# RNA-Seq Tutorial 1

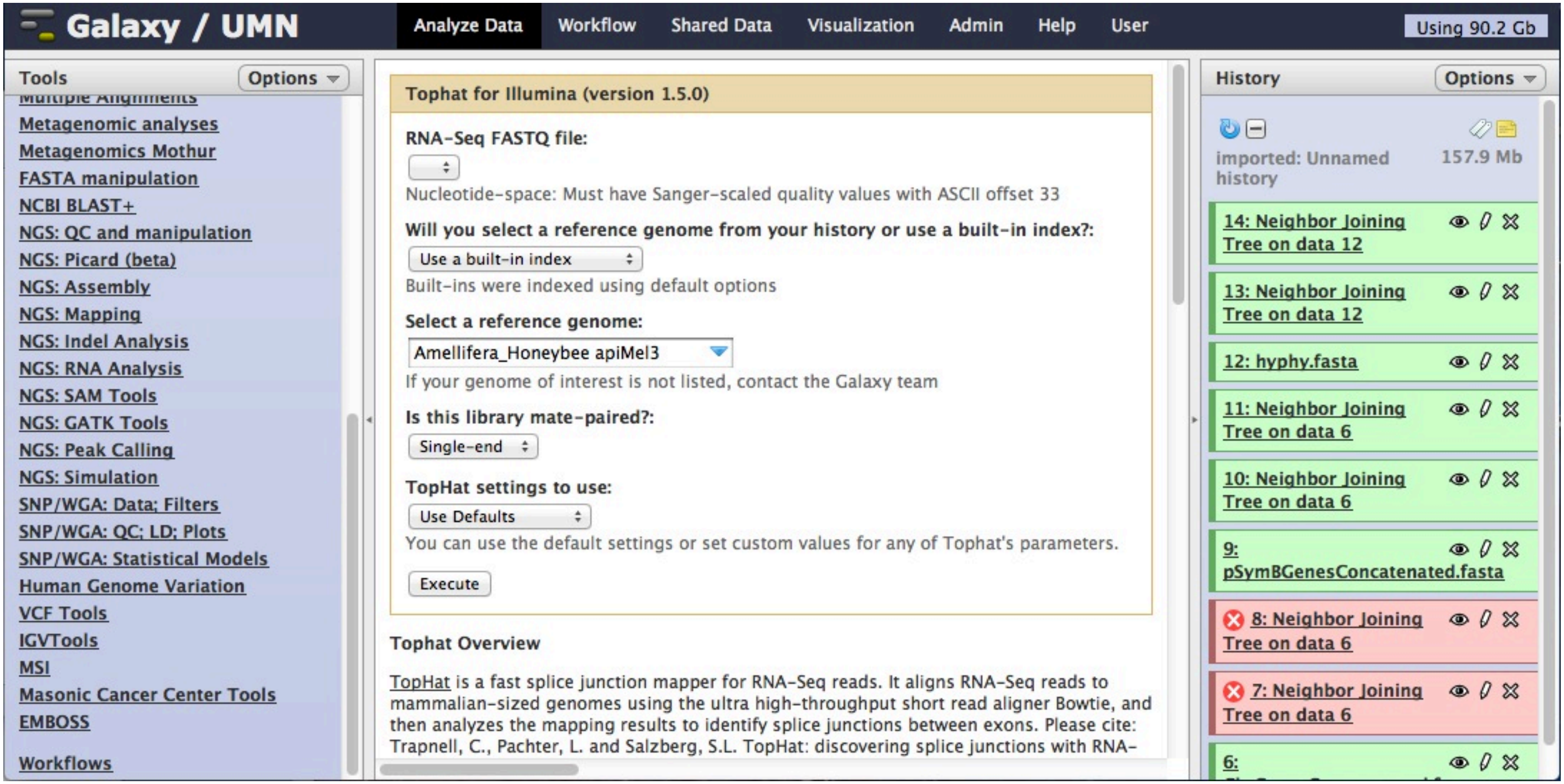
John Garbe
Research Informatics Support Systems, MSI

March 19, 2012

# RNA-Seq Tutorials

- Tutorial 1
  - RNA-Seq experiment design and analysis
  - Instruction on individual software will be provided in other tutorials
- Tutorial 2
  - Hands-on using TopHat and Cufflinks in Galaxy
- Tutorial 3
  - Advanced RNA-Seq Analysis topics

# Galaxy.msi.umn.edu



**Tophat for Illumina (version 1.5.0)**

**RNA-Seq FASTQ file:**

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

**Will you select a reference genome from your history or use a built-in index?:** Use a built-in index

Built-ins were indexed using default options

**Select a reference genome:** Amellifera_Honeybee apiMel3

If your genome of interest is not listed, contact the Galaxy team

**Is this library mate-paired?:** Single-end

**TopHat settings to use:** Use Defaults

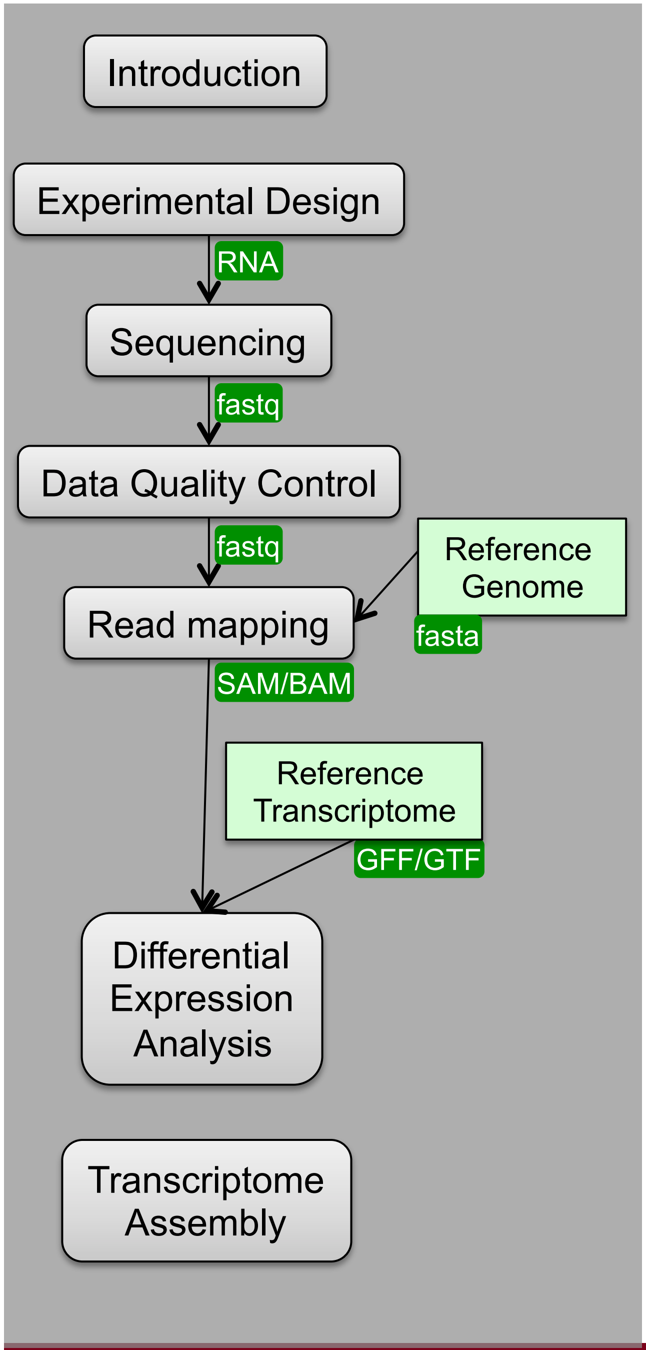
You can use the default settings or set custom values for any of Tophat's parameters.

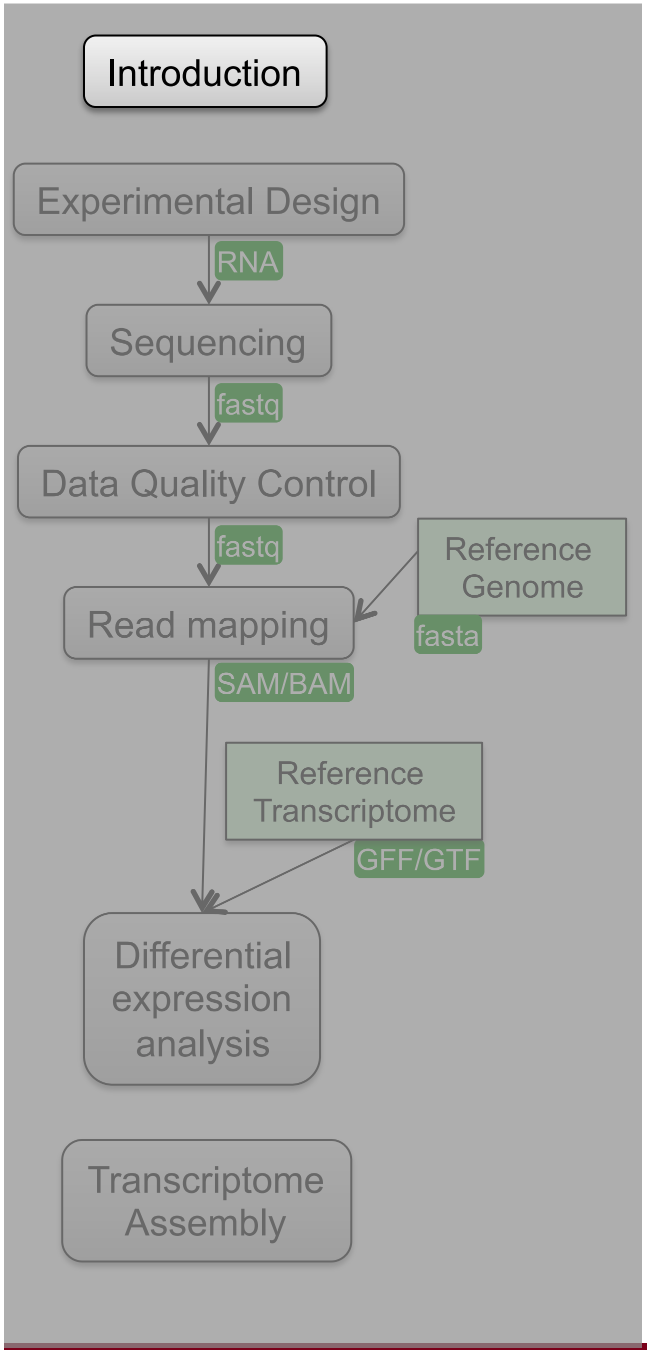
**Tophat Overview**

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. Please cite: Trapnell, C., Pachter, L. and Salzberg, S.L. TopHat: discovering splice junctions with RNA-

Web-base platform for bioinformatic analysis

# Outline

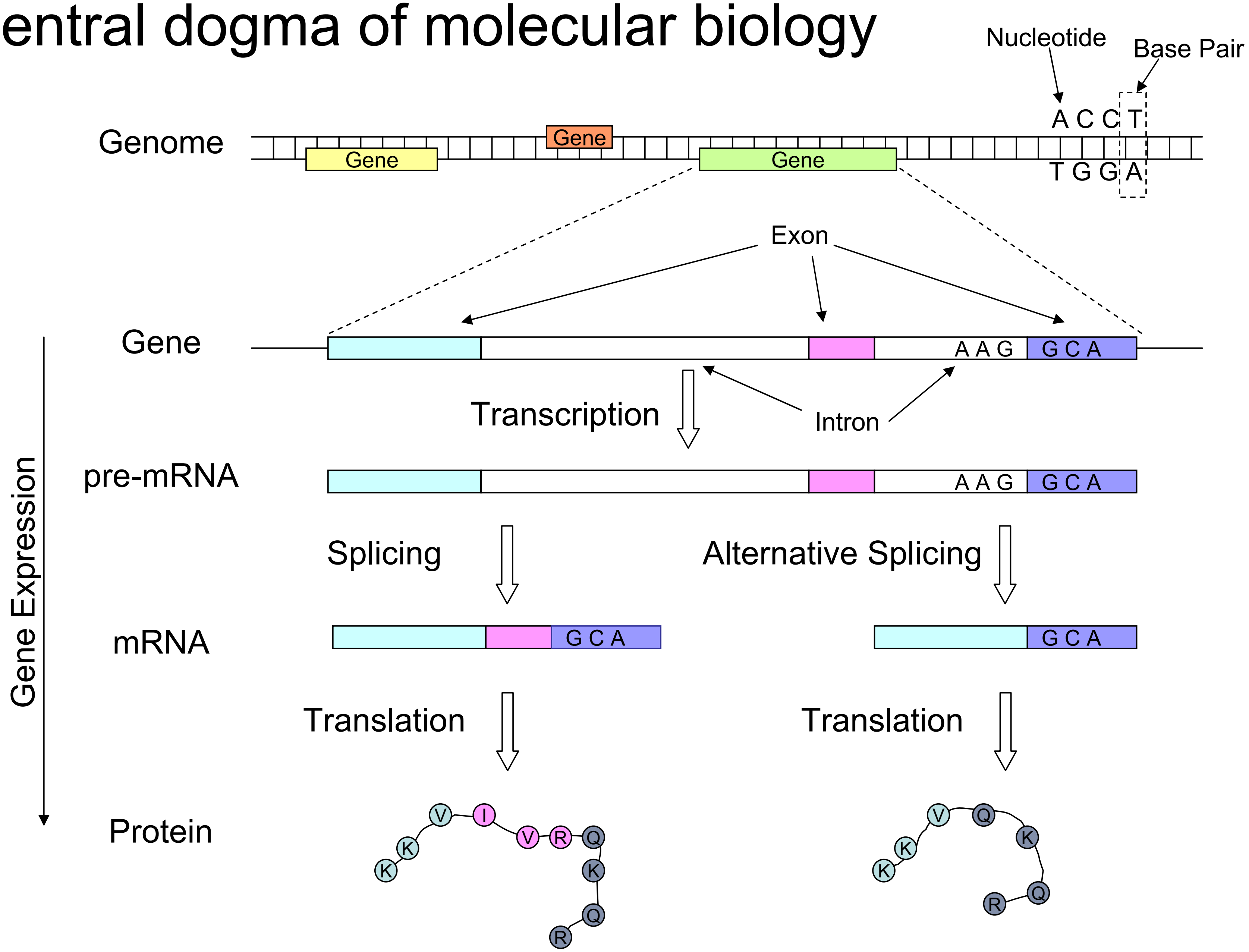

Introduction
Experimental Design
RNA
Sequencing
fastq
Data Quality Control
fastq
Reference Genome
fasta
Read mapping
SAM/BAM
Reference Transcriptome
GFF/GTF
Differential Expression Analysis
Transcriptome Assembly



# Introduction

- Gene expression
- RNA-Seq
- Platform characteristics
- Microarray comparison

# Central dogma of molecular biology



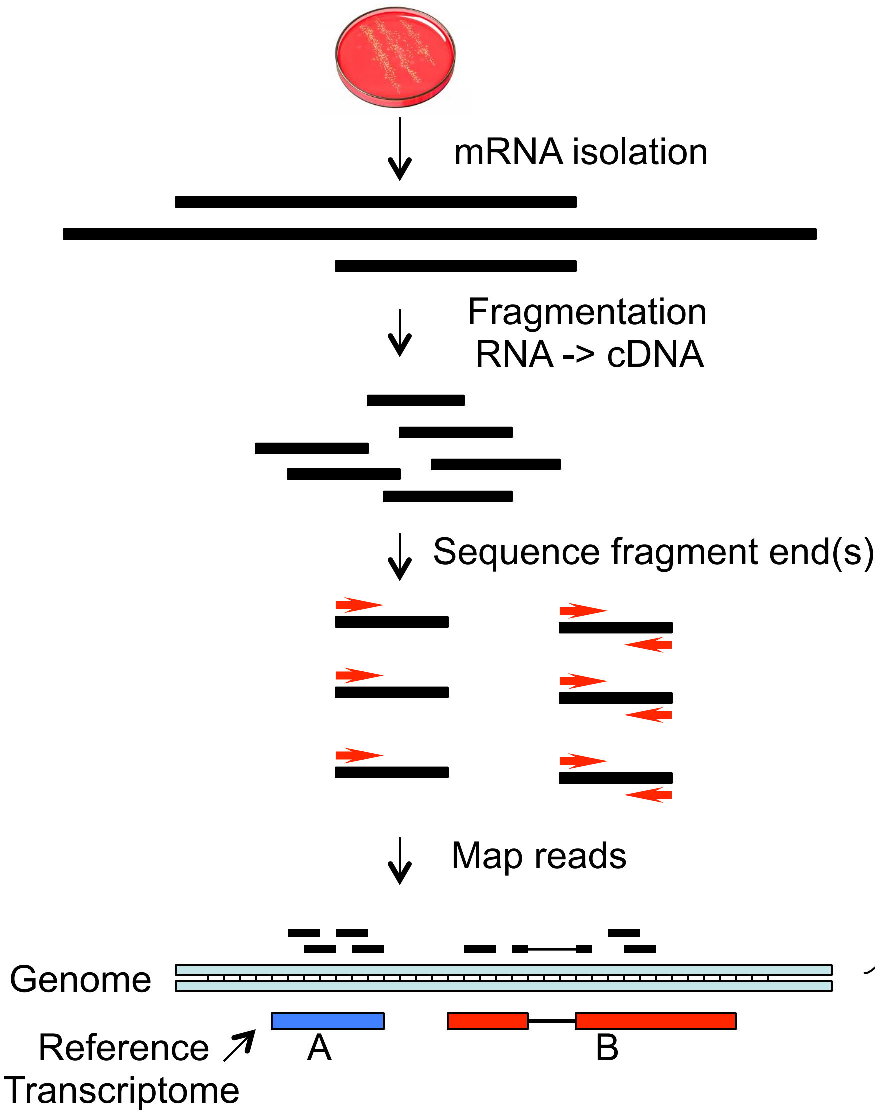
92–94% of human genes undergo alternative splicing,
86% with a minor isoform frequency of 15% or more

*E.T. Wang, et al, Nature 456, 470-476 (2008)*

# Introduction

- RNA-Seq
  - High-throughput sequencing of RNA
  - Transcriptome assembly
    - Qualitative identification of expressed sequence
  - Differential expression analysis
    - Quantitative measurement of transcript expression

## Sample 1



mRNA isolation

Fragmentation
RNA -> cDNA

Sequence fragment end(s)

Map reads

Genome

Reference
Transcriptome

Calculate transcript abundance

| | Gene A | Gene B |
|---|---|---|
| Sample 1 | 4 | 4 |

# of Reads

| | Gene A | Gene B |
|---|---|---|
| Sample 1 | 4 | 2 |

Reads per kilobase of exon

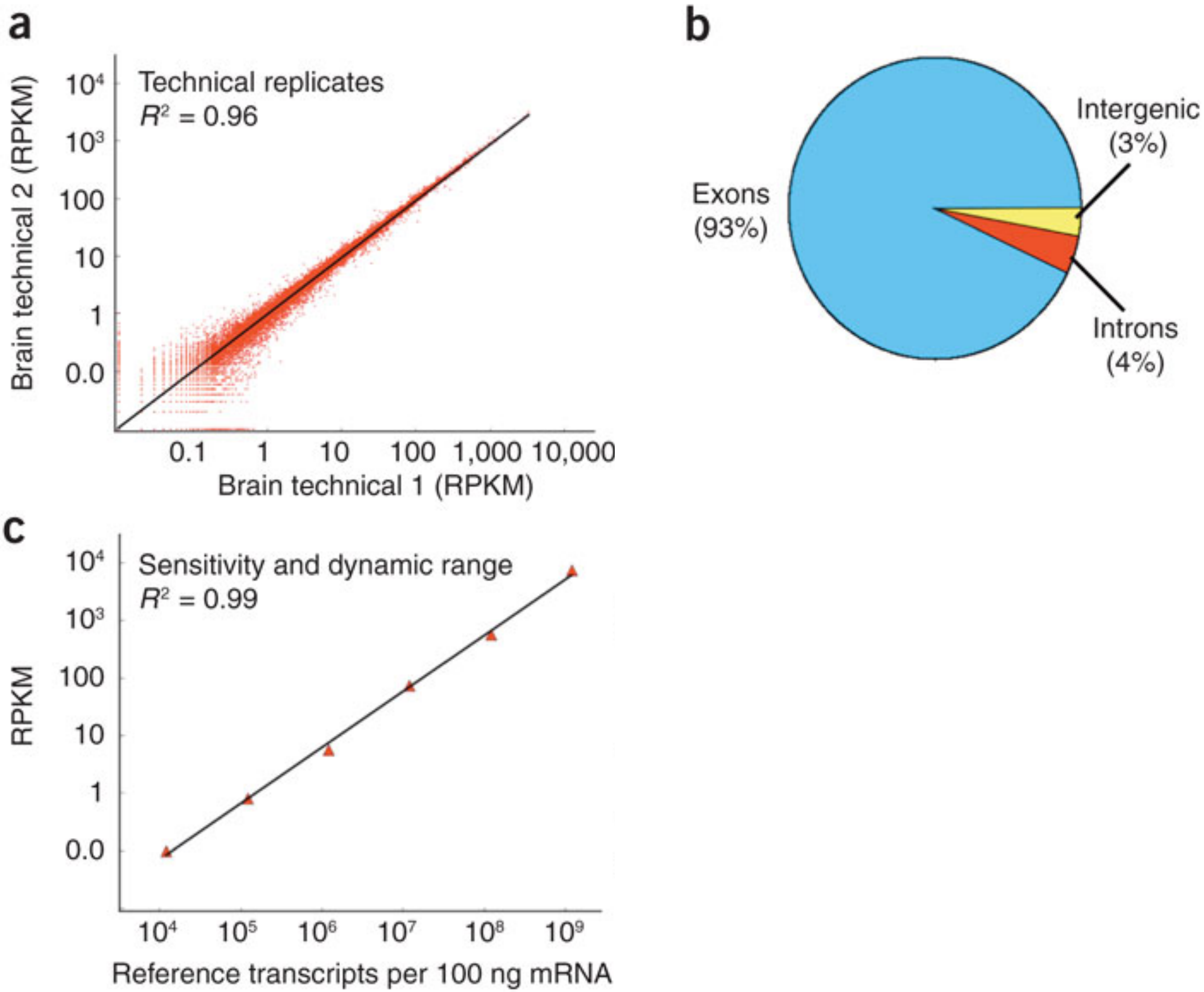
| | Gene A | Gene B | Total |
|---|---|---|---|
| Sample 1 | 4 | 2 | 6 |
| Sample 2 | 7 | 5 | 12 |

Reads per kilobase of exon

| | Gene A | Gene B | Total |
|---|---|---|---|
| Sample 1 | .7 | .3 | 6 |
| Sample 2 | .6 | .3 | 12 |

Reads per kilobase of exon per million mapped reads

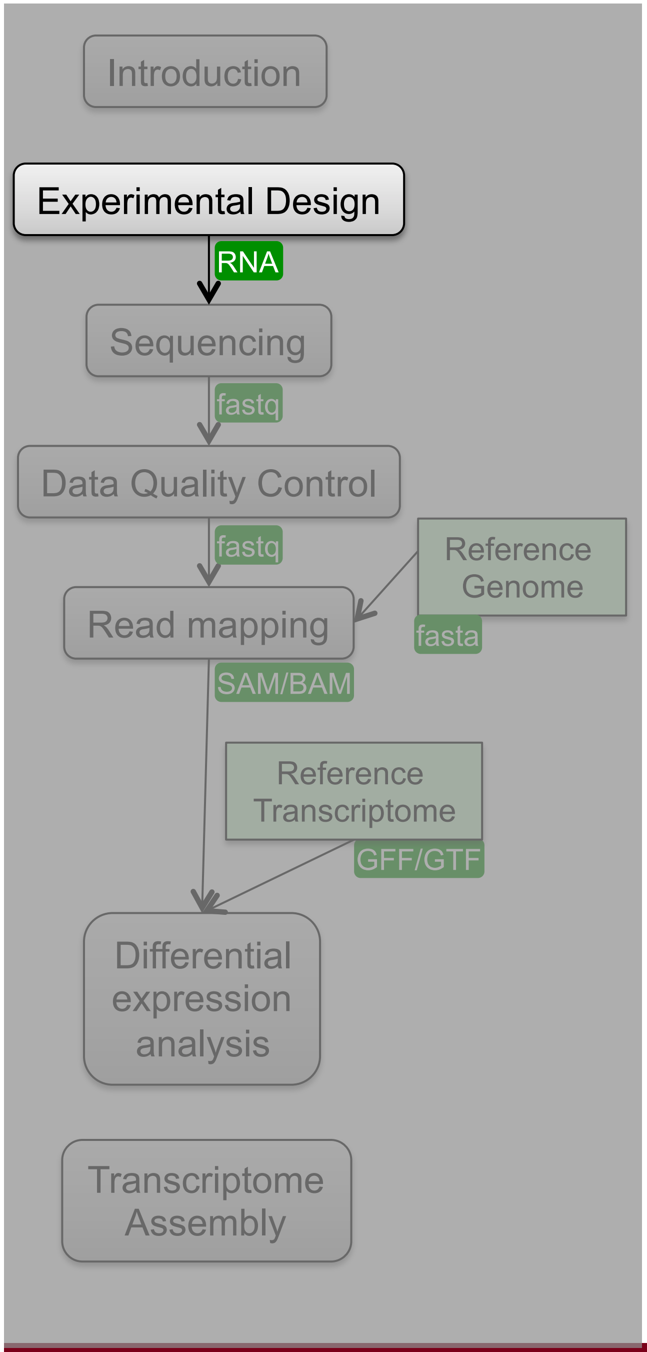
RPKM



Ali Mortazavi et al., *Nature Methods* - **5**, 621 - 628 (2008)

# Introduction

- RNA-Seq (vs Microarray)
  - Strong concordance between platforms
  - Higher sensitivity and dynamic range
  - Lower technical variation
  - Available for all species
  - Novel transcribed regions
  - Alternative splicing
  - Allele-specific expression
  - Fusion genes
  - Higher informatics cost

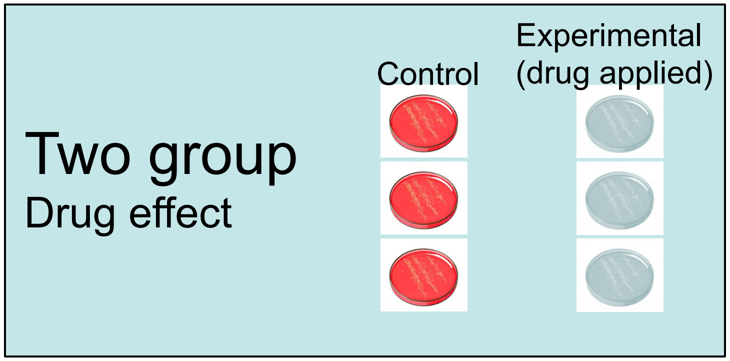


# Experimental Design

- Biological comparison(s)
- Paired-end vs single end reads
- Read length
- Read depth
- Replicates
- Pooling

# Experimental design

- Simple designs (Pairwise comparisons)

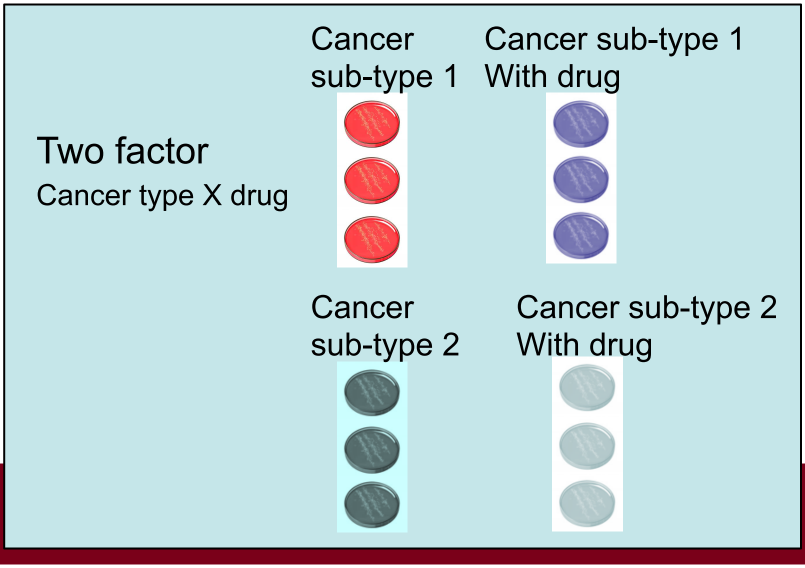


Two group
Drug effect

Control

Experimental
(drug applied)

- Complex designs – ⚠ Consult a statistician



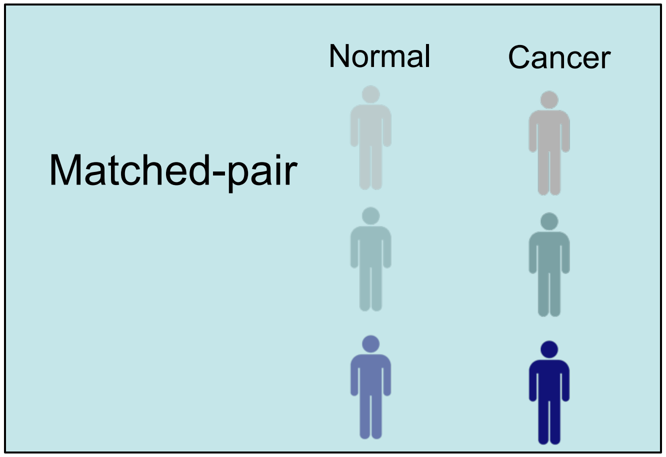
Two factor
Cancer type X drug

Cancer
sub-type 1

Cancer sub-type 1
With drug

Cancer
sub-type 2

Cancer sub-type 2
With drug



Matched-pair

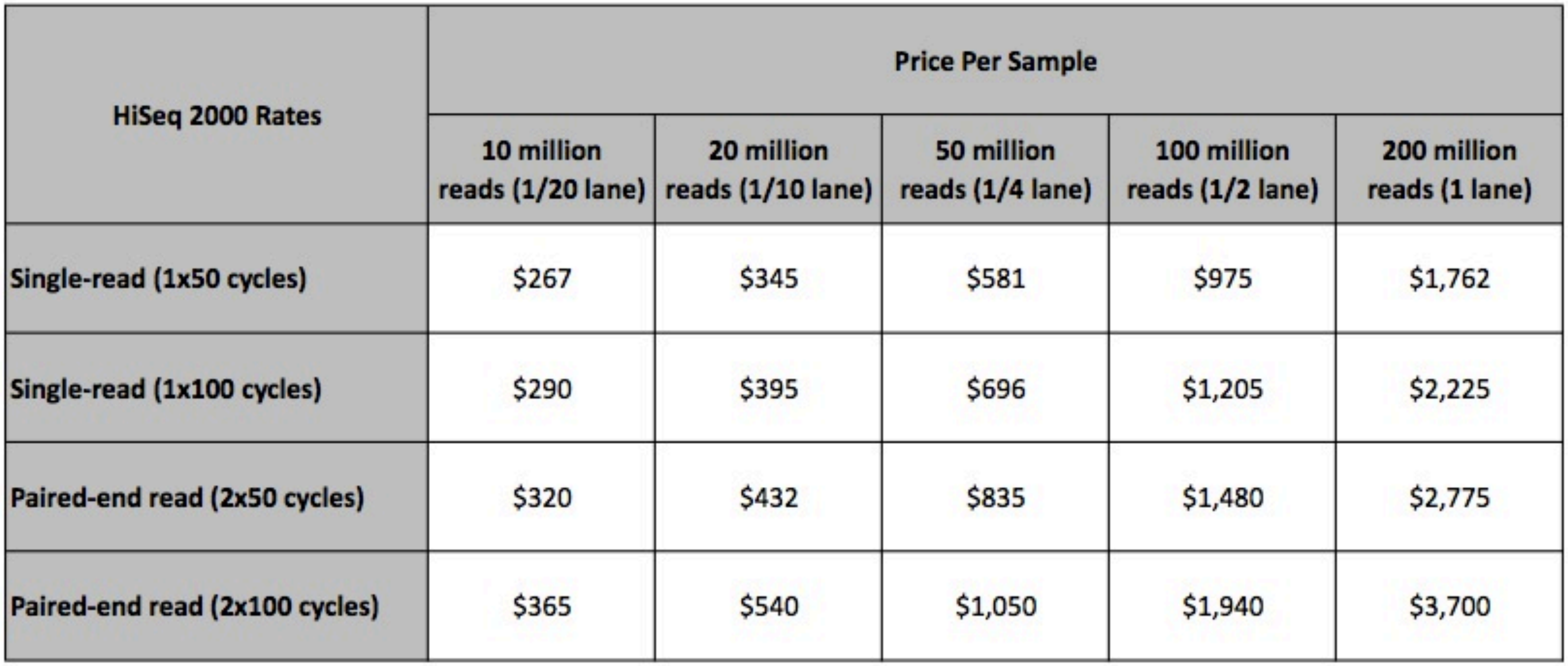
Normal

Cancer

# Experimental design

- What are my goals?
  - Transcriptome assembly?
  - Differential expression analysis?
  - Identify rare transcripts?
- What are the characteristics of my system?
  - Large, complex genome?
  - Introns and high degree of alternative splicing?
  - No reference genome or transcriptome?

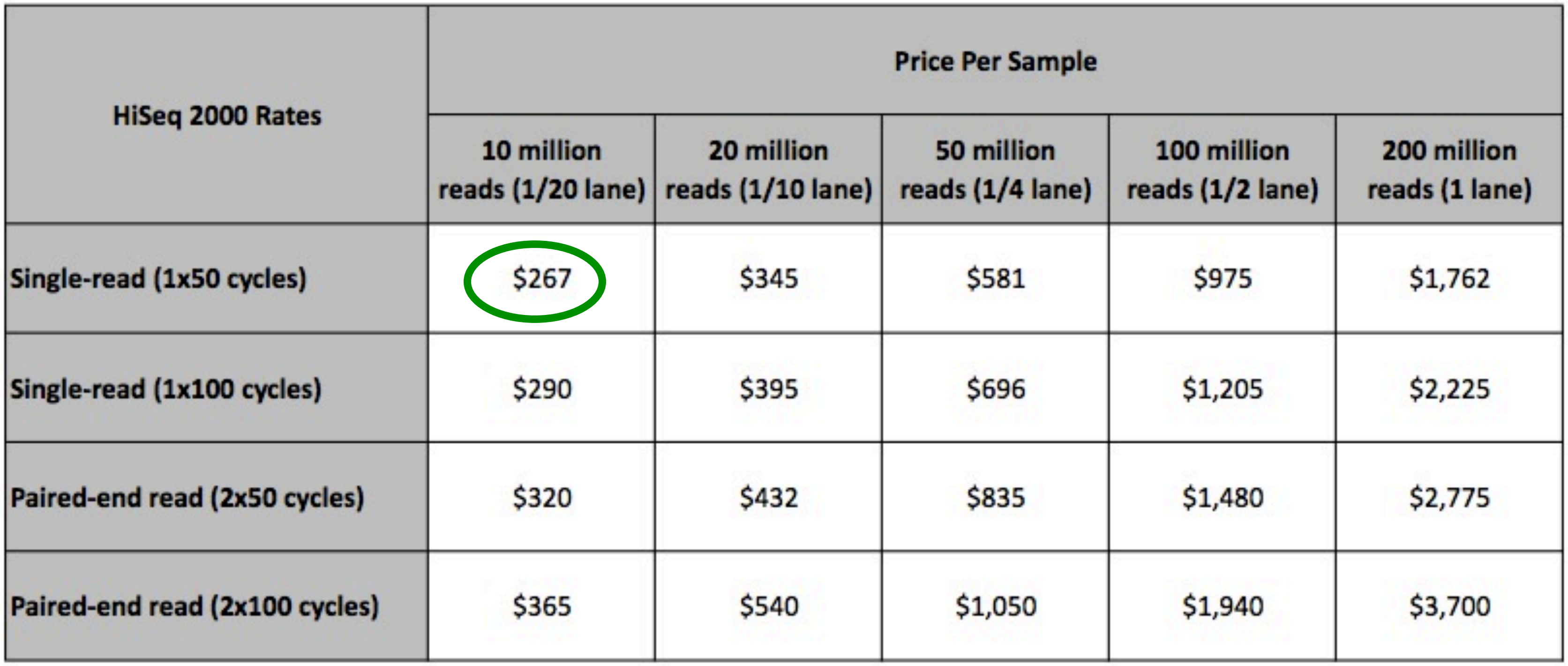
# Experimental design



| HiSeq 2000 Rates | Price Per Sample | | | | |
|---|---|---|---|---|---|
| | 10 million reads (1/20 lane) | 20 million reads (1/10 lane) | 50 million reads (1/4 lane) | 100 million reads (1/2 lane) | 200 million reads (1 lane) |
| Single-read (1x50 cycles) | $267 | $345 | $581 | $975 | $1,762 |
| Single-read (1x100 cycles) | $290 | $395 | $696 | $1,205 | $2,225 |
| Paired-end read (2x50 cycles) | $320 | $432 | $835 | $1,480 | $2,775 |
| Paired-end read (2x100 cycles) | $365 | $540 | $1,050 | $1,940 | $3,700 |

BMGC RNA-Seq Price list (Jan 2012)

# Experimental design

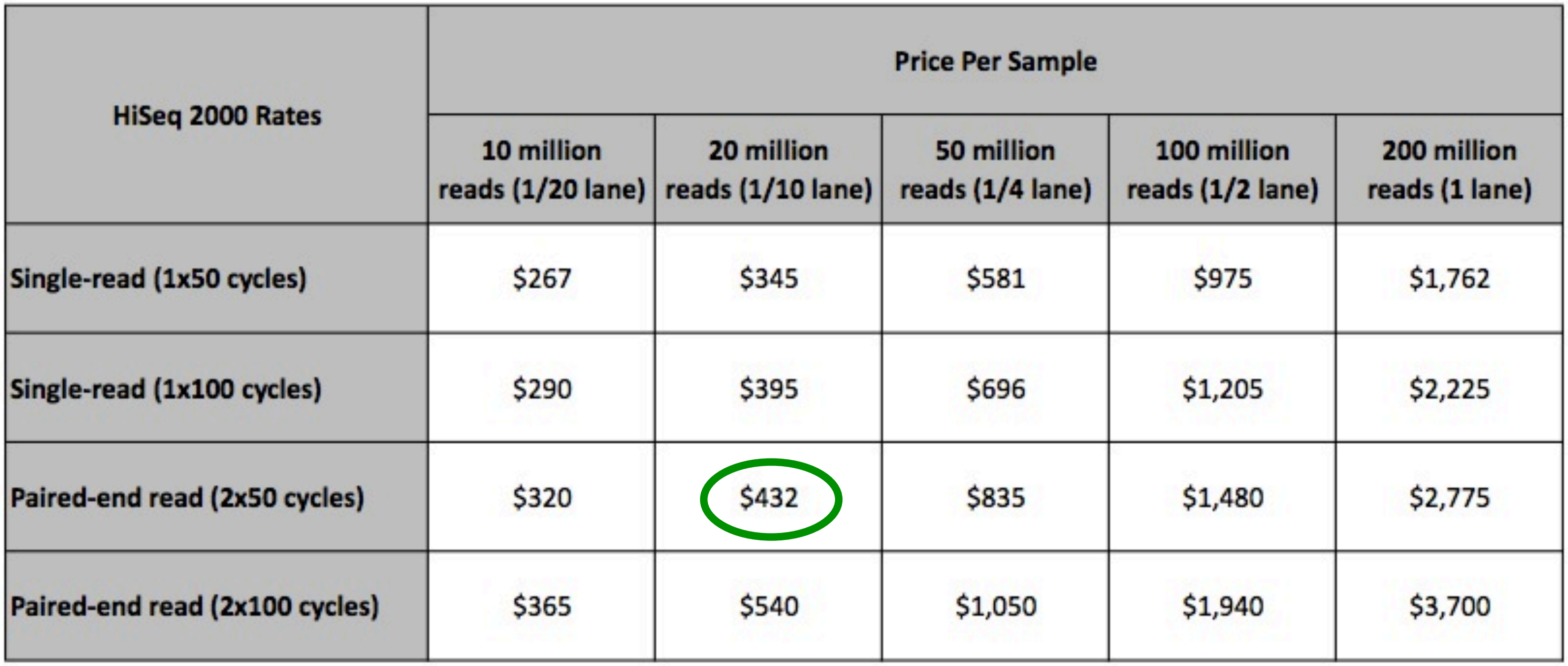


| HiSeq 2000 Rates | Price Per Sample | | | | |
|---|---|---|---|---|---|
| | 10 million reads (1/20 lane) | 20 million reads (1/10 lane) | 50 million reads (1/4 lane) | 100 million reads (1/2 lane) | 200 million reads (1 lane) |
| Single-read (1x50 cycles) | $267 | $345 | $581 | $975 | $1,762 |
| Single-read (1x100 cycles) | $290 | $395 | $696 | $1,205 | $2,225 |
| Paired-end read (2x50 cycles) | $320 | $432 | $835 | $1,480 | $2,775 |
| Paired-end read (2x100 cycles) | $365 | $540 | $1,050 | $1,940 | $3,700 |

10 million reads per sample, 50bp single-end reads

- Small genomes with no alternative splicing

# Experimental design

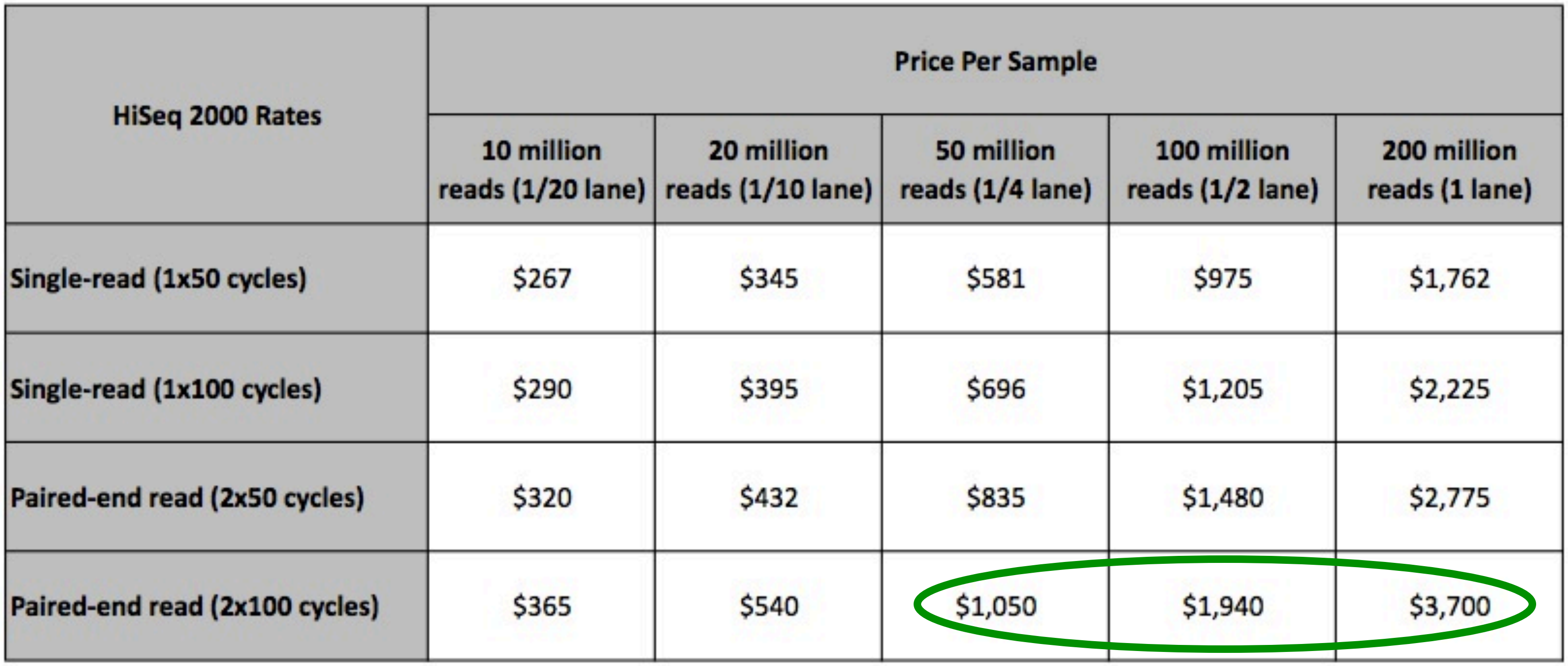


| HiSeq 2000 Rates | Price Per Sample | | | | |
|---|---|---|---|---|---|
| | 10 million reads (1/20 lane) | 20 million reads (1/10 lane) | 50 million reads (1/4 lane) | 100 million reads (1/2 lane) | 200 million reads (1 lane) |
| Single-read (1x50 cycles) | $267 | $345 | $581 | $975 | $1,762 |
| Single-read (1x100 cycles) | $290 | $395 | $696 | $1,205 | $2,225 |
| Paired-end read (2x50 cycles) | $320 | [✓ $432] | $835 | $1,480 | $2,775 |
| Paired-end read (2x100 cycles) | $365 | $540 | $1,050 | $1,940 | $3,700 |

20 million reads per sample, 50bp paired-end reads

- Mammalian genomes (large transcriptome, alternative splicing, gene duplication)

# Experimental design



| HiSeq 2000 Rates | Price Per Sample | | | | |
|---|---|---|---|---|---|
| | 10 million reads (1/20 lane) | 20 million reads (1/10 lane) | 50 million reads (1/4 lane) | 100 million reads (1/2 lane) | 200 million reads (1 lane) |
| Single-read (1x50 cycles) | $267 | $345 | $581 | $975 | $1,762 |
| Single-read (1x100 cycles) | $290 | $395 | $696 | $1,205 | $2,225 |
| Paired-end read (2x50 cycles) | $320 | $432 | $835 | $1,480 | $2,775 |
| Paired-end read (2x100 cycles) | $365 | $540 | $1,050 | $1,940 | $3,700 |

50-200 million reads per sample, 100bp paired-end reads

- Transcriptome Assembly (100X coverage of transcriptome)

50bp Paired-end >> 100bp Single-end

# Experimental design

- Technical replicates
  - Not needed: low technical variation
    - Minimize batch effects
    - Randomize sample order ⚠
- Biological replicates
  - Not needed for transcriptome assembly
  - Essential for differential expression analysis
  - Difficult to estimate
    - 3+ for cell lines
    - 5+ for inbred lines
    - 20+ for human samples

# Experimental design

- Pooling samples
  - Limited RNA obtainable
    - Multiple pools per group required
  - Transcriptome assembly

# Experimental design

## RNA-seq: technical variability and sampling

Lauren M McIntyre, Kenneth K Lopiano, Alison M Morse, Victor Amin, Ann L Oberg, Linda J Young and Sergey V Nuzhdin

BMC Genomics 2011, 12:293

## Statistical Design and Analysis of RNA Sequencing Data

Paul L. Auer and R. W. Doerge

Genetics. 2010 June; 185(2): 405–416.

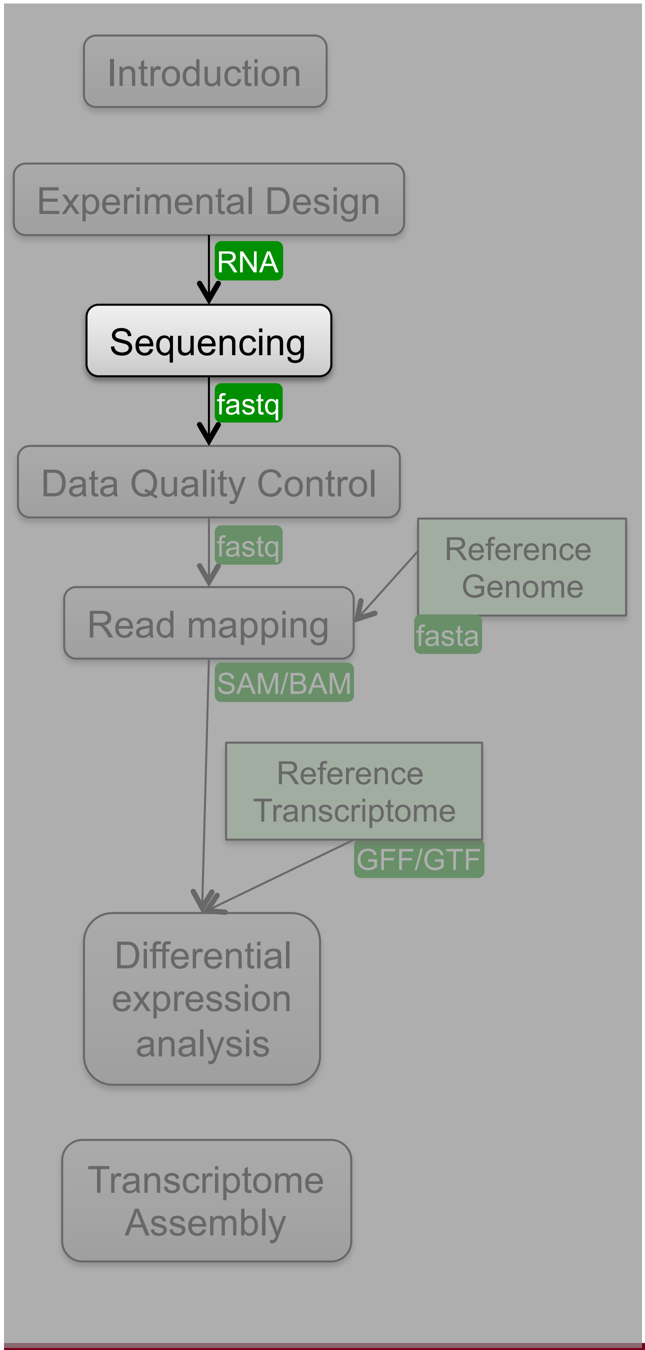
## Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries

Daniel Aird, Michael G Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B Jaffe, Chad Nusbaum and Andreas Gnirke

Genome Biology 2011, 12:R18

## ENCODE RNA-Seq guidelines

http://www.encodeproject.org/ENCODE/experiment_guidelines.html



# Sequencing

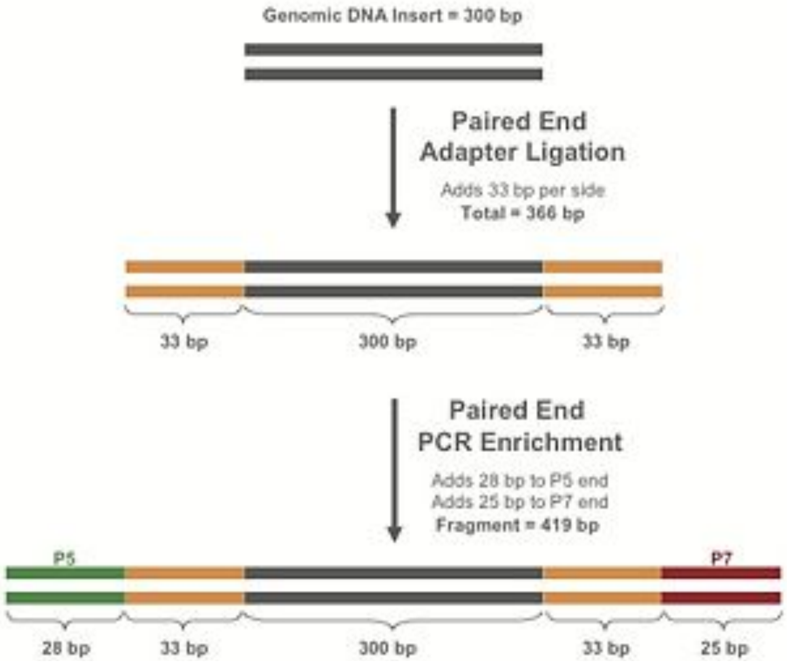
- Platforms
- Library preparation
- Multiplexing
- Sequence reads

# Sequencing

- Illumina sequencing by synthesis
  - GAIIx
    - replaced by HiSeq
  - HiSeq2000
  - MiSeq
    - low throughput, fast turnaround
- SOLiD (not available at BMGC)
  - "Color-space" reads (require special mapping software)
  - Low error rate
- 454 pyrosequencing
  - Longer reads, lower throughput

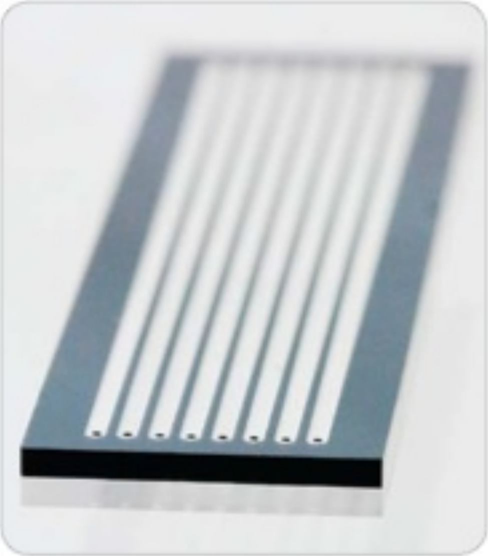
# Sequencing

- Library preparation (Illumina TruSeq protocol for HiSeq)
  - RNA isolation
  - Ploy-A purification
  - Fragmentation
  - cDNA synthesis using random primers
  - Adapter ligation
  - Size selection
  - PCR amplification

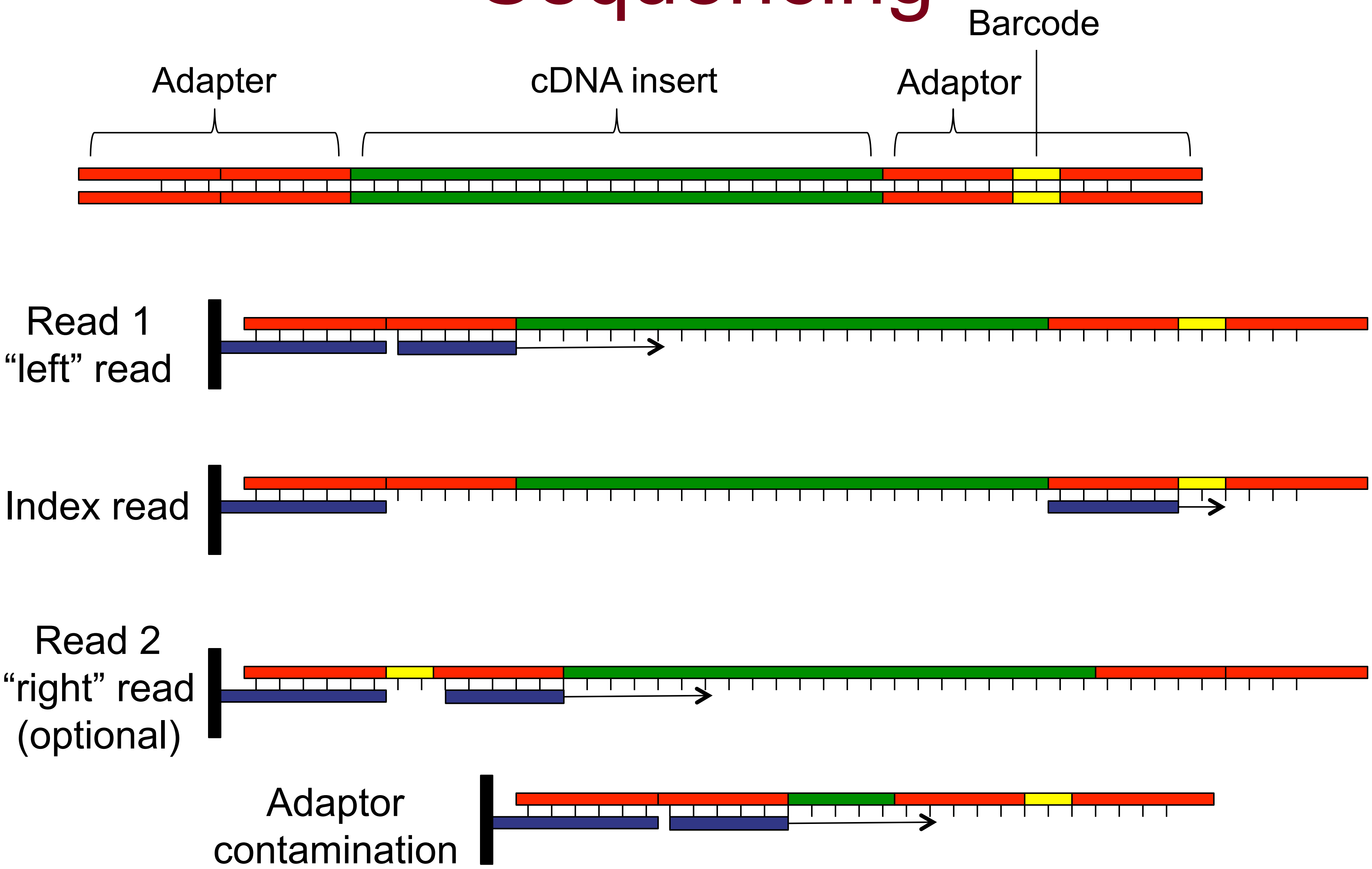


# Sequencing

- Flowcell
  - 8 lanes
  - 200 Million reads per lane
  - Multiplex up to 24 samples on one lane using barcodes



# Sequencing



Barcode

Adapter

cDNA insert

Adaptor

Read 1
“left” read

Index read

Read 2
“right” read
(optional)

Adaptor
contamination

# Sequencing

- Library types
  - Polyadenylated RNA > 200bp (standard method)
  - Total RNA
  - Small RNA
  - Strand-specific
    - Gene-dense genomes (bacteria, archaea, lower eukaryotes)
    - Antisense transcription (higher eukaryotes)
  - Low input
  - Library capture

# Sequence Data Format

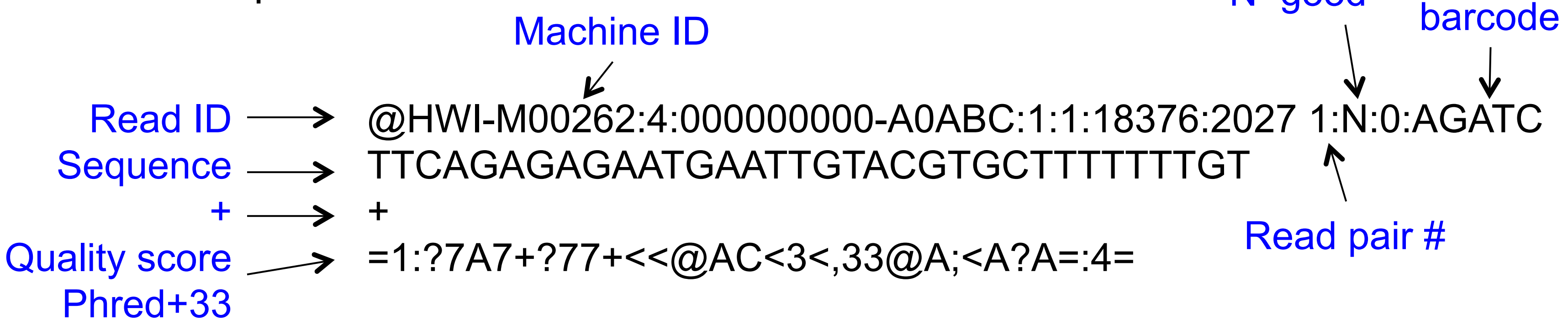
- Data delivery
  - /project/PI-groupname/120318_SN261_0348_A81JUMABXX
    - fastq_flt/ Bad reads removed by Illumina software, for use in data analysis
    - fastq/ Raw sequence output for submission to public archives, contains bad reads
  - Upload to Galaxy

⚠ Don't use in analysis

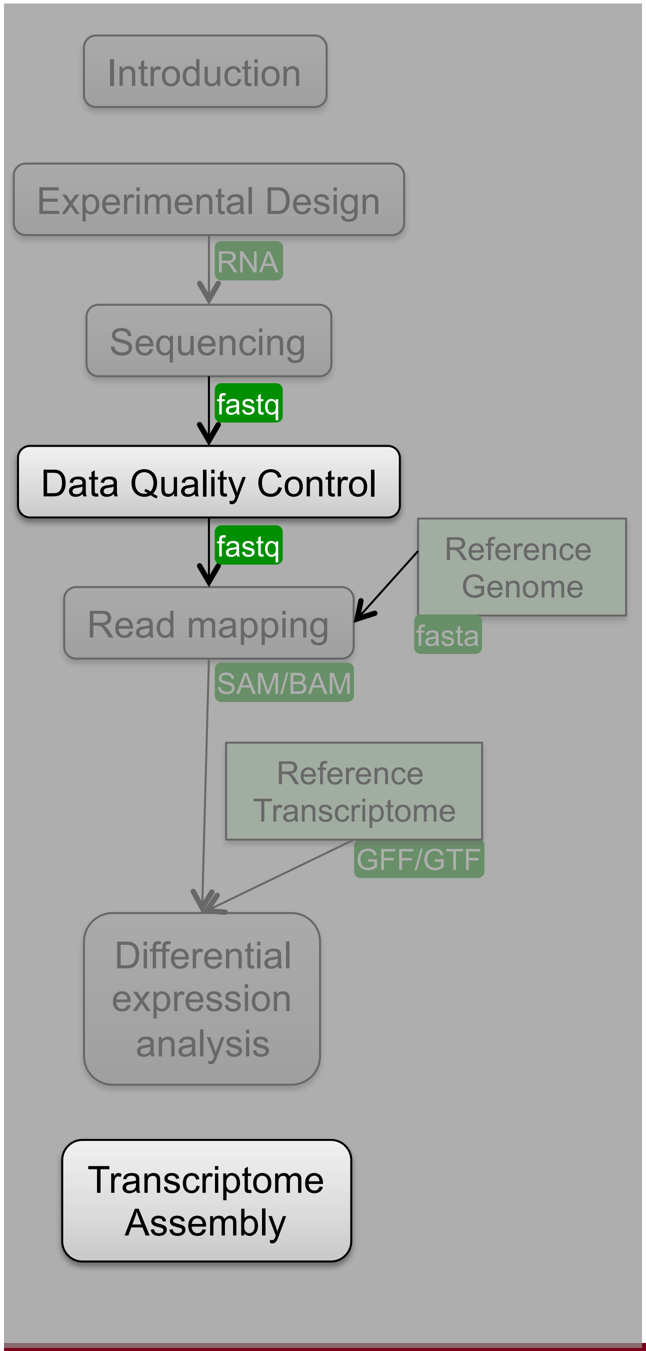
- File names
  - L1_R1_CCAAT_cancer1.fastq
  - L1_R2_CCAAT_cancer1.fastq
- Fastq format (Illumina Casava 1.8.0)
  - 4 lines per read

⚠ Formats vary

QC Filter flag
Y=bad
N=good



```
@HWI-M00262:4:000000000-A0ABC:1:1:18376:2027 1:N:0:AGATC
TTCAGAGAGAATGAATTGTACGTGCTTTTTTTGT
+
=1:?7A7+?77+<<@AC<3<,33@A;<A?A=:4=
```



# Data Quality Control

- Quality assessment
- Trimming and filtering

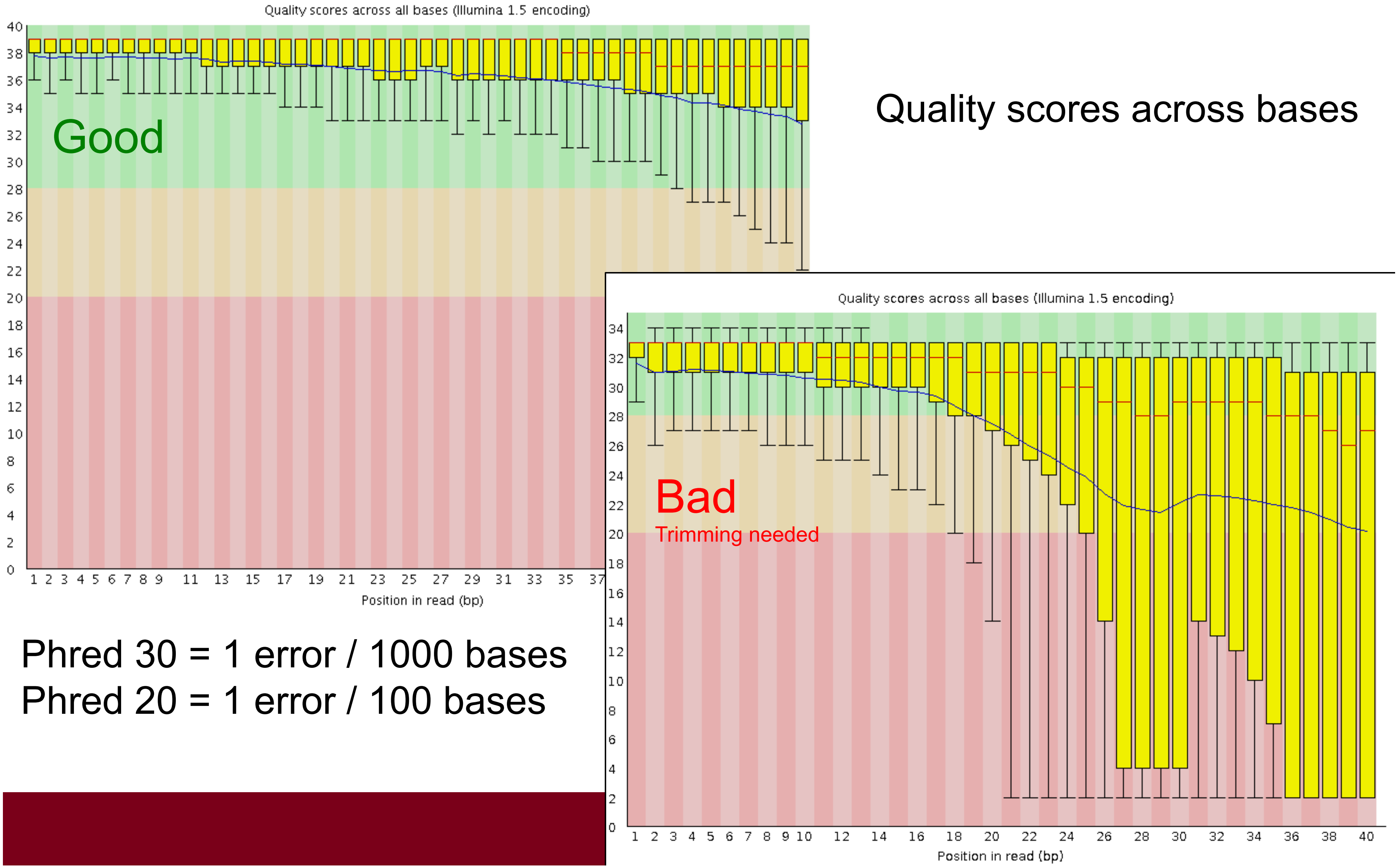
# Data Quality Assessment

- Evaluate read library quality
  - Identify contaminants
  - Identify poor/bad samples
- Software
  - FastQC (recommended)
    - Command-line, Java GUI, or Galaxy
  - SolexaQC
    - Command-line
    - Supports quality-based read trimming and filtering
  - SAMStat
    - Command-line
    - Also works with bam alignment files

# Data Quality Assessment

- Trimming: remove bad bases from (end of) read
  - Adaptor sequence
  - Low quality bases
- Filtering: remove bad reads from library
  - Low quality reads
  - Contaminating sequence
  - Low complexity reads (repeats)
  - Short reads
    - Short (< 20bp) reads slow down mapping software
    - Only needed if trimming was performed
- Software
  - Galaxy, many options (NGS: QC and manipulation)
  - Tagdust
  - Many others: http://seqanswers.com/wiki/Software/list

# Data Quality Assessment - FastQC

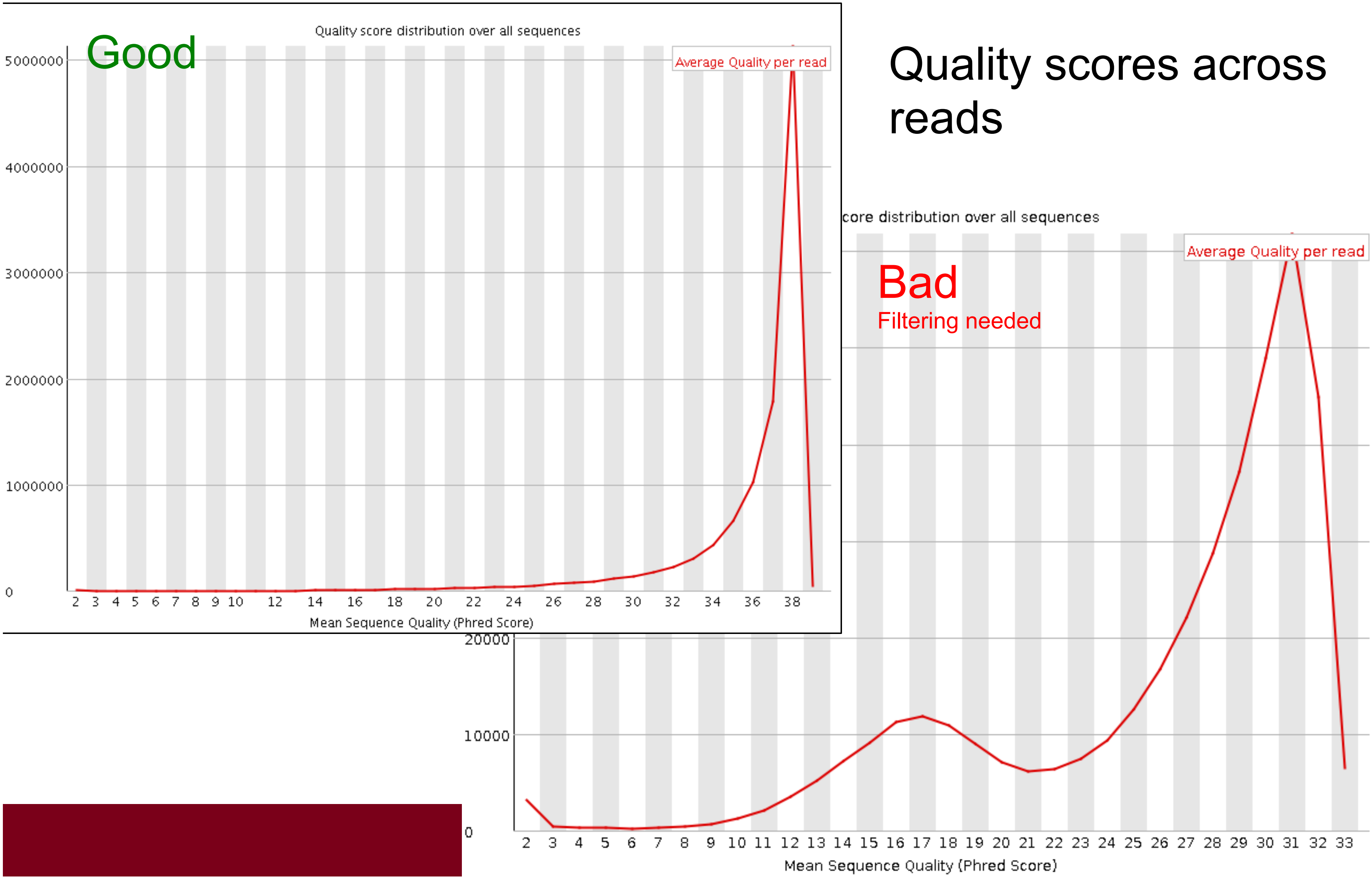


Quality scores across bases

Phred 30 = 1 error / 1000 bases
Phred 20 = 1 error / 100 bases

# Data Quality Assessment - FastQC

Quality scores across reads



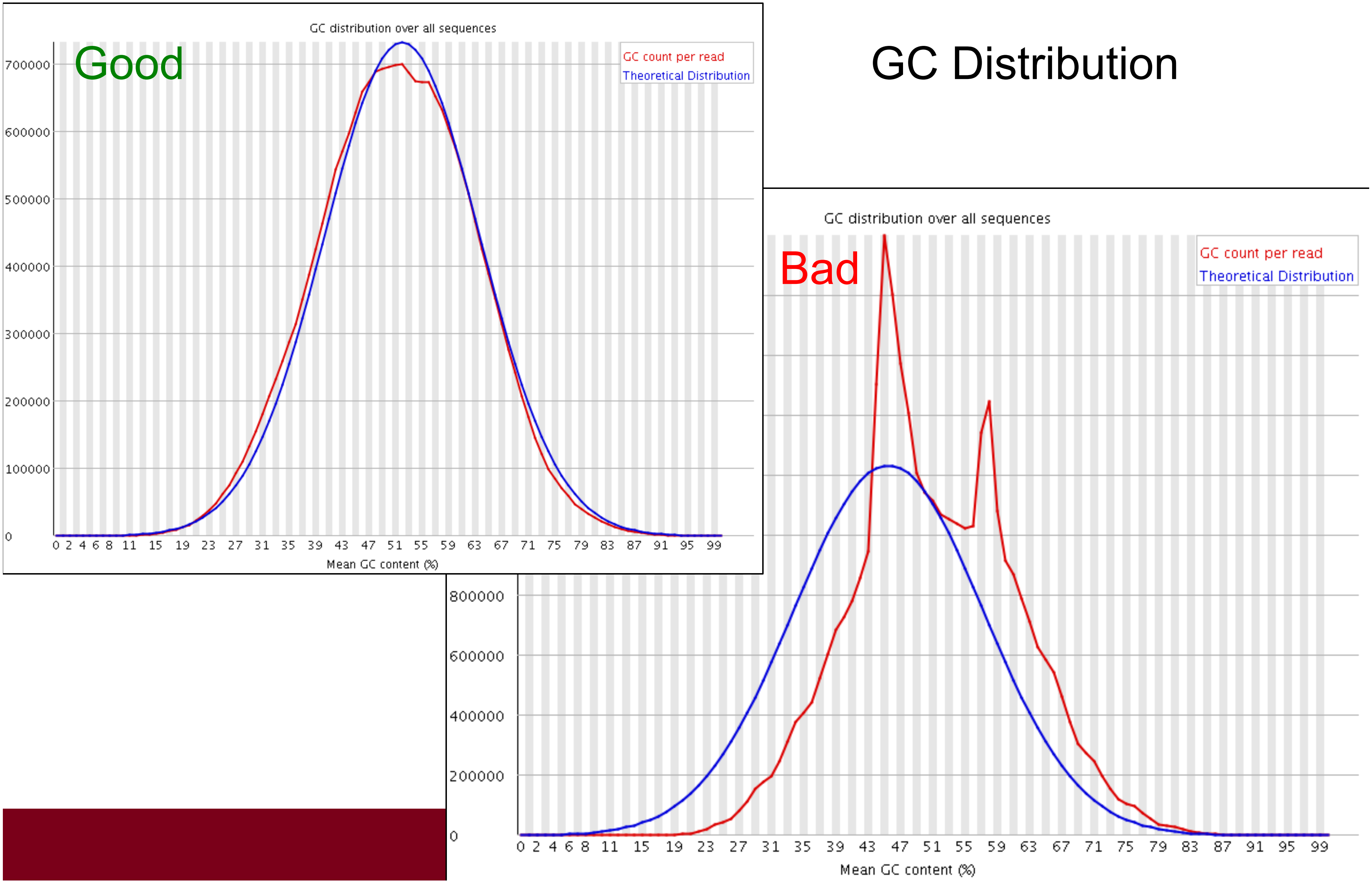
Good

Bad

Filtering needed

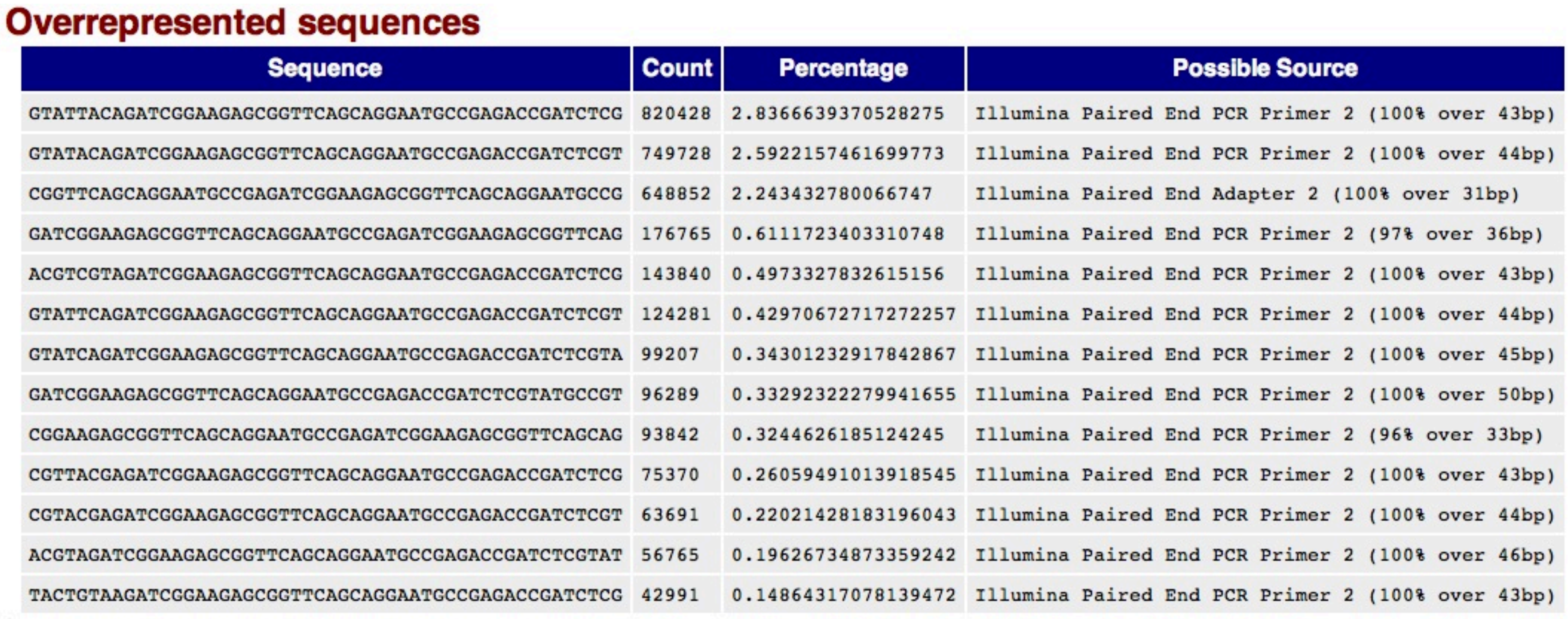
# Data Quality Assessment - FastQC

GC Distribution



# Data Quality Assessment - FastQC

High level of sequencing adapter contamination, trimming needed

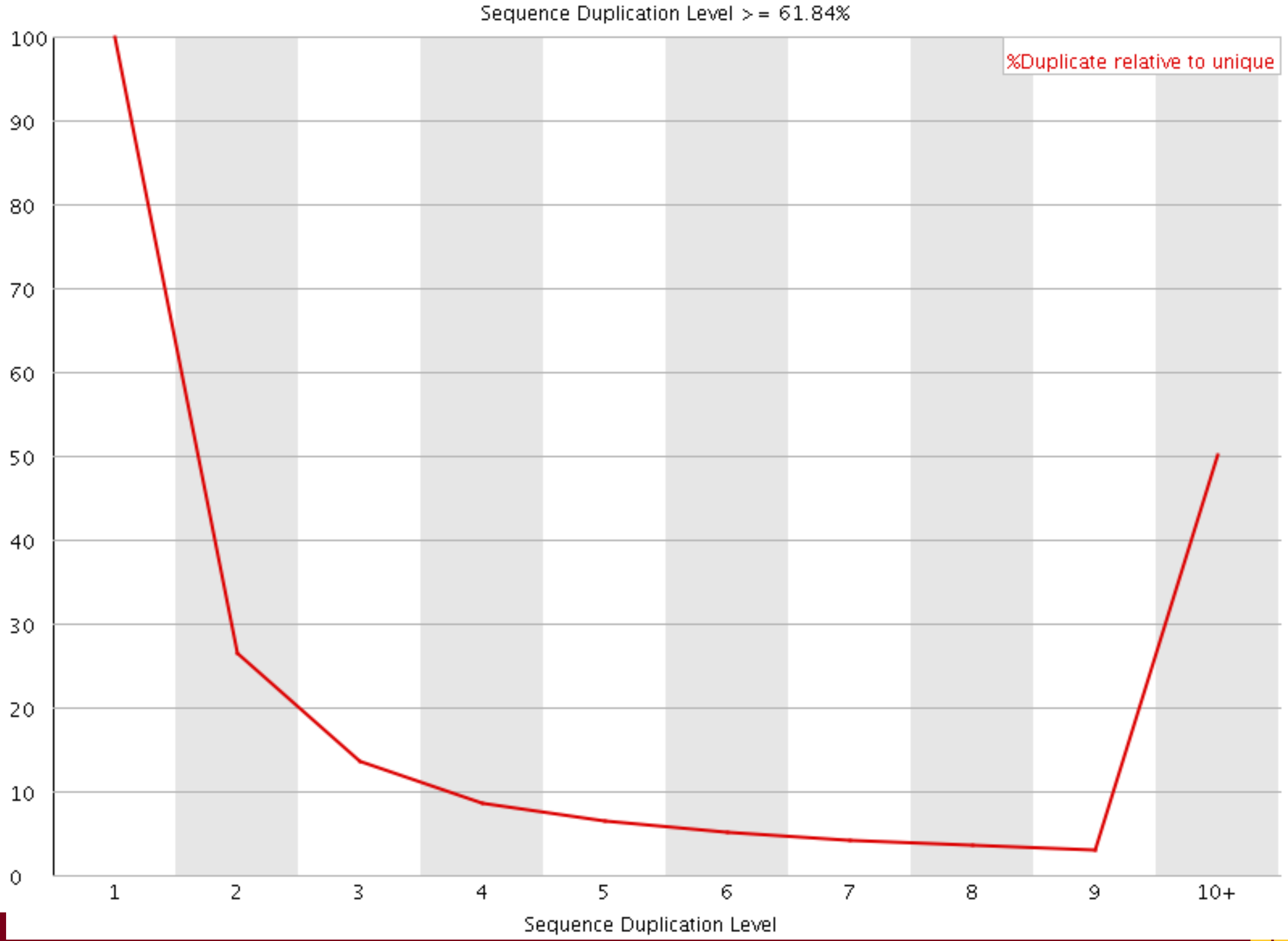


## Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| GTATTACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCG | 820428 | 2.8366639370528275 | Illumina Paired End PCR Primer 2 (100% over 43bp) |
| GTATACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGT | 749728 | 2.592215746169973 | Illumina Paired End PCR Primer 2 (100% over 44bp) |
| CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGCAGGAATGCCG | 648852 | 2.243432780066747 | Illumina Paired End Adapter 2 (100% over 31bp) |
| GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAG | 176765 | 0.6111723403310748 | Illumina Paired End PCR Primer 2 (97% over 36bp) |
| ACGTCGTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCG | 143840 | 0.4973327832615156 | Illumina Paired End PCR Primer 2 (100% over 43bp) |
| GTATTCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGT | 124281 | 0.42970672717272257 | Illumina Paired End PCR Primer 2 (100% over 44bp) |
| GTATCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTA | 99207 | 0.34301232917842867 | Illumina Paired End PCR Primer 2 (100% over 45bp) |
| GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGT | 96289 | 0.33292322279941655 | Illumina Paired End PCR Primer 2 (100% over 50bp) |
| CGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGCAG | 93842 | 0.3244626185124245 | Illumina Paired End PCR Primer 2 (96% over 33bp) |
| CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCG | 75370 | 0.26059491013918545 | Illumina Paired End PCR Primer 2 (100% over 43bp) |
| CGTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGT | 63691 | 0.22021428183196043 | Illumina Paired End PCR Primer 2 (100% over 44bp) |
| ACGTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTAT | 56765 | 0.19626734873359242 | Illumina Paired End PCR Primer 2 (100% over 46bp) |
| TACTGTAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCG | 42991 | 0.14864317078139472 | Illumina Paired End PCR Primer 2 (100% over 43bp) |

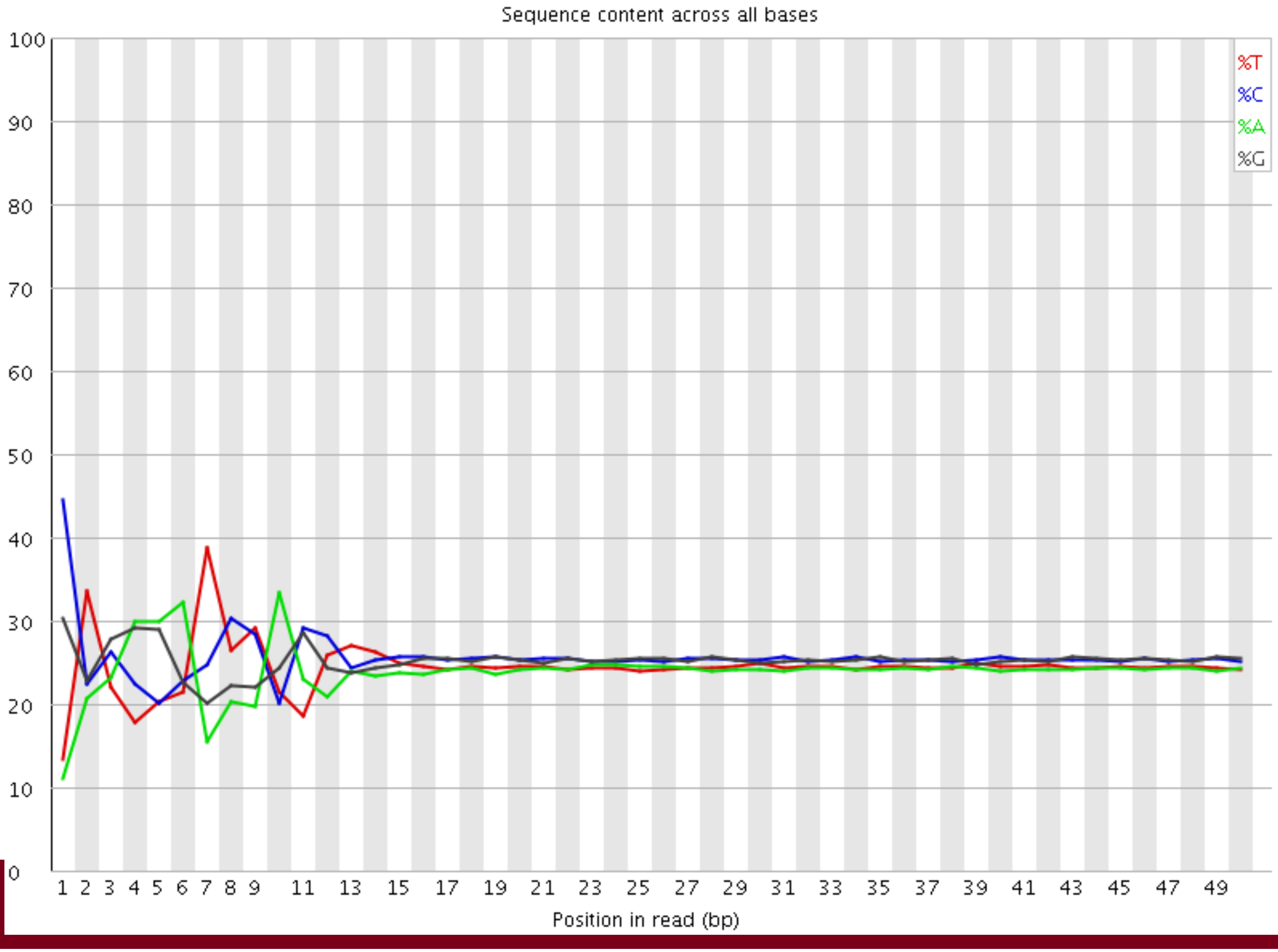
# Data Quality Assessment - FastQC

Normal level of sequence duplication in 20 million read mammalian sample



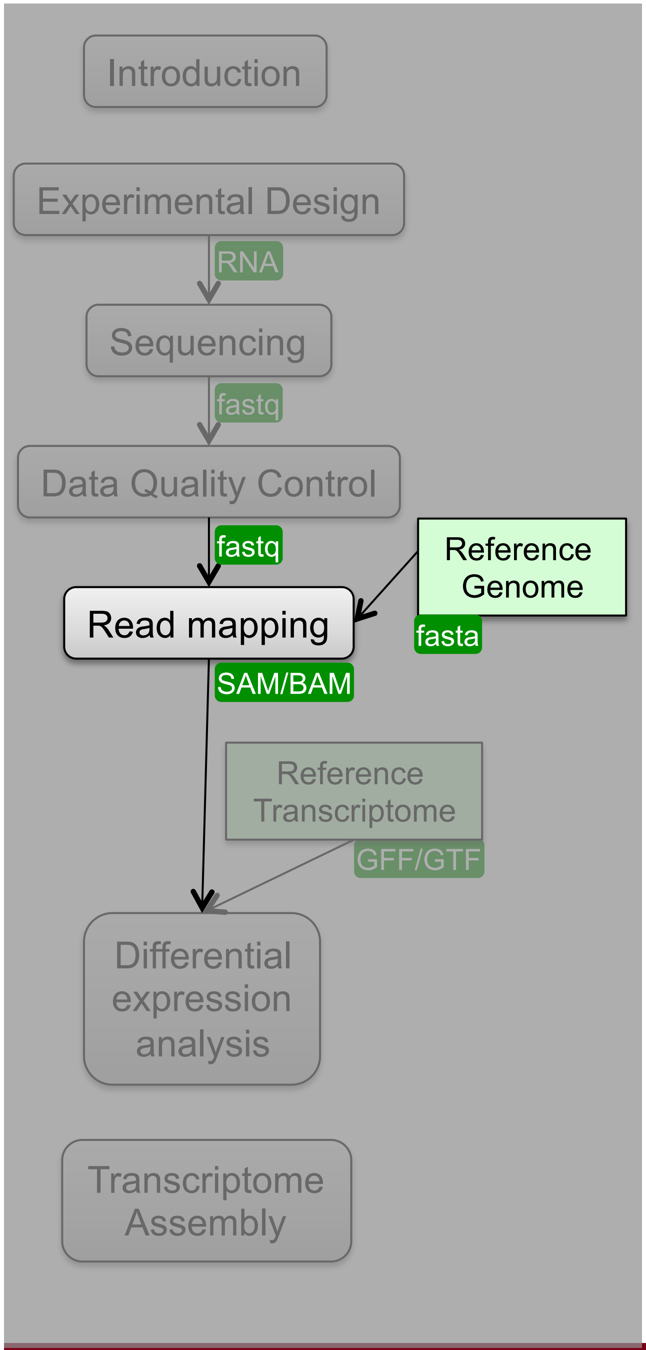
# Data Quality Assessment - FastQC

Normal sequence bias at beginning of reads due to non-random hybridization of random primers



# Data Quality Assessment

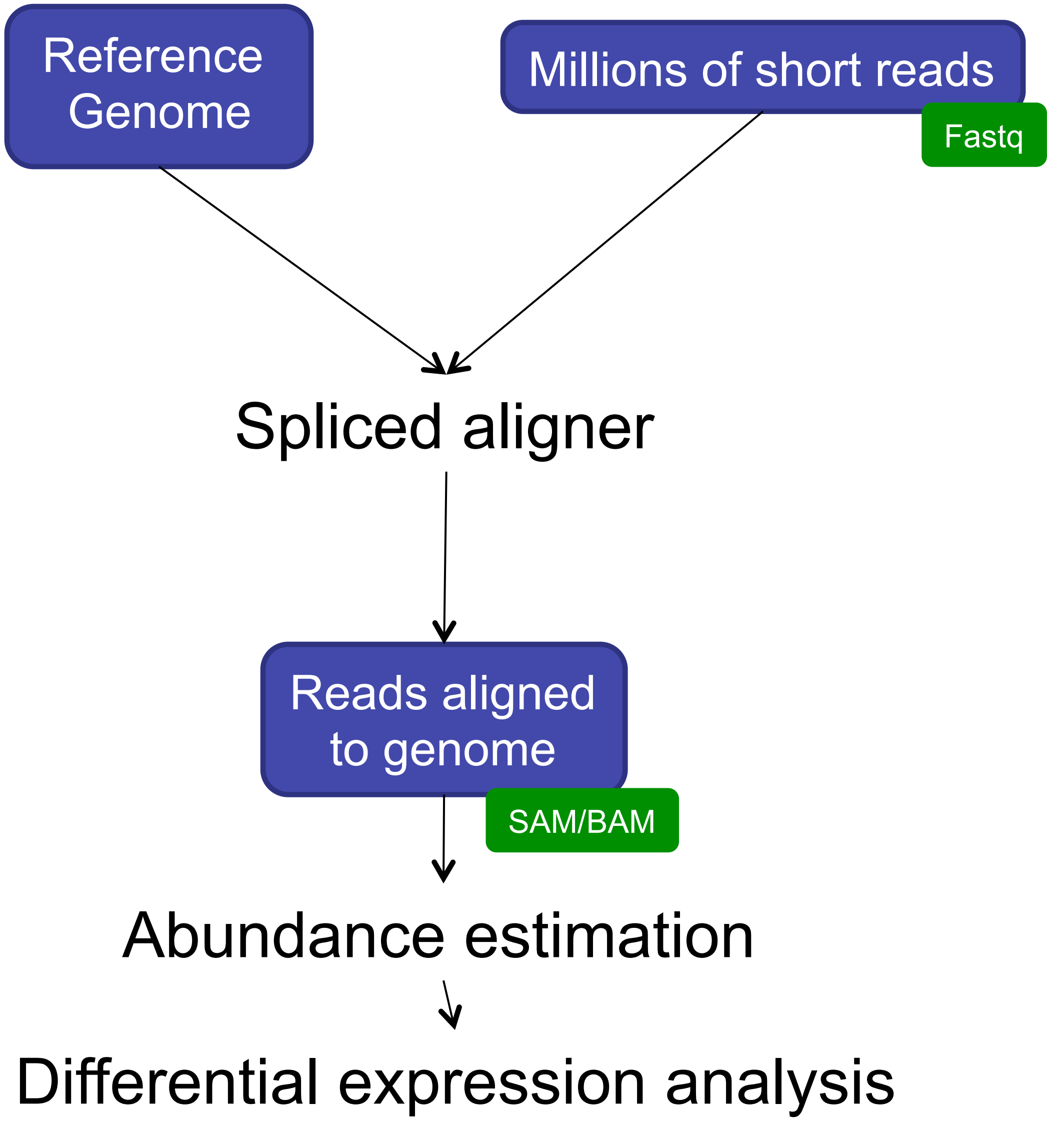
- Recommendations
  - Generate quality plots for all read libraries
  - Trim and/or filter data if needed
    - Always trim and filter for de novo transcriptome assembly
  - Regenerate quality plots after trimming and filtering to determine effectiveness



# Read Mapping

- Pipeline
- Software
- Input
- Output

# Mapping – with reference genome



Reference Genome

Millions of short reads

Fastq

Spliced aligner

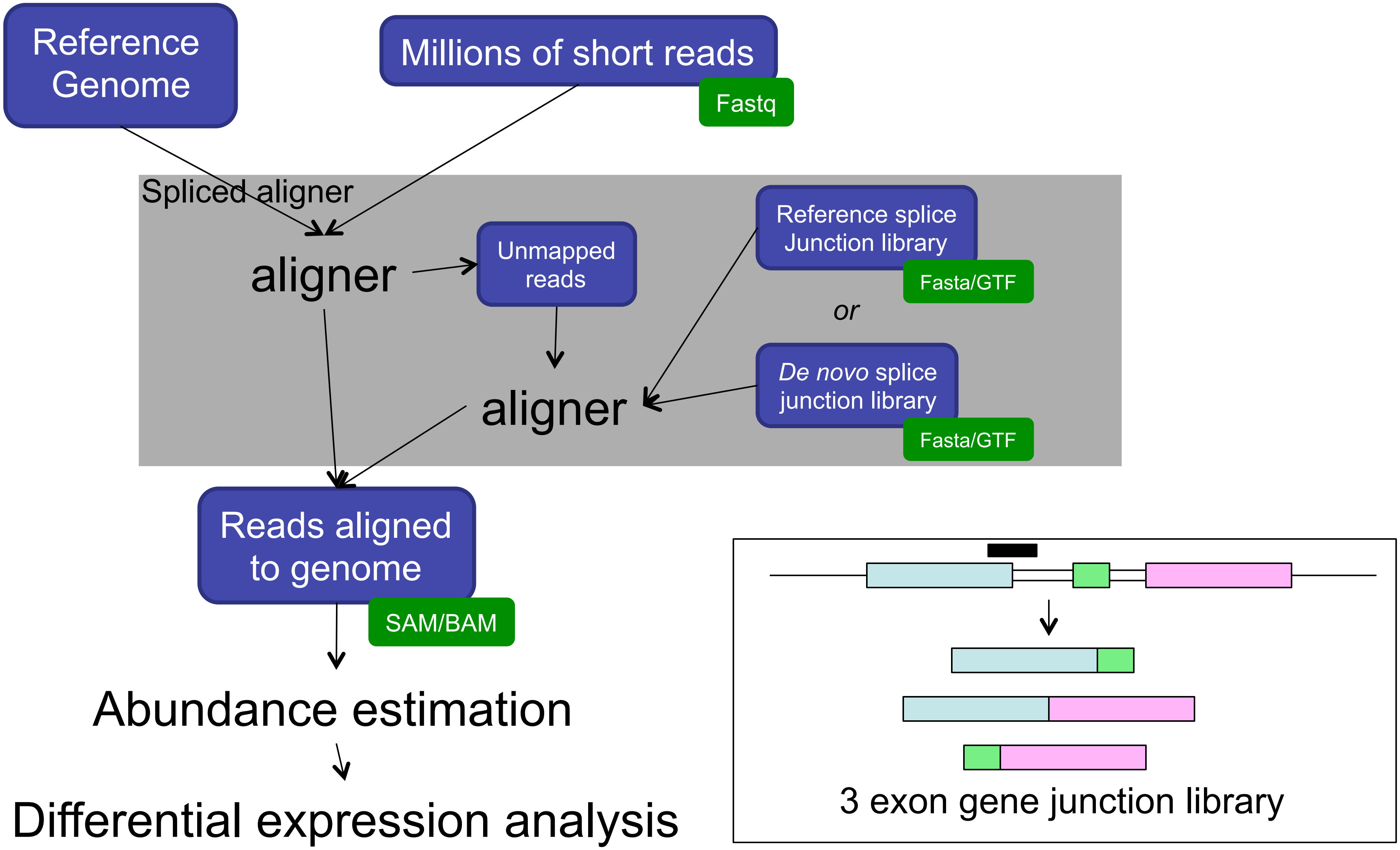
Reads aligned to genome

SAM/BAM

Abundance estimation

Differential expression analysis

# Mapping – with reference genome



Reference Genome

Millions of short reads

Fastq

Spliced aligner

aligner

Unmapped reads

Reference splice Junction library

Fasta/GTF

or

*De novo* splice junction library

Fasta/GTF

aligner

Reads aligned to genome

SAM/BAM

Abundance estimation

Differential expression analysis

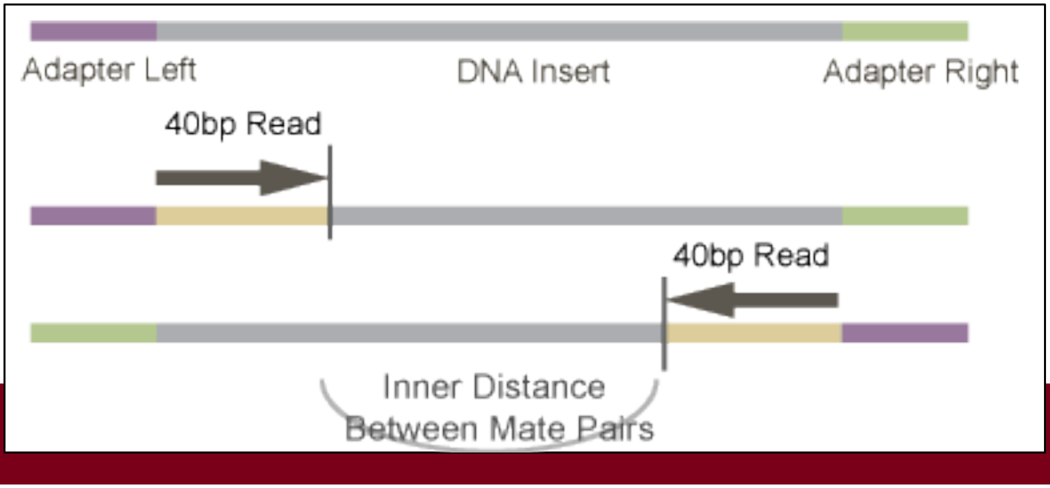
3 exon gene junction library

# Mapping

- Alignment algorithm must be
  - Fast
  - Able to handle SNPs, indels, and sequencing errors
  - Allow for introns for reference genome alignment (spliced alignment)
- Burrows Wheeler Transform (BWT) mappers
  - Faster
  - Few mismatches allowed (< 3)
  - Limited indel detection
  - Spliced: Tophat, MapSplice
  - Unspliced: BWA, Bowtie
- Hash table mappers
  - Slower
  - More mismatches allowed
  - Indel detection
  - Spliced: GSNAP, MapSplice
  - Unspliced: SHRiMP, Stampy

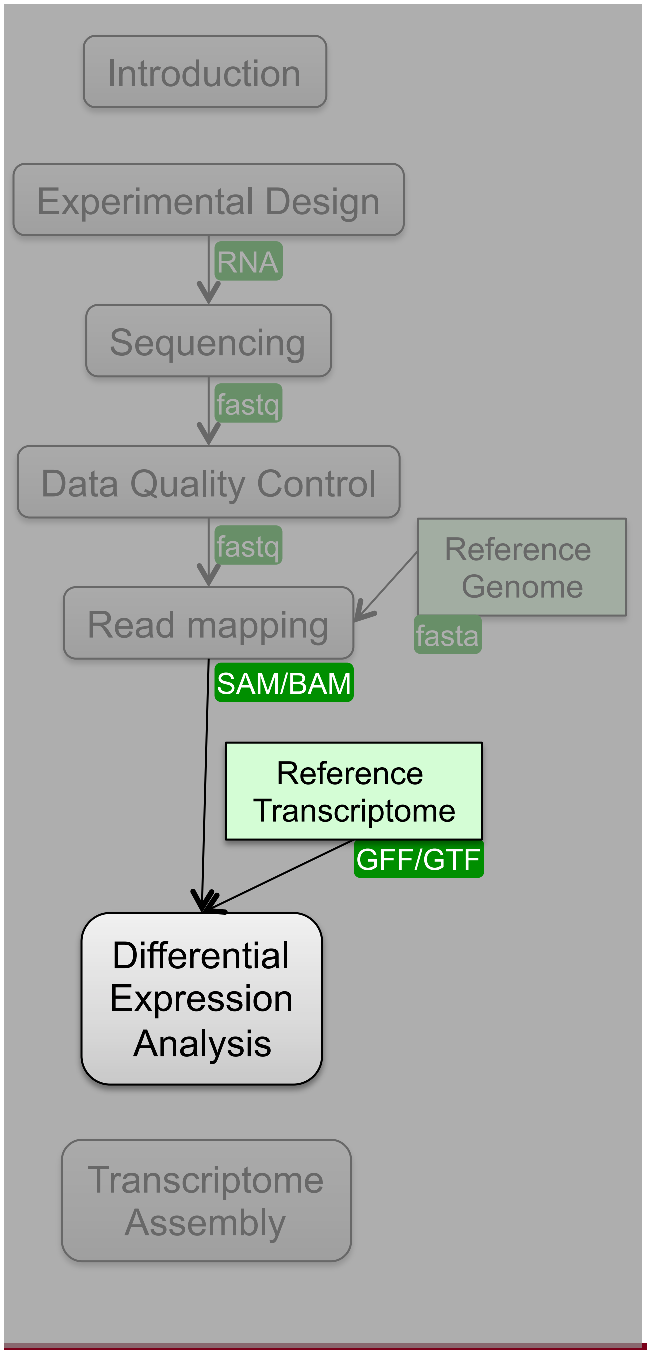
# Mapping

- Input
  - Fastq read libraries
  - Reference genome index (software-specific: /project/db/genomes)
  - Insert size mean and stddev (for paired-end libraries)
    - Map library (or a subset) using estimated mean and stddev
    - Calculate empirical mean and stddev
      - Galaxy: NGS Picard: insertion size metrics
      - Cufflinks standard error
    - Re-map library using empirical mean and stddev



# Mapping

- Output
  - SAM (text) / BAM (binary) alignment files
    - SAMtools – SAM/BAM file manipulation
  - Summary statistics (per read library)
    - % reads with unique alignment
    - % reads with multiple alignments
    - % reads with no alignment
    - % reads properly paired (for paired-end libraries)



# Differential Expression

- Discrete vs continuous data
- Cuffdiff and EdgeR

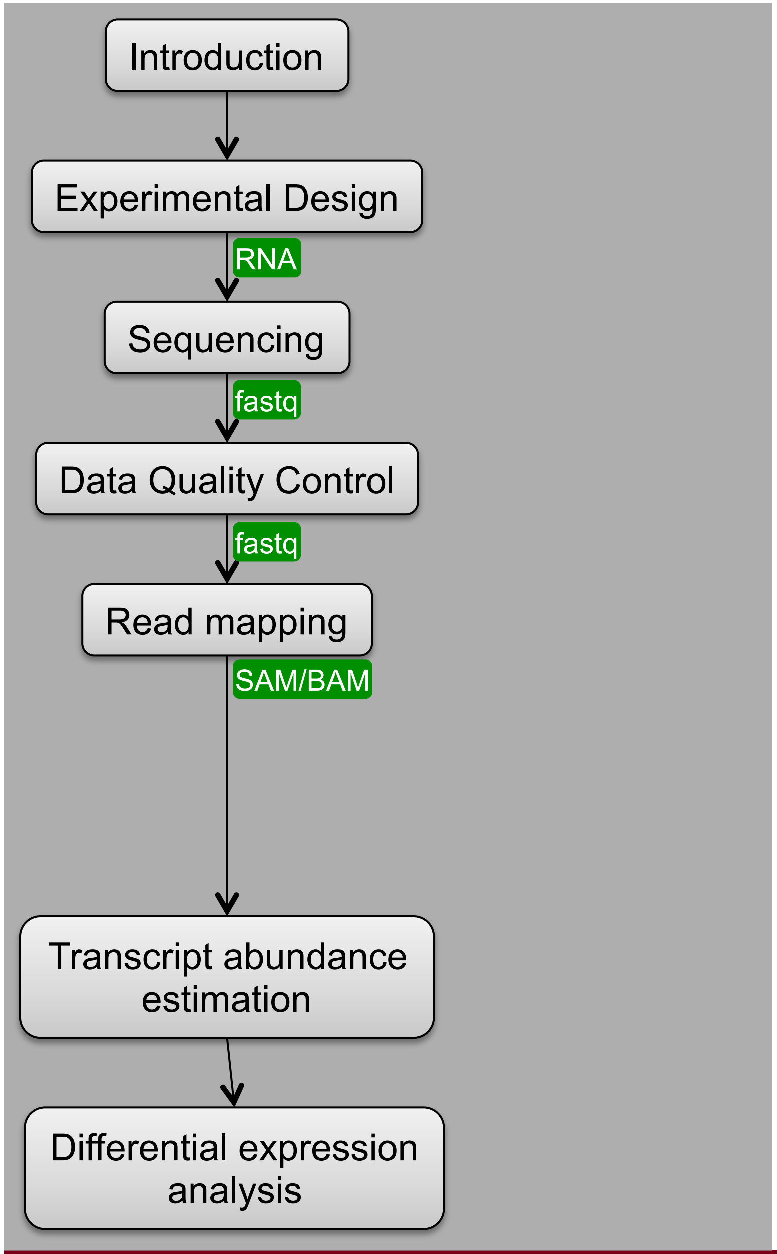
# Differential Expression

- Discrete vs Continuous data
  - Microarray florescence intensity data: continuous
    - Modeled using normal distribution
  - RNA-Seq read count data: discrete
    - Modeled using negative binomial distribution

Microarray software cannot be used to analyze RNA-Seq data

# Differential Expression

- Cuffdiff (Cufflinks package)
  - Pairwise comparisons
  - Differential gene, transcript, and primary transcript expression; differential splicing and promoter use
  - Easy to use, well documented
  - Input: transcriptome, SAM/BAM read alignments (abundance estimation built-in)
- EdgeR
  - Complex experimental designs using generalized linear model
  - Information sharing among genes (Bayesian gene-wise dispersion estimation)
  - Difficult to use R package — ⚠ Consult a statistician
  - Input: raw gene/transcript read counts (calculate abundance using separate software)



Introduction

Experimental Design

RNA

Sequencing

fastq

Data Quality Control

fastq

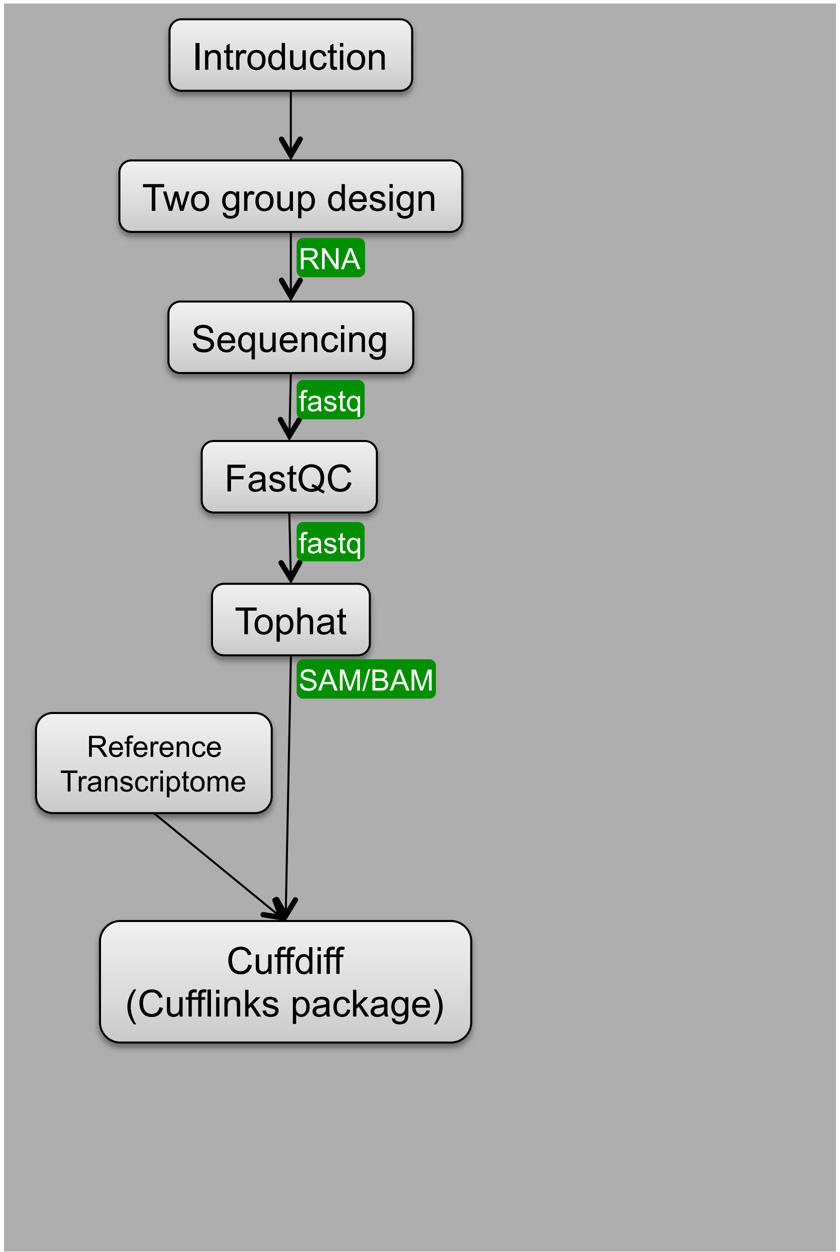
Read mapping

SAM/BAM

Transcript abundance estimation

Differential expression analysis

Suggested Implementation
(reference genome available)



Introduction

Two group design

RNA

Sequencing

fastq

FastQC

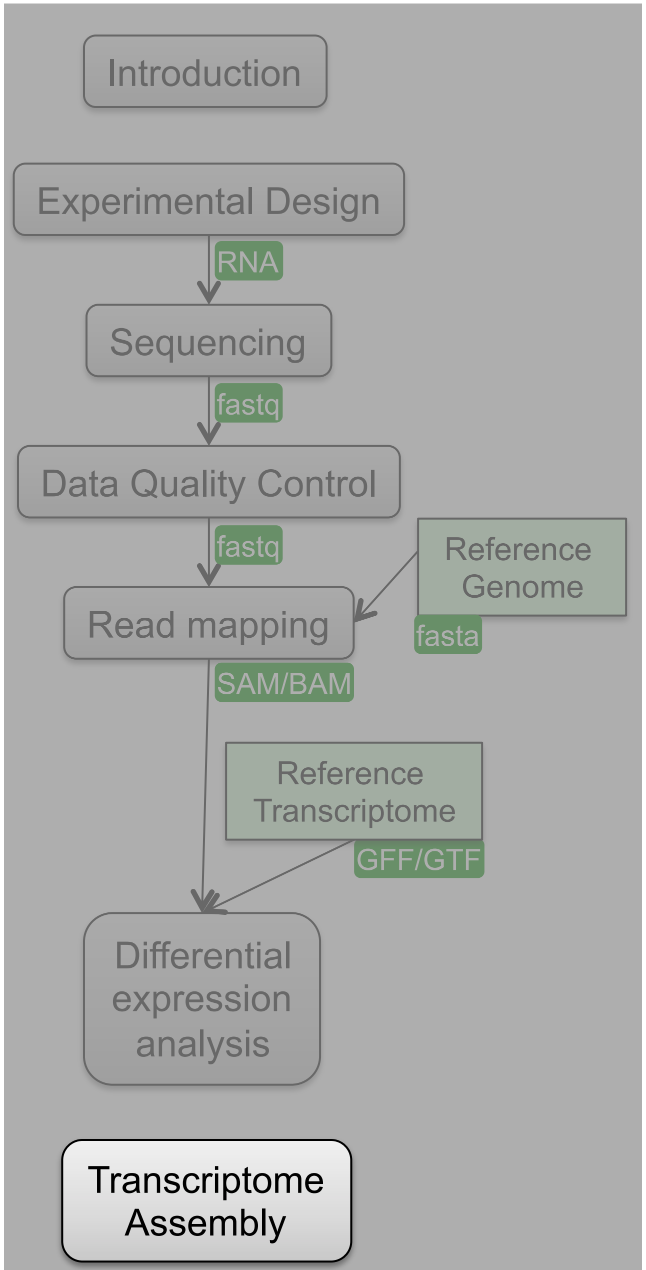
fastq

Tophat

SAM/BAM

Reference Transcriptome

Cuffdiff
(Cufflinks package)



Introduction

Experimental Design

RNA

Sequencing

fastq

Data Quality Control

fastq

Reference Genome

Read mapping

fasta

SAM/BAM

Reference Transcriptome

GFF/GTF

Differential expression analysis

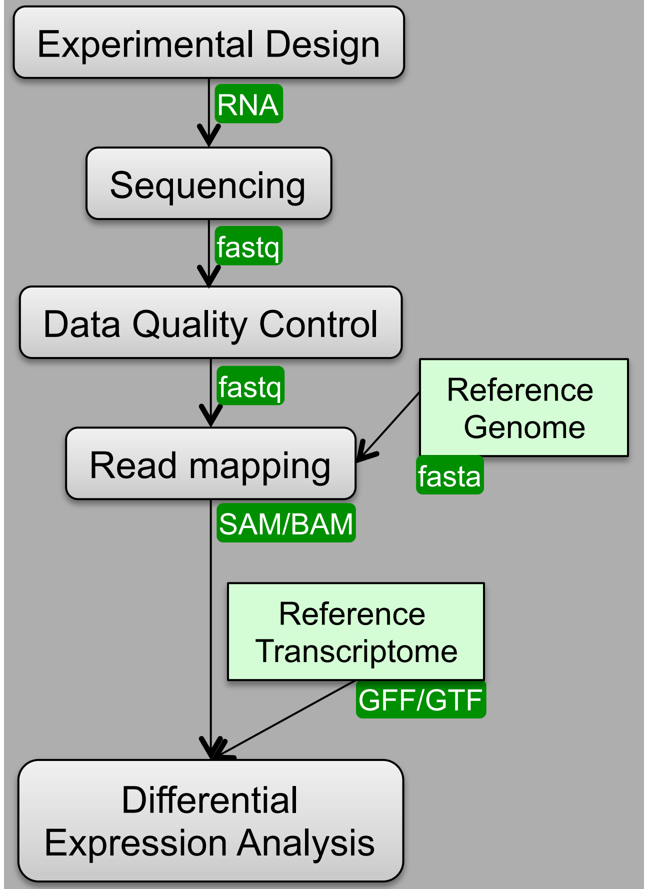
Transcriptome Assembly

# Transcriptome Assembly

- Pipeline
- Software
- Input
- Output

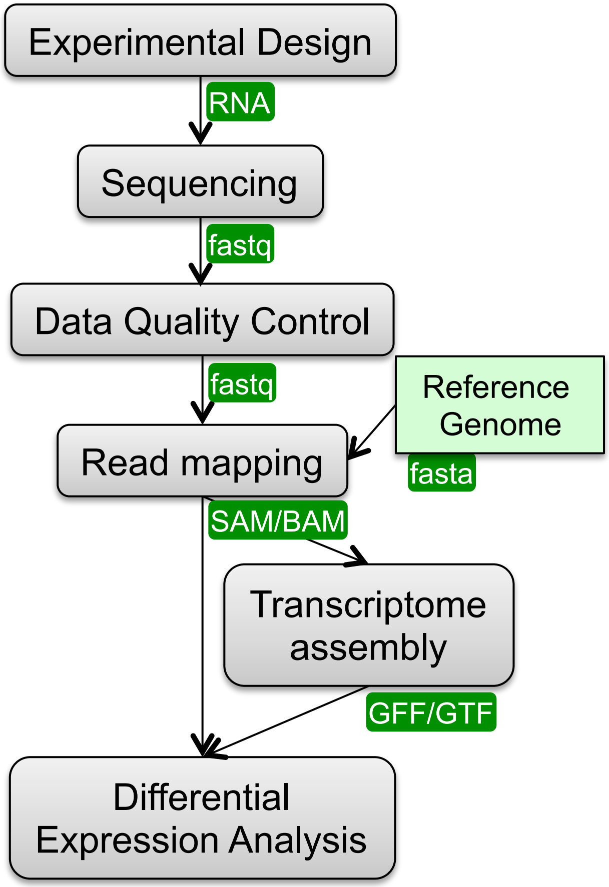
## RNA-Seq

- Reference genome
- Reference transcriptome


Experimental Design
RNA
Sequencing
fastq
Data Quality Control
fastq
Read mapping
Reference Genome
fasta
SAM/BAM
Reference Transcriptome
GFF/GTF
Differential Expression Analysis

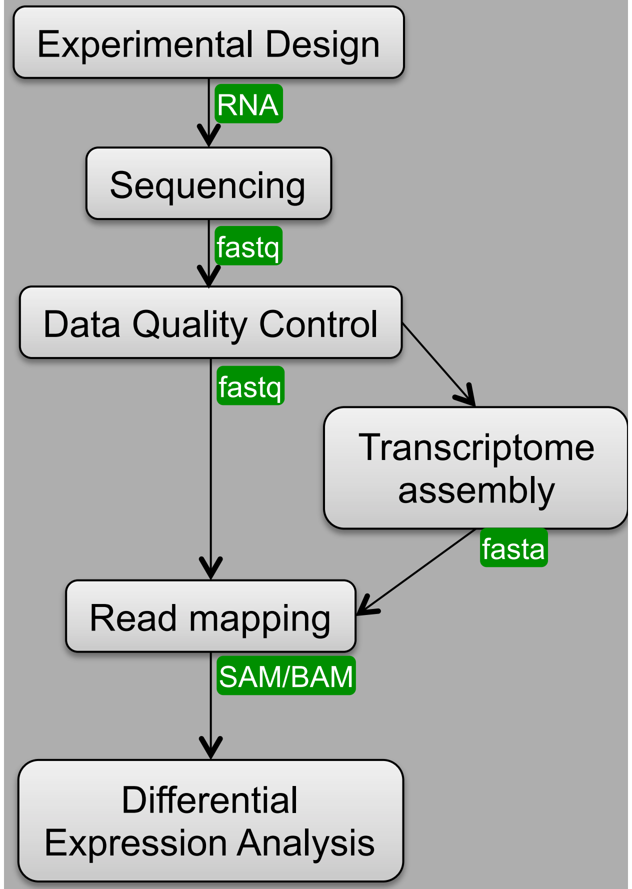
## RNA-Seq

- Reference genome
- No reference transcriptome

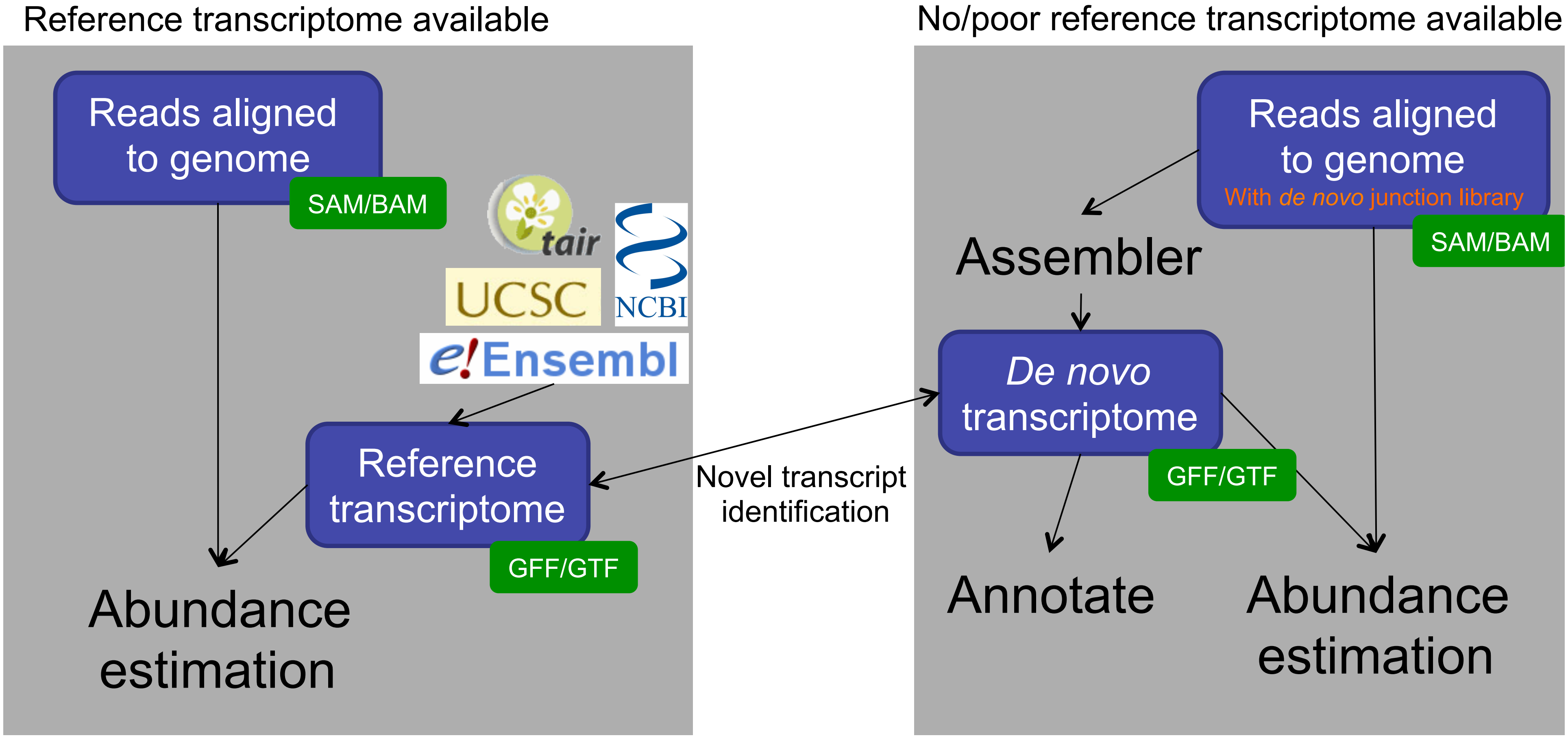

Experimental Design
RNA
Sequencing
fastq
Data Quality Control
fastq
Read mapping
Reference Genome
fasta
SAM/BAM
Transcriptome assembly
GFF/GTF
Differential Expression Analysis

## RNA-Seq

- No reference genome
- No reference transcriptome


Experimental Design
RNA
Sequencing
fastq
Data Quality Control
fastq
Transcriptome assembly
fasta
Read mapping
SAM/BAM
Differential Expression Analysis

# Transcriptome Assembly -with reference genome



Reference transcriptome available

Reads aligned to genome

SAM/BAM

Reference transcriptome

GFF/GTF

Abundance estimation

Novel transcript identification

No/poor reference transcriptome available

Reads aligned to genome

With *de novo* junction library

SAM/BAM

Assembler

*De novo* transcriptome

GFF/GTF

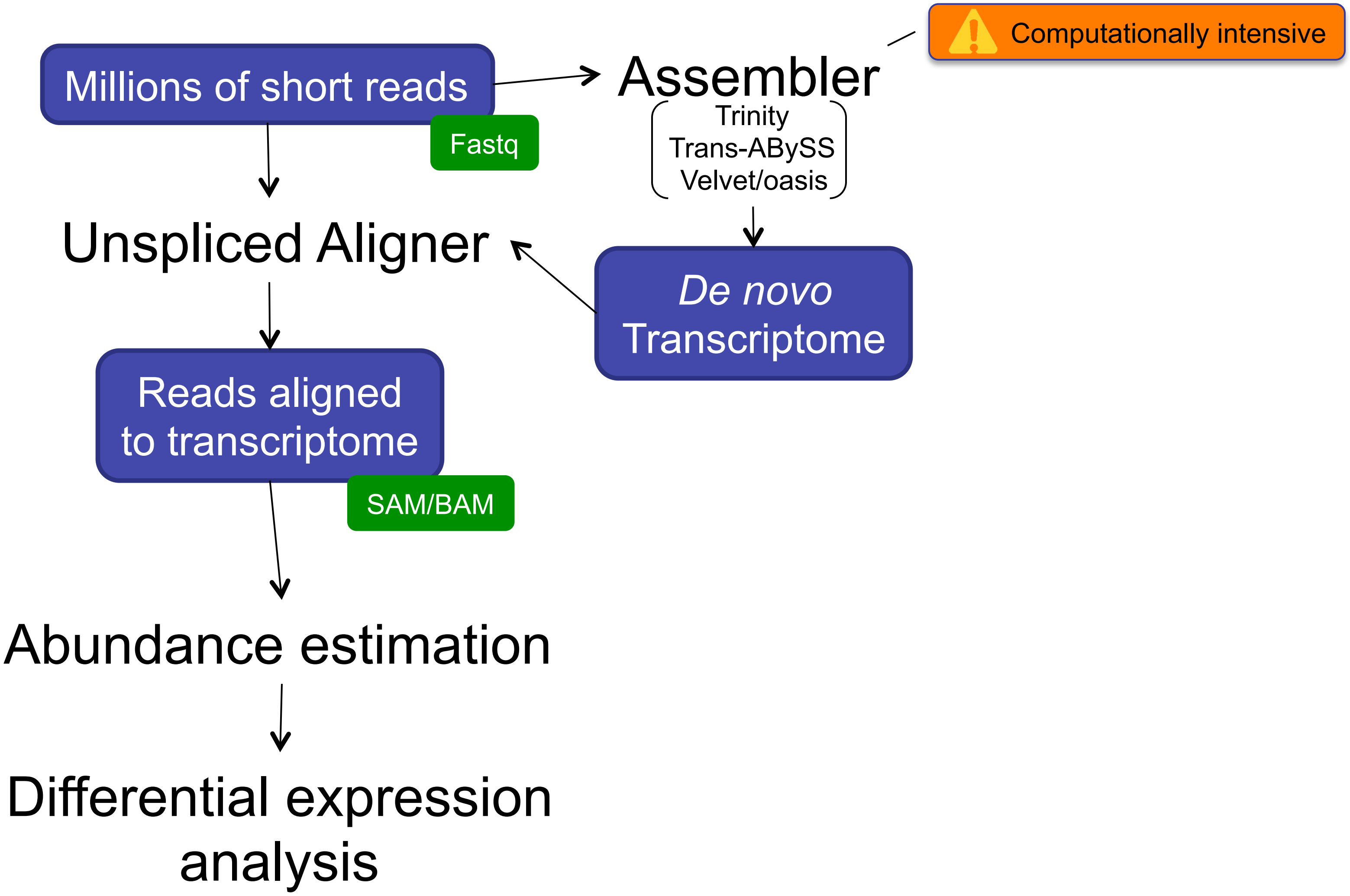
Annotate

Abundance estimation

# Transcriptome Assembly -with reference genome

- Reference genome based assembly
  - Cufflinks, Scripture
- Reference annotation based assembly
  - Cufflinks
- Transcriptome comparison
  - Cuffcompare
- Transcriptome Annotation
  - Generate cDNA fasta from annotation (Cufflinks' gffread program)
  - Align to library of known cDNA (RefSeq, GenBank)

# Transcriptome Assembly – no reference genome



# Further Reading

Bioinformatics for High Throughput Sequencing

Online access through U library

Rodríguez-Ezpeleta, Naiara.; Hackenberg, Michael.; Aransay, Ana M.;
SpringerLink New York, NY : Springer c2012

RNA sequencing: advances, challenges and opportunities

Fatih Ozsolak1 & Patrice M. Milos1
Nature Reviews Genetics 12, 87-98 (February 2011)

Computational methods for transcriptome annotation and quantification using RNA-seq

Manuel Garber, Manfred G Grabherr, Mitchell Guttman & Cole Trapnell
Nature Methods 8, 469–477 (2011)

Table of RNA-Seq software

Next-generation transcriptome assembly

Jeffrey A. Martin & Zhong Wang
Nature Reviews Genetics 12, 671-682 (October 2011)

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David Kelley, Harold Pimentel, Steven Salzberg, John L Rinn & Lior Pachter
Nature Protocols 7, 562–578 (2012)

SEQanswers.com
biostar.stackexchange.com

Popular bioinformatics forums

# Questions / Discussion