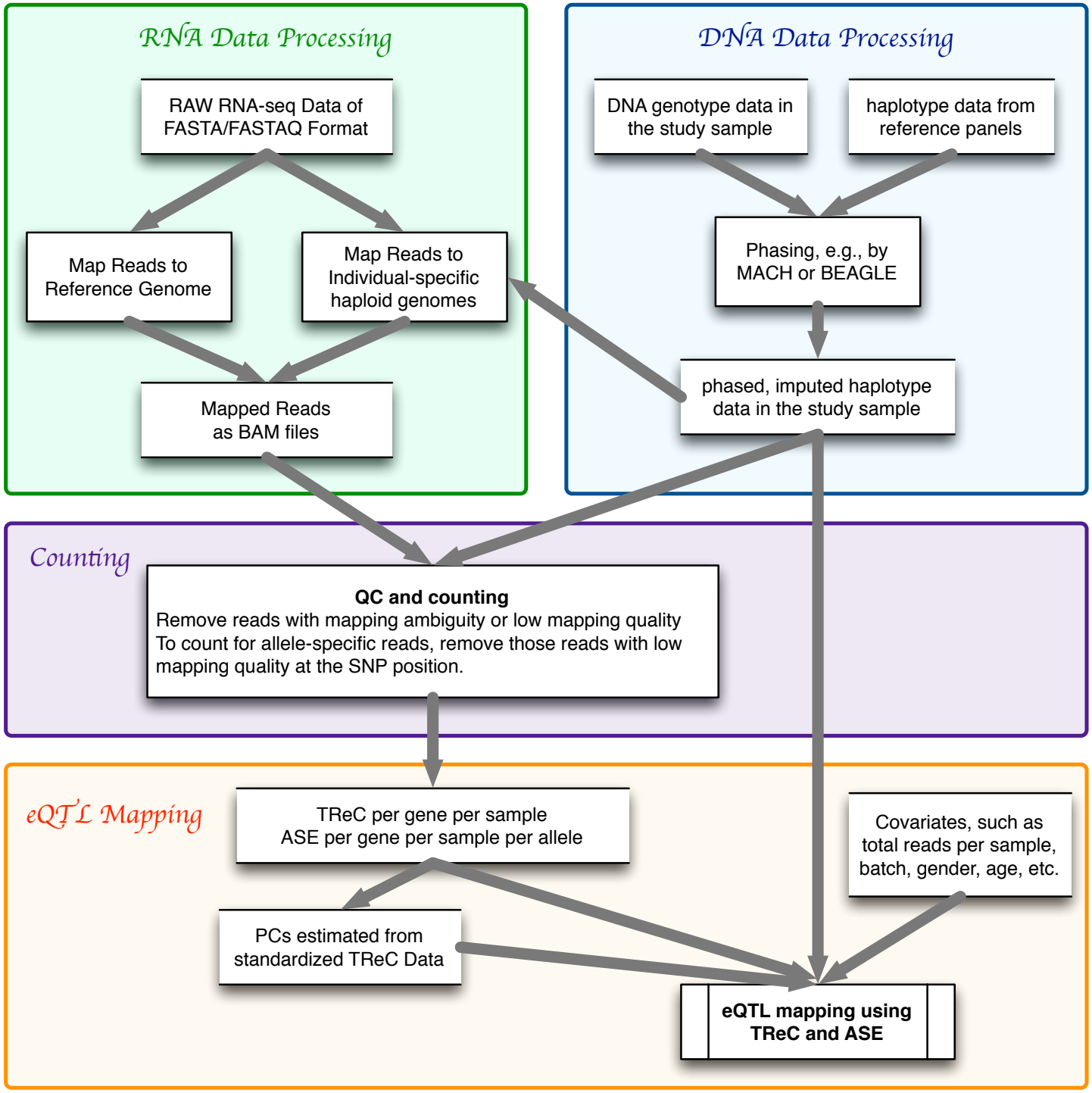
Real data analysis of 65 HapMap YRI samples (Pickrell et al. (2010) Nature 464, 768–772)

TReC: Total Read Count, TReCASE=TReC+ASE





- Sun (2012), A Statistical Framework for eQTL mapping using RNA-seq, *Biometrics, 68(1):1-11*


RNA Data Processing
RAW RNA-seq Data of FASTA/FASTAQ Format
Map Reads to Reference Genome
Map Reads to Individual-specific haploid genomes
Mapped Reads as BAM files
DNA Data Processing
DNA genotype data in the study sample
haplotype data from reference panels
Phasing, e.g., by MACH or BEAGLE
phased, imputed haplotype data in the study sample
Counting
QC and counting
Remove reads with mapping ambiguity or low mapping quality
To count for allele-specific reads, remove those reads with low mapping quality at the SNP position.
eQTL Mapping
TReC per gene per sample
ASE per gene per sample per allele
Covariates, such as total reads per sample, batch, gender, age, etc.
PCs estimated from standardized TReC Data
eQTL mapping using TReC and ASE

http://www.bios.unc.edu/~weisun/software/asSeq.pdf

asSeq: A set of tools for the study of allele-specific RNA-seq data

Wei Sun and Vasyl Zhabotynsky

February 22, 2013

1 Overview

```
> library(asSeq)
```

This vignette describes how to use `R/asSeq` to perform eQTL mapping using total expression and/or allele-specific expression, including a pipeline for input data preparation.

```
# ------------------------------------------------------------------------
# 3. getUnique and filtering
# ------------------------------------------------------------------------
prepareBAM(bamF, sprintf("%s_sorted_by_name", sami), sortIt=FALSE)
```

5.3 Extract the allele-specific reads

R function `extractAsReads` from our R package `asSeq` can be used to extract allele-specific sequence reads. For example, the command can be

```
extractAsReads(input,  snpList, outTag="myOutput", prop.cut=.5, min.snpQ=10, phred=33)
```

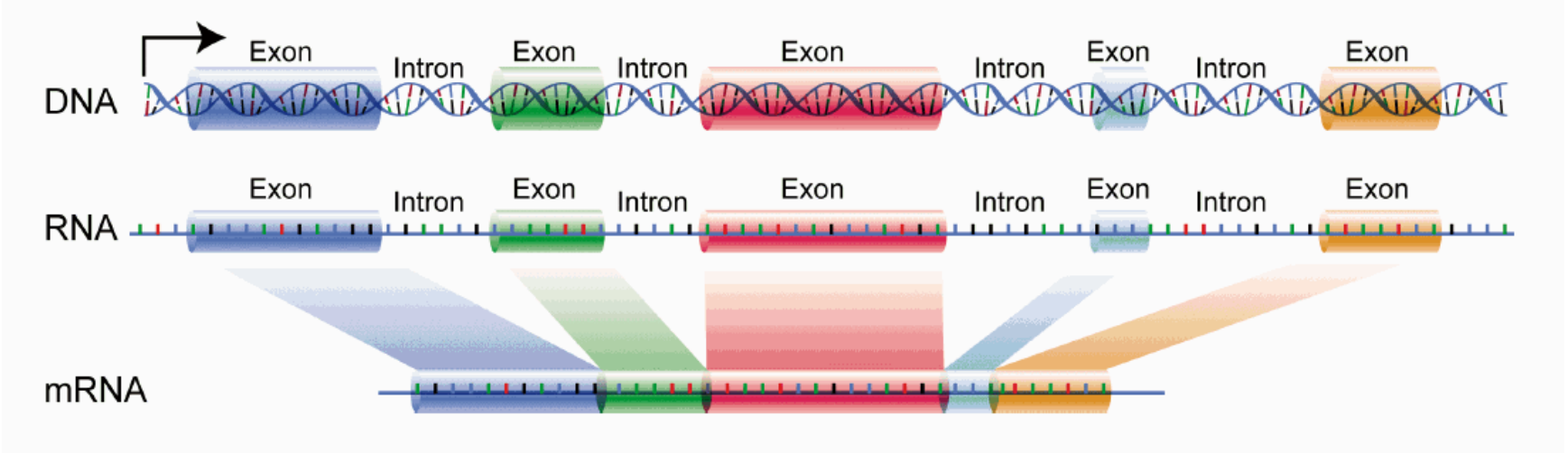
```
chr1 1019668 G A
chr1 1020428 T C
chr1 1020496 A G
chr1 1029889 C T
```

Outline

- 1. An brief introduction to RNAseq
- 2. Allele-specific expression
- 3. Isoform-specific expression

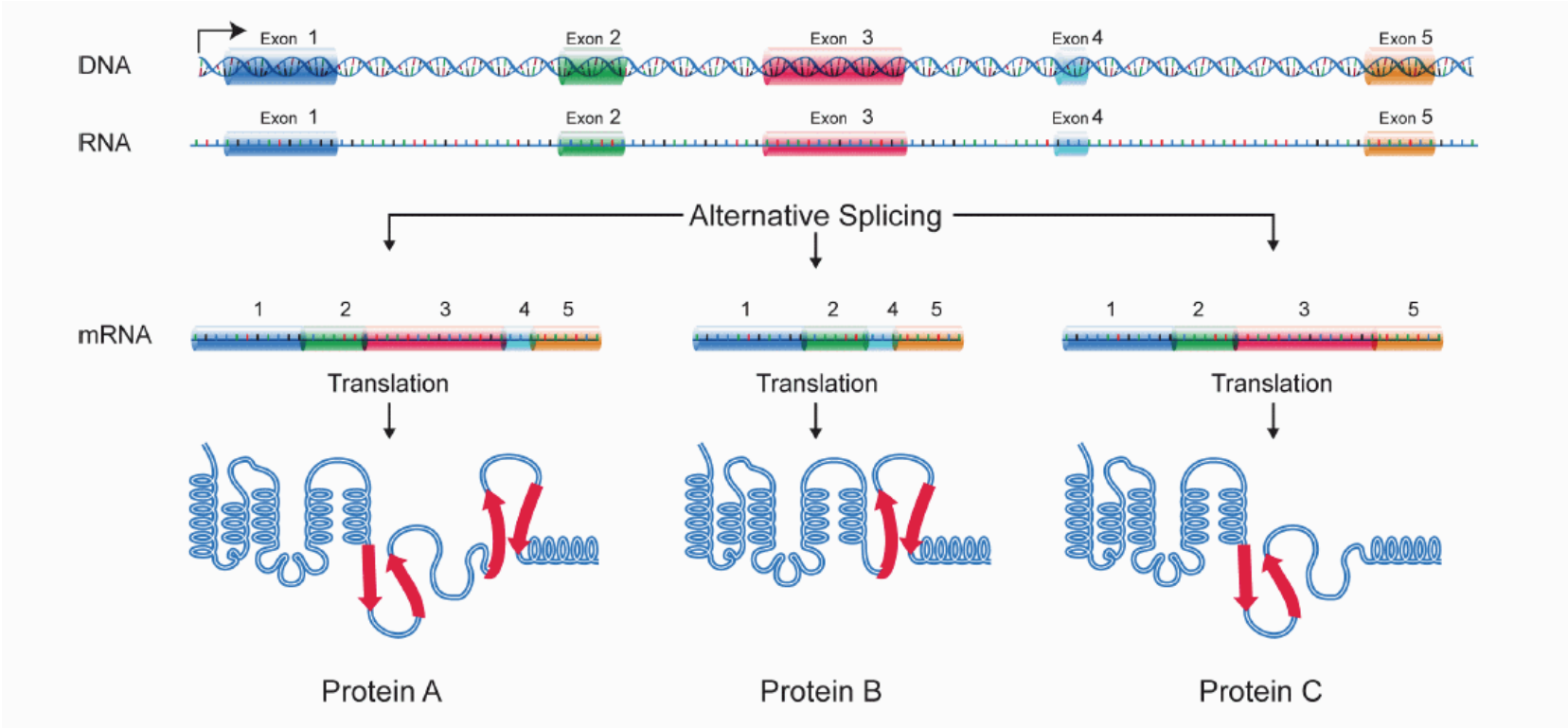
Transcription - splicing

Transcription: DNA ➔ RNA
Splicing: RNA ➔ messenger RNA



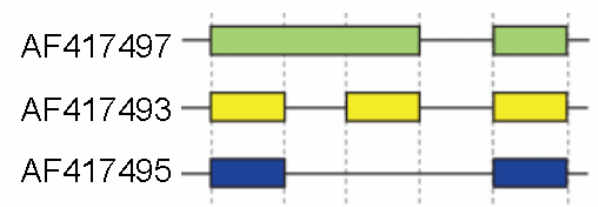
Alternative Splicing

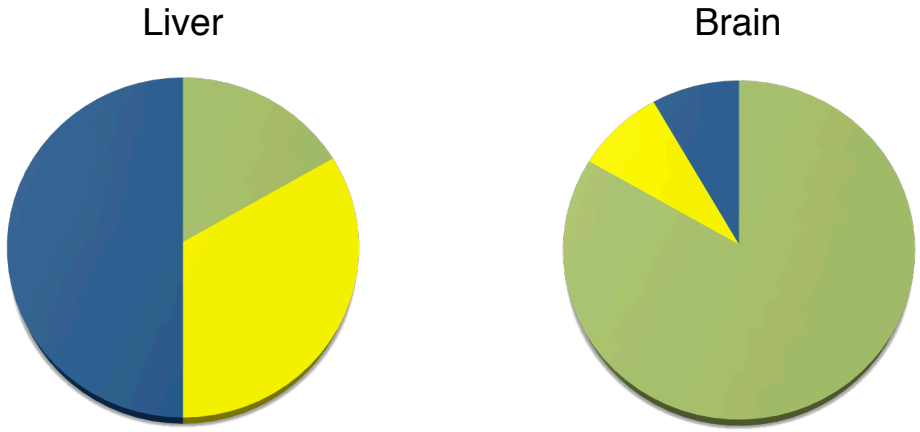
Alternative splicing is a process by which the exons of the RNA are reconnected in multiple ways during RNA splicing. The resulting different mRNAs are **RNA Isoforms** and they may be translated into different protein isoforms.



Isoform-specific expression

- Differential Isoform usage
  - the relative expression of an isoform with respect to the total expression of the corresponding gene





Differential Isoform Usage

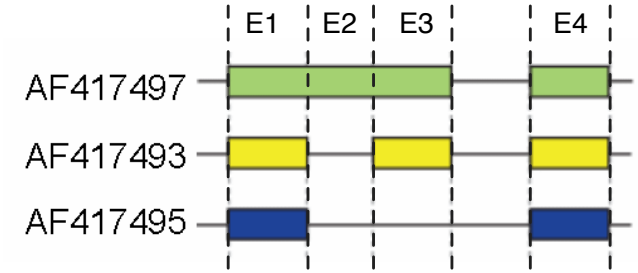
- Cufflinks, CuffDiff (Trapnell et al. 2010, NatBiotechnol. 28(5):511-5)
  - Estimate RNA-isoform expression, followed by differential isoform usage testing
- FDM (Singh et al. 2011 bioinformatics vol. 27 2633-40)
  - Testing for differential isoform usage using spliced read alignment, without estimating RNA-isoform expression
- Both CuffDiff and FDM perform two sample test
  - Case vs. Control, treatment vs. placebo
  - Testing for differential isoform usage vs. a continuous variable? e.g., additive genotype, dosage, patient survival time.

- Sun et al. (2012), manuscript

Isoform-specific expression

Construct non-overlapping exons

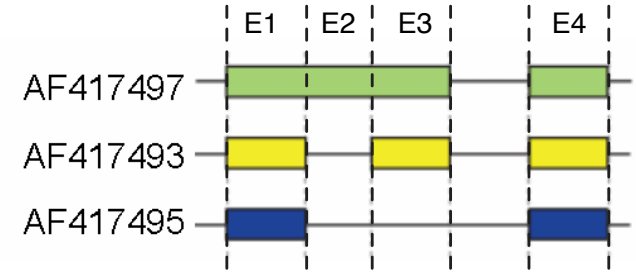
Assume the exon locations and lengths are known, but they may be overlapping with each other. Construct a new set of exons that are non-overlapping



For each gene, count the number of RNA-seq fragments for each exon set

Exon Set	Count
E1	50
E2	20
E3	30
E4	50
E1:E2	10
E2:E3	10
E1:E3	5
E3:E4	18
E1:E4	15

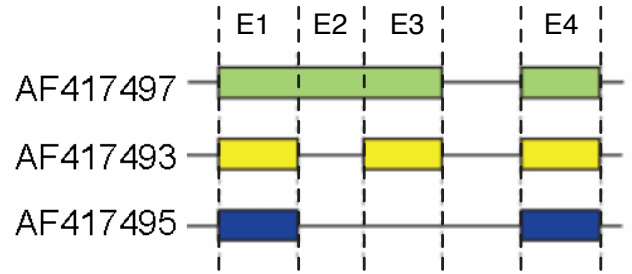
RNA-isoform selection and abundance estimation



Exon Set	Count
E1	50
E2	20
E3	30
E4	50
E1:E2	10
E2:E3	10
E1:E3	5
E3:E4	18
E1:E4	15

If the expression are all from isoform AF417497,
What is the pattern of read count across exon sets?

RNA-isoform selection and abundance estimation



$$y = b_1 x_1 + b_2 x_2 + b_3 x_3$$

Exon Set	Count
E1	50
E2	20
E3	30
E4	50
E1:E2	10
E2:E3	10
E1:E3	5
E3:E4	18
E1:E4	15

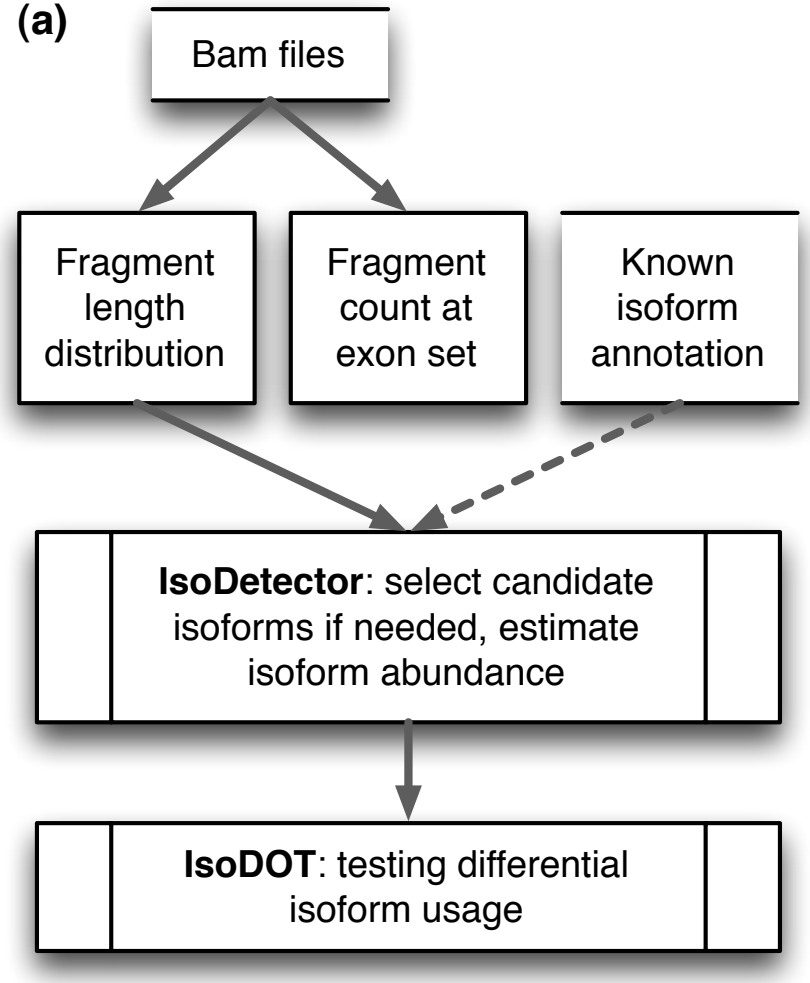
AF417497	AF417493	AF417495
1	1	1
1	0	0
1	1	0
1	1	1
1	0	0
1	0	0
0.1	1	0
1	1	0
0.01	0.02	1

IsoDOT

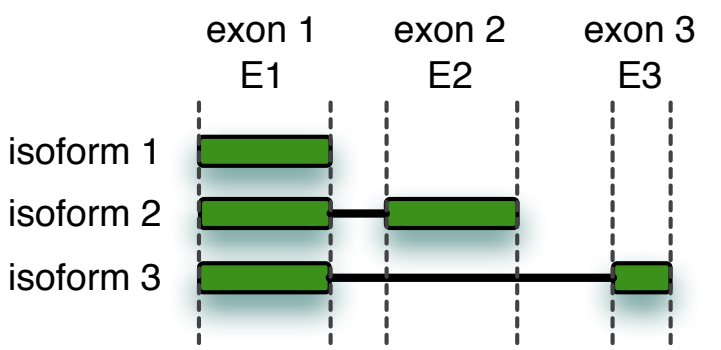
- Challenges
  - This is a very challenging (negative binomial) regression problem because
  - There is a huge number of possible isoforms, especially when there is no complete annotation on the RNA isoforms of the gene of interest.
  - There can be strong correlations of the expression pattern of several isoforms.
- Solutions
  - Select a initial group of candidate isoforms.
  - Apply penalized regression.

An overview of IsoDOT

(a)



(b)



(c)

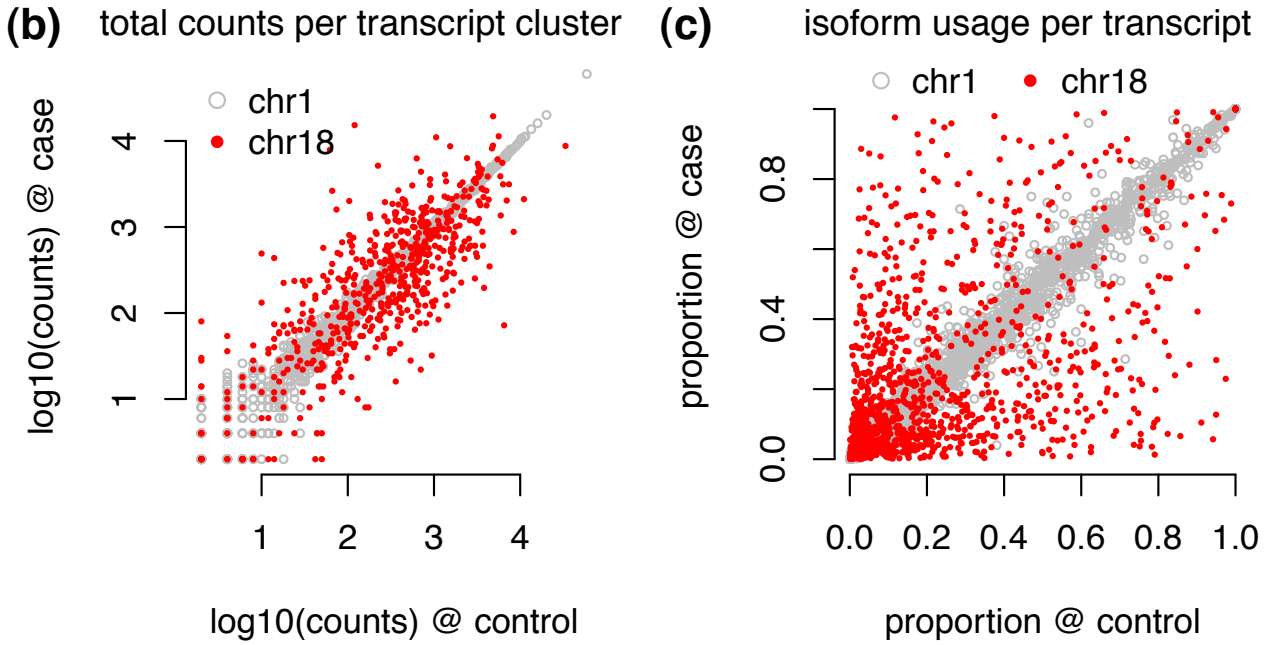
Exon Set	Count	Isoform 1	Isoform 2	Isoform 3
E1	50	$\eta_{1;1}$	$\eta_{1;2}$	$\eta_{1;3}$
E2	20	0	$\eta_{2;2}$	0
E3	10	0	0	$\eta_{3;3}$
E1:E2	10	0	$\eta_{1,2;2}$	0
E1:E3	5	0	0	$\eta_{1,3;3}$
E2:E3	0	0	0	0
E1:E2:E3	0	0	0	0

$$\mu = x_1 b_1 + x_2 b_2 + x_3 b_3$$
$$y \sim NB(\mu, \phi)$$

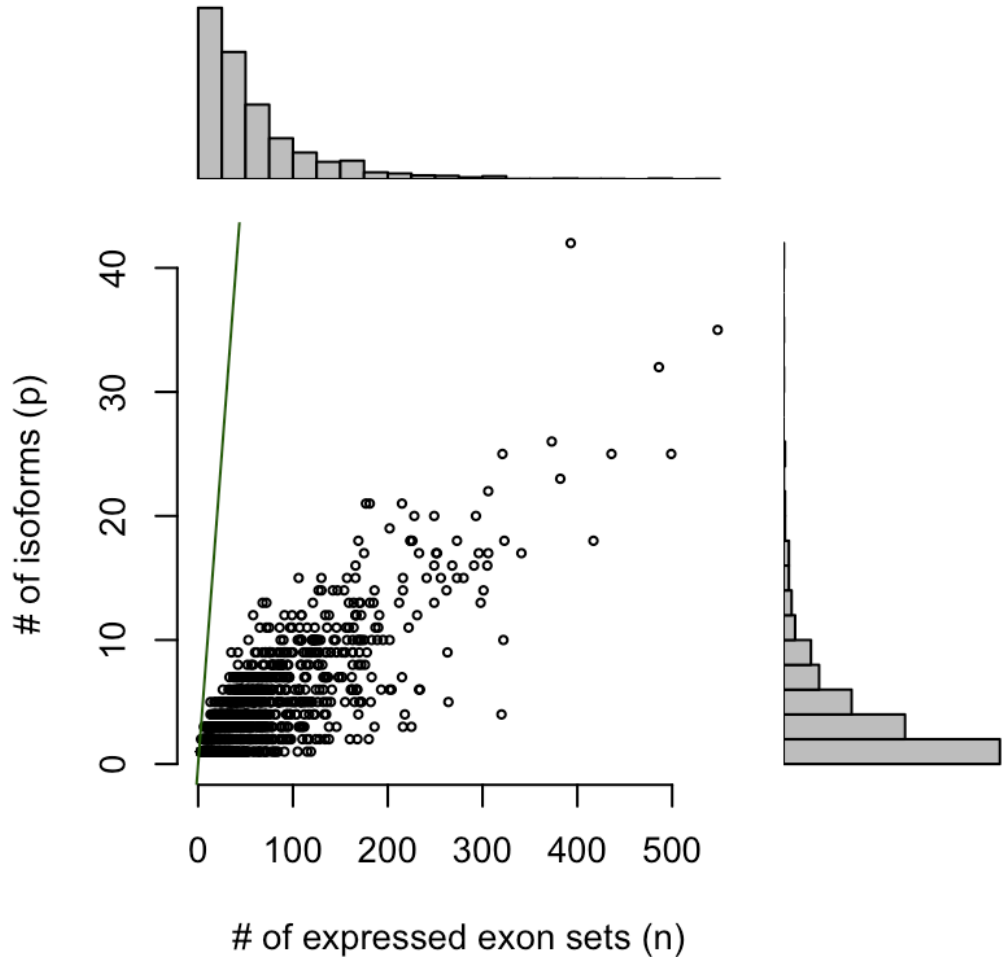
Simulation

We simulated two independent sets of ~2 million 76bp paired-end RNA-seq reads by Flux Simulator, using the transcriptome annotation of chromosome 1 and 18 of mouse genome.

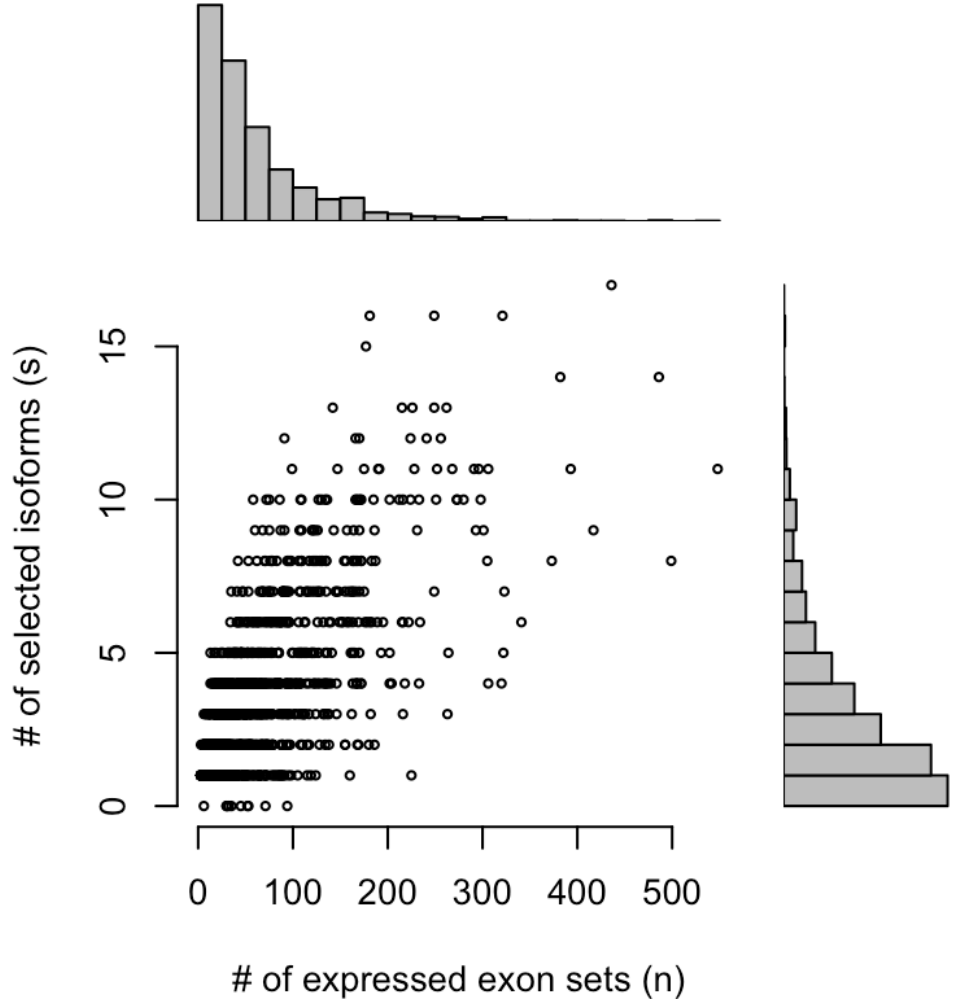
A case and a control sample were generates such as all the genes of chr1 are equivalently expressed between the case and the control and all the genes of chr18 are differentially expressed between the case and the control.



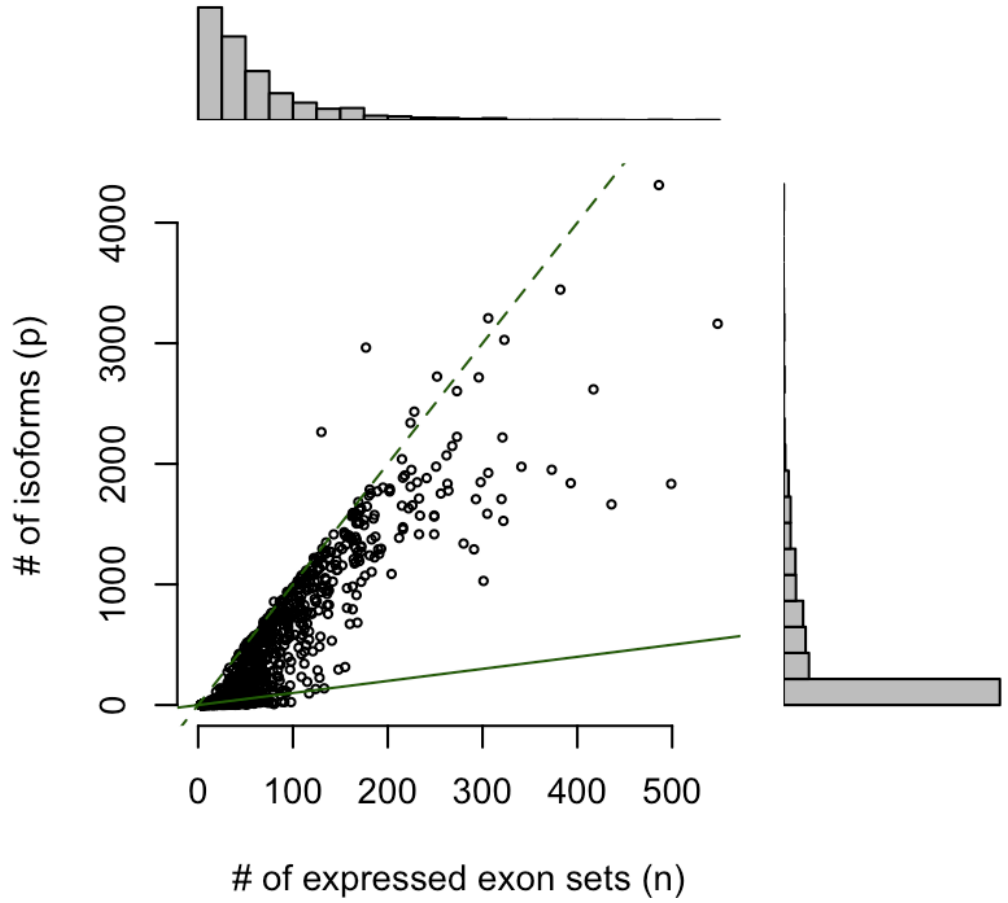
The dimension of this problem: known isoforms



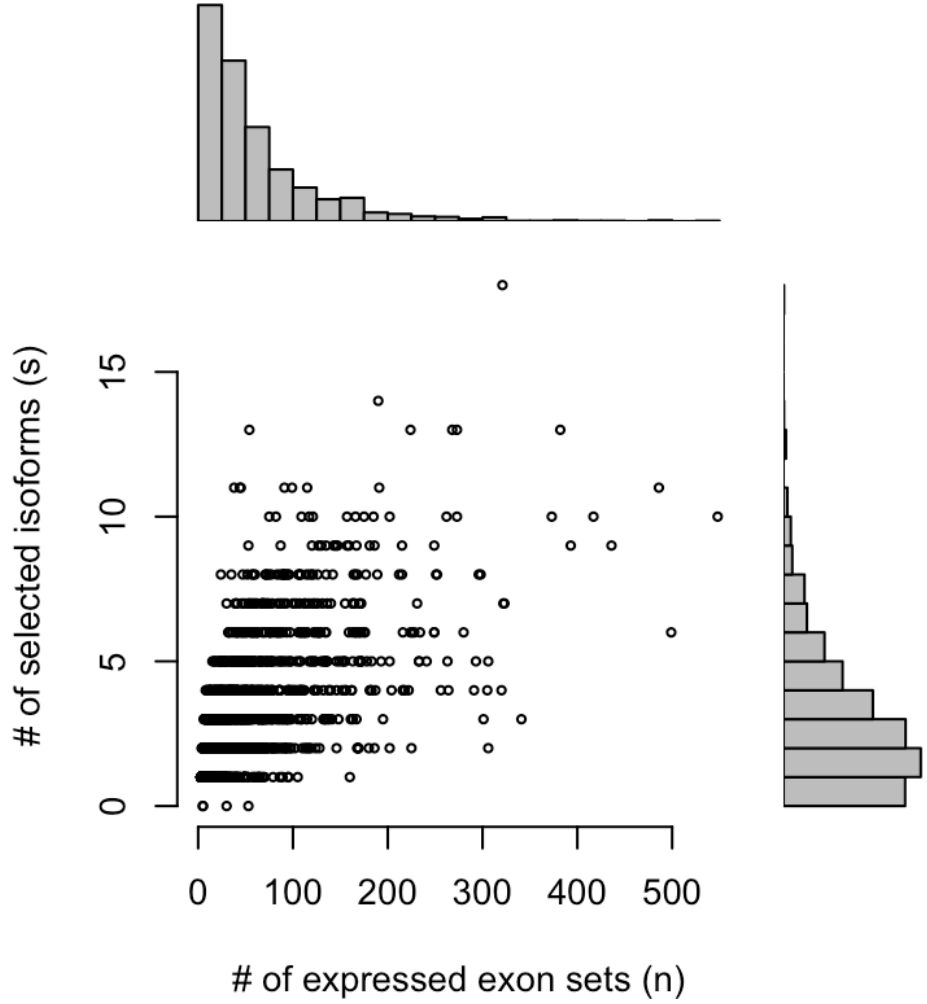
The dimension of this problem: known isoforms, after isoform selection



The dimension of this problem: unknown isoforms



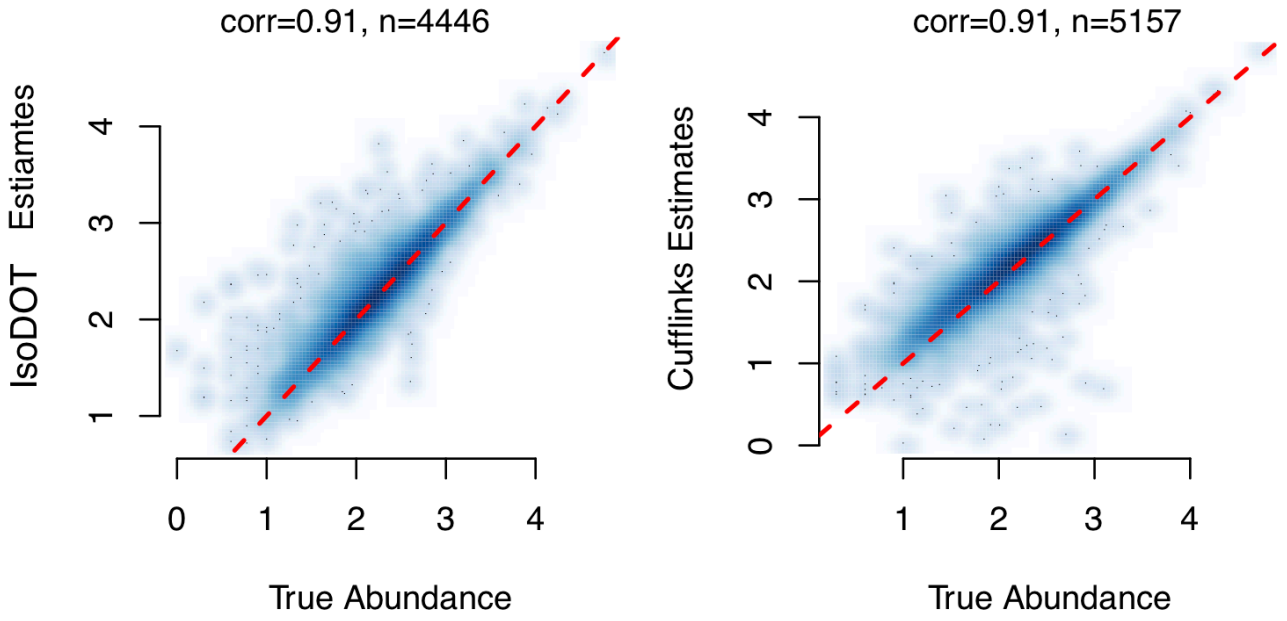
The dimension of this problem: unknown isoforms, after isoform selection



IsoDOT

Compare IsoDOT with Cufflinks in terms of RNA isoform abundance estimation, using simulated data from flux-simulator, **with isoform annotation**

IsoDOT and cufflinks have comparable performance

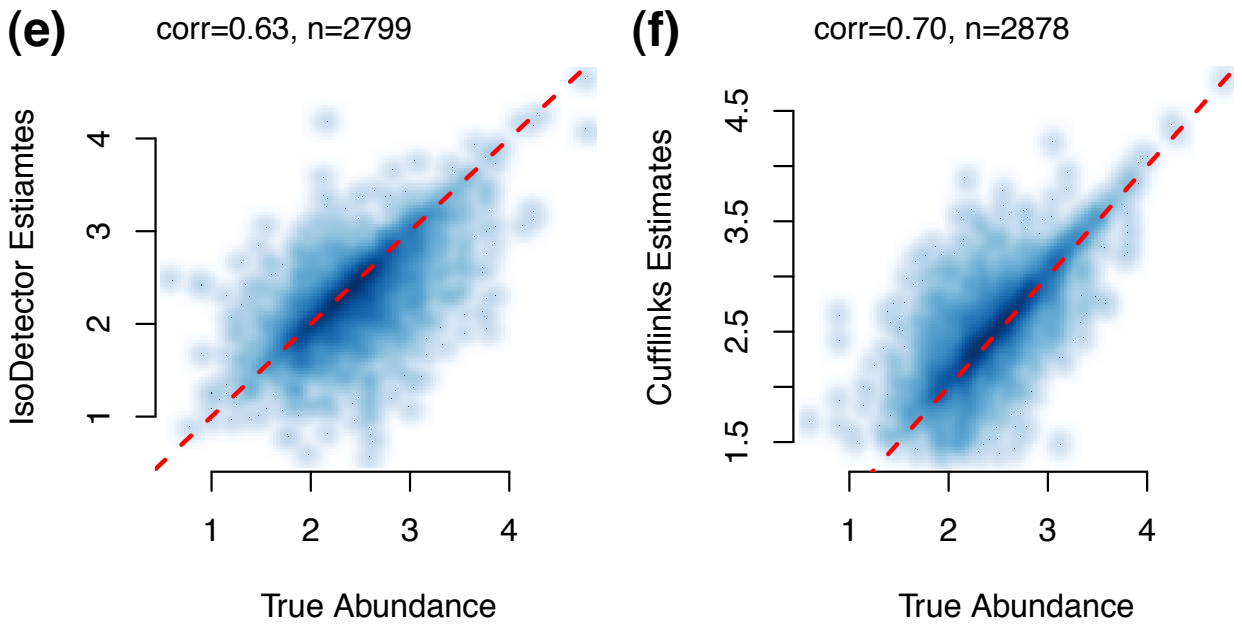


- Sun et al. (2012), manuscript

IsoDOT

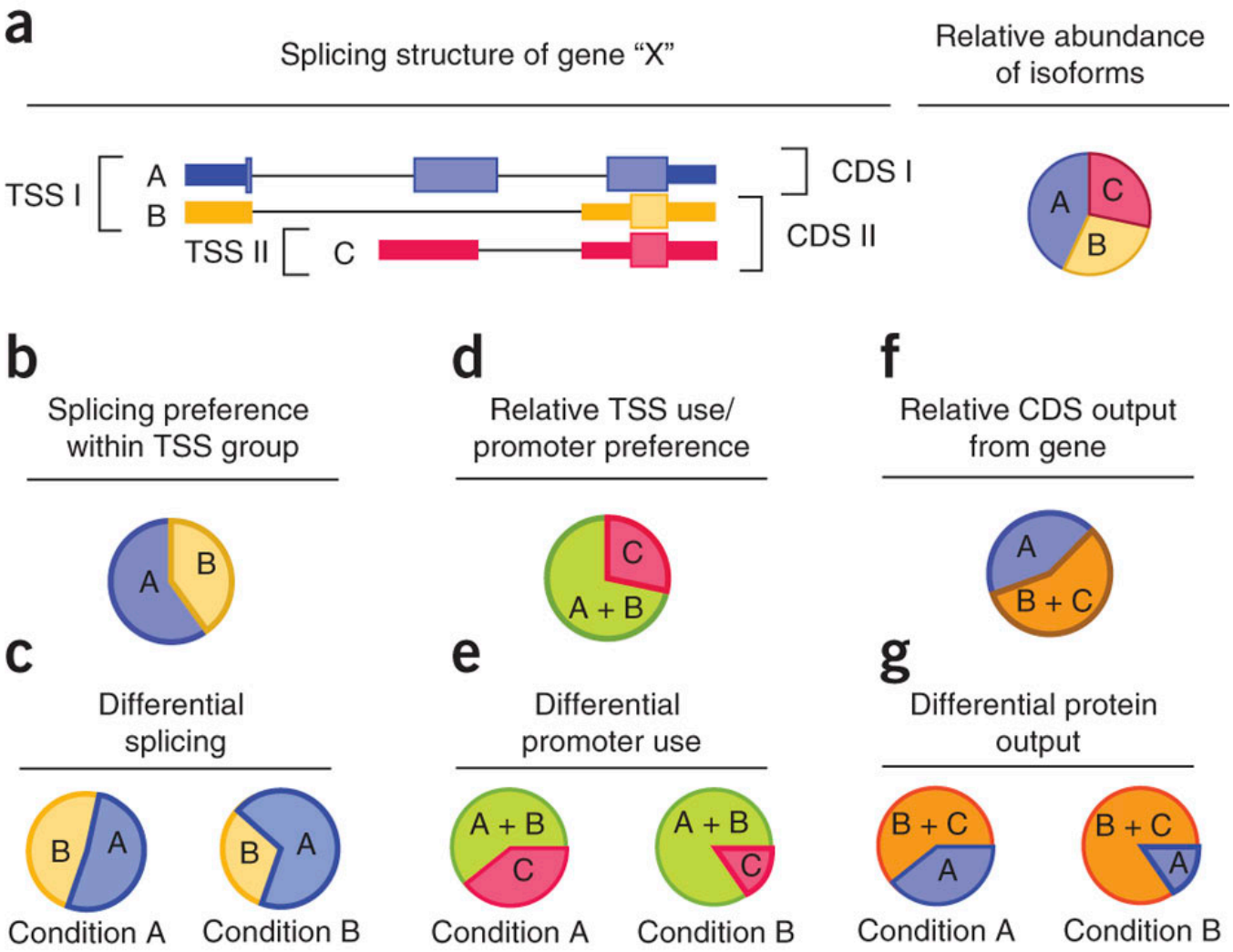
Compare IsoDOT with Cufflinks in terms of RNA isoform abundance estimation, using simulated data from flux-simulator, **without isoform annotation**

IsoDOT and cufflinks have comparable performance

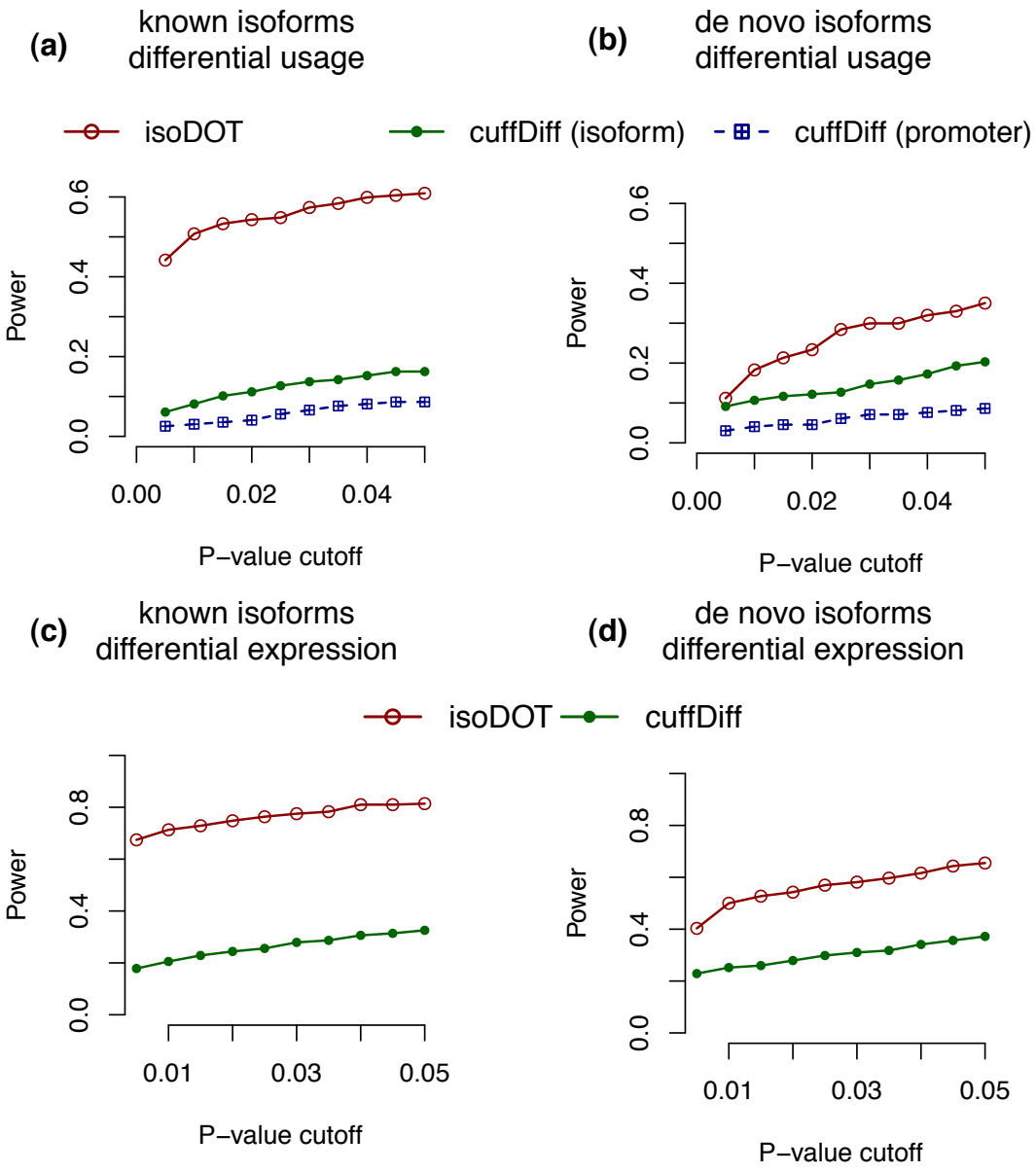


- Sun et al. (2012), manuscript

Different tests performed by Cufflinks



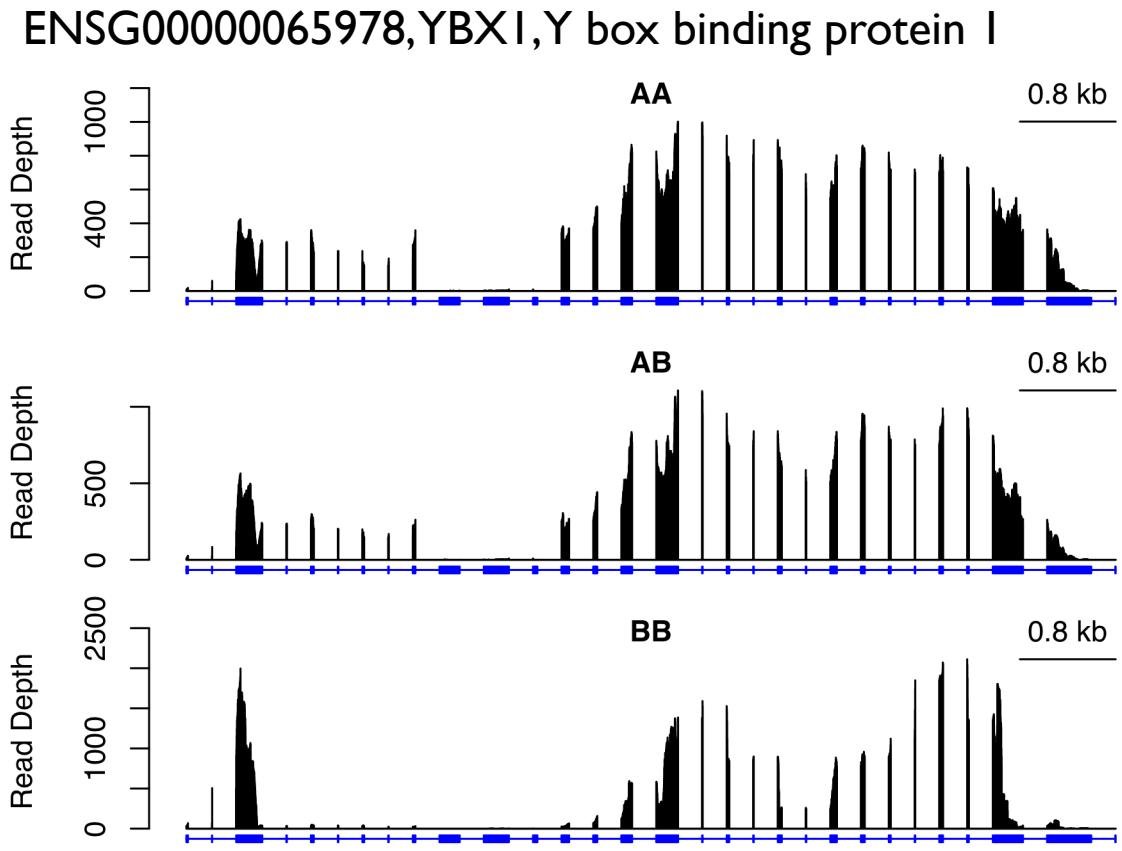
Nat Protoc. 2012 Mar1;7(3):562-78.



Test for differential isoform usage using simulated RNA-seq data from one case vs. one control

Isoform-specific eQTL

Real data analysis on 60 CEU HapMap samples



- Sun et al. (2012), manuscript

http://www.bios.unc.edu/~weisun/software/isoform_files/isoform.pdf

isoform: A set of tools for RNA isoform study using RNA-seq data.

Wei Sun

February 22, 2013

1 Overview

```
> library(isoform)
```

This vignette describes how to use `R/isoform` to estimate isoform abundance and assess differential isoform usage between cases and controls using an example. We are working on applying this method to assess differential isoform usage with respect to continuous covariate, e.g., isoform-specific eQTL mapping. Such examples will be added soon.

sorting and some basic QC using function prepareBAM from R package `asSeq`.

```
cmd2 = sprintf("samtools sort -n %s %s_sorted_by_name", bami, sami)
system(cmd2)
bamF = sprintf("%s_sorted_by_name.bam", sami)

prepareBAM(bamF, sprintf("%s_sorted_by_name", sami), sortIt=FALSE)
```

The following set of codes can be used to count the number of reads per exon set.

```
library(isoform)

bedFile = "Mus_musculus.NCBIM37.67.nonoverlap.exon.bed"
bamFile = "mm9_simu_set1_paired_reads_sorted_by_name.bam"
ctFile  = "mm9_simu_set1_counts_paired_reads.txt"
```

```
countReads(bamFile, bedFile, ctFile)
```