系统生物学

天津医科大学
生物医学工程与技术学院

2016-2017 学年上学期（秋）
2013 级生信班

第三章　转录组学

伊现富（Yi Xianfu）

天津医科大学（TIJMU）
生物医学工程与技术学院

2016 年 10 月

教学提纲

教学提纲

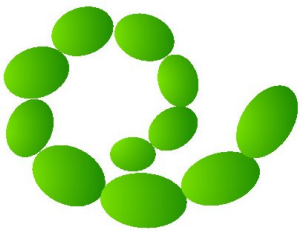
教学提纲

DNA → Genome

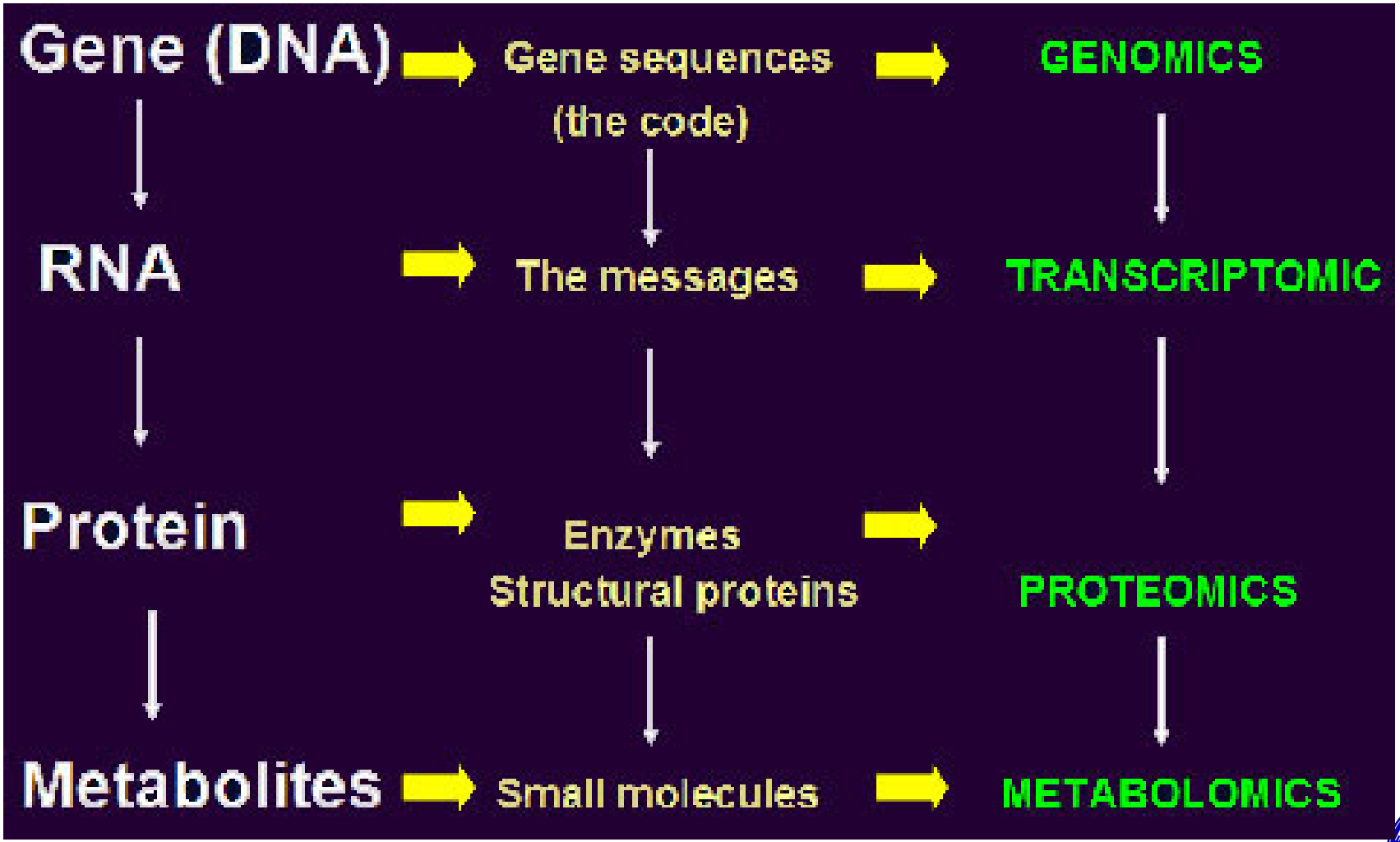
Transcription

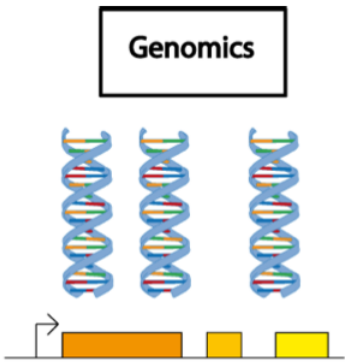
mRNA → Transcriptome

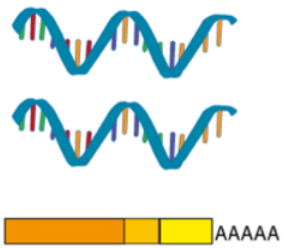
Translation

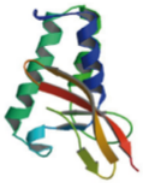


Protein → Proteome

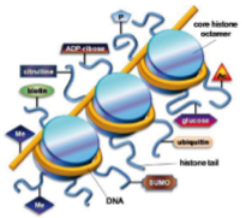

Gene (DNA)
Gene sequences
(the code)
GENOMICS
RNA
The messages
TRANSCRIPTOMIC
Protein
Enzymes
Structural proteins
PROTEOMICS
Metabolites
Small molecules
METABOLOMICS

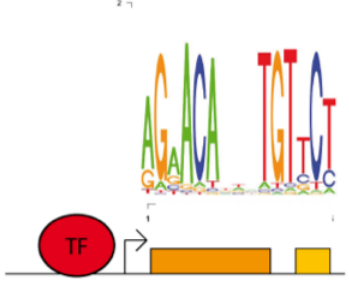




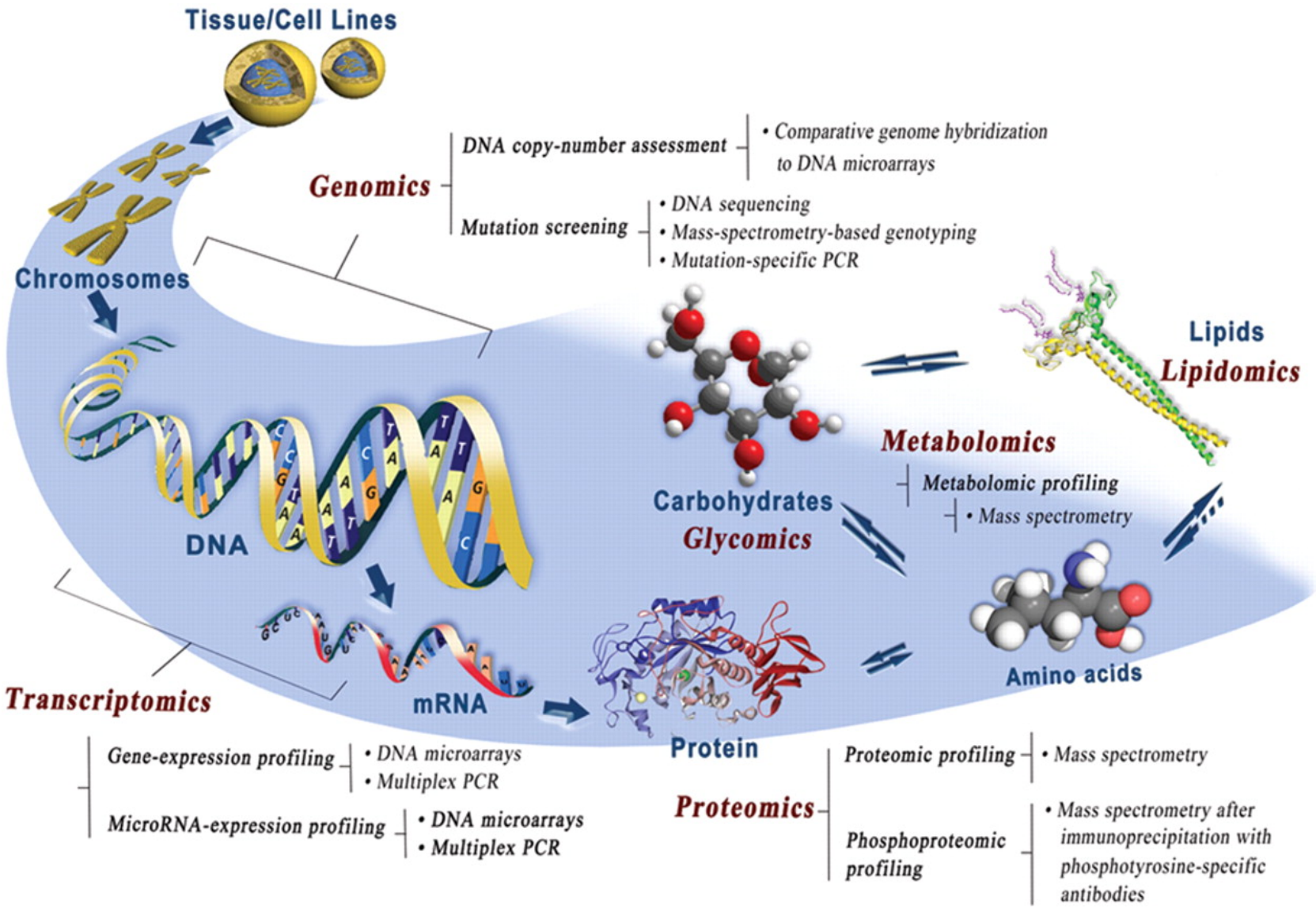


> Replicase polyprotein 1ab, P19811
MATFSATGFG GSFVRDWSLD LPDACE-
HGAG LCCEVDGSTL CAECFRGCEG M
.....





Genomics	Transriptomics	Proteomics	Epigenomics	Cistromics
- genomic DNA Sequencing - annotations	-microarray hybridization -RNA-seq	- liquid chromatography + - Mass Spectrometry	- ChIP-Seq - bisulfite seqencing	- ChIP-Seq
Genes sequences	mRNA sequences mRNA expression level	Protein sequences	Histones modifications DNA methylation	Genomic repertoire of the binding sequences of a transcriptome factor

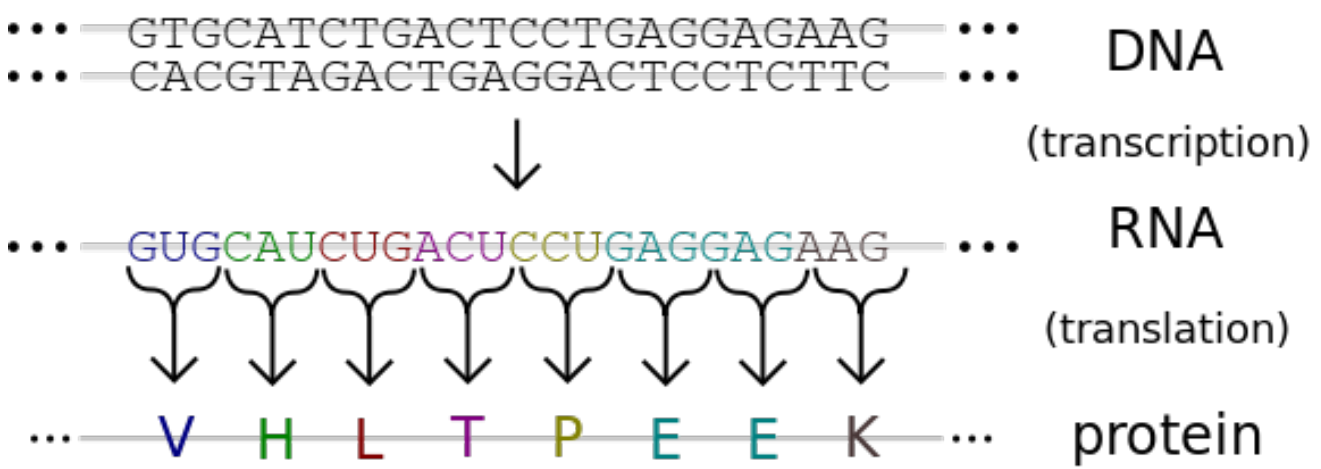

Tissue/Cell Lines
Genomics
DNA copy-number assessment
• Comparative genome hybridization to DNA microarrays
Mutation screening
• DNA sequencing
• Mass-spectrometry-based genotyping
• Mutation-specific PCR
Chromosomes
DNA
Transcriptomics
Gene-expression profiling
• DNA microarrays
• Multiplex PCR
MicroRNA-expression profiling
• DNA microarrays
• Multiplex PCR
mRNA
Protein
Proteomics
Proteomic profiling
• Mass spectrometry
Phosphoproteomic profiling
• Mass spectrometry after immunoprecipitation with phosphotyrosine-specific antibodies
Carbohydrates
Glycomics
Metabolomics
Metabolomic profiling
• Mass spectrometry
Lipids
Lipidomics
Amino acids

教学提纲

基因表达

基因表达（gene expression）是用基因中的信息来合成基因产物的过程。产物通常是蛋白质，但对于非蛋白质编码基因，如转运 RNA（tRNA）和小核 RNA（snRNA），产物则是 RNA。

基因表达的过程可分为转录、RNA 剪接、翻译、蛋白质的翻译后修饰这几步。基因表达调控控制细胞的结构与功能，同时也是细胞分化、形态发生及生物体适应性的基础。不同的时间、不同的环境，以及不同部位的细胞，或是基因在细胞中的含量差异，皆可能使基因产生不同的表现。



基因表达谱

基因表达谱（gene expression profile）是一种在分子生物学领域，借助cDNA、表达序列标签（EST）或寡核苷酸芯片来测定细胞基因表达情况(包括特定基因是否表达、表达丰度、不同组织、不同发育阶段以及不同生理状态下的表达差异）的方法。

通过一次性测定大量基因构建起细胞功能的总体态势图，可以从图谱中区分出正在分裂的细胞，以及细胞对于特征性治疗的反应。基因表达谱还有助于了解疾病的发病机制、药物的生理反应和治疗效果。

基因表达图谱从逻辑上说是基因测序的下一个步骤：基因序列包含细胞可能存在的功能的信息，而基因表达谱则包含细胞实际上正在完成的工作的信息。

Technique

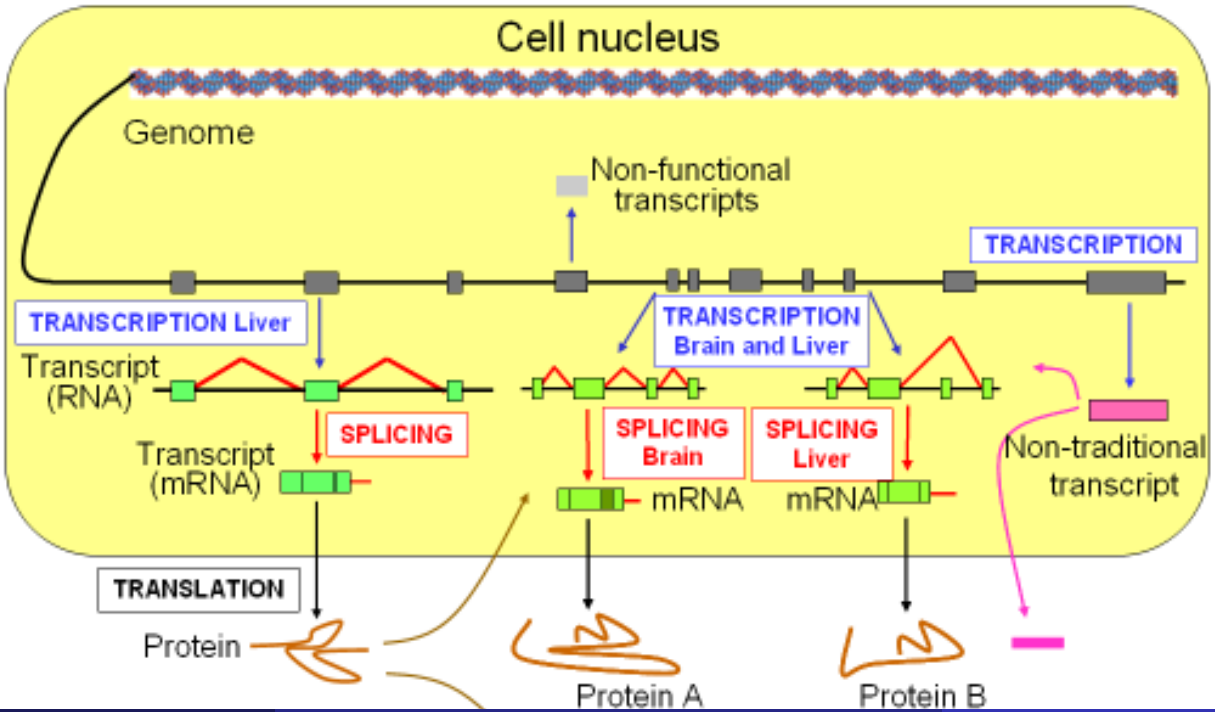
DNA microarray technology measures the relative activity of previously identified target genes.

Sequence based techniques, like serial analysis of gene expression (SAGE, SuperSAGE) are also used for gene expression profiling. SuperSAGE is especially accurate and can measure any active gene, not just a predefined set.

The advent of next-generation sequencing has made sequence based expression analysis an increasingly popular, “digital” alternative to microarrays called RNA-Seq.

转录组

转录组（transcriptome），也称为“转录物组”，广义上指在相同环境（或生理条件）下的在一个细胞、或一群细胞中所能转录出的所有 RNA 的总和，包括信使 RNA（mRNA）、核糖体 RNA（rRNA）、转运 RNA（tRNA）及非编码 RNA；狭义上则指细胞所能转录出的所有信使 RNA（mRNA）。



转录组

转录组这个术语可用于指代给定有机体中的转录本总集，或者是特定细胞类型中的特定转录本子集。

不考虑突变，固定给定细胞株的基因组数量基本上是不变的；与之不同，转录物组可以随外部环境条件而有所变化。由于转录物组包括了所有在细胞里的 mRNA 的转录，除却异常的 mRNA 降解现象（例如转录衰减）以外，转录组反映了在任何给定时间内活跃表达的基因。

转录物组学的研究，也被称为“基因表达谱”，检测了在一个特定的细胞群内的 mRNA 表达水平，通常采用基于 DNA 微阵列技术的高通量技术。使用新一代测序技术在核苷酸水平上来研究转录物组，被称为“RNA 测序（RNA-Seq）”。

Methods of construction

There are two general methods of inferring transcriptomes.

- One approach maps sequence reads onto a reference genome, either of the organism itself (whose transcriptome is being studied) or of a closely related species.
- The other approach, *de novo* transcriptome assembly, uses software to infer transcripts directly from short sequence reads.

转录组学

转录组学（或“转录物组学”，transcriptomics）是分子生物学的分支，负责研究在单个细胞或一个细胞群的特定细胞类型内所生产的 mRNA 分子。

转录组测定的是表达的基因数目。这个数目包括了在各种不同水平上表达的基因。

丰度

每个细胞里每一种 mRNA 分子的平均数称为该种 mRNA 的丰度。根据丰度可把 mRNA 群体分为两大类：

- 高丰度 mRNA 组分。通常由每个细胞里不到 100 种的 mRNA 而每种 mRNA 分子由 1000-10,000 份备份所组成，通常占总 mRNA 的 50% 左右。
- 除高丰度 mRNA 组分外，另一半 mRNA 由长约 10,000nt 的种类繁多的序列所组成，每种序列在 mRNA 中只有少量备份。称为“稀有 mRNA”或“复杂 mRNA”。

转录组学

大量表达的基因之间是有很大区别的。例如，卵清蛋白只在输卵管细胞里合成而不在肝脏内合成。它占输卵管内 mRNA 总量的一半。但是高丰度的 mRNA 只占表达基因数的很小一部分。根据生物体的基因数目以及不同类型细胞出现转录变化的基因数，人们需要了解不同表型细胞稀有 mRNA 基因相同的程度。

稀有 mRNA 是普遍共有的。一个细胞中的 mRNA 序列只有 10% 左右是该细胞独有的，大部分序列是许多（甚至是所有）类型的细胞所共有的。这提示哺乳动物中共有的基因数也许达 10,000 个，包括了各种类型细胞所需的功能。编码这种类型的功能的基因，有时被称为“持家基因”或“组成型基因”。这类功能不同于特定细胞表现型所需的特定功能（例如卵清蛋白或珠蛋白的功能）。编码特殊功能的基因称为“奢侈基因”。

问题

- 一个细胞、组织或生物体的全部 RNA 集合体中包括多少种 RNA，各种 RNA 的数量有多少？
- 在不同发育时期和不同外界环境作用下 RNA 集合体会出现怎样的变化？
- 在细胞中转录是怎样被调节的？
- ……

转录组学

转录组学（transcriptomics）是对转录水平上发生的事件及其相互关系和意义进行整体研究的一门学科。

问题

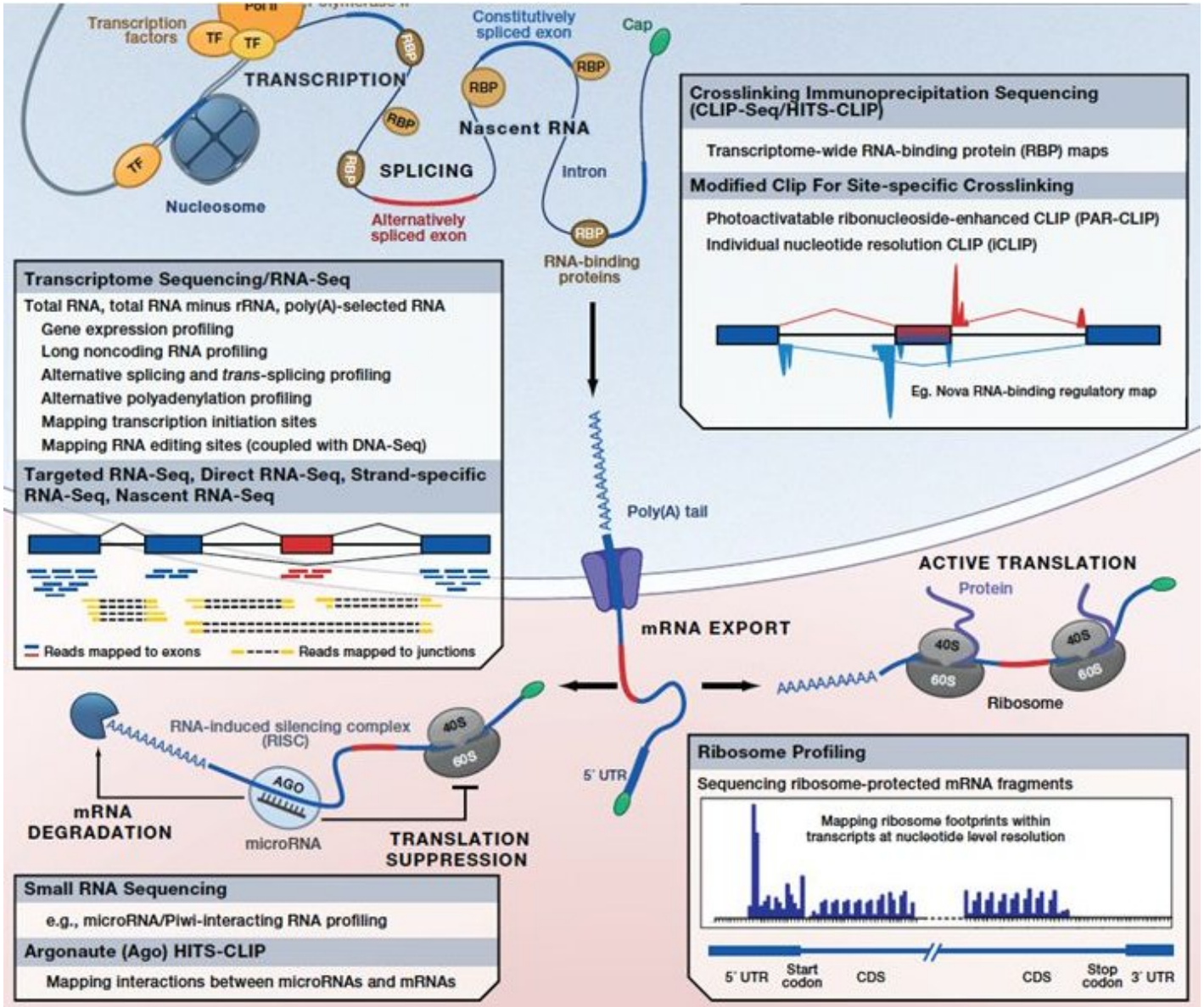
- 一个细胞、组织或生物体的全部 RNA 集合体中包括多少种 RNA，各种 RNA 的数量有多少？
- 在不同发育时期和不同外界环境作用下 RNA 集合体会出现怎样的变化？
- 在细胞中转录是怎样被调节的？
- ……

转录组学

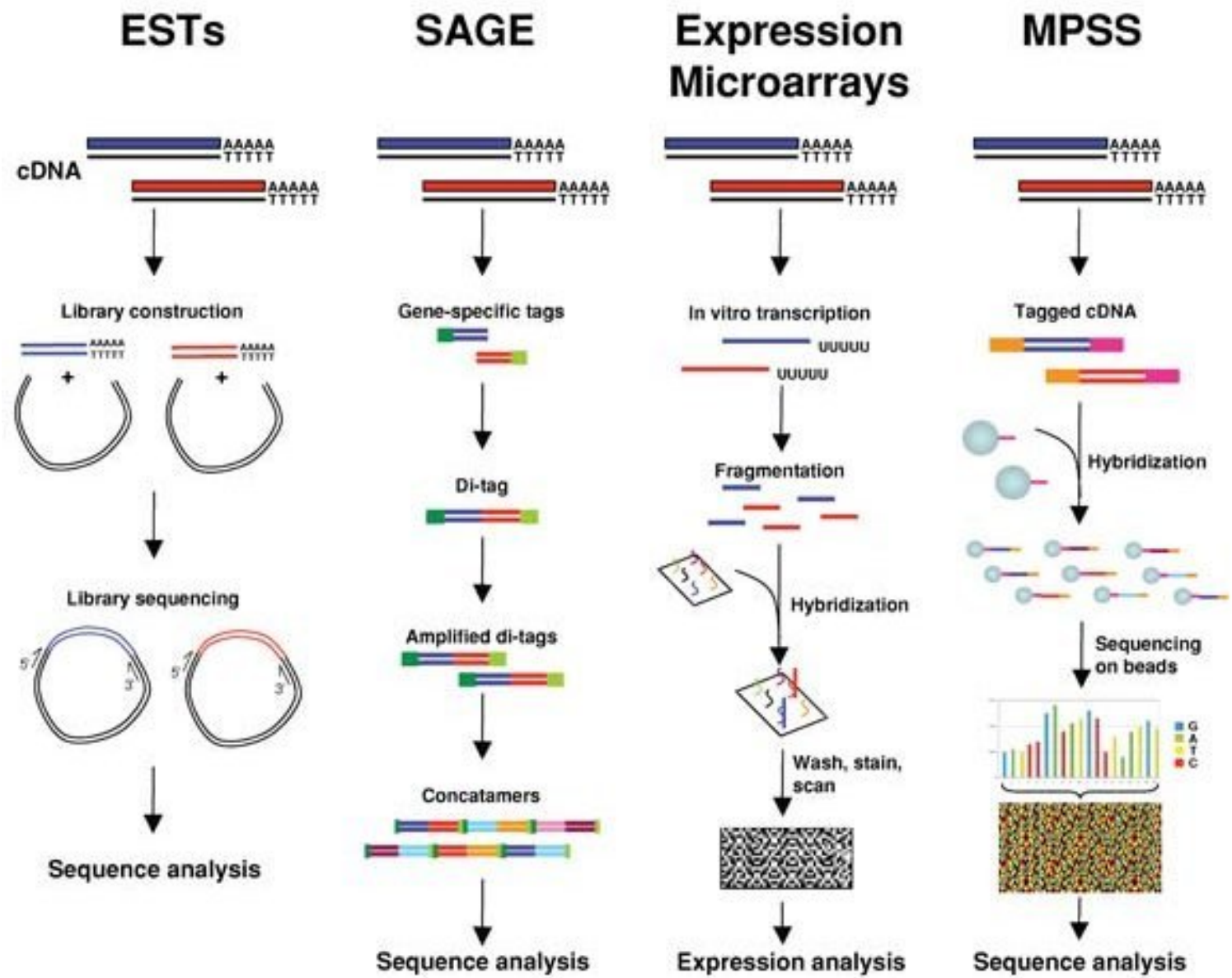
转录组学（transcriptomics）是对转录水平上发生的事件及其相互关系和意义进行整体研究的一门学科。

研究内容（4 个水平）

- 对特定细胞的转录与加工机制进行研究
- 对转录物编制目录便于进一步归类研究
- 绘制动态的转录物图形
- 转录物调控网络


Transcription factors
TF
TF
TRANSCRIPTION
Nucleosome
Constitutively spliced exon
Cap
RBP
Nascent RNA
SPLICING
Intron
Alternatively spliced exon
RNA-binding proteins
Crosslinking Immunoprecipitation Sequencing (CLIP-Seq/HITS-CLIP)
Transcriptome-wide RNA-binding protein (RBP) maps
Modified Clip For Site-specific Crosslinking
Photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP)
Individual nucleotide resolution CLIP (iCLIP)
Eg. Nova RNA-binding regulatory map
Transcriptome Sequencing/RNA-Seq
Total RNA, total RNA minus rRNA, poly(A)-selected RNA
Gene expression profiling
Long noncoding RNA profiling
Alternative splicing and trans-splicing profiling
Alternative polyadenylation profiling
Mapping transcription initiation sites
Mapping RNA editing sites (coupled with DNA-Seq)
Targeted RNA-Seq, Direct RNA-Seq, Strand-specific RNA-Seq, Nascent RNA-Seq
Reads mapped to exons
Reads mapped to junctions
Poly(A) tail
mRNA EXPORT
ACTIVE TRANSLATION
Protein
40S
60S
Ribosome
RNA-induced silencing complex (RISC)
AGO
microRNA
mRNA DEGRADATION
TRANSLATION SUPPRESSION
5' UTR
Ribosome Profiling
Sequencing ribosome-protected mRNA fragments
Mapping ribosome footprints within transcripts at nucleotide level resolution
5' UTR
Start codon
CDS
CDS
Stop codon
3' UTR
Small RNA Sequencing
e.g., microRNA/Piwi-interacting RNA profiling
Argonaute (Ago) HITS-CLIP
Mapping interactions between microRNAs and mRNAs

教学提纲


ESTs
SAGE
Expression Microarrays
MPSS
cDNA
AAAAA
TTTTT
Library construction
Library sequencing
Sequence analysis
Gene-specific tags
Di-tag
Amplified di-tags
Concatamers
Sequence analysis
In vitro transcription
UUUUU
Fragmentation
Hybridization
Wash, stain, scan
Expression analysis
Tagged cDNA
Hybridization
Sequencing on beads
G
A
T
C
Sequence analysis

ADVANTAGES:

- Can detect novel genes and exons
- No hybridization required
- High specificity

- Can detect novel genes and exons
- No hybridization required

- Powerful method to find specific sequences
- Relatively fast and inexpensive

- Can detect novel genes and exons
- Tags are longer and more unique
- Identifies genes with lower expression levels
- Creates digital data that is easy to share and compare

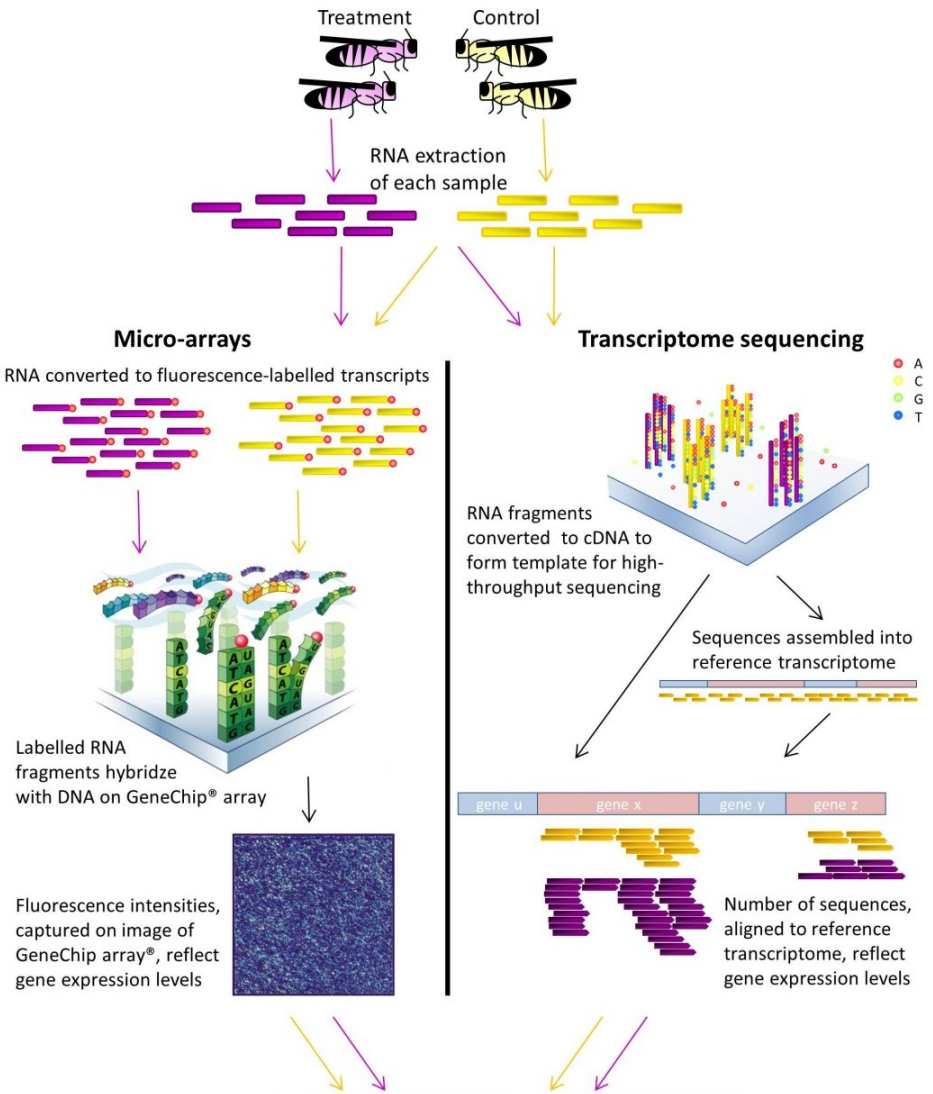
DISADVANTAGES:

- Can not detect genes with low expression levels

- Costly and time-consuming
- Ambiguous tag assignment

- Can not detect novel genes
- Requires hybridization - false positives and negatives
- Difficult to compare data from different platforms

- Costly and time-consuming

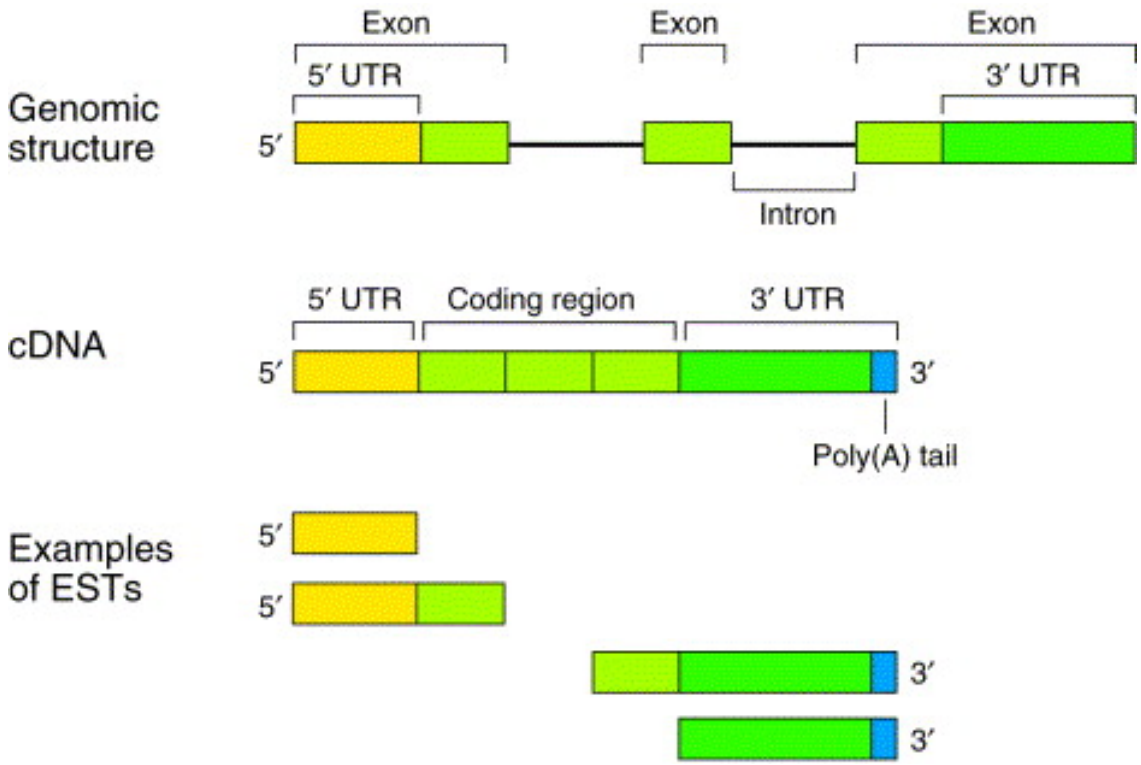


Gene ID	Transcript abundance in Treatment sample	Transcript abundance in Control sample
ABc00001	3778	3894
ABc00002	768	189

EST

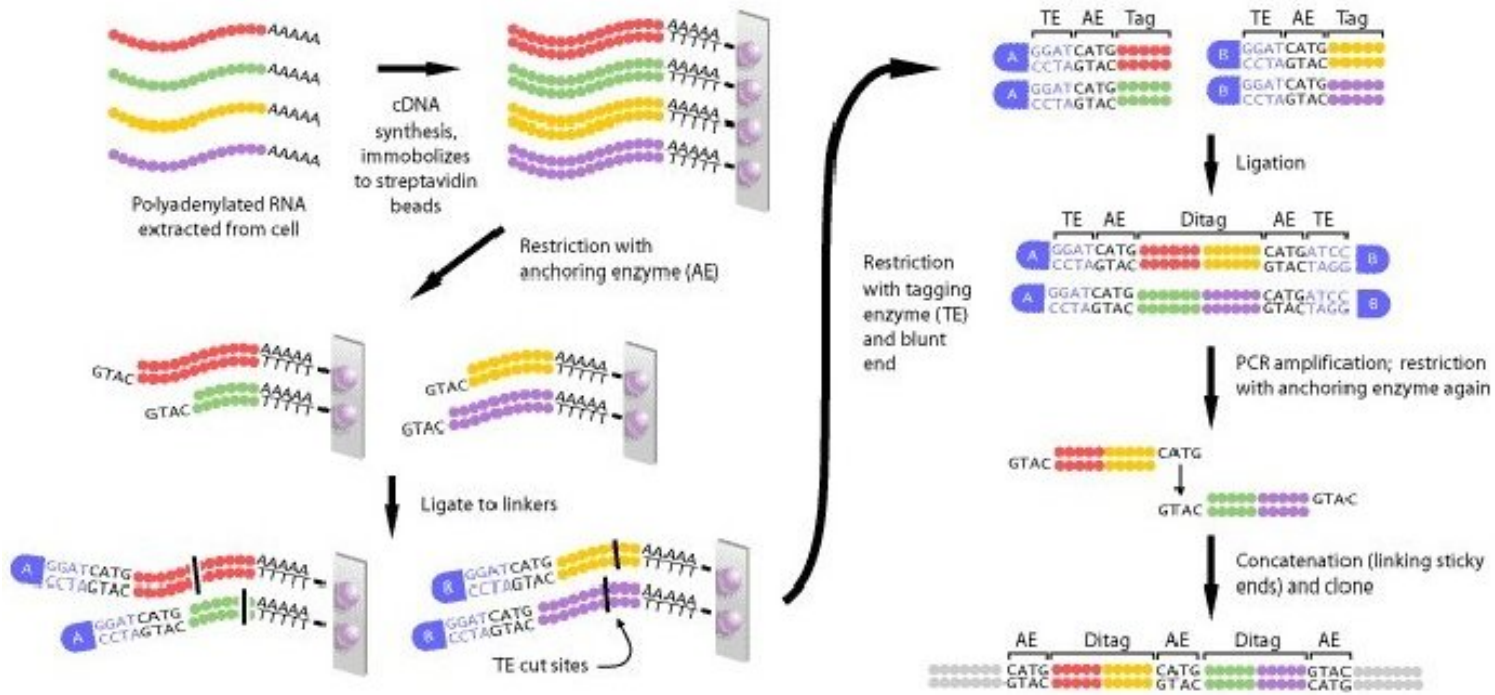
EST（expressed sequence tag，表达序列标签）是从 cDNA 文库中生成的一些很短的序列（500 ～ 800bp），代表在特定组织或发育阶段表达的基因。

In genetics, an expressed sequence tag or EST is a short sub-sequence of a cDNA sequence. ESTs may be used to identify gene transcripts, and are instrumental in gene discovery and in gene-sequence determination.



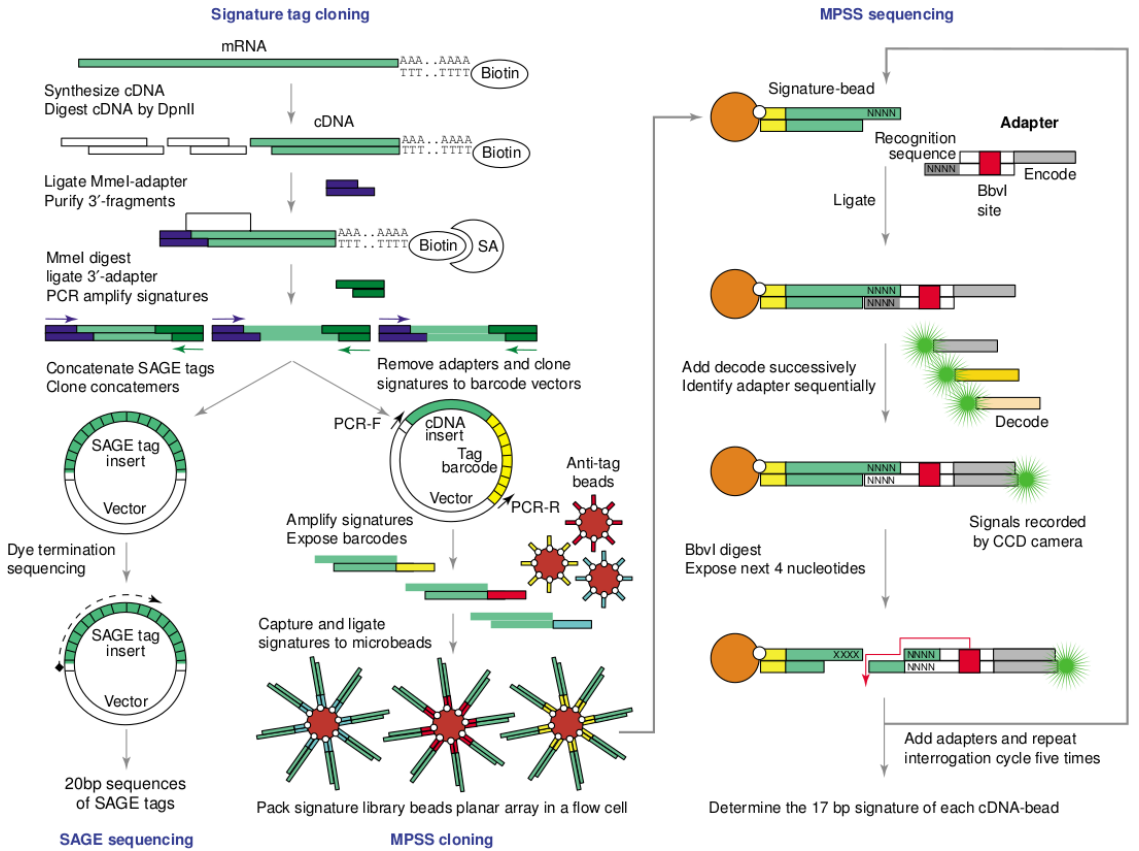
SAGE

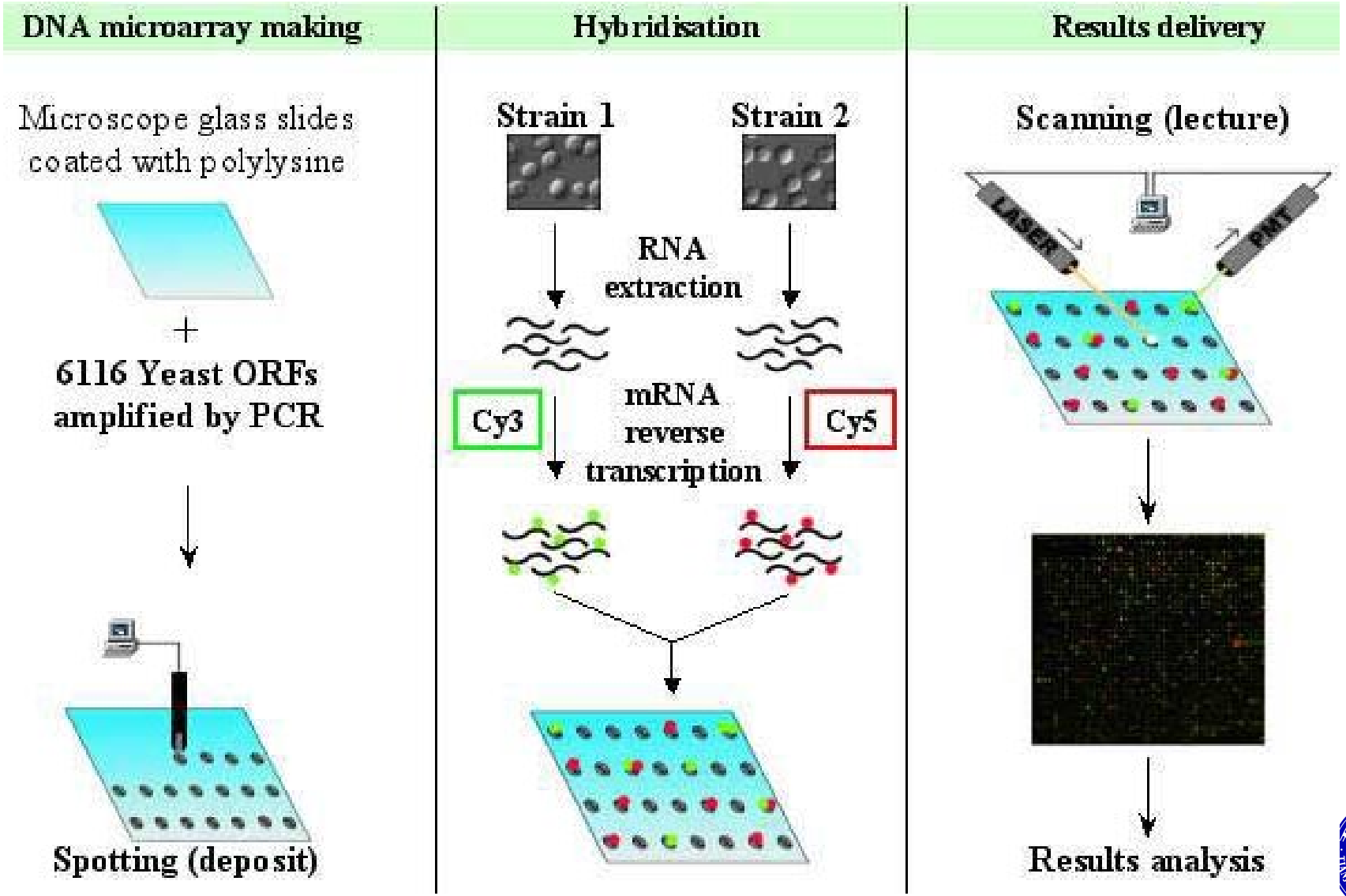
Serial analysis of gene expression (SAGE) is a technique used by molecular biologists to produce a snapshot of the messenger RNA population in a sample of interest in the form of small tags that correspond to fragments of those transcripts.



MPSS

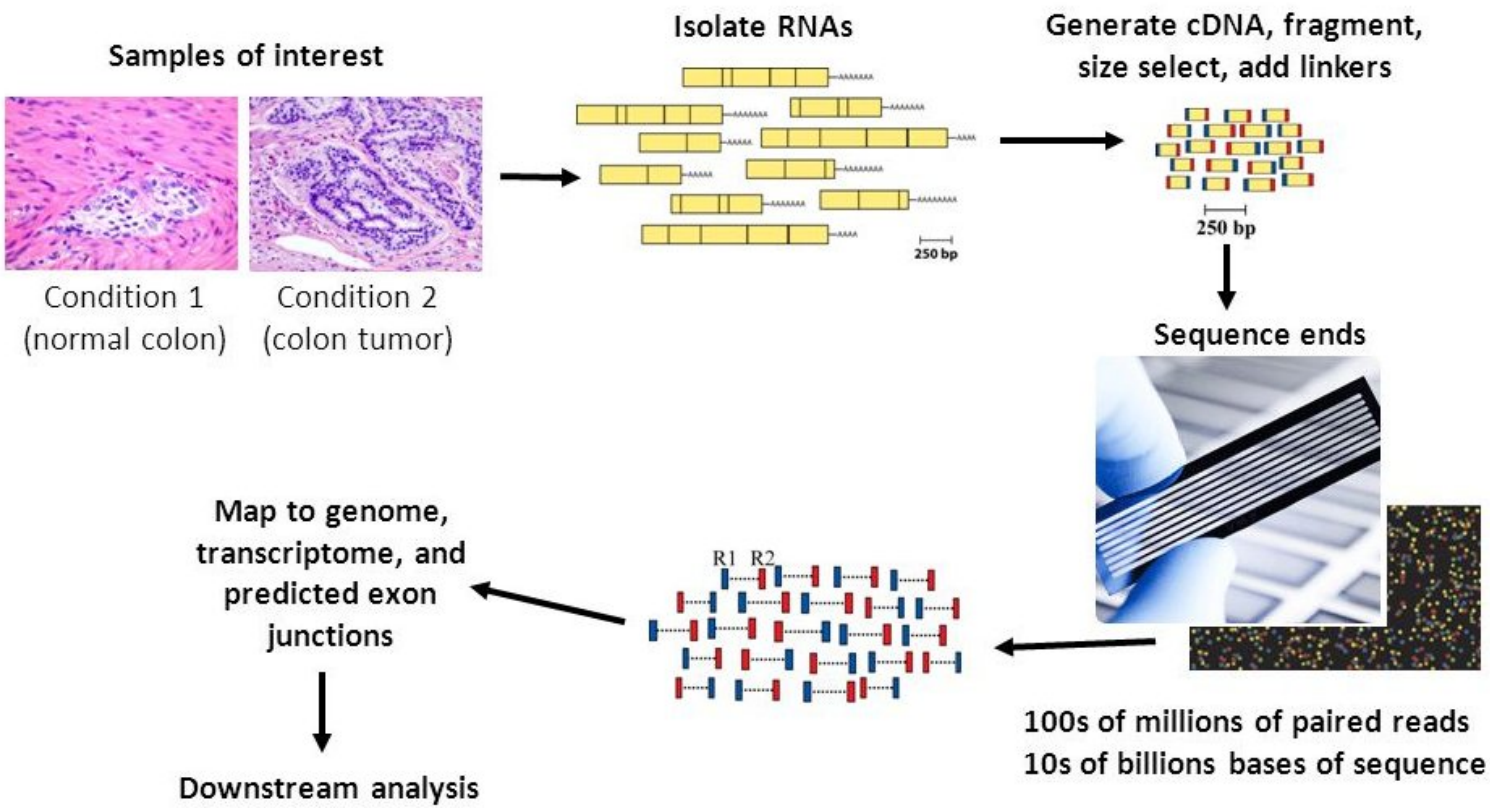
Massive parallel signature sequencing (MPSS) is a procedure that is used to identify and quantify mRNA transcripts, resulting in data similar to serial analysis of gene expression (SAGE), although it employs a series of biochemical and sequencing steps that are substantially different.




DNA microarray making
Hybridisation
Results delivery
Microscope glass slides coated with polylysine
+
6116 Yeast ORFs amplified by PCR
Spotting (deposit)
Strain 1
Strain 2
RNA extraction
Cy3
Cy5
mRNA reverse transcription
Scanning (lecture)
LASER
PMT
Results analysis

RNA-Seq

RNA-seq (RNA sequencing), also called whole transcriptome shotgun sequencing (WTSS), uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment in time.



Choose the right technology

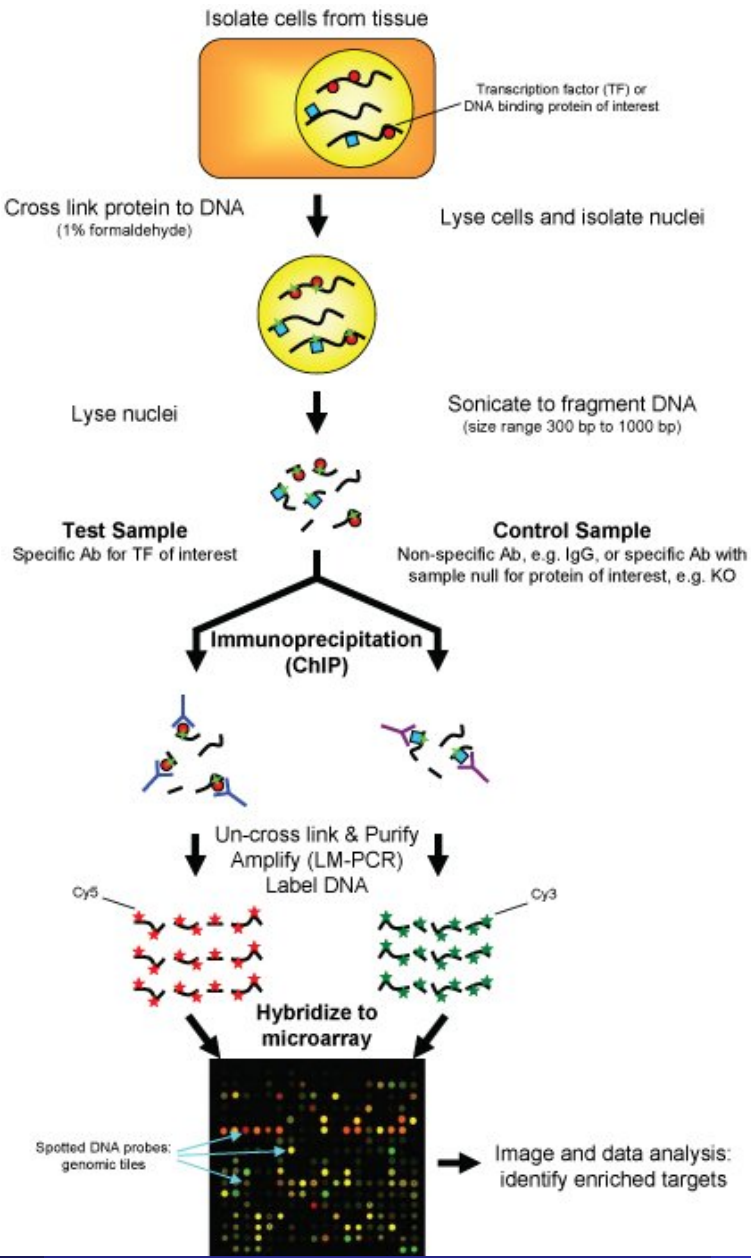
RNA-seq	Microarray
Identification of novel genes, transcripts & exons	Well validated QC and analysis methods
Greater dynamic range	Well characterized biases
Less bias due to genetic variation	Quick turnaround from established core facilities
Repeatable	Currently less expensive (for model organisms)
No species-specific primer/probe design	
More accurate relative to qPCR	
Many more applications	

ChIP-on-chip

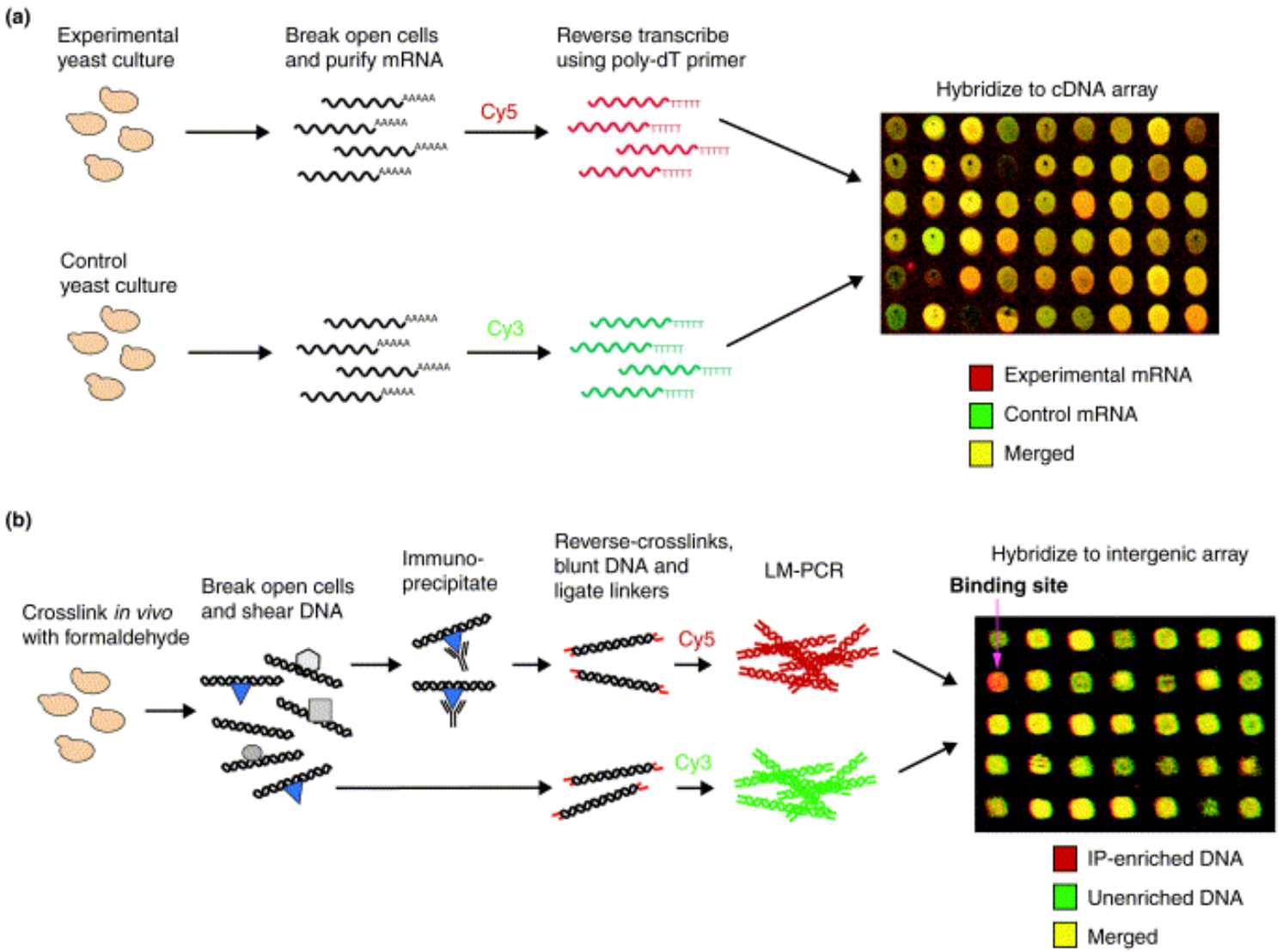
ChIP-on-chip (also known as ChIP-chip) is a technology that combines chromatin immunoprecipitation (“ChIP”) with DNA microarray (“chip”).

Like regular ChIP, ChIP-on-chip is used to investigate interactions between proteins and DNA *in vivo*. Specifically, it allows the identification of the cistrome, sum of binding sites, for DNA-binding proteins on a genome-wide basis. Whole-genome analysis can be performed to determine the locations of binding sites for almost any protein of interest. As the name of the technique suggests, such proteins are generally those operating in the context of chromatin. The most prominent representatives of this class are transcription factors, replication-related proteins, like Origin Recognition Complex Protein(ORC), histones, their variants, and histone modifications.

转录组学 | 研究方法 | ChIP-on-chip

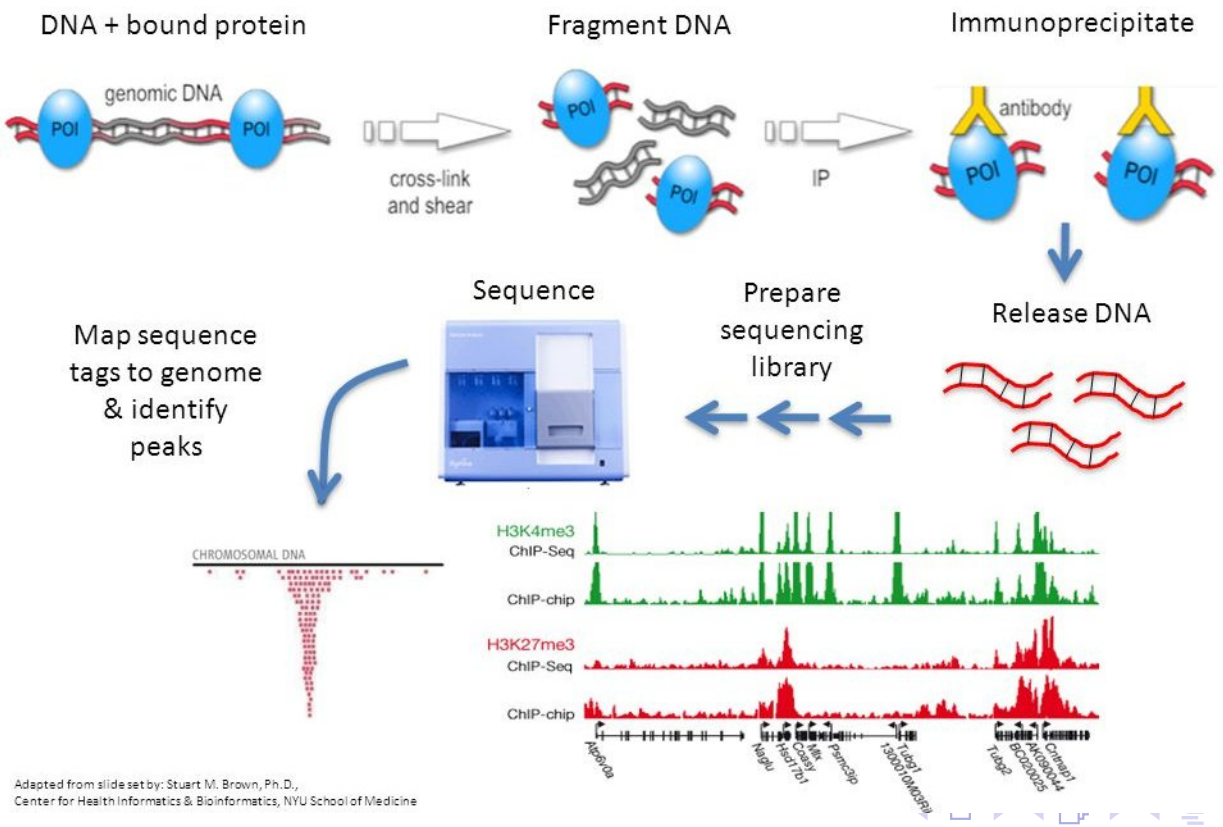


转录组学 | 研究方法 | ChIP-on-chip

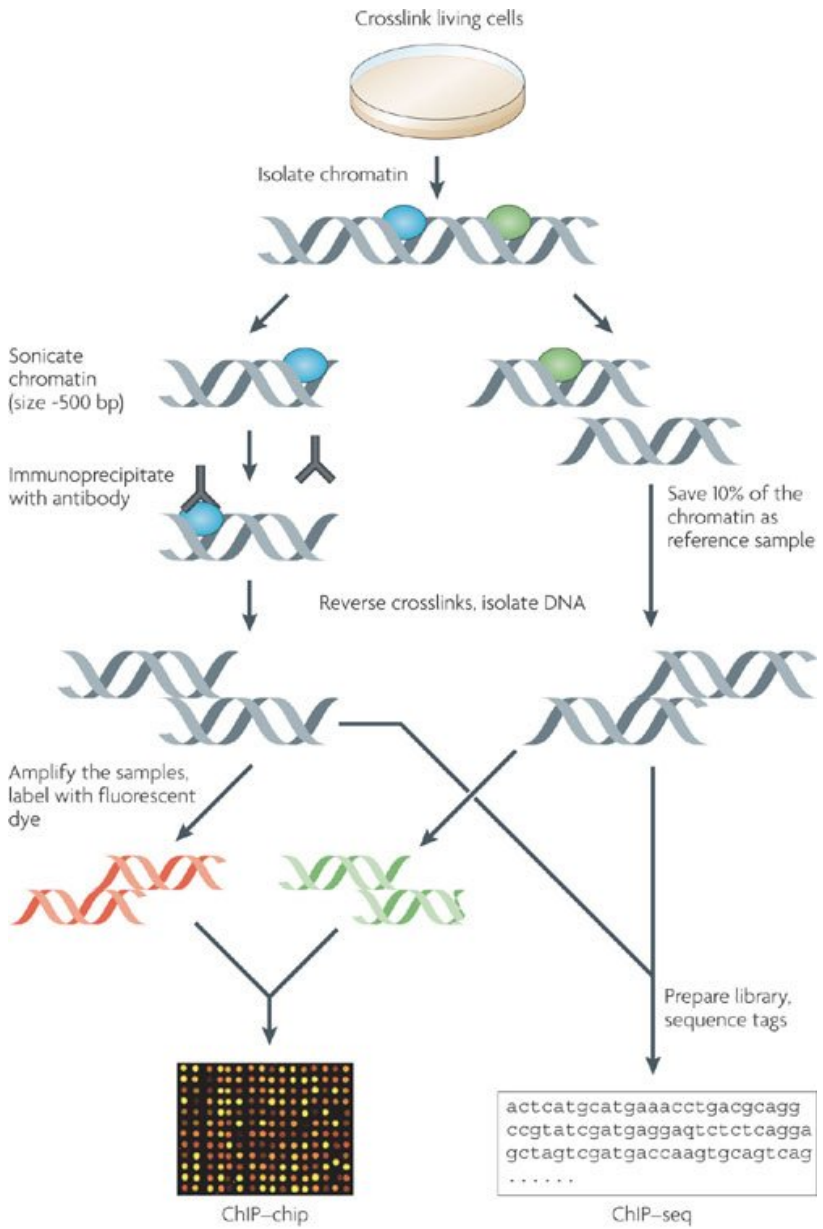


ChIP-Seq

染色质免疫沉淀-测序（ChIP-sequencing，简称为 ChIP-seq）被用于分析蛋白质与 DNA 的交互作用。该技术将染色质免疫沉淀（ChIP）与大规模并行 DNA 测序结合起来以鉴定与 DNA 相关蛋白的结合部位。其可被用于精确绘制任意目的蛋白在全基因组上的结合位点。在此之前，ChIP-on-chip 是研究这些蛋白-DNA 联系的最常用的技术。



转录组学 | 研究方法 | ChIP-Seq vs. ChIP-chip

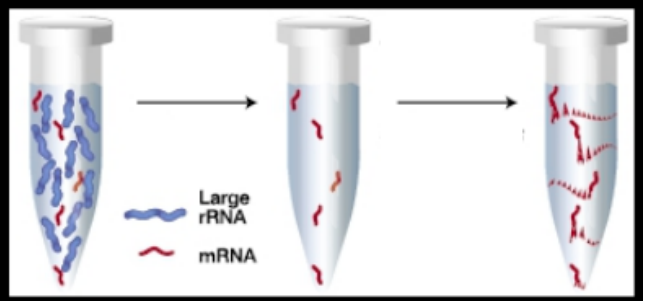


教学提纲

教学提纲

RNA

- RNA in cells consists of
  - 95% ribosomal rRNA and tRNA
  - other non-coding ncRNA
  - protein coding mRNA



- Sequence is transcribed from genome but
  - Introns spliced out
  - mRNA is polyadenylated (“A”s added to end)

RNA-Seq

RNA 测序（RNA sequencing，简称 RNA-Seq，也被称为全转录物组鸟枪法测序，Whole Transcriptome Shotgun Sequencing，简称 WTSS）是基于第二代测序技术的转录组学研究方法。RNA 测序是使用第二代测序的能力，在给定时刻从一个基因组中，揭示 RNA 的存在和数量的一个快照的技术。

RNA-seq (RNA sequencing), also called whole transcriptome shotgun sequencing (WTSS), uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment in time.

RNA-Seq

首先提取生物样品的全部转录的 RNA，然后反转录为 cDNA 后进行二代高通量测序，在此基础上进行片段的重叠组装，从而可得到一个个的转录本。

进而可以形成对该生物样品当前发育状态的基因表达状况的全局了解。

进一步说，若和下一阶段的生物样品的 RNA-Seq 转录组进行比较，则可以得到全部的（在转录层面）基因表达的上调及下调——这就形成了表达谱，针对关键基因则可以形成你想要的通路 (pathway) 的构建。

RNA-Seq

相较于一个静态的染色体而言，细胞内的转录物组是一个处于不断变化的动态过程。随着现在的下一代基因测序（NGS）技术的发展，使得可测得的 DNA 碱基覆盖面增加且样本输出的吞吐量增大。

有助于对细胞内 RNA 转录物进行测序，提供包括选择性剪接转录本、转录后修饰、基因融合、突变/SNPs 以及基因表达量改变等细节。

RNA 测序不仅能检测 mRNA 的转录，还能观测到包括总 RNA 和小 RNA（miRNA、tRNA 和核糖体 RNA）在内不同 RNA 群体的表达谱。RNA 测序还能用来确定外显子/内含子的边界，修正之前注释的 5’ 和 3’ 端基因边界。未来的 RNA 测序研究还包括观察感染时细胞传导路径的变化和癌症中不同基因表达程度。

技术

下一代基因测序之前，对转录物组学和基因表达的研究主要基于基因表达芯片（微阵列），后者包含数以千计用于探测靶向序列的 DNA 探针，可以得到所有表达出转录物的表达谱。基因表达芯片之后，基因表达的系列分析（SAGE）是主要的基因分析技术。

Prior to RNA-Seq, gene expression studies were done with hybridization-based microarrays. Issues with microarrays include cross-hybridization artifacts, poor quantification of lowly and highly expressed genes, and needing to know the sequence *a priori*. Because of these technical issues, transcriptomics transitioned to sequencing-based methods. These progressed from Sanger sequencing of Expressed Sequence Tag libraries, to chemical tag-based methods (e.g., serial analysis of gene expression), and finally to the current technology, NGS of cDNA (notably RNA-Seq).

教学提纲

poly(A)-poly(T)

Frequently, in mRNA analysis the 3' polyadenylated (poly(A)) tail is targeted in order to ensure that coding RNA is separated from noncoding RNA. This can be accomplished simply with poly (T) oligos covalently attached to a given substrate. Presently many studies utilize magnetic beads for this step.

poly(T) & rRNA

Studies including portions of the transcriptome outside poly(A) RNAs have shown that when using poly(T) magnetic beads, the flow-through RNA (non-poly(A) RNA) can yield important noncoding RNA gene discovery which would have otherwise gone unnoticed.

Also, since ribosomal RNA represents over 90% of the RNA within a given cell, studies have shown that its removal via probe hybridization increases the capacity to retrieve data from the remaining portion of the transcriptome.

poly(A)-poly(T)

Frequently, in mRNA analysis the 3' polyadenylated (poly(A)) tail is targeted in order to ensure that coding RNA is separated from noncoding RNA. This can be accomplished simply with poly (T) oligos covalently attached to a given substrate. Presently many studies utilize magnetic beads for this step.

poly(T) & rRNA

Studies including portions of the transcriptome outside poly(A) RNAs have shown that when using poly(T) magnetic beads, the flow-through RNA (non-poly(A) RNA) can yield important noncoding RNA gene discovery which would have otherwise gone unnoticed.

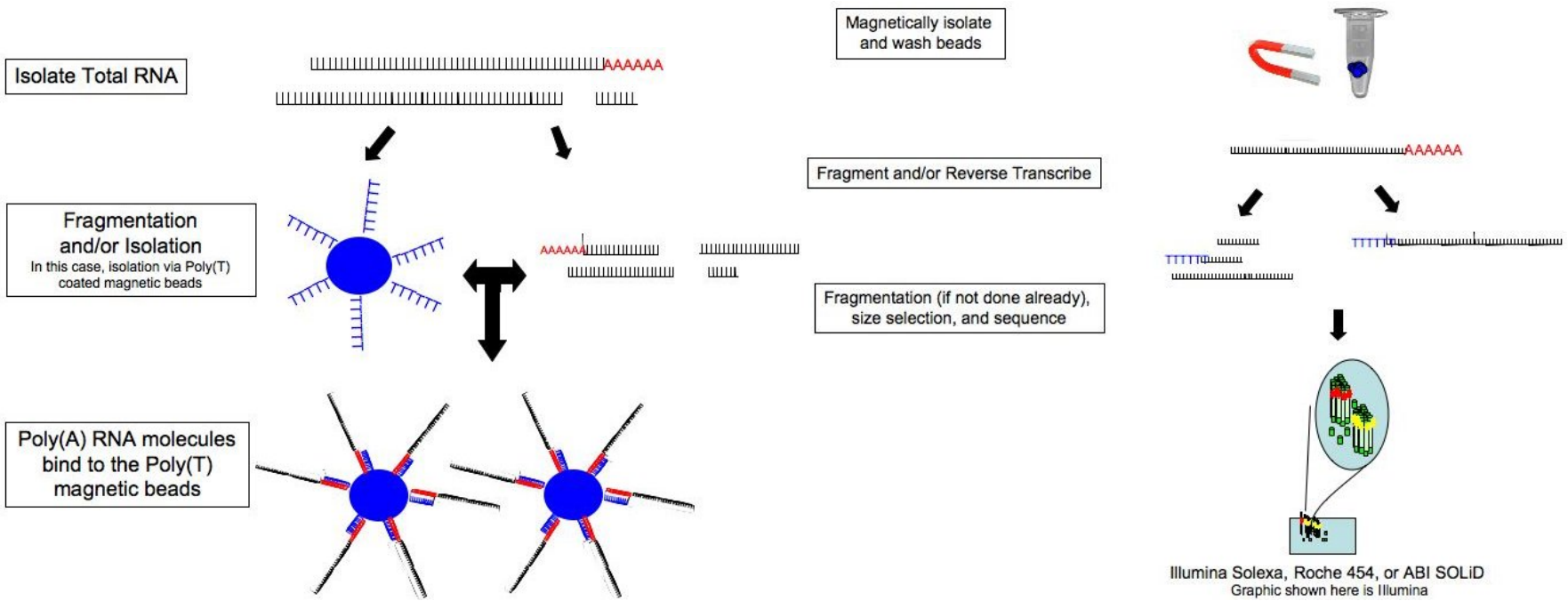
Also, since ribosomal RNA represents over 90% of the RNA within a given cell, studies have shown that its removal via probe hybridization increases the capacity to retrieve data from the remaining portion of the transcriptome.

reverse transcription

Due to the 5' bias of randomly primed-reverse transcription as well as secondary structures influencing primer binding sites, hydrolysis of RNA into 200-300 nucleotides prior to reverse transcription reduces both problems simultaneously. However, there are trade-offs with this method where although the overall body of the transcripts are efficiently converted to DNA, the 5' and 3' ends are less so. Depending on the aim of the study, researchers may choose to apply or ignore this step.

Once the cDNA is synthesized it can be further fragmented to reach the desired fragment length of the sequencing system.

转录组学 | RNA-Seq | Method | RNA Poly(A) library



size selection

When sequencing RNA other than mRNA, the library preparation is modified. The cellular RNA is selected based on the desired size range. For small RNA targets, such as miRNA, the RNA is isolated through size selection. This can be performed with a size exclusion gel, through size selection magnetic beads, or with a commercially developed kit.

Once isolated, linkers are added to the 3' and 5' end then purified.

The final step is cDNA generation through reverse transcription.

转录组学 | RNA-Seq | Method | Direct RNA sequencing

DRSTM

As converting RNA into cDNA using reverse transcriptase has been shown to introduce biases and artifacts that may interfere with both the proper characterization and quantification of transcripts, single molecule Direct RNA Sequencing (DRSTM) technology was under development by Helicos (now bankrupt). DRSTM sequences RNA molecules directly in a massively-parallel manner without RNA conversion to cDNA or other biasing sample manipulations such as ligation and amplification.

Transcriptome Assembly

- RNA-Seq
  - Reference genome
  - Reference transcriptome
- RNA-Seq
  - Reference genome
  - No reference transcriptome
- RNA-Seq
  - No reference genome
  - No reference transcriptome

2 methods

Two different assembly methods are used for producing a transcriptome from raw sequence reads: genome-guided and *de-novo*.

genome-guided

An “easier” and relatively computationally cheaper approach is that of aligning the millions of reads to a “reference genome”. There are many tools available for aligning genomic reads to a reference genome, however, special attention is needed when aligning a transcriptome to a genome, mainly when dealing with genes having intronic regions. Several software packages exist for short read alignment, and recently specialized algorithms for transcriptome alignment have been developed, e.g. Bowtie for RNA-seq short read alignment, TopHat for aligning reads to a reference genome to discover splice sites, Cufflinks to assemble the transcripts and compare/merge them with others, or FANSe. These tools can also be combined to form a comprehensive system.

de novo

This approach does not rely on the presence of a reference genome in order to reconstruct the nucleotide sequence. Due to the small size of the short reads, *de novo* assembly may be difficult, though some software does exist (Velvet (algorithm), Oases, and Trinity to mention a few), as there cannot be large overlaps between each read needed to easily reconstruct the original sequences. The deep coverage also makes the computing power to track all the possible alignments prohibitive. This deficit can be improved using longer sequences obtained from the same sample using other techniques such as Sanger sequencing, and using larger reads as a “skeleton” or a “template” to help assemble reads in difficult regions (e.g. regions with repetitive sequences).

Notes

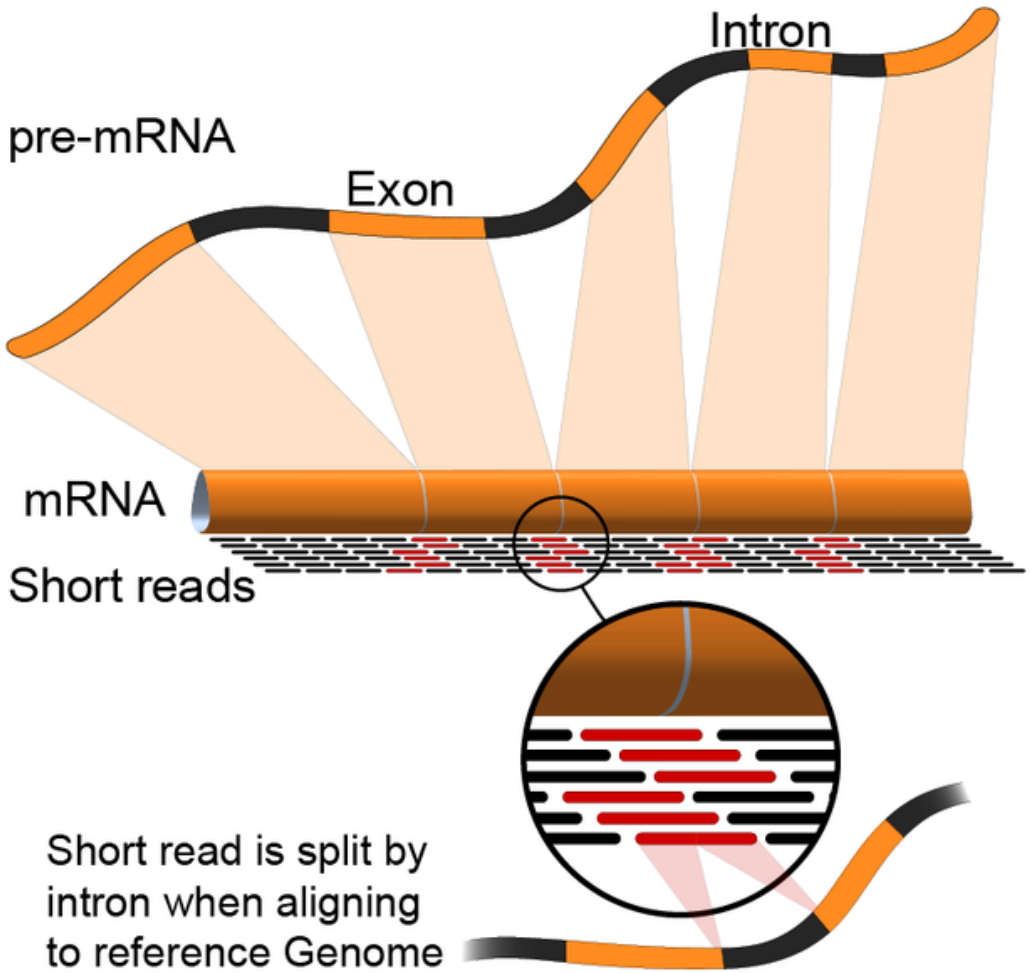
- assembly quality can vary a lot depending on which metric is used
- assemblies that scored well in one species did not really perform well in the other species
- the “most reliable” assembly could be then obtained by combining different approaches

junction

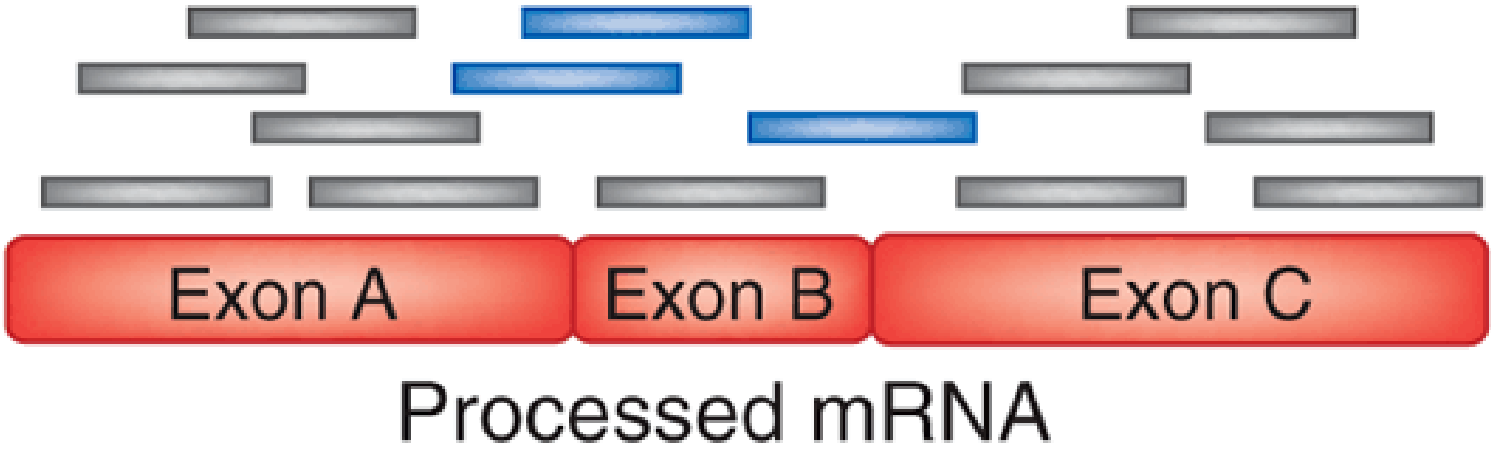
The created library and the short reads obtained cannot come from intronic sequences, so library reads spanning the junction of two or more exons will not align to the genome.

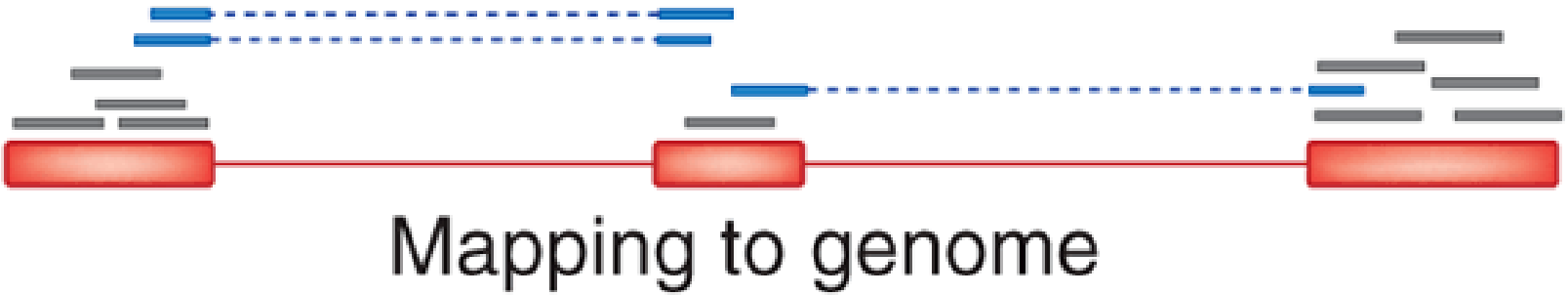
A possible method to work around this is to try to align the unaligned short reads using a proxy genome generated with known exonic sequences. This need not cover whole exons, only enough so that the short reads can match on both sides of the exon-exon junction with minimum overlap.

转录组学 | RNA-Seq | Method | Transcriptome assembly



转录组学 | RNA-Seq | Method | Transcriptome assembly



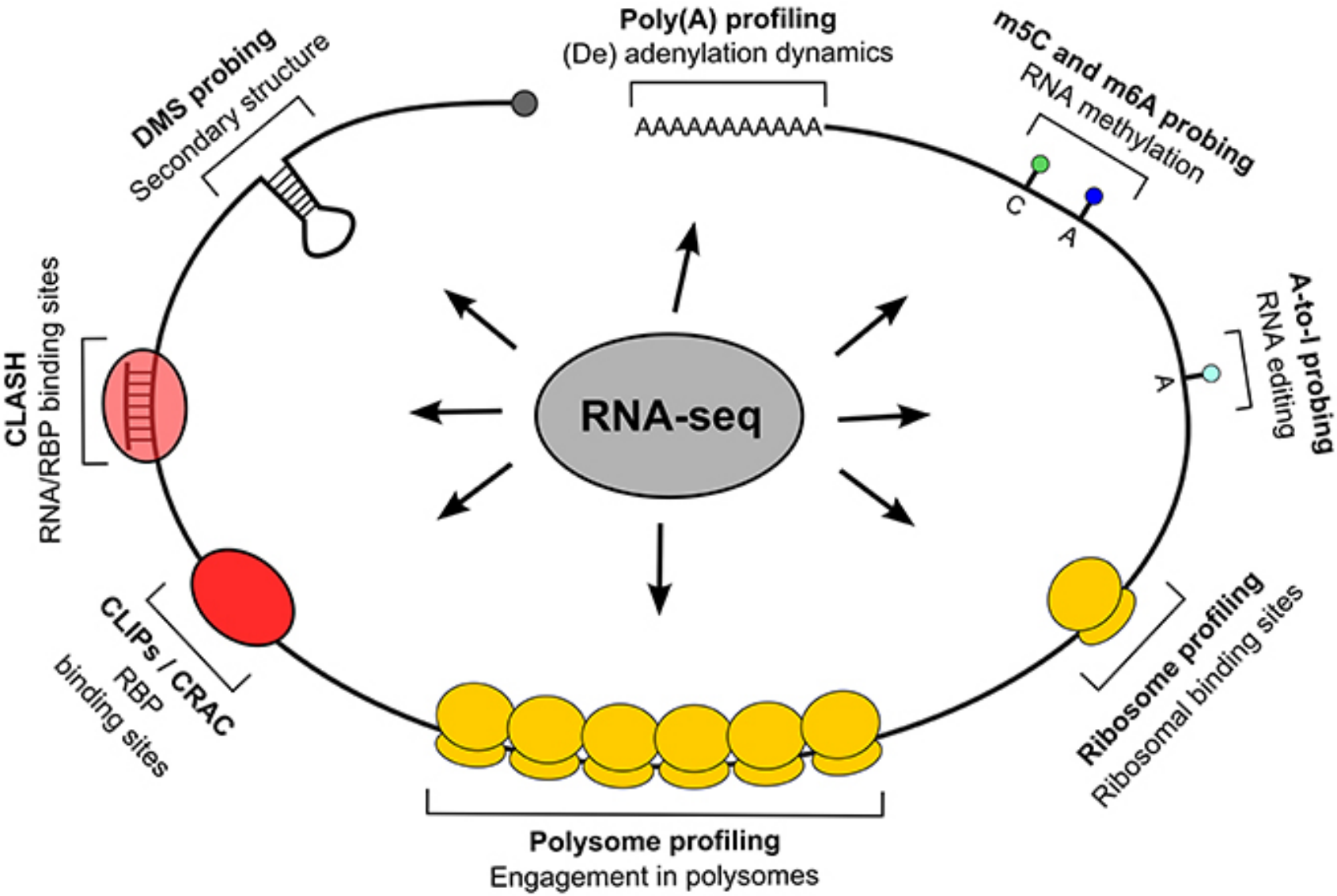


转录组学 | RNA-Seq | Method | Experimental considerations

pros & cons

- Tissue specificity: Gene expression is not uniform throughout an organism's cells, it is strongly dependent on the tissue type being measured. RNA-Seq can provide a complete snapshot of all the transcripts being available at that precise moment in the cell.
- Time dependent: During a cell's lifetime and context, its gene expression levels change. Any single sequencing experiment will offer information regarding one point in time.
- Coverage: coverage/depth can affect the mutations seen.
- Subjectivity of the analysis: Numerous attempts have been taken to uniformly analyze the data. However, the results can vary due to the multitude of algorithms and pipelines available.
- Data management: The main issue with NGS data is the volume of data produced.
- Downstream interpretation of the data: Different layers of interpretations have to be considered when analyzing RNA-Seq data.

- RNA-Seq
  - Transcriptome assembly
    - Qualitative identification of expressed sequence
  - Differential expression analysis
    - Quantitative measurement of transcript expression
- RNA-Seq Applications
  - **Annotation**: Identify novel genes, transcripts, exons, splicing events, ncRNAs
  - Detecting RNA editing and SNPs
  - **Measurements: RNA quantification and differential gene expression**


Poly(A) profiling
(De) adenylation dynamics
AAAAAAAAAAA
DMS probing
Secondary structure
m5C and m6A probing
RNA methylation
C
A
A-to-I probing
RNA editing
A
CLASH
RNA/RBP binding sites
RNA-seq
CLIPs / CRAC
RBP
binding sites
Ribosome profiling
Ribosomal binding sites
Polysome profiling
Engagement in polysomes

RNA-Seq applications

- Sequencing RNA transcripts, applications:
  - *de novo* transcriptome
  - **gene-wise differential expression**
  - novel splice variants
  - differential splicing
  - fusion genes
  - non-coding RNA
  - polyadenylation length
  - post-transcriptional modification
  - ...
- Differential expression analysis most common
  - Commonly called "DGE"

RNA-seq Applications

- **Annotation**
  - ➢ Identify novel genes, transcripts, exons, splicing events, ncRNAs.
- Detecting RNA editing and SNPs.
- **Measurements: RNA quantification and differential gene expression**
  - ➢ Abundance of transcripts between different conditions

Why do an RNA-seq experiment?

Detect differential expression

Assess allele-specific expression

Quantify alternative transcript usage

Discover novel genes/transcripts, gene fusions, circRNA

Profile transcriptome

Ribosome profiling to measure translation

Why do an RNA-seq experiment?

Detect differential expression

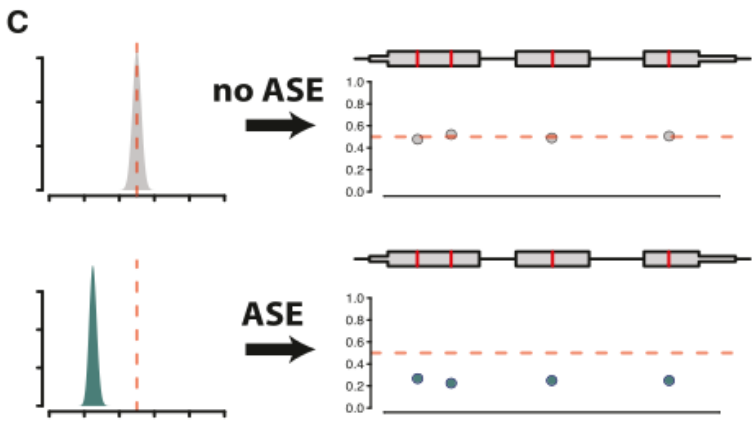
Assess allele-specific expression

Quantify alternative transcript usage

Discover novel genes/transcripts, gene fusions, circRNA

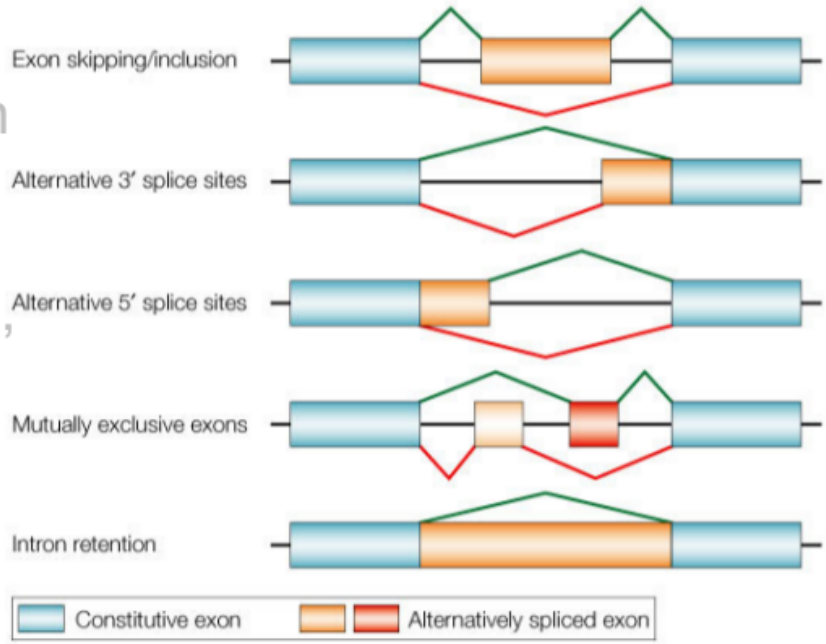
Profile transcriptome

Ribosome profiling to measure translation



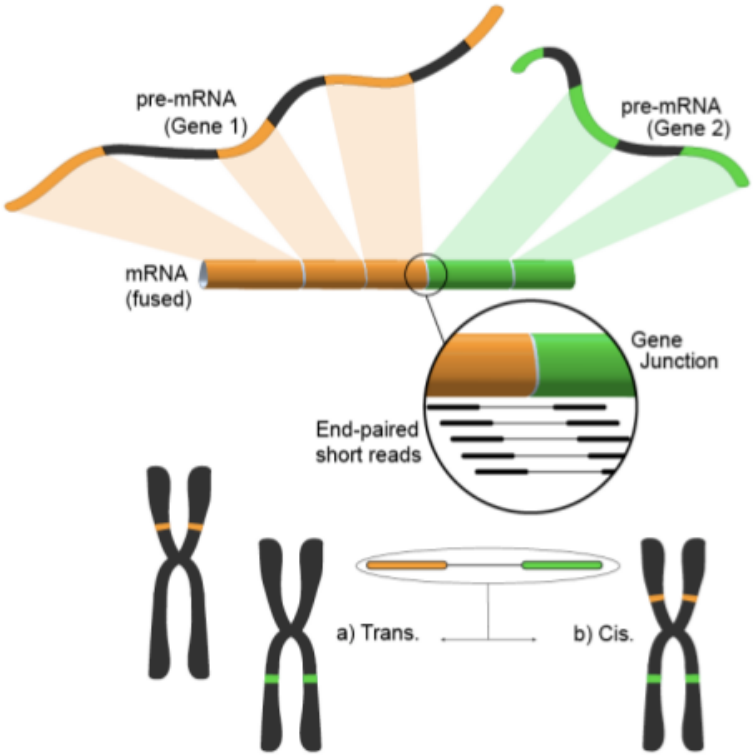
Why do an RNA-seq experiment?

Detect differential expression
Assess allele-specific expression
Quantify alternative transcript usage
Discover novel genes/transcripts, gene fusions, circRNA
Profile transcriptome
Ribosome profiling to measure translation



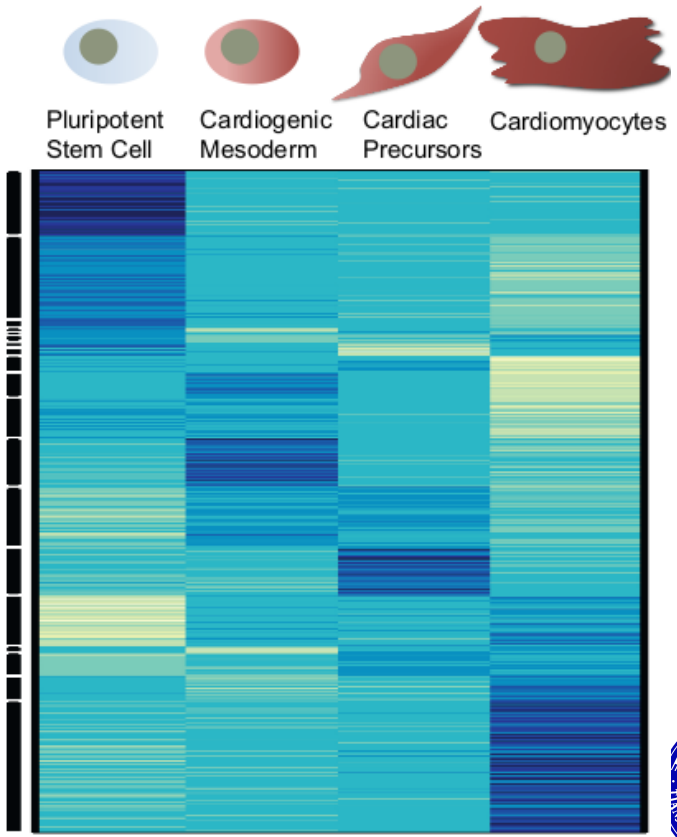
Why do an RNA-seq experiment?

Detect differential expression
Assess allele-specific expression
Quantify alternative transcript usage
Discover novel genes/transcripts, gene fusions, circRNA
Profile transcriptome
Ribosome profiling to measure translation



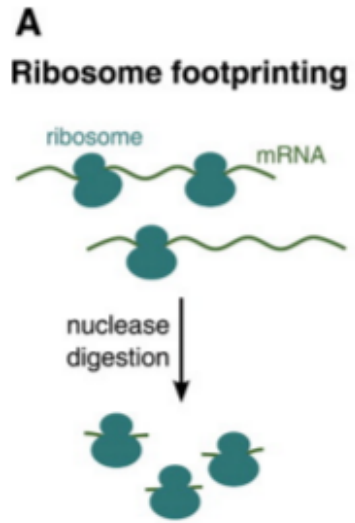
Why do an RNA-seq experiment?

Detect differential expression
Assess allele-specific expression
Quantify alternative transcript usage
Discover novel genes/transcripts, gene fusions, circRNA
Profile transcriptome
Ribosome profiling to measure translation



Why do an RNA-seq experiment?

Detect differential expression
Assess allele-specific expression
Quantify alternative transcript usage
Discover novel genes/transcripts, gene fusions, circRNA
Profile transcriptome
Ribosome profiling to measure translation (Ribo-seq)



Experimental design

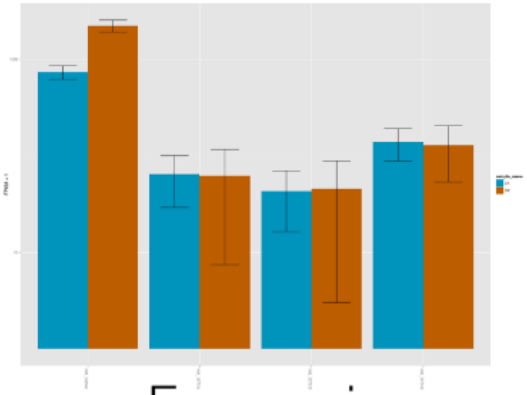
- What are my goals?
  - Transcriptome assembly?
  - Differential expression analysis?
  - Identify rare transcripts?
- What are the characteristics of my system?
  - Large, complex genome?
  - Introns and high degree of alternative splicing?
  - No reference genome or transcriptome?

Experimental Outputs

Left reads:
Input : 11607291
Mapped : 11606003 (100.0% of input)
of these: 57648 (0.5%) have multiple alignments (104 have >20)
Right reads:
Input : 11607291
Mapped : 11606001 (100.0% of input)
of these: 57647 (0.5%) have multiple alignments (104 have >20)
100.0% overall read mapping rate.

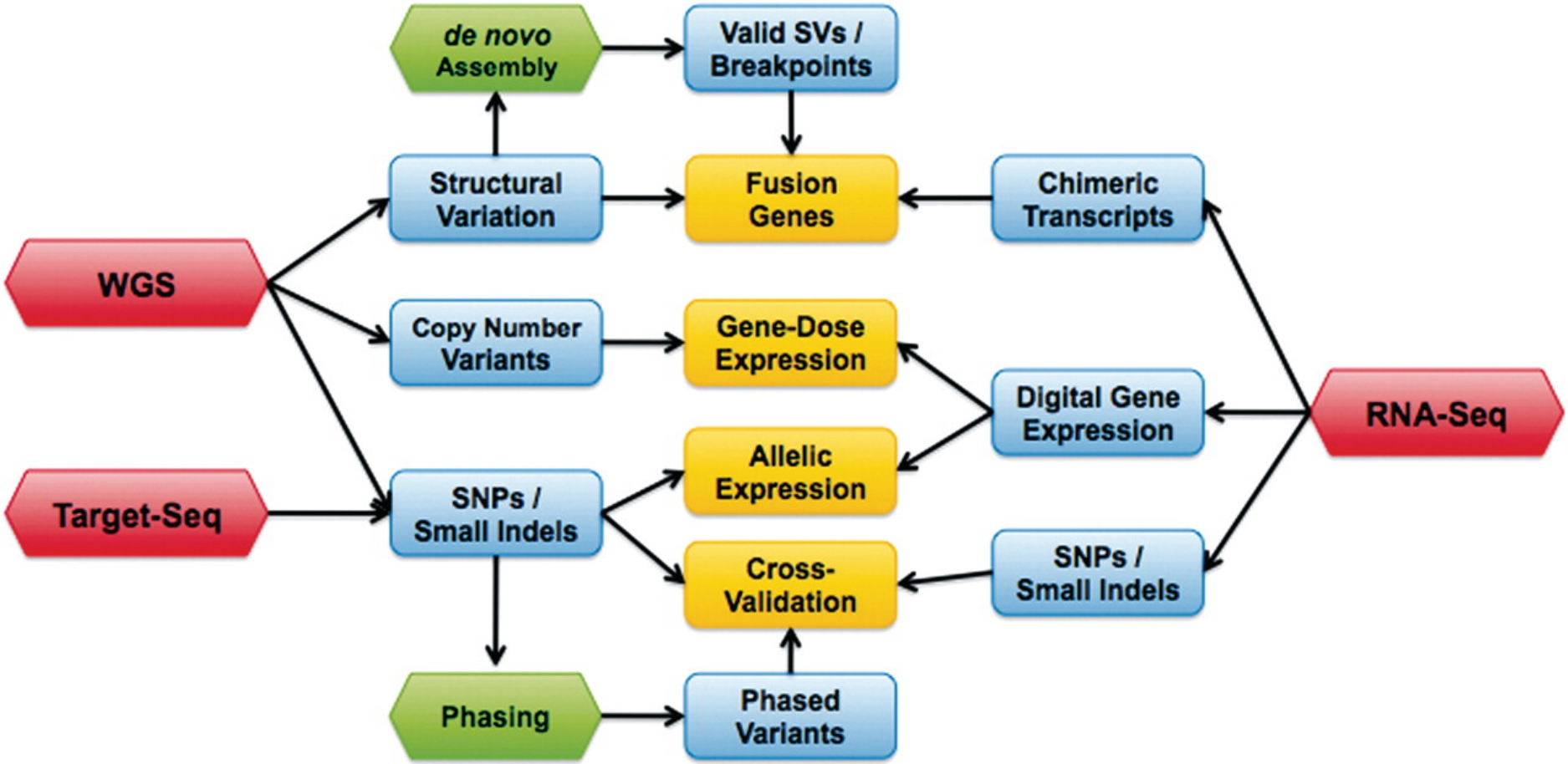
Aligned pairs: 11604800
of these: 57644 (0.5%) have multiple alignments
324 (0.0%) are discordant alignments
100.0% concordant pair alignment rate.

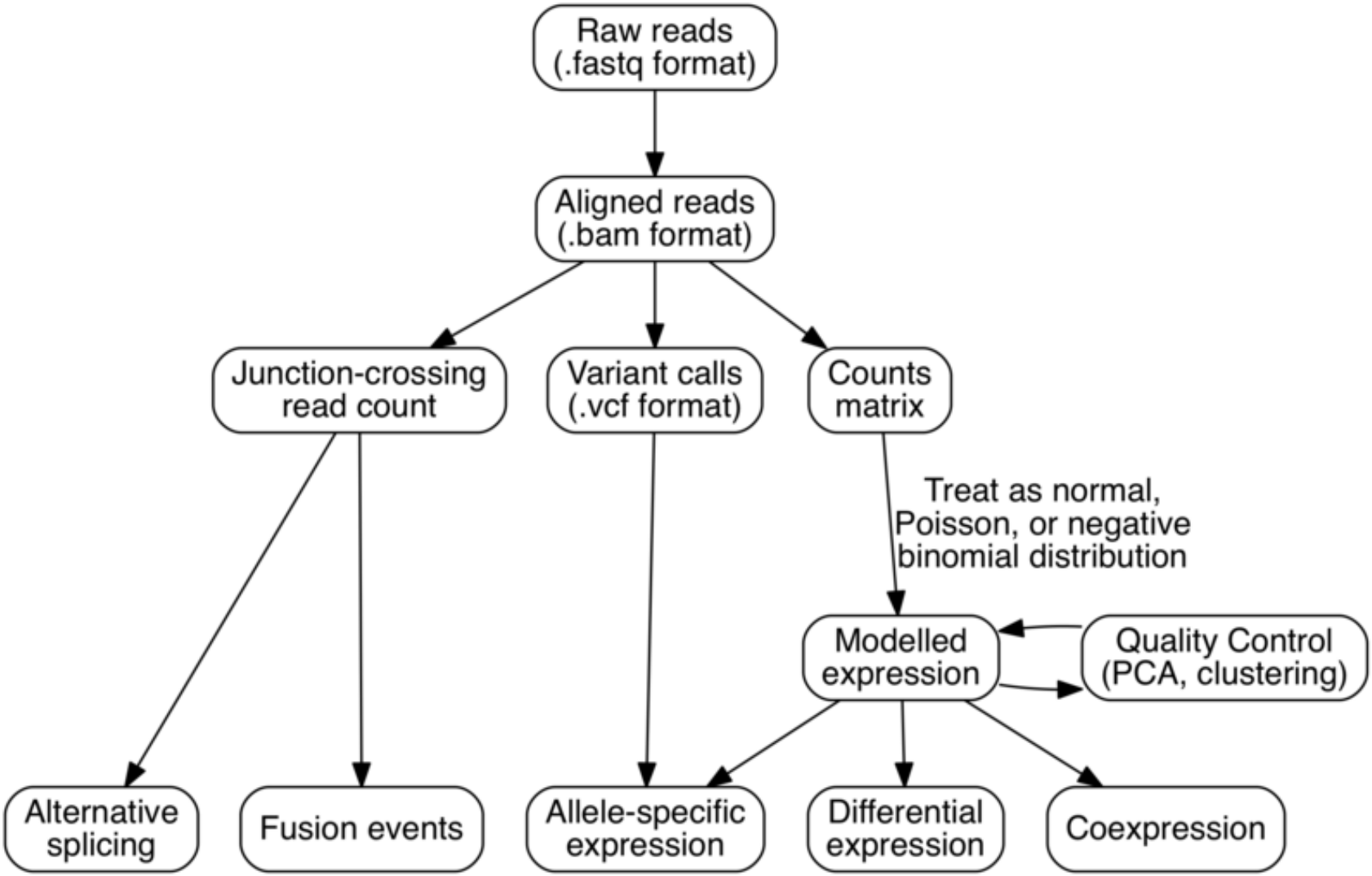
Assembly



Expression
Differentially expressed

Splicing


de novo Assembly
Valid SVs / Breakpoints
Structural Variation
Fusion Genes
Chimeric Transcripts
WGS
Copy Number Variants
Gene-Dose Expression
Digital Gene Expression
RNA-Seq
Allelic Expression
Target-Seq
SNPs / Small Indels
Cross-Validation
SNPs / Small Indels
Phasing
Phased Variants

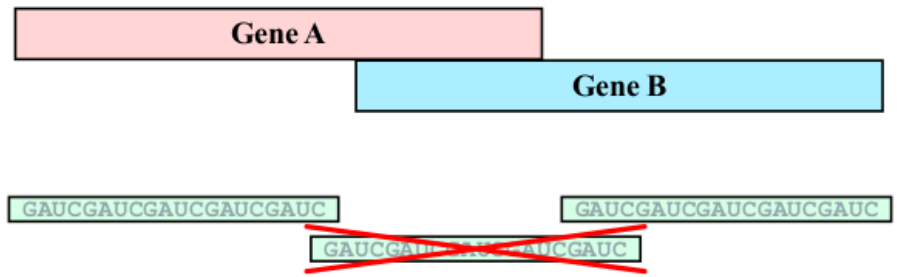

Raw reads
(.fastq format)
Aligned reads
(.bam format)
Junction-crossing
read count
Variant calls
(.vcf format)
Counts
matrix
Treat as normal,
Poisson, or negative
binomial distribution
Modelled
expression
Quality Control
(PCA, clustering)
Alternative
splicing
Fusion events
Allele-specific
expression
Differential
expression
Coexpression

Gene expression

- Which genes are expressed in what tissues, and at what levels.
- Measuring mRNA concentration levels is a useful tool in determining how the transcriptional machinery of the cell is affected in the presence of external signals (e.g. drug treatment), or how cells differ between a healthy state and a diseased state.
- Expression levels are expressed as Fragments Per Kilobase of transcript per Million mapped reads (FPKM).
- An R-based statistical package known as CummeRbund can be used to generate expression comparison charts for visual analysis.

Measure expression levels in RNA-Seq data

1. Align read to reference genome
2. Measuring expression = counting aligned reads
   - Count in annotated exons
   - Positive integers (read counts of 3.1415 or −42 are impossible)
   - Quantitative (read count has an absolute meaning)
- Observation (read count) must be statistically independent
  - No multi-map reads
  - Skip overlapping gene annotations



Differential expression

RNA-Seq is generally used to compare gene expression between conditions, such as a drug treatment vs non-treated, and find out which genes are up- or down regulated in each condition.

Differently expressed genes can be identified by using tools that count the sequencing reads per gene and compare them between samples. The most commonly used tools for this type of analysis are DESeq and edgeR, packages from Bioconductor. Both these tools use a model based on the negative binomial distribution.

Differential gene expression analysis tools

- Alignment
  - TopHat [15, 5]
  - STAR [3]
  - ... many many more
- Measuring expression (quantification)
  - HTSeq-count [2]
  - Cufflinks [16]
  - featureCounts [8]
- Group-wise comparison (hypothesis testing)
  - EdgeR [12]
  - DESeq2 [11]
  - Cuffdiff [14]

Measure expression levels in RNA-Seq data

In Galaxy

- featureCounts [8]
  - Pro's
    - Fast
    - Flexible
    - Free (GPL)
    - Accepts both BAM and SAM files
    - Only requires name-sorted files when mate-pairs are counted together (name-sorting is slow)
  - Con's
    - Built-in name-sorting supports no threading – rather do this with samtools [7]

Differential gene expression analysis

edgeR

- edgeR [12]
  - Differential gene expression analysis
  - Free R Package (GPL2)
  - Galaxy wrapper does normalizations for you
    - Use raw reads, do NOT use FPKM/RPKM!
- "Limma" for count data
  - Not Gaussian (normal) distributed like e.g. micro-array data – but negative binomial

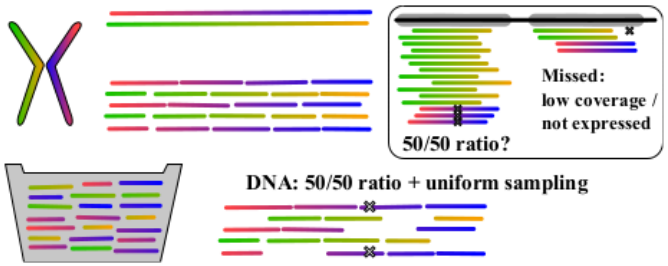
Absolute quantification of transcripts

It is not possible to do absolute quantification using the common RNA-Seq pipeline, because it only provides RNA levels relative to all transcripts. If the total amount of RNA in the cell changes between conditions, relative normalization will misrepresent the changes for individual transcripts. Absolute quantification of mRNAs is possible by performing RNA-Seq with added spike ins, samples of RNA at known concentrations. After sequencing, the read count of the spike ins sequences is used to determine the direct correspondence between read count and biological fragments.

SNV discovery

RNA-seq is limited to transcribed regions however, since it will only discover sequence variations in exon regions. This misses many subtle but important intron alleles that affect disease such as transcription regulators, leaving analysis to only large effectors. While some correlation exists between exon to intron variation, only whole genome sequencing would be able to capture the source of all relevant SNPs.

Single Nucleotide Polymorphisms in RNA-Seq



- Major difference(s) between DNA-Seq:
  - Detected SNPs are expressed
    - Biological context
    - SNPs RNA-Seq only within exons and ncRNAs
    - Allele specific expression profiles
- Detection:
  - Expression affects coverage; in DNA-seq coverage should be uniform

Single Nucleotide Polymorphisms in RNA-Seq

Using: Samtools, VarScan

	A	C	T	G	A	
reference						
read1	a	c	c	g	c	Alignment
read2	a	c	t	g	a	
read3	a	c	c	g	a	
read4	a	c	c	g	a	
read5	a	c	t	a	a	
read6		c	c	g	a	
read7			c	g	a	
read quality (q)	0.99	0.99	0.85	0.8	0.99	
aligned	5	6	7	7	7	Expected (based on quality)
q*aligned	4.95	5.94	5.95	5.6	6.93	
(1-q)*aligned	0.05	0.06	1.05	1.4	0.07	
exp match (abs)	5	6	6	6	7	
exp mismatch (abs)	0	0	1	1	0	
obs match	5	6	2	6	6	Observed
obs mismatch	0	0	5	1	1	
P(obs|exp) fisher exact	1.000	1.000	0.049	0.538	0.500	Hypothesis testing
P < 0.05	REF	REF	SNP	REF	REF	

Single Nucleotide Polymorphisms in RNA-Seq

Detection tools

- Alignment
  - TopHat [15, 5]
  - STAR [3]
  - ... many many more
- SNV calling
  - VarScan2 [6]
  - samtools [7]
  - exactSNP *(part of subread [9] package)*
  - GATK [17]

Post-transcriptional edits

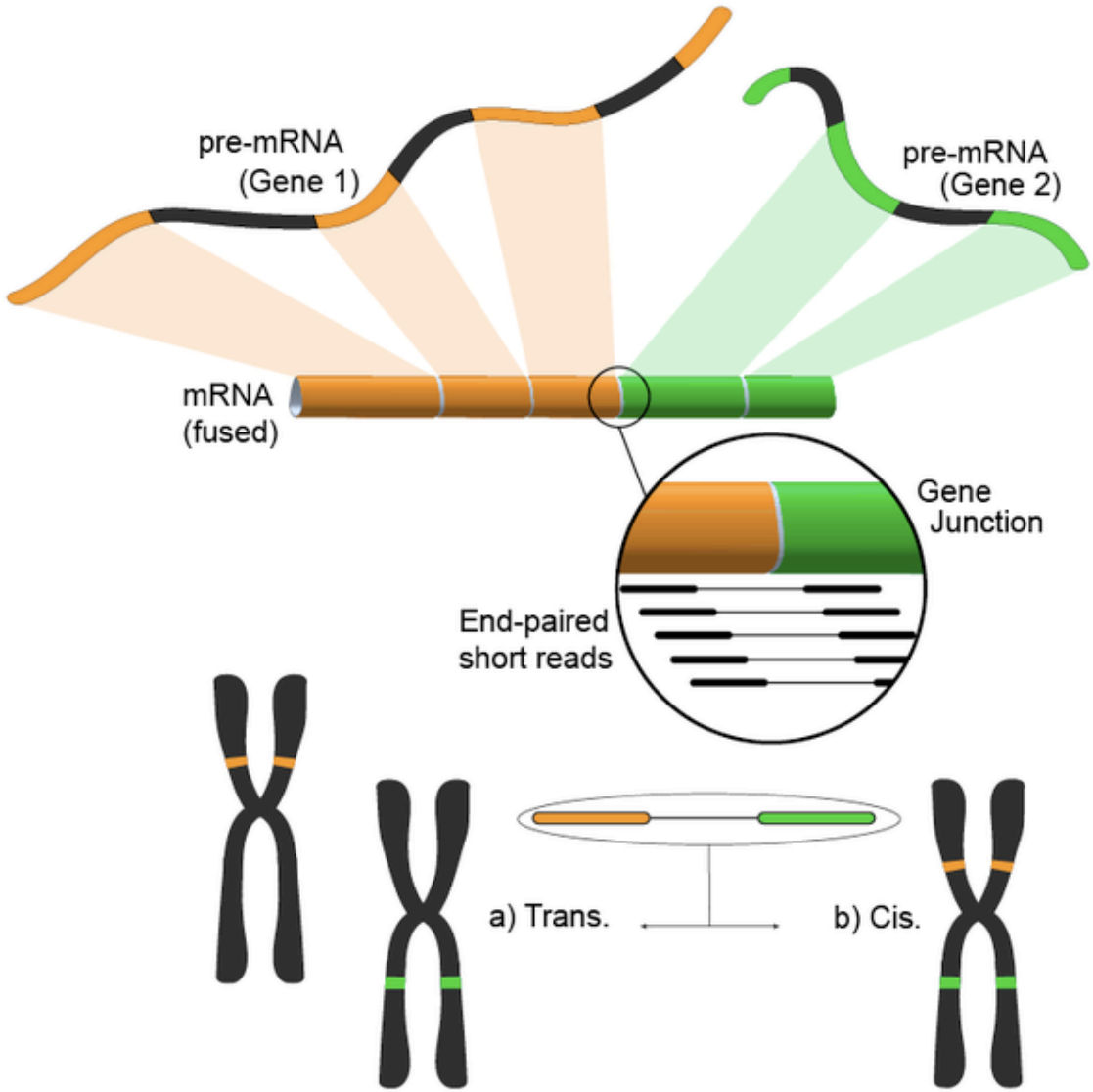
Having the matching genomic and transcriptomic sequences of an individual can also help in detecting post-transcriptional edits, where, if the individual is homozygous for a gene, but the gene's transcript has a different allele, then a post-transcriptional modification event is determined.

mRNA centric single nucleotide variants (SNVs) are generally not considered as a representative source of functional variation in cells, mainly due to the fact that these mutations disappear with the mRNA molecule, however the fact that efficient DNA correction mechanisms do not apply to RNA molecules can cause them to appear more often. This has been proposed as the source of certain prion diseases, also known as TSE or transmissible spongiform encephalopathies.

Fusion gene detection

Caused by different structural modifications in the genome, fusion genes have gained attention because of their relationship with cancer. The ability of RNA-seq to analyze a sample's whole transcriptome in an unbiased fashion makes it an attractive tool to find these kinds of common events in cancer.

The idea follows from the process of aligning the short transcriptomic reads to a reference genome. Most of the short reads will fall within one complete exon, and a smaller but still large set would be expected to map to known exon-exon junctions. The remaining unmapped short reads would then be further analyzed to determine whether they match an exon-exon junction where the exons come from different genes. This would be evidence of a possible fusion event, however, because of the length of the reads, this could prove to be very noisy. An alternative approach is to use pair-end reads, when a potentially large number of paired reads would map each end to a different exon, giving better coverage of these events. Nonetheless, the end result consists of multiple and potentially novel combinations of genes providing an ideal starting point for further validation.


pre-mRNA
(Gene 1)
pre-mRNA
(Gene 2)
mRNA
(fused)
Gene
Junction
End-paired
short reads
a) Trans.
b) Cis.

Coexpression networks

Coexpression networks are data-derived representations of genes behaving in a similar way across tissues and experimental conditions. Their main purpose lies in hypothesis generation and guilt-by-association approaches for inferring functions of previously unknown genes. RNA-Seq data has been recently used to infer genes involved in specific pathways based on Pearson correlation, both in plants and mammals. The main advantage of RNA-Seq data in this kind of analysis over the microarray platforms is the capability to cover the entire transcriptome, therefore allowing the possibility to unravel more complete representations of the gene regulatory networks.

Co-expression modules may corresponds to cell types or pathways. Highly connected intramodular hubs can be interpreted as representatives of their respective module.

Application to genomic medicine

RNA-Seq data could help researchers interpreting the “personalized transcriptome” so that it will help understanding the transcriptomic changes happening therefore, ideally, identifying gene drivers for a disease. The feasibility of this approach is however dictated by the costs in terms of money and time.

RNA-Seq applications to the clinic have the potentials to significantly affect patient’s life and, on the other hand, requires a team of specialists (bioinformaticians, physicians/clinicians, basic researchers, technicians) to fully interpret the huge amount of data generated by this analysis.

Application to genomic medicine

Compared with microarrays, NGS technology has identified novel and low frequency RNAs associated with disease processes. This advantage aids in the diagnosis and possible future treatments of diseases, including cancer. Numerous studies have demonstrated NGS's ability to detect aberrant mRNA and small non-coding RNA expression in disease processes above that provided by microarrays. The lower cost and higher throughput offered by NGS confers another advantage to researchers.

The role of small non-coding RNAs in disease processes has also been explored in recent years.

ENCODE & TCGA

A lot of emphasis has been given to RNA-Seq data after the Encyclopedia of the regulatory elements (ENCODE) and The Cancer Genome Atlas (TCGA) projects have used this approach to characterize dozens of cell lines and thousands of primary tumor samples, respectively.

RNA-Seq data provide a unique snapshot of the transcriptomic status of the disease and look at an unbiased population of transcripts that allows the identification of novel transcripts, fusion transcripts and non-coding RNAs that could be undetected with different technologies.

ENCODE

ENCODE aimed to identify genome-wide regulatory regions in different cohort of cell lines and transcriptomic data are paramount in order to understand the downstream effect of those epigenetic and genetic regulatory layers.

TCGA

TCGA aimed to collect and analyze thousands of patient's samples from 30 different tumor types in order to understand the underlying mechanisms of malignant transformation and progression.

ENCODE

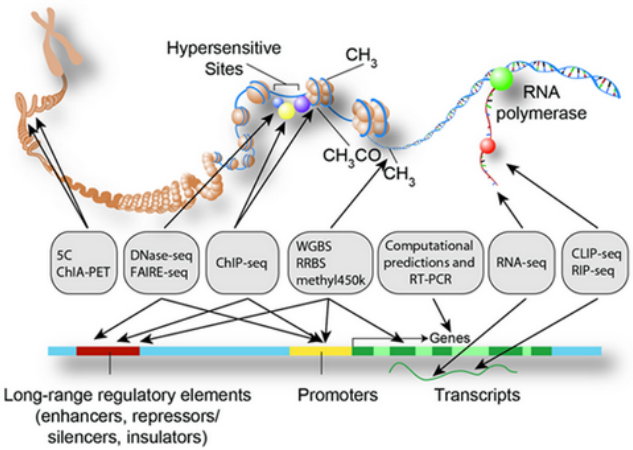
ENCODE aimed to identify genome-wide regulatory regions in different cohort of cell lines and transcriptomic data are paramount in order to understand the downstream effect of those epigenetic and genetic regulatory layers.

TCGA

TCGA aimed to collect and analyze thousands of patient's samples from 30 different tumor types in order to understand the underlying mechanisms of malignant transformation and progression.

ENCODE Data Encyclopedia Materials & Methods Help Search...

ENCODE: Encyclopedia of DNA Elements



The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Quick Start

To find and download ENCODE Consortium data:

- Click the Data toolbar above and browse data
  - By assay
  - By biosample
  - By genomic annotations
- Enter search terms like "skin", "ChIP-seq", or "CTCF"

Additional help using the ENCODE Portal:

- Getting Started

News Follow @EncodeDCC

August 3rd, 2016: With collaborations from the DAC (Data Analysis Center), a total of 336 Annotation File Sets of promoter-like and enhancer-like regions have been released! [read more]

August 1st, 2016: 191 new ENCODE experiments released in July - check them all out here.

July 21st, 2016: Check out the latest GGR release from the Reddy lab, here on the portal, 84 ChIP-seq expts in A549 +/- dex treatment.

July 20th, 2016: Mouse e10.5 histone ChIP-seq dataset from from the Ren lab available on the portal here

ENCODE

DNA 元件百科全书（Encyclopedia of DNA Elements，简称为 ENCODE 项目）是一个由美国国家人类基因组研究所 (NHGRI) 在 2003 年 9 月发起的一项公共联合研究项目，旨在找出人类基因组中所有功能组件。这是继完成人类基因组计划后国家人类基因组研究所开始的最重要的项目之一。

三个阶段

- 试验阶段：测试和比较现有方法以便严格分析一个所定义的人类基因组序列的一部分
- 技术发展阶段：分析整个基因组，并进行“额外中试规模研究”
- 生产阶段：2012 年 9 月 5 日，该项目的初步结果被整理为 30 篇论文并同时发表于多个刊物，包括 6 篇论文在《自然》、6 篇论文在《基因组生物学》及 18 篇论文在《基因组研究》上

ENCODE

DNA 元件百科全书（Encyclopedia of DNA Elements，简称为 ENCODE 项目）是一个由美国国家人类基因组研究所 (NHGRI) 在 2003 年 9 月发起的一项公共联合研究项目，旨在找出人类基因组中所有功能组件。这是继完成人类基因组计划后国家人类基因组研究所开始的最重要的项目之一。

三个阶段

- 试验阶段：测试和比较现有方法以便严格分析一个所定义的人类基因组序列的一部分
- 技术发展阶段：分析整个基因组，并进行“额外中试规模研究”
- 生产阶段：2012 年 9 月 5 日，该项目的初步结果被整理为 30 篇论文并同时发表于多个刊物，包括 6 篇论文在《自然》、6 篇论文在《基因组生物学》及 18 篇论文在《基因组研究》上

Important features about the organization and function of the human genome

- The vast majority (80.4%) of the human genome participates in at least one biochemical RNA and/or chromatin associated event in at least one cell type. Much of the genome lies close to a regulatory event: 95% of the genome lies within 8kb of a DNA-protein interaction (as assayed by bound ChIP-seq motifs or DNaseI footprints), and 99% is within 1.7kb of at least one of the biochemical events measured by ENCODE.
- Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus some of them are expected to be functional.
- Classifying the genome into seven chromatin states suggests an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoters-like features, as well hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands of narrow states with distinct functional properties.

Important features about the organization and function of the human genome

- It is possible to quantitatively correlate RNA sequence production and processing with both chromatin marks and transcription factor (TF) binding at promoters, indicating that promoter functionality can explain the majority of RNA expression variation.
- Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions; this number is at least as large as those that lie in protein coding genes.
- SNPs associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein coding genes. In many cases, the disease phenotypes can be associated with a specific cell type or TF.

Most striking finding

The most striking finding was that the fraction of human DNA that is biologically active is considerably higher than even the most optimistic previous estimates. In an overview paper, the ENCODE Consortium reported that its members were able to assign biochemical functions to over 80% of the genome. Much of this was found to be involved in controlling the expression levels of coding DNA, which makes up less than 1% of the genome.

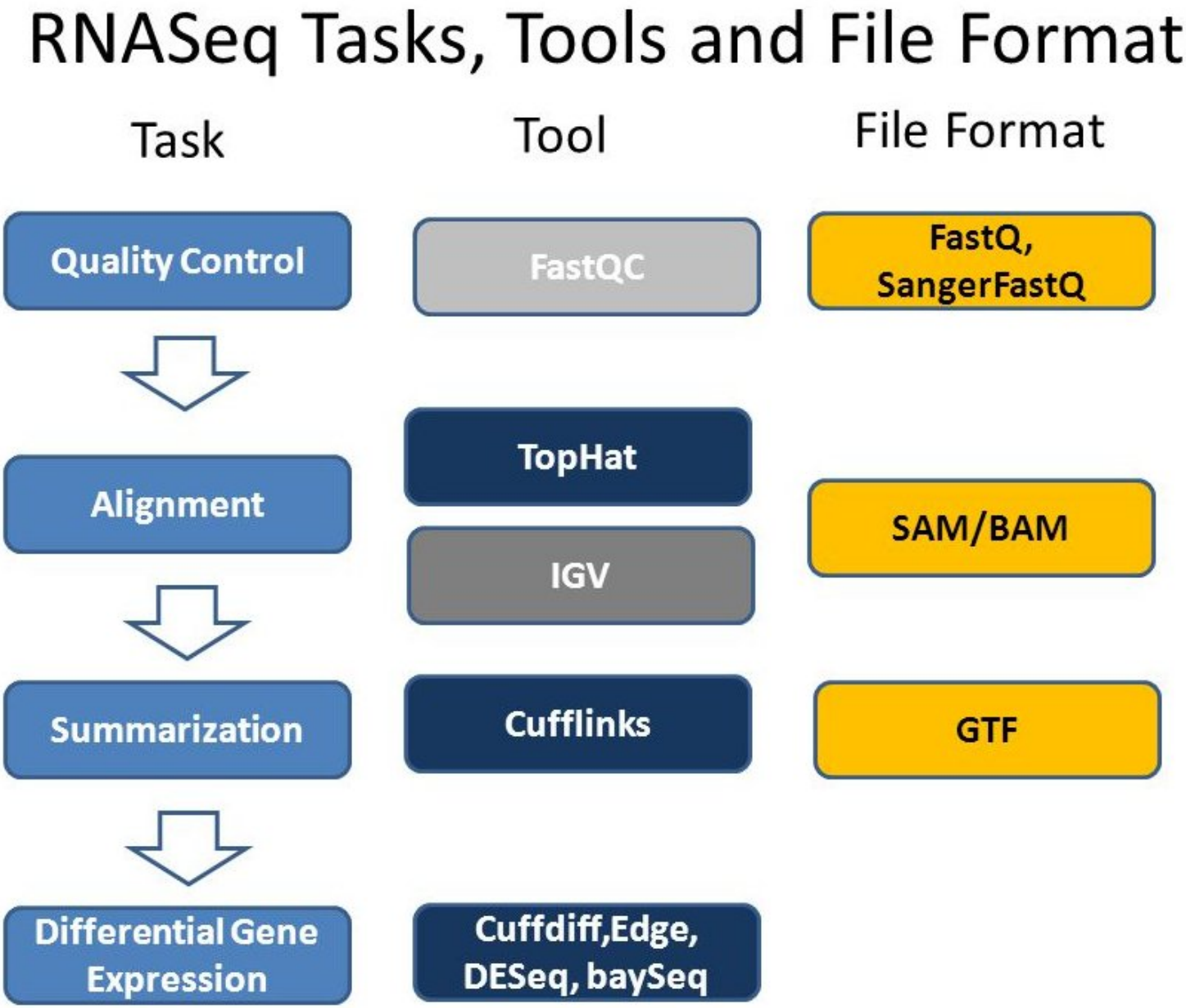
Most important new elements of the “encyclopedia”

- A comprehensive map of DNase 1 hypersensitive sites, which are markers for regulatory DNA that is typically located adjacent to genes and allows chemical factors to influence their expression. The map identified nearly 3 million sites of this type, including nearly all that were previously known and many that are novel.
- A lexicon of short DNA sequences that form recognition motifs for DNA-binding proteins. Approximately 8.4 million such sequences were found, comprising a fraction of the total DNA roughly twice the size of the exome. Thousands of transcription promoters were found to make use of a single stereotyped 50-base-pair footprint.

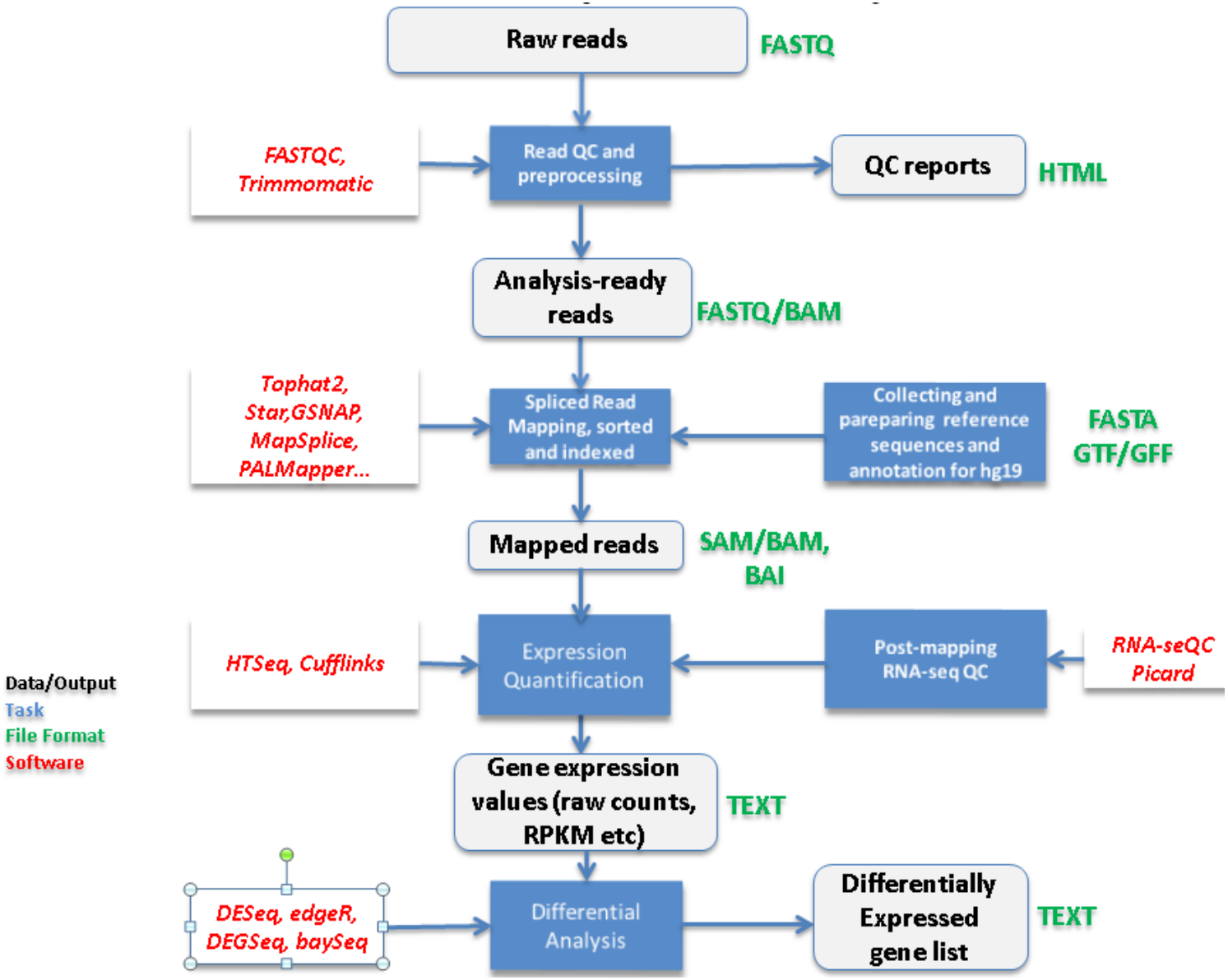
Most important new elements of the “encyclopedia”

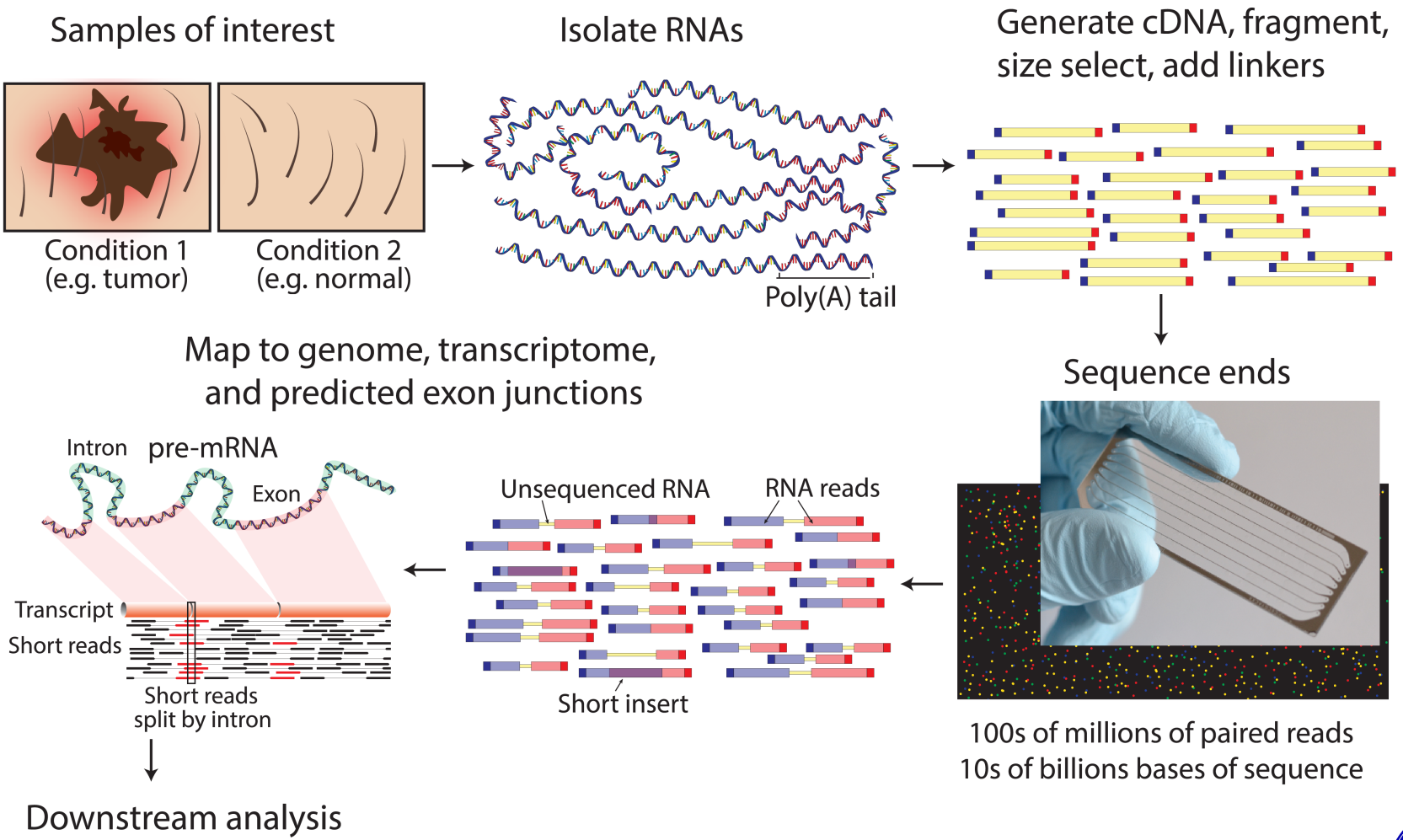
- A preliminary sketch of the architecture of the network of human transcription factors, that is, factors that bind to DNA in order to promote or inhibit the expression of genes. The network was found to be quite complex, with factors that operate at different levels as well as numerous feedback loops of various types.
- A measurement of the fraction of the human genome that is capable of being transcribed into RNA. This fraction was estimated to add up to more than 75% of the total DNA, a much higher value than previous estimates. The project also began to characterize the types of RNA transcripts that are generated at various locations.

教学提纲

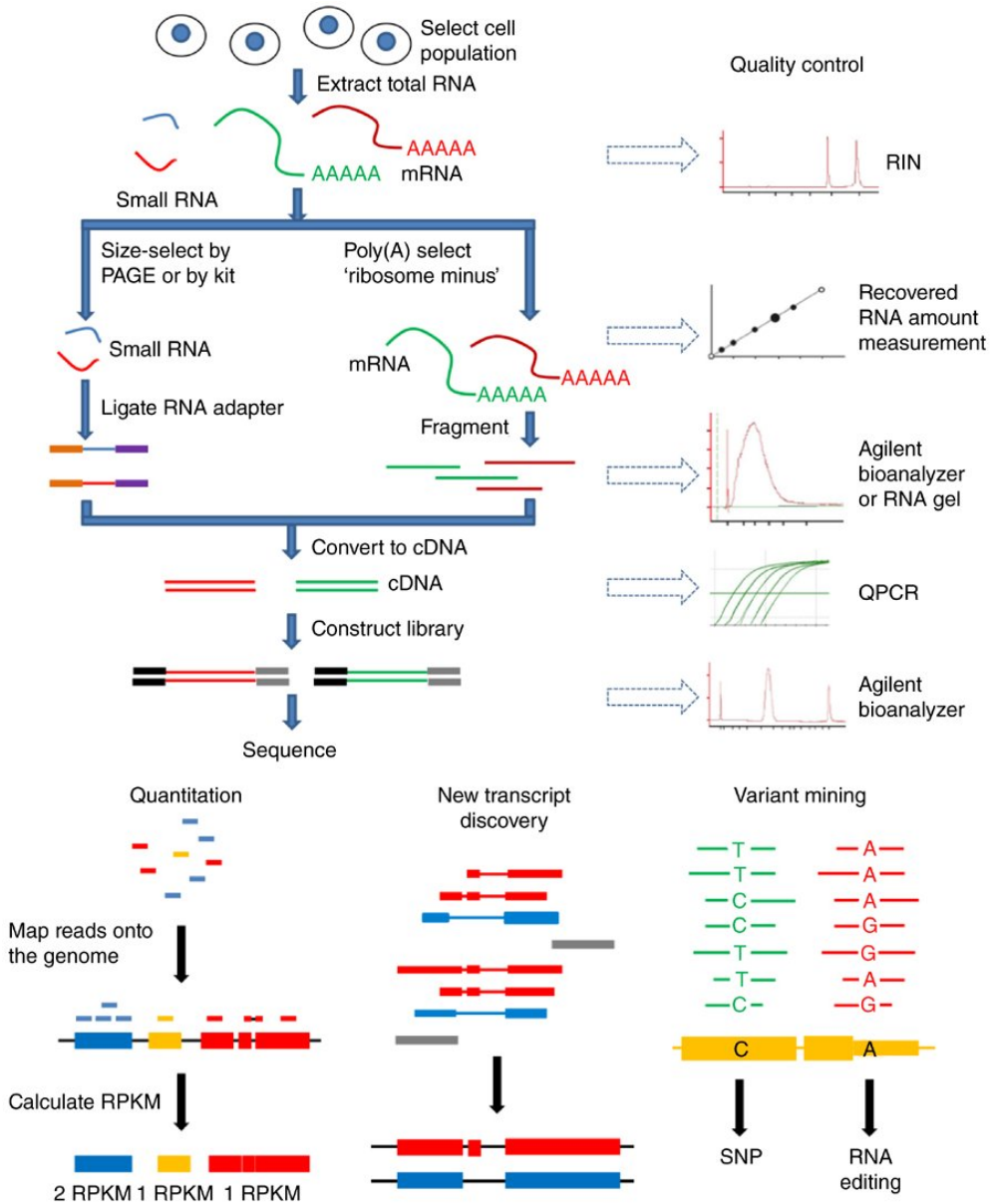

RNASeq Tasks, Tools and File Formats
Task
Tool
File Format
Quality Control
FastQC
FastQ, SangerFastQ
Alignment
TopHat
IGV
SAM/BAM
Summarization
Cufflinks
GTF
Differential Gene Expression
Cuffdiff,Edge, DESeq, baySeq

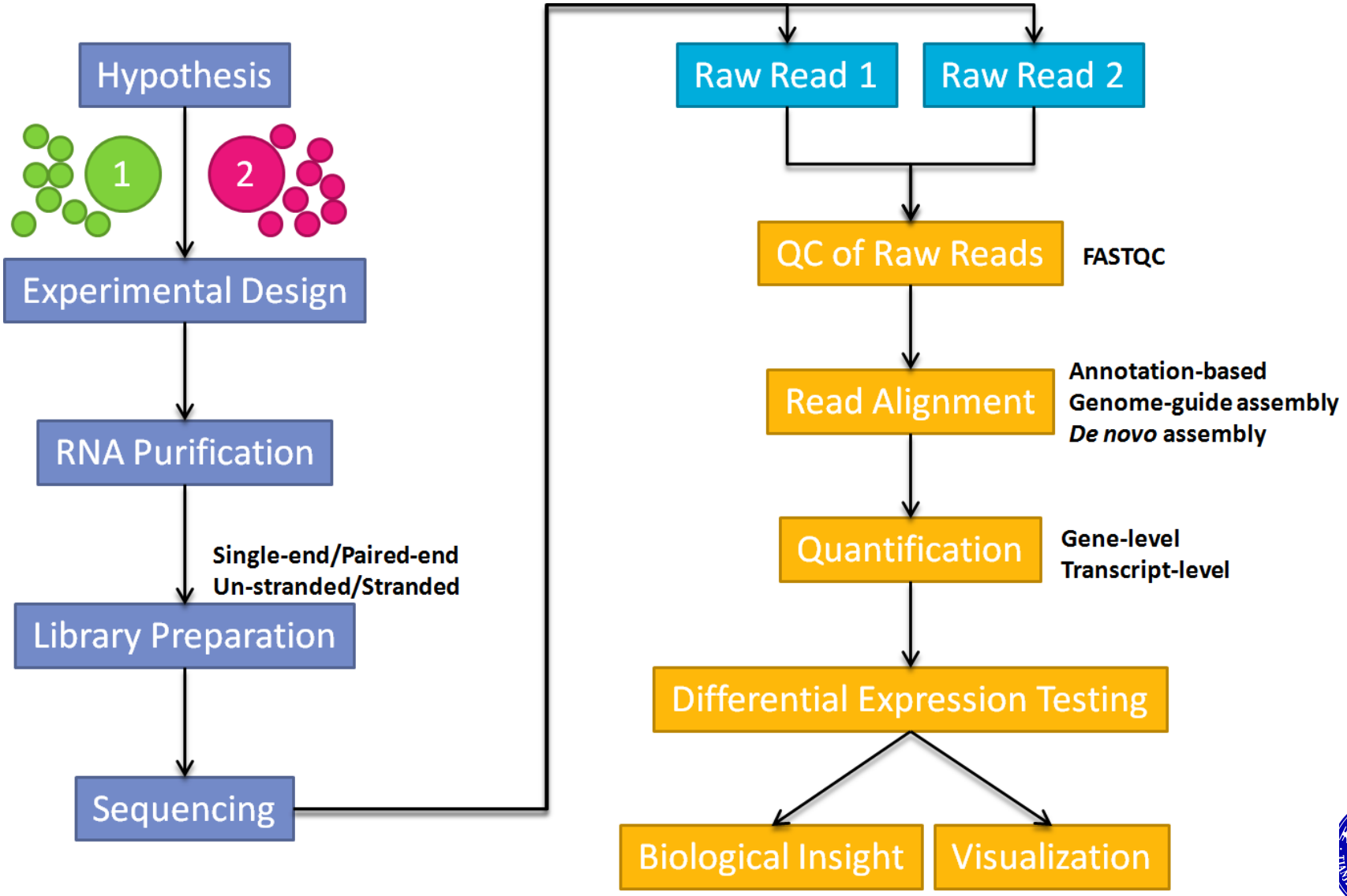
转录组学 | RNA-Seq | 分析 | 概览




Samples of interest
Condition 1
(e.g. tumor)
Condition 2
(e.g. normal)
Isolate RNAs
Poly(A) tail
Generate cDNA, fragment,
size select, add linkers
Sequence ends
100s of millions of paired reads
10s of billions bases of sequence
Unsequenced RNA
RNA reads
Short insert
Map to genome, transcriptome,
and predicted exon junctions
Intron
pre-mRNA
Exon
Transcript
Short reads
Short reads
split by intron
Downstream analysis

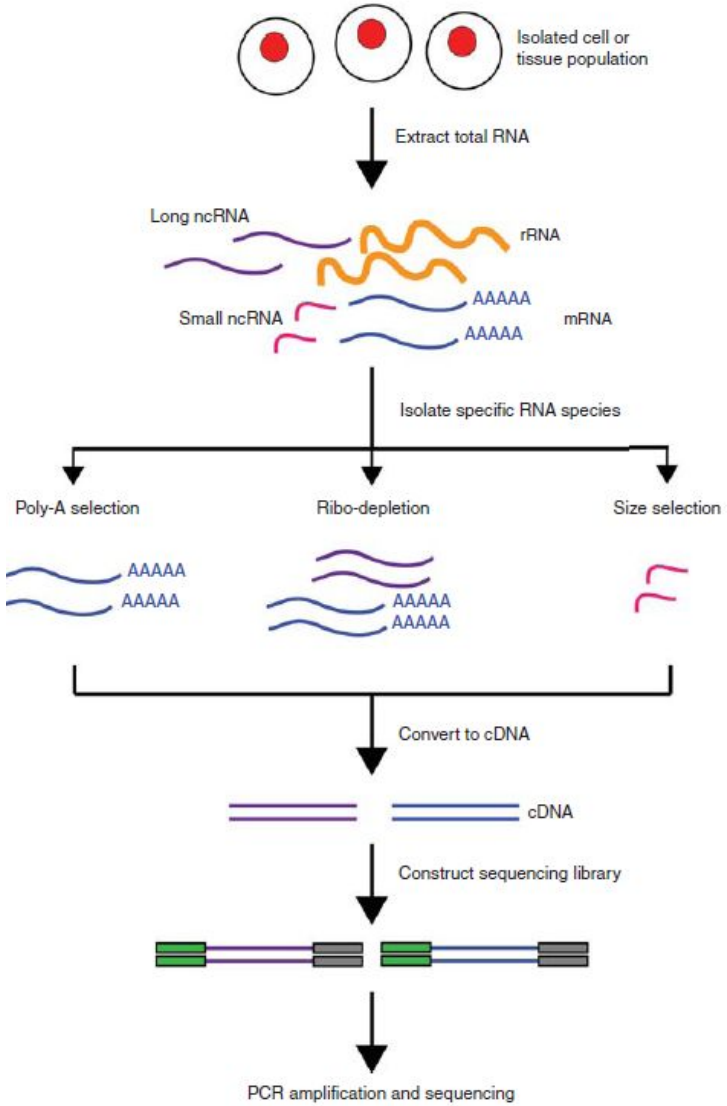
转录组学 | RNA-Seq | 分析 | 流程




Hypothesis
1
2
Experimental Design
RNA Purification
Single-end/Paired-end
Un-stranded/Stranded
Library Preparation
Sequencing
Raw Read 1
Raw Read 2
QC of Raw Reads
FASTQC
Read Alignment
Annotation-based
Genome-guide assembly
De novo assembly
Quantification
Gene-level
Transcript-level
Differential Expression Testing
Biological Insight
Visualization

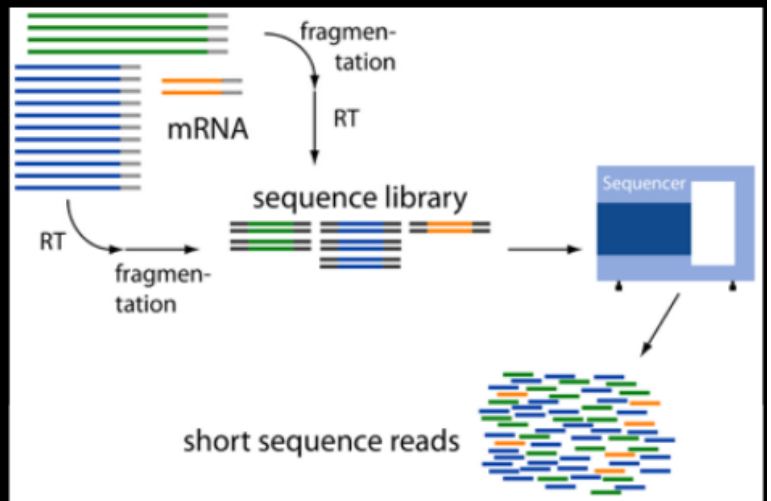
Steps in Preparing an RNA-Seq Library

1. Purify RNA
2. Bind polyA fraction (mRNA)
3. Fragment RNA (200 bp)
4. Convert to cDNA by random priming
5. Apply adaptors and sequence
6. Analyze millions of 25 bp reads


Isolated cell or
tissue population
Extract total RNA
Long ncRNA
rRNA
AAAAA
Small ncRNA
mRNA
AAAAA
Isolate specific RNA species
Poly-A selection
Ribo-depletion
Size selection
AAAAA
AAAAA
AAAAA
AAAAA
Convert to cDNA
cDNA
Construct sequencing library
PCR amplification and sequencing

RNA-Seq

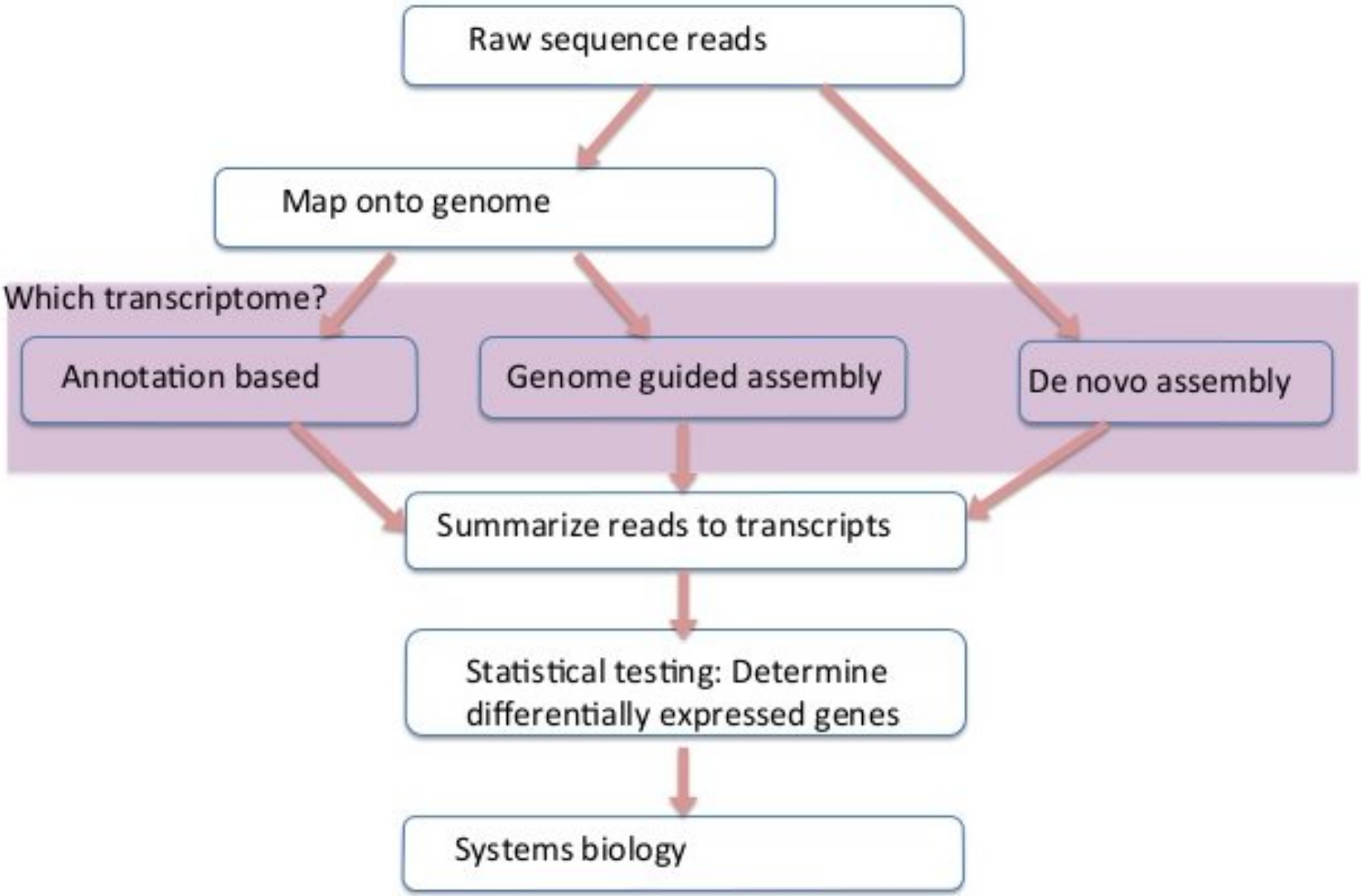
- Deplete rRNA
- or select for polyadenylated RNA
- Fragmentation
- Reverse transcribe to cDNA
- Attach adaptor sequences
- Size selection
- Amplify by PCR
- High-throughput sequencing (eg Illumina)

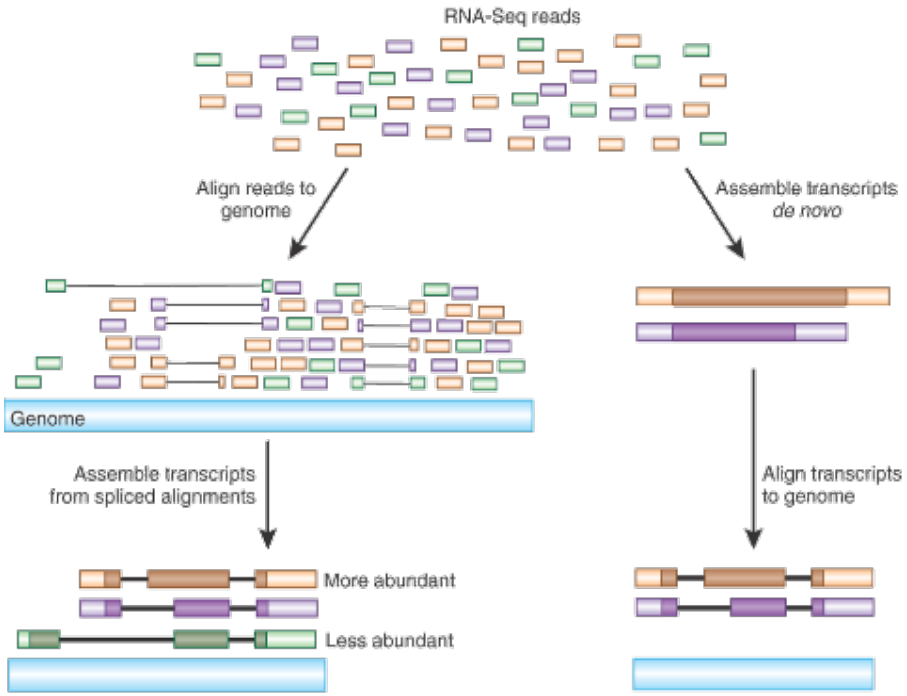


Pepke *et al*, Nature Methods, 2009

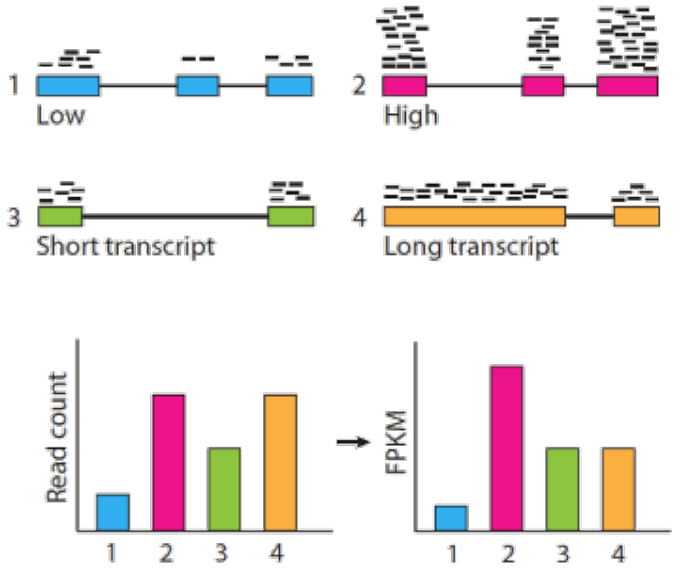
RNA-Seq output

- Several million "reads" per sample
- Reads are RNA sequence starting from random locations within the original mRNA
- May read through into adaptor sequence
- Current typical length ~150 bases (really just needs to be long enough to locate in genome)

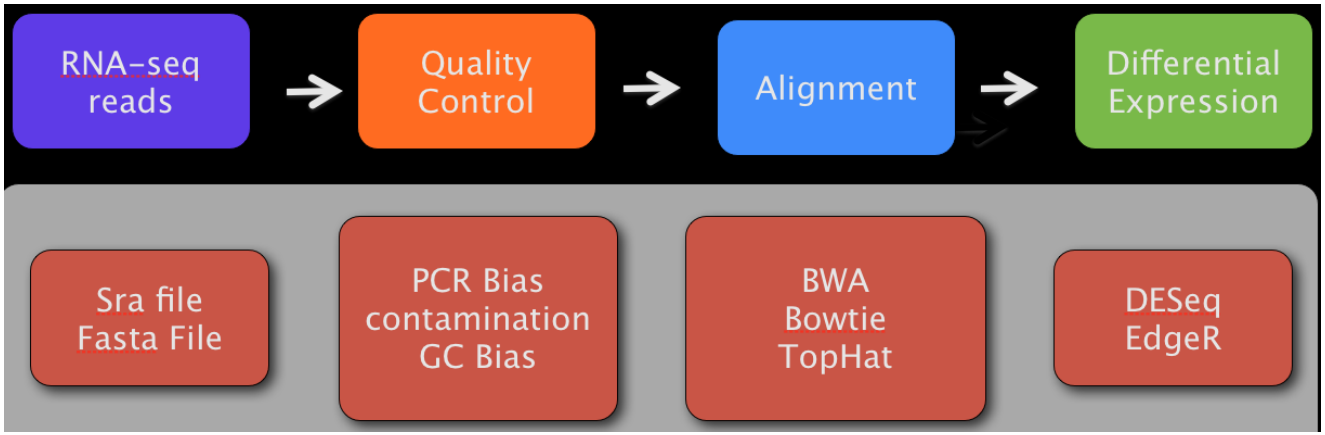

Raw sequence reads
Map onto genome
Which transcriptome?
Annotation based
Genome guided assembly
De novo assembly
Summarize reads to transcripts
Statistical testing: Determine differentially expressed genes
Systems biology

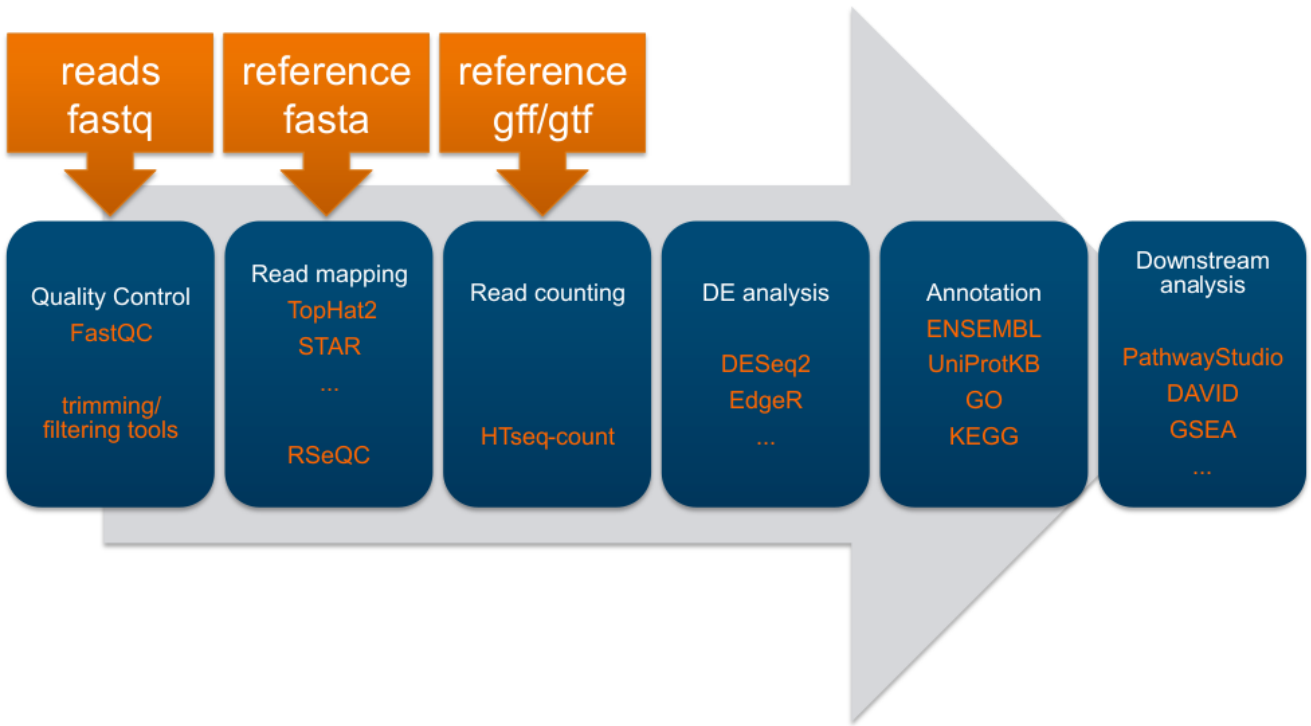


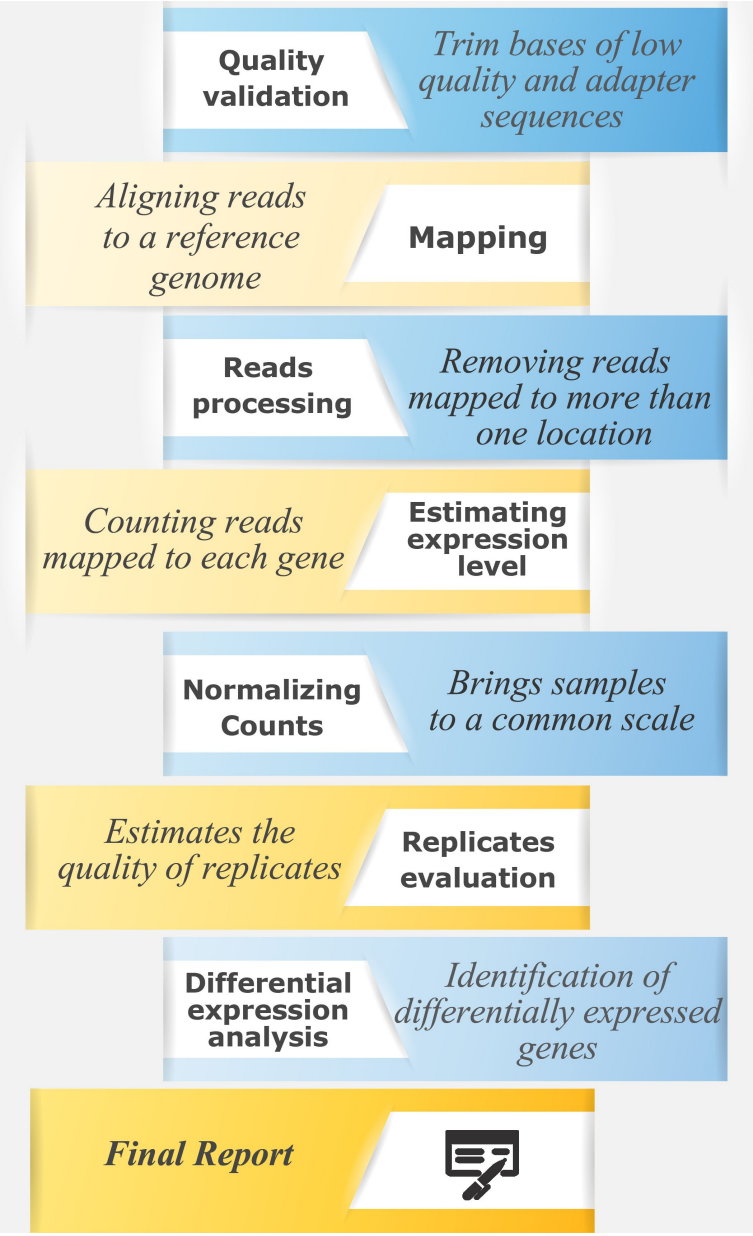
Nature Biotechnology . 2010 (28): 421–423.

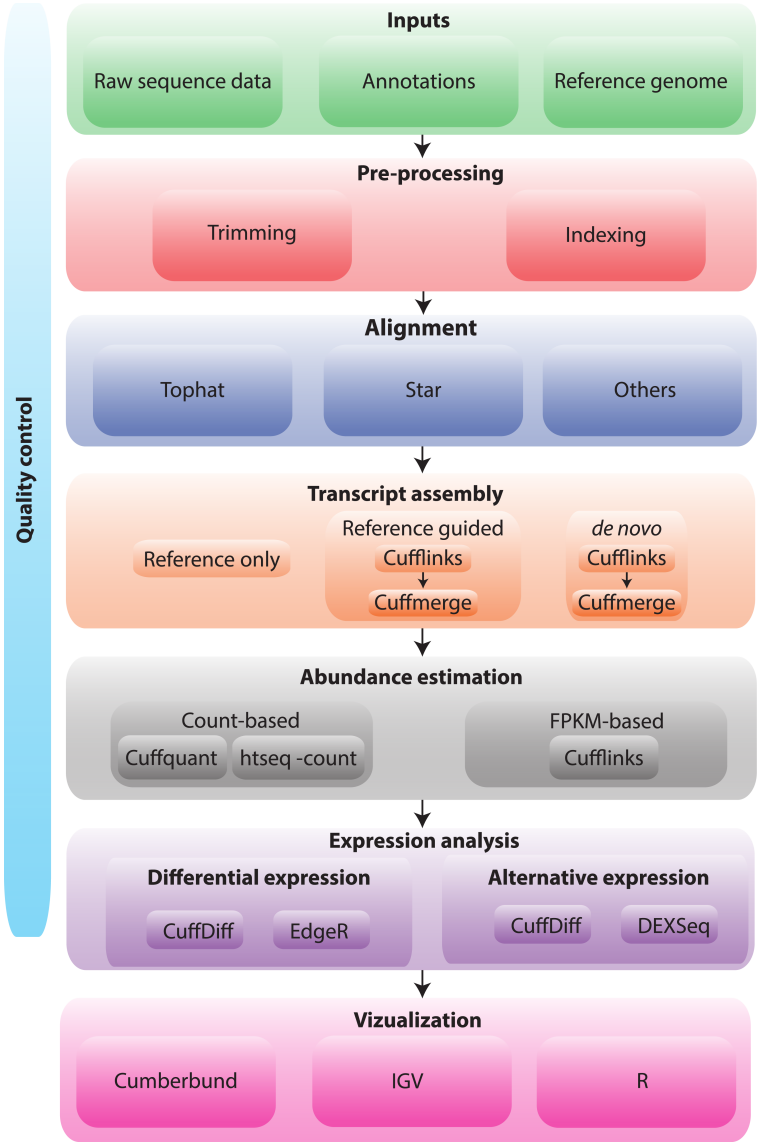


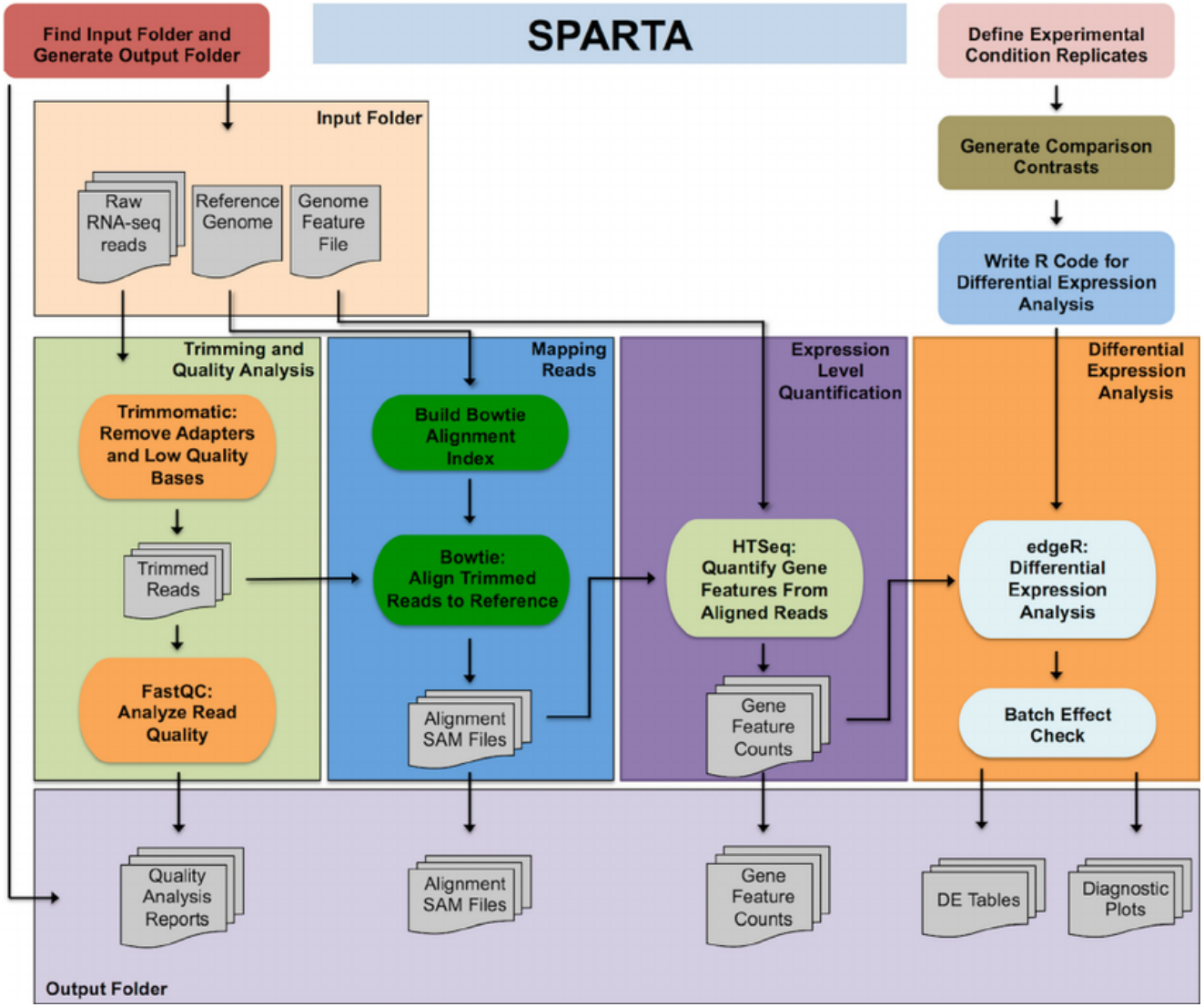
Nature Methods. 2011 (8): 469–477.

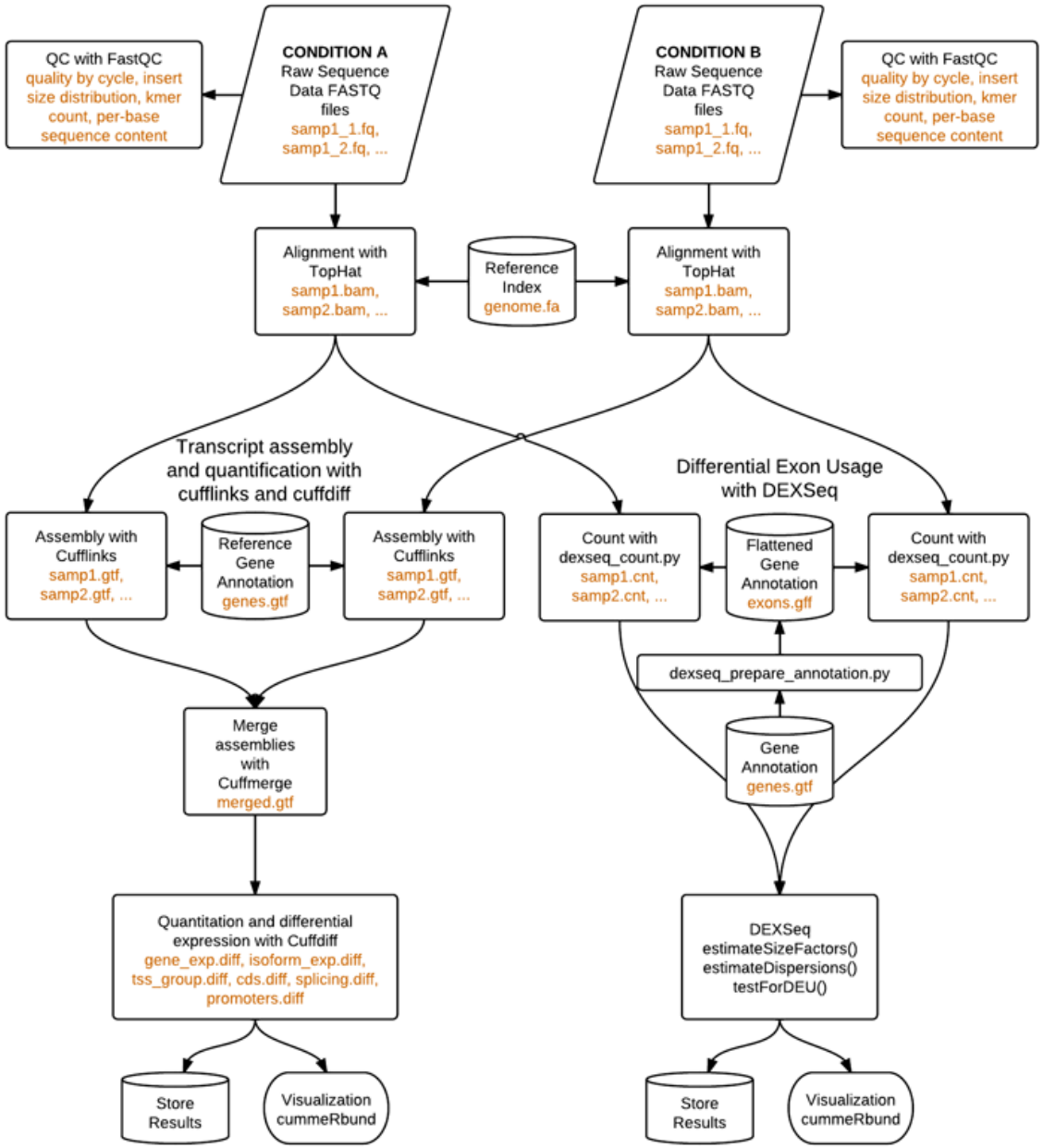

RNA-seq reads
Quality Control
Alignment
Differential Expression
Sra file
Fasta File
PCR Bias
contamination
GC Bias
BWA
Bowtie
TopHat
DESeq
EdgeR


reads
fastq
reference
fasta
reference
gff/gtf
Quality Control
FastQC
trimming/
filtering tools
Read mapping
TopHat2
STAR
...
RSeQC
Read counting
HTseq-count
DE analysis
DESeq2
EdgeR
...
Annotation
ENSEMBL
UniProtKB
GO
KEGG
Downstream
analysis
PathwayStudio
DAVID
GSEA
...


Quality validation
Trim bases of low quality and adapter sequences
Aligning reads to a reference genome
Mapping
Reads processing
Removing reads mapped to more than one location
Counting reads mapped to each gene
Estimating expression level
Normalizing Counts
Brings samples to a common scale
Estimates the quality of replicates
Replicates evaluation
Differential expression analysis
Identification of differentially expressed genes
Final Report


Quality control
Inputs
Raw sequence data
Annotations
Reference genome
Pre-processing
Trimming
Indexing
Alignment
Tophat
Star
Others
Transcript assembly
Reference only
Reference guided
Cufflinks
Cuffmerge
de novo
Cufflinks
Cuffmerge
Abundance estimation
Count-based
Cuffquant
htseq -count
FPKM-based
Cufflinks
Expression analysis
Differential expression
CuffDiff
EdgeR
Alternative expression
CuffDiff
DEXSeq
Vizualization
Cumberbund
IGV
R


SPARTA
Find Input Folder and Generate Output Folder
Define Experimental Condition Replicates
Generate Comparison Contrasts
Write R Code for Differential Expression Analysis
Input Folder
Raw RNA-seq reads
Reference Genome
Genome Feature File
Trimming and Quality Analysis
Trimmomatic: Remove Adapters and Low Quality Bases
Trimmed Reads
FastQC: Analyze Read Quality
Mapping Reads
Build Bowtie Alignment Index
Bowtie: Align Trimmed Reads to Reference
Alignment SAM Files
Expression Level Quantification
HTSeq: Quantify Gene Features From Aligned Reads
Gene Feature Counts
Differential Expression Analysis
edgeR: Differential Expression Analysis
Batch Effect Check
Output Folder
Quality Analysis Reports
Alignment SAM Files
Gene Feature Counts
DE Tables
Diagnostic Plots

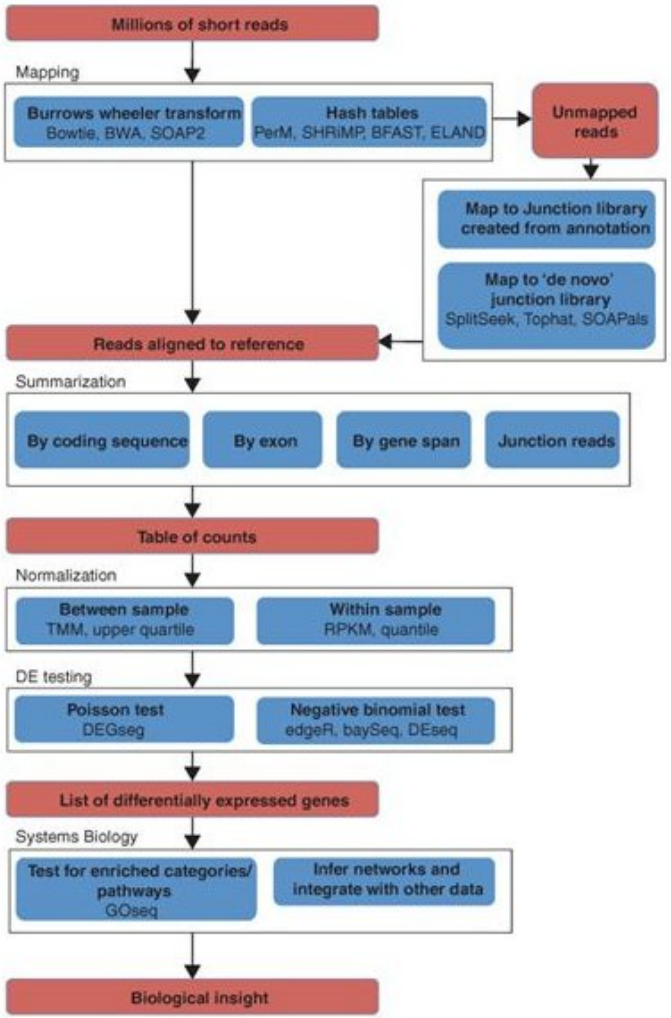

QC with FastQC
quality by cycle, insert size distribution, kmer count, per-base sequence content
CONDITION A
Raw Sequence Data FASTQ files
samp1_1.fq, samp1_2.fq, ...
CONDITION B
Raw Sequence Data FASTQ files
samp1_1.fq, samp1_2.fq, ...
QC with FastQC
quality by cycle, insert size distribution, kmer count, per-base sequence content
Alignment with TopHat
samp1.bam, samp2.bam, ...
Reference Index
genome.fa
Alignment with TopHat
samp1.bam, samp2.bam, ...
Transcript assembly and quantification with cufflinks and cuffdiff
Differential Exon Usage with DEXSeq
Assembly with Cufflinks
samp1.gtf, samp2.gtf, ...
Reference Gene Annotation
genes.gtf
Assembly with Cufflinks
samp1.gtf, samp2.gtf, ...
Count with dexseq_count.py
samp1.cnt, samp2.cnt, ...
Flattened Gene Annotation
exons.gff
Count with dexseq_count.py
samp1.cnt, samp2.cnt, ...
dexseq_prepare_annotation.py
Merge assemblies with Cuffmerge
merged.gtf
Gene Annotation
genes.gtf
Quantitation and differential expression with Cuffdiff
gene_exp.diff, isoform_exp.diff, tss_group.diff, cds.diff, splicing.diff, promoters.diff
DEXSeq
estimateSizeFactors()
estimateDispersions()
testForDEU()
Store Results
Visualization cummeRbund
Store Results
Visualization cummeRbund

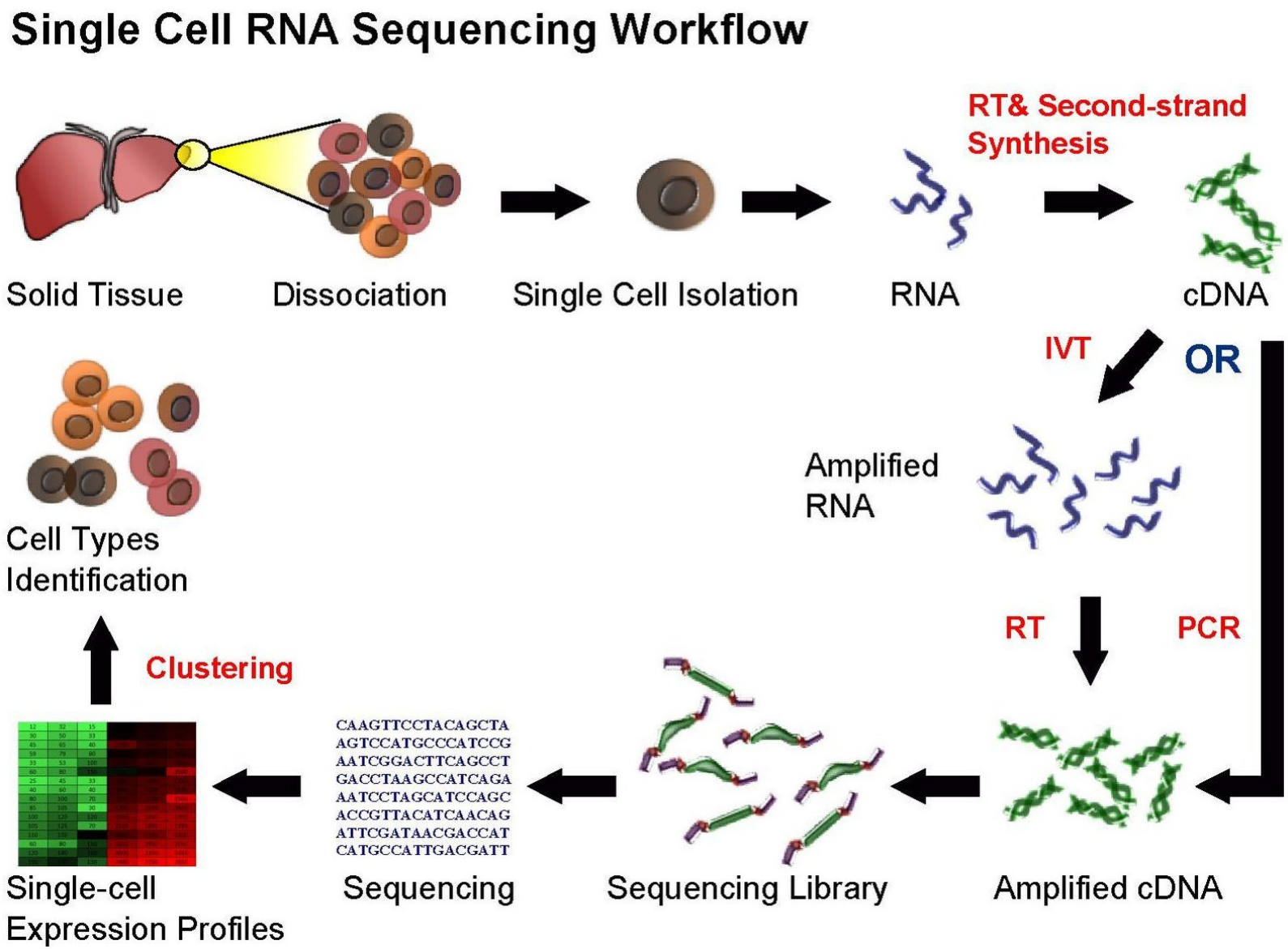
RNA-seq workflows

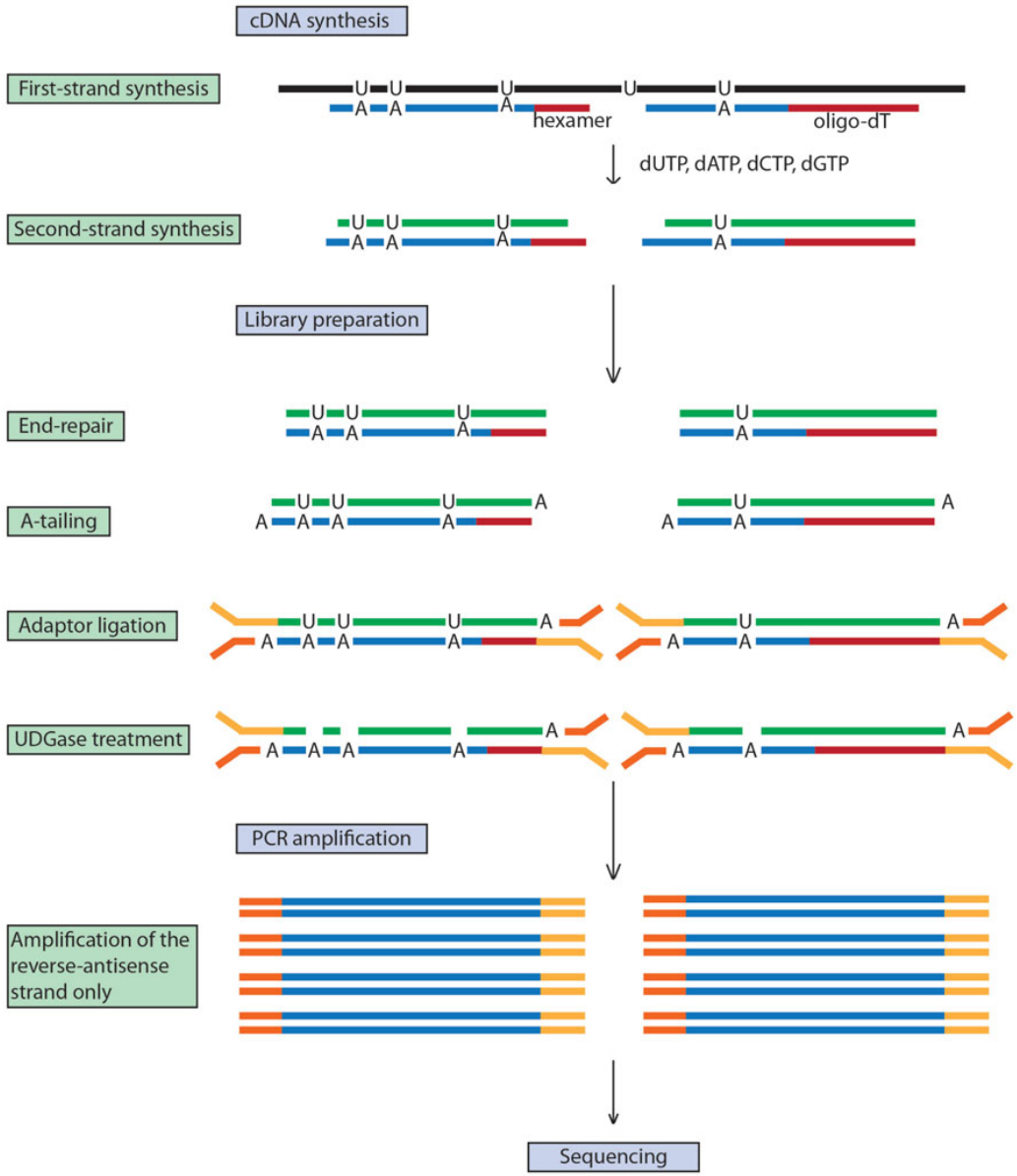
- Sequencing: obtain raw data (fastq format)
- Quality control (optional): FASTX

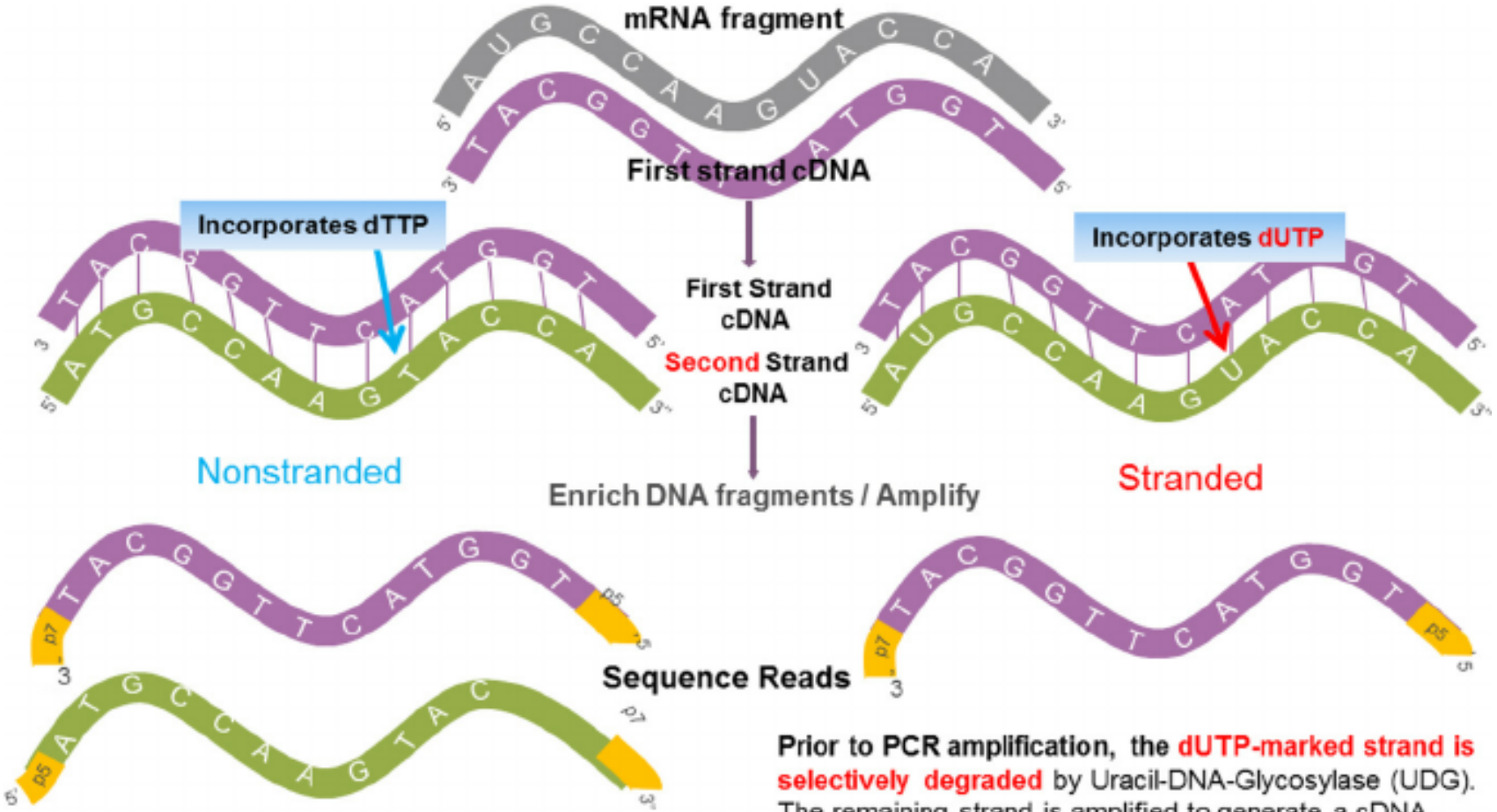
- Workflow 1: tophat2 (align) -> cufflinks (transcript assembly) -> cuffdiff (DEGs), cuffmerge (merge assemblies)

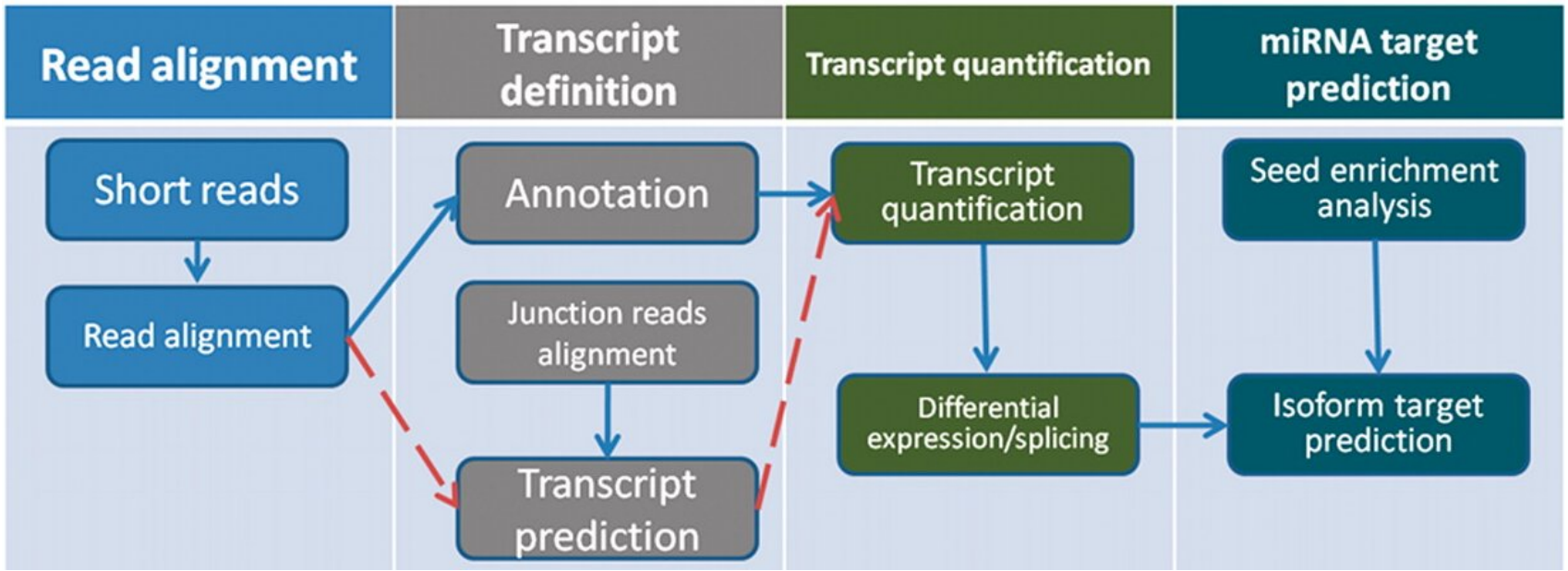
  Workflow 2: bowtie2 (align) -> HTSeq-count (count by gene) -> edgeR or DESeq (DEGs)
- Fusion detection (optional): “chimerascan” or “defuse”

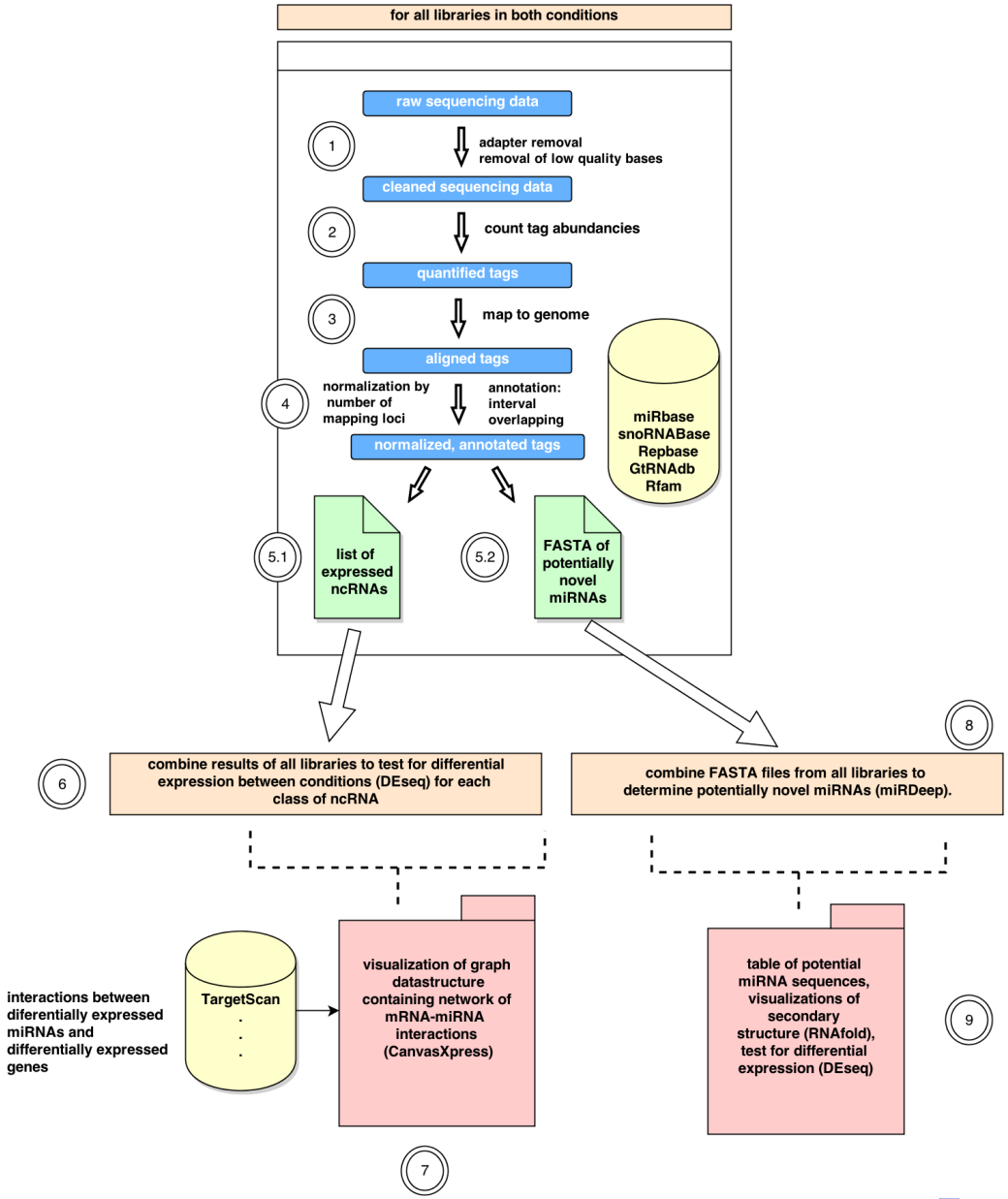


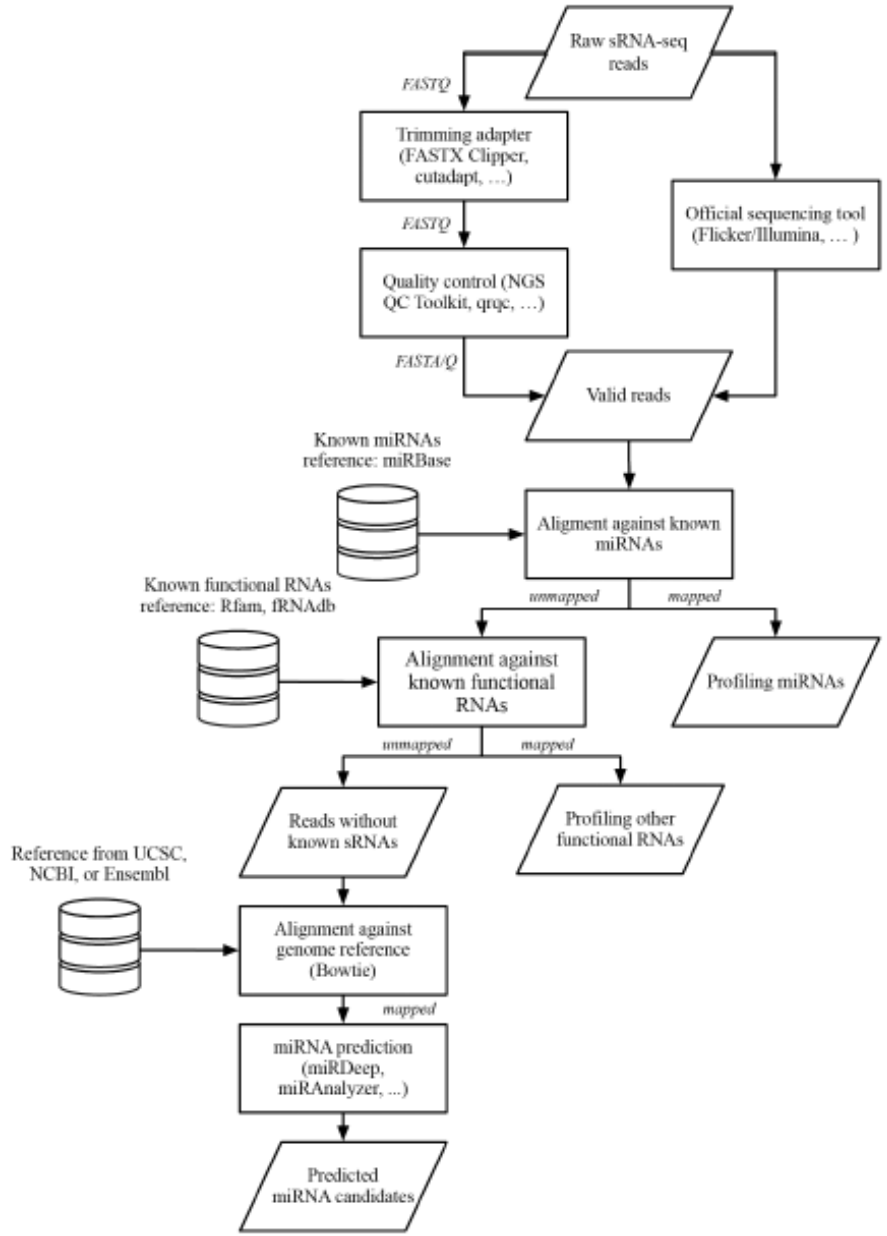

Single Cell RNA Sequencing Workflow
RT& Second-strand Synthesis
Solid Tissue
Dissociation
Single Cell Isolation
RNA
cDNA
IVT
OR
Amplified RNA
Cell Types Identification
RT
PCR
Clustering
CAAGTTCCTACAGCTA
AGTCCATGCCCATCCG
AATCGGACTTCAGCCT
GACCTAAGCCATCAGA
AATCCTAGCATCCAGC
ACCGTTACATCAACAG
ATTCGATAACGACCAT
CATGCCATTGACGATT
Single-cell Expression Profiles
Sequencing
Sequencing Library
Amplified cDNA


cDNA synthesis
First-strand synthesis
hexamer
oligo-dT
dUTP, dATP, dCTP, dGTP
Second-strand synthesis
Library preparation
End-repair
A-tailing
Adaptor ligation
UDGase treatment
PCR amplification
Amplification of the reverse-antisense strand only
Sequencing


mRNA fragment
First strand cDNA
Incorporates dTTP
Incorporates dUTP
First Strand cDNA
Second Strand cDNA
Nonstranded
Stranded
Enrich DNA fragments / Amplify
Sequence Reads
Prior to PCR amplification, the dUTP-marked strand is selectively degraded by Uracil-DNA-Glycosylase (UDG). The remaining strand is amplified to generate a cDNA library suitable for sequencing.


Read alignment
Transcript definition
Transcript quantification
miRNA target prediction
Short reads
Read alignment
Annotation
Junction reads alignment
Transcript prediction
Transcript quantification
Differential expression/splicing
Seed enrichment analysis
Isoform target prediction


for all libraries in both conditions
raw sequencing data
1
adapter removal
removal of low quality bases
cleaned sequencing data
2
count tag abundancies
quantified tags
3
map to genome
aligned tags
4
normalization by number of mapping loci
annotation: interval overlapping
normalized, annotated tags
miRbase
snoRNABase
Repbase
GtRNAdb
Rfam
5.1
list of expressed ncRNAs
5.2
FASTA of potentially novel miRNAs
6
combine results of all libraries to test for differential expression between conditions (DEseq) for each class of ncRNA
8
combine FASTA files from all libraries to determine potentially novel miRNAs (miRDeep).
interactions between diferentially expressed miRNAs and differentially expressed genes
TargetScan
visualization of graph datastructure containing network of mRNA-miRNA interactions (CanvasXpress)
table of potential miRNA sequences, visualizations of secondary structure (RNAfold), test for differential expression (DEseq)
9
7


Raw sRNA-seq reads
FASTQ
Trimming adapter (FASTX Clipper, cutadapt, ...)
FASTQ
Quality control (NGS QC Toolkit, qrqc, ...)
FASTA/Q
Official sequencing tool (Flicker/Illumina, ...)
Valid reads
Known miRNAs reference: miRBase
Aligment against known miRNAs
unmapped
mapped
Known functional RNAs reference: Rfam, fRNAdb
Alignment against known functional RNAs
Profiling miRNAs
unmapped
mapped
Reads without known sRNAs
Profiling other functional RNAs
Reference from UCSC, NCBI, or Ensembl
Alignment against genome reference (Bowtie)
mapped
miRNA prediction (miRDeep, miRAnalyzer, ...)
Predicted miRNA candidates

Read counting options

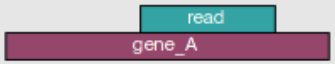
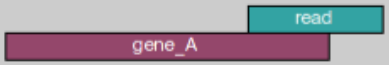
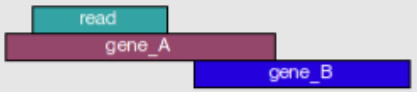
Count per gene
analysis with DESeq2, edgeR

Count per transcript
analysis with CummeRbund, DESeq2, edgeR

Count per exon
analysis with DEXSeq

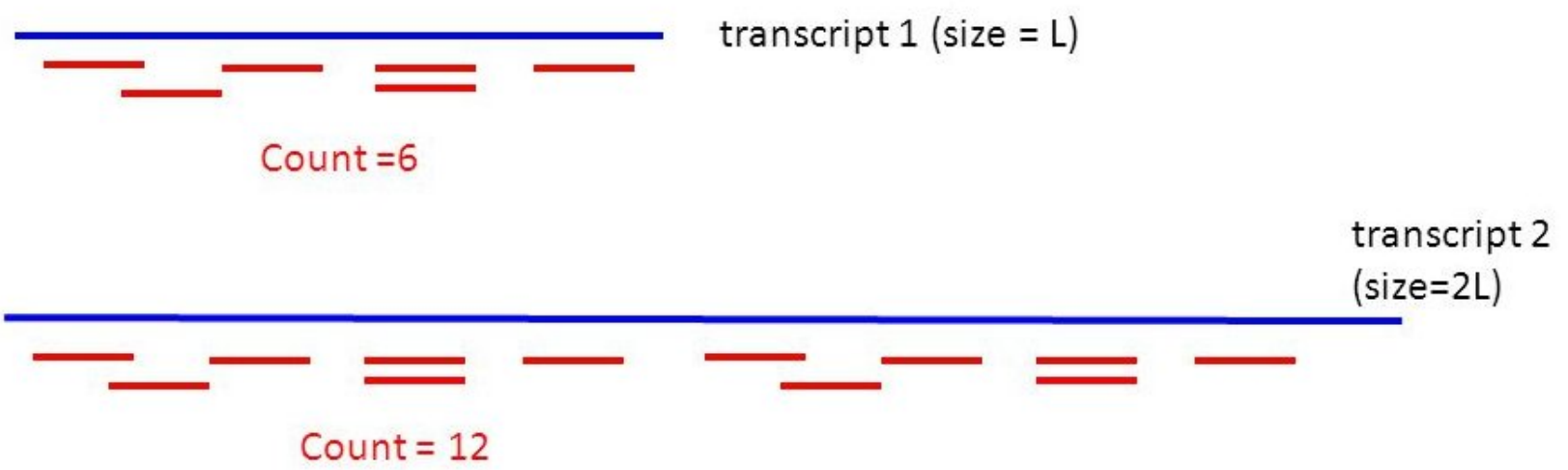
Tools: HTSeq-count

http://www-huber.embl.de/users/anders/HTSeq/doc/count.html

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Normalization for gene length and library size: RPKM / FPKM

One sample, two transcripts



You can't conclude that gene 2 has a higher expression than gene 1!

Normalization

"Data don't make any sense,
we will have to resort to statistics."

Some samples are sequenced at higher depth than others

Normalization using the total counts:

total counts in sample *j* / total counts in a reference sample

But... longer genes/transcripts will generate more reads

One proposed solution:

use RPKM = reads per kilobase per million mapped reads

Normalization: the problem

Number of reads (coverage) will not be exactly the same for each sample

Problem: Need to scale RNA counts per gene to total sample coverage

- **Solution** – divide counts per million reads

Problem: Longer genes have more reads, gives better chance to detect DE

- **Solution** – divide counts by gene length

Result = **RPKM** (Reads Per KB per Million)

Normalization: classic (used by Cufflinks)

RPKM: Reads per kilobases per million mapped reads

1kbp transcript with 2000 alignments in a sample of 10 million reads (out of which 8 millions are mapped)

RPKM = 2000 / (1 * 8) = 250

by extension:

FPKM: Fragment per kilobases per million mapped reads

a fragment is a pair of reads (Paired-end)

Other normalization methods

Sequencing depth, gene length, and count distribution are the main biases that must be accounted for in the normalization and/or differential expression calculations.
Biological Replicates are essential to increase the robustness of statistics

FPKM (Trapnell et al., 2010) - Fragments per Kilobase of exon per Million mapped reads, analogous to RPKM.
Upper-quartile (Bullard et al., 2010) - Counts are divided by upper-quartile of counts for transcripts with at least one read.
TMM (EdgeR) (Robinson and Oshlack, 2010) - Trimmed mean of M values.
RLE (DESeq) (Anders & Huber, 2010) – Relative Log Expression (Median of ratios)
Quantiles, as in microarray normalization (Irizarry et al., 2003).

Many methods which is the best?

Table 3: Summary of comparison results for the seven normalization methods under consideration

Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	−	+	+	−	−
UQ	++	++	+	++	−
Med	++	++	−	++	−
DESeq	++	++	++	++	++
TMM	++	++	++	++	++
Q	++	−	+	++	−
RPKM	−	+	+	−	−

A '−' indicates that the method provided unsatisfactory results for the given criterion, while a '+' and '++' indicate satisfactory and very satisfactory results for the given criterion.

Both methods are found in Bioconductor packages (DESeq2 and EdgeR)

Normalization: different goals

- **R/FPKM:** (Mortazavi et al. 2008)
  - Correct for: differences in sequencing depth and transcript length
  - Aiming to: compare a gene across samples and diff genes within sample
- **TMM**: (Robinson and Oshlack 2010)
  - Correct for: differences in transcript pool composition; extreme outliers
  - Aiming to: provide better across-sample comparability
- **TPM**: (Li et al 2010, Wagner et al 2012)
  - Correct for: transcript length distribution in RNA pool
  - Aiming to: provide better across-sample comparability
- **Limma voom (logCPM)**: (Lawet al 2013)
  - Aiming to: stabilize variance; remove dependence of variance on the mean

Differential Expression

- Should we do differential expression on RPKM/FPKM or TPM?

- Cufflinks: RPKM/FPKM
- LIMMA-VOOM and DESeq: TPM
- Power to detect DE is proportional to length
- Continued development and updates

RPKM / FPKM / TPM

- RPKM (Reads per kilobase of transcript per million reads of library)
  - Corrects for total library coverage
  - Corrects for gene length
  - Comparable between different genes within the same dataset
- FPKM (Fragments per kilobase of transcript per million reads of library)
  - Only relevant for paired end libraries
  - Pairs are not independent observations
  - RPKM/2
- TPM (transcripts per million)
  - Normalises to transcript copies instead of reads
  - Corrects for cases where the average transcript length differs between samples

RPKM:
Reads Per Kilobase and Million mapped reads

- Unit of measurement

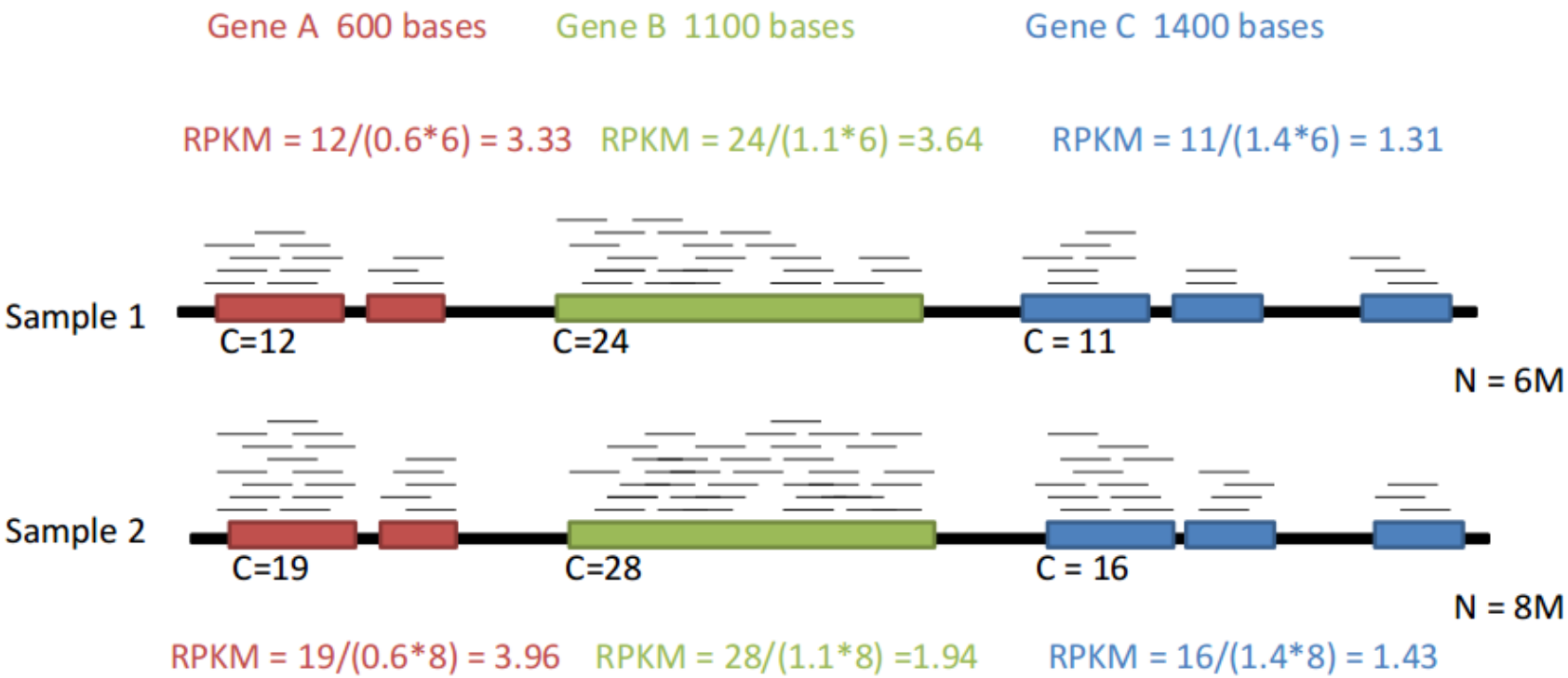
$$RPKM = \#MappedReads * \frac{1000bases * 10^6}{length\ of\ transcript * Total\ number\ of\ mapped\ reads}$$

- RPKM reflects the molar concentration of a transcript in the starting sample by normalizing for
  - RNA length
  - Total read number in the measurement

- This facilitates transparent comparison of transcript levels within and between samples

$$RPKM = \frac{number\ of\ reads\ of\ the\ region}{\frac{total\ reads}{1{,}000{,}000}\ x\ \frac{region\ length}{1{,}000}}$$

RPKM Example



Reporting quantitative expression: FPKM/RPKM

- In NGS RNA-seq experiments, quantitative gene expression data is normalized for total gene/transcript length and the number of sequencing reads, and reported as
  - RPKM: Reads Per Kilobase of exon per Million mapped reads. Used for reporting data based on single-end reads
  - FPKM: Fragments Per Kilobase of exon per Million fragments. Used for reporting data based on paired-end fragments

FPKM: Fragments per K per M

What's the difference between FPKM and RPKM?

- Paired-end RNA-Seq experiments produce two reads per fragment, but that doesn't necessarily mean that both reads will be mappable. For example, the second read is of poor quality.
- If we were to count reads rather than fragments, we might double-count some fragments but not others, leading to a skewed expression value.
- Thus, FPKM is calculated by counting fragments, not reads.

$$\mathrm{RPKM} = \frac{total\ exon\ reads}{mapped\ reads\ (millions) * exon\ length\ (KB)}$$

$$\mathrm{FPKM} = \frac{total\ fragments}{mapped\ reads\ (millions) * exon\ length\ (KB)}$$

TPM

当我们进行 RNA-Seq 时，会使用 RPKM 或者 FPKM 来代表某个 gene 或是 isoform 的表达量多寡。可是当我们想要比较不同次实验内的某个基因，其表达量相比于“整体基因表达”而言，是否维持在“固定比例”时，便无法使用这样的计算方式。因此 Wagner *et. al.* 在 2012 年的时候提出 TPM（Transcript Per Million）的概念来弥补这个缺点。

TPM – Transcripts Per Million

Theory Biosci.
DOI 10.1007/s12064-012-0162-3

SHORT COMMUNICATION

Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples

Günter P. Wagner · Koryu Kin · Vincent J. Lynch

A slightly modified RPKM measure that accounts for differences in gene length distribution in the transcript population

TPM

$$= \frac{\frac{\text{total exon reads}}{\text{exon length (KB)}}}{\left(\frac{\text{GeneA mapped reads (millions)}}{\text{exon length (KB)}} + \frac{\text{GeneB mapped reads (millions)}}{\text{exon length (KB)}} + \frac{\text{GeneC mapped reads (millions)}}{\text{exon length (KB)}} + \cdots\right)}$$

$$\mathrm{TPM}_i = \left(\frac{\mathrm{FPKM}_i}{\sum_j \mathrm{FPKM}_j}\right) \cdot 10^6$$

假設某個生物具有4個基因，分別為A，B，C，D。然後我們做了3次的RNA-Seq實驗，將獲得的reads與每個基因進行比對，其比對的數目如下表所示。

	Replicate 1	Replicate 2	Replicate 3
Gene A (2kb)	10,000,000	12,000,000	30,000,000
Gene B (4kb)	20,000,000	25,000,000	60,000,000
Gene C (1kb)	5,000,000	8,000,000	15,000,000
Gene D (10kb)	0	0	1,000,000
Sum	35,000,000	45,000,000	106,000,000

根據RPKM的公式,

$$\text{RPKM} = \frac{total\ exon\ reads}{\text{mapped reads (millions)} * \text{exon length (KB)}}$$

我們來計算實驗1的Gene A的RPKM,

$$\text{RPKM} = \frac{10{,}000{,}000}{(10+20+5)*2} = 142857$$

然後再將表格內所有的數值都轉換成RPKM之後,我們得到下方表格

	Replicate 1 (RPKM)	Replicate 2 (RPKM)	Replicate 3 (RPKM)
Gene A (2kb)	142857	133333	141509
Gene B (4kb)	142857	138889	141509
Gene C (1kb)	142857	177778	141509
Gene D (10kb)	0	0	943
Sum	428,571	450,000	425,470

我們再根據TPM的公式

$$\mathrm{TPM} = \frac{\frac{\text{total exon reads}}{\text{exon length (KB)}}}{\left(\frac{\text{GeneA mapped reads (millions)}}{\text{exon length (KB)}} + \frac{\text{GeneB mapped reads (millions)}}{\text{exon length (KB)}} + \frac{\text{GeneC mapped reads (millions)}}{\text{exon length (KB)}} + \cdots\right)}$$

來計算實驗1的Gene A的TPM，

$$\mathrm{TPM} = \frac{\frac{10{,}000{,}000}{2}}{\left(\frac{10}{2} + \frac{20}{4} + \frac{5}{1} + \frac{0}{10}\right)} = 333333$$

然後再將表格內所有的數值都轉換成TPM之後，我們得到下方表格

	Replicate 1 (TPM)	Replicate 2 (TPM)	Replicate 3 (TPM)
Gene A (2kb)	333333	296296	332594
Gene B (4kb)	333333	308642	332594
Gene C (1kb)	333333	395062	332594
Gene D (10kb)	0	0	2217
Sum	1,000,000	1,000,000	1,000,000

比较

比较 RPKM 和 TPM 的表格之后，可以发现 RPKM 在“Sum”这一行，三个实验中的数值均不同，因此我们很难直接利用 RPKM 数值去比较每个基因“相对于整体”的表达量多寡。然而 TMP 则一律为 1,000,000，也就是 1 million，意即 TPM 的定义（Transcript **Per Million**）。因此这些数值便可以直接比较，了解基因之间的相对表达量。

Metrics

- RPKM (Reads Per Kilobase Million)
- FPKM (Fragments Per Kilobase Million)
- TPM (Transcripts Per Kilobase Million)

These three metrics attempt to normalize for sequencing depth and gene length.

RPKM

Here's how you do it for RPKM:

1. Count up the total reads in a sample and divide that number by 1,000,000 − this is our "per million" scaling factor.
2. Divide the read counts by the "per million" scaling factor. This normalizes for sequencing depth, giving you reads per million (RPM)
3. Divide the RPM values by the length of the gene, in kilobases. This gives you RPKM.

FPKM

FPKM is very similar to RPKM. RPKM was made for single-end RNA-seq, where every read corresponded to a single fragment that was sequenced. FPKM was made for paired-end RNA-seq. With paired-end RNA-seq, two reads can correspond to a single fragment, or, if one read in the pair did not map, one read can correspond to a single fragment. The only difference between RPKM and FPKM is that FPKM takes into account that two reads can map to one fragment (and so it doesn't count this fragment twice).

TPM

TPM is very similar to RPKM and FPKM. The only difference is the order of operations. Here's how you calculate TPM:

1. Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK).
2. Count up all the RPK values in a sample and divide this number by 1,000,000. This is your "per million" scaling factor.
3. Divide the RPK values by the "per million" scaling factor. This gives you TPM.

So you see, when calculating TPM, the only difference is that you normalize for gene length first, and then normalize for sequencing depth second. However, the effects of this difference are quite profound.

Compare

When you use TPM, the sum of all TPMs in each sample are the same. This makes it easier to compare the proportion of reads that mapped to a gene in each sample. In contrast, with RPKM and FPKM, the sum of the normalized reads in each sample may be different, and this makes it harder to compare samples directly.

Here's an example. If the TPM for gene A in Sample 1 is 3.33 and the TPM in sample B is 3.33, then I know that the exact same proportion of total reads mapped to gene A in both samples. This is because the sum of the TPMs in both samples always add up to the same number (so the denominator required to calculate the proportions is the same, regardless of what sample you are looking at.)

With RPKM or FPKM, the sum of normalized reads in each sample can be different. Thus, if the RPKM for gene A in Sample 1 is 3.33 and the RPKM in Sample 2 is 3.33, I would not know if the same proportion of reads in Sample 1 mapped to gene A as in Sample 2. This is because the denominator required to calculate the proportion could be different for the two samples.

Workflow	Category	Package	Reference
Preprocessing of raw data	Raw data QC	FastQC	[8]
		HTQC	[9]
	Read trimming	FASTX-Toolkit	[10]
		FLEXBAR	[11]
Read alignment	Unspliced aligner	MAQ	[13]
		BWA	[14]
		Bowtie	[15]
	Spliced aligner	TopHat	[16]
		MapSplice	[17]
		STAR	[18]
		GSNAP	[19]
RNA-seq specific quality control		RNA-SeQC	[20]
		RSeQC	[21]
		Qualimap 2	[22]
Transcriptome reconstruction	Reference-guided	Cufflinks	[24]
		Scripture	[25]
		StringTie	[26]
	Reference-independent	Trinity	[27]
		Oases	[28]
		transABySS	[29]
Expression quantification	Gene-level quantification	ALEXA-seq	[32]
		Enhanced read analysis of gene expression (ERANGE)	[33]
		Normalization by expected uniquely mappable area (NEUMA)	[34]
	Isoform-level quantification	Cufflinks	[24]
		StringTie	[26]
		RSEM	[35]
		Sailfish	[36]
Differential expression	Gene-level	NOIseq	[23]
		edgeR	[39]
		DESeq	[40]
		SAMseq	[41]
	Isoform-level	Cuffdiff	[24]
		EBSeq	[42]
		Ballgown	[45]

Quality control

- FastQC: FastQC is a quality control tool for high-throughput sequence data (Babraham Institute) and is developed in Java. Import of data is possible from FastQ files, BAM or SAM format. This tool provides an overview to inform about problematic areas, summary graphs and tables to rapid assessment of data. Results are presented in HTML permanent reports. FastQC can be run as a stand-alone application or it can be integrated into a larger pipeline solution.
- NGSQC: cross-platform quality analysis pipeline for deep sequencing data.

Quality control

- RNA-SeQC: RNA-SeQC is a tool with application in experiment design, process optimization and quality control before computational analysis. Essentially, provides three types of quality control: read counts (such as duplicate reads, mapped reads and mapped unique reads, rRNA reads, transcript-annotated reads, strand specificity), coverage (like mean coverage, mean coefficient of variation, 5'/3' coverage, gaps in coverage, GC bias) and expression correlation (the tool provides RPKM-based estimation of expression levels). RNA-SeQC is implemented in Java and is not required installation, however can be run using the GenePattern web interface. The input could be one or more BAM files. HTML reports are generated as output.
- RSeQC: RSeQC analyzes diverse aspects of RNA-Seq experiments: sequence quality, sequencing depth, strand specificity, GC bias, read distribution over the genome structure and coverage uniformity. The input can be SAM, BAM, FASTA, BED files or Chromosome size file (two-column, plain text file). Visualization can be performed by genome browsers like UCSC, IGB and IGV. However, R scripts can also be used to visualization.

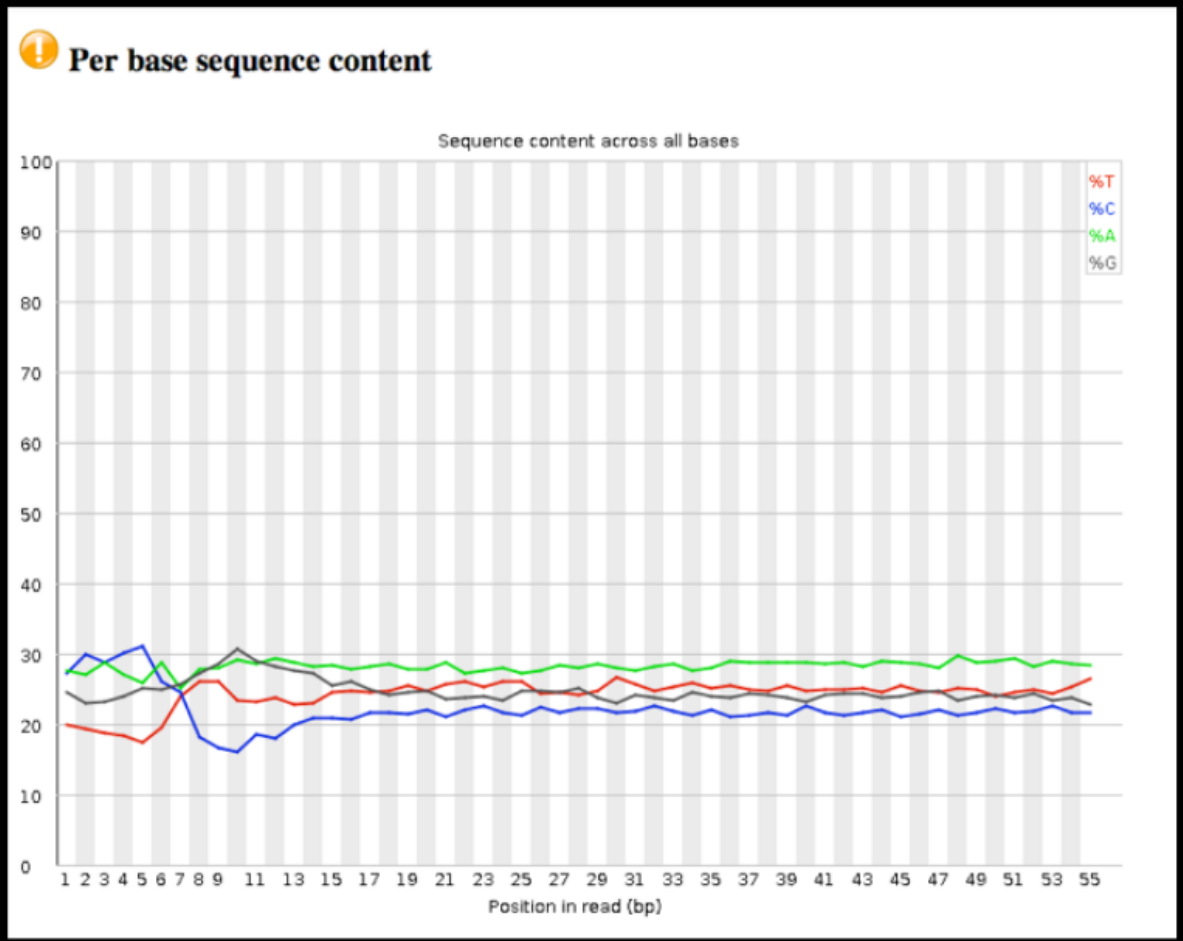
NGS Data Quality Control

- FASTQ format
- Examine quality in an RNA-Seq dataset
- Trim/filter as we see fit, hopefully without breaking anything.

Quality Control is not sexy.

It is vital.

NGS Data Quality: Sequence bias at front of reads?



From a sequence specific bias that is caused by use of random hexamers in library preparation.

Hansen, *et al.*, "Biases in Illumina transcriptome sequencing caused by random hexamer priming" *Nucleic Acids Research*, Volume 38, Issue 12 (2010)

Trimming and adapters removal

- FASTX: FASTX Toolkit is a set of command line tools to manipulate reads in files FASTA or FASTQ format. These commands make possible preprocess the files before mapping with tools like Bowtie. Some of the tasks allowed are: conversion from FASTQ to FASTA format, information about statistics of quality, removing sequencing adapters, filtering and cutting sequences based on quality or conversion DNA/RNA.
- PRINSEQ: PRINSEQ generates statistics of your sequence data for sequence length, GC content, quality scores, n-plicates, complexity, tag sequences, poly-A/T tails, odds ratios. Filter the data, reformat and trim sequences.
- cutadapt: Cutadapt finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads.

Read Mapping

- Alignment algorithm must be
  - fast
  - able to handle SNPs, indels, and sequencing errors
  - allow for introns for reference genome alignment
- Input
  - fastq read library
  - reference genome index
  - insert size mean and stddev(for paired-end libraries)
- Output
  - SAM (text) / BAM (binary) alignment files

De novo Splice Aligners that also use annotation optionally

- TopHat: TopHat is prepared to find de novo junctions. TopHat aligns reads in two steps. Firstly, unspliced reads are aligned with Bowtie. After, the aligned reads are assembled with Maq resulting islands of sequences. Secondly, the splice junctions are determined based on the initially unmapped reads and the possible canonical donor and acceptor sites within the island sequences.

转录组学 | RNA-Seq | 分析 | 工具 | Transcriptome assemblers

Genome-Guided assemblers

- Cufflinks: Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols.
- Scripture: Scripture is a method for transcriptome reconstruction that relies solely on RNA-Seq reads and an assembled genome to build a transcriptome *ab initio*. The statistical methods to estimate read coverage significance are also applicable to other sequencing data. Scripture also has modules for ChIP-Seq peak calling.

Normalization, Quantitative analysis and Differential Expression

- Cufflinks/Cuffdiff: Cufflinks is appropriate to measure global *de novo* transcript isoform expression. It performs assembly of transcripts, estimation of abundances and determines differential expression (Cuffdiff) and regulation in RNA-Seq samples.
- DESeq: DESeq is a Bioconductor package to perform differential gene expression analysis based on negative binomial distribution.
- EdgeR: EdgeR is a R package for analysis of differential expression of data from DNA sequencing methods, like RNA-Seq, SAGE or ChIP-Seq data. edgeR employs statistical methods supported on negative binomial distribution as a model for count variability.
- DEGseq: DEGseq is an R package to identify differentially expressed genes from RNA-Seq data.
- baySeq: This package identifies differential expression in high-throughput 'count' data, such as that derived from next-generation sequencing machines, calculating estimated posterior likelihoods of differential expression (or more complex hypotheses) via empirical Bayesian methods.

Analysis pipeline/Integrated solutions

- easyRNASeq: easyRNASeq calculates the coverage of high-throughput short-reads against a genome of reference and summarizes it per feature of interest (e.g. exon, gene, transcript). The data can be normalized as 'RPKM' or by the 'DESeq' or 'edgeR' package.
- Galaxy: Galaxy is a general purpose workbench platform for computational biology. There are several publicly accessible Galaxy servers that support RNA-Seq tools and workflows, including NBIC's Andromeda, the CBIIT-Giga server, the Galaxy Project's public server, the GeneNetwork Galaxy server, the University of Oslo's Genomic Hyperbrowser, URGI's server (which supports S-MART), and many others.
- GenePattern: GenePattern offers integrated solutions to RNA-Seq analysis.
- Taverna: Taverna is an open source and domain-independent Workflow Management System –a suite of tools used to design and execute scientific workflows and aid *in silico* experimentation.

Visualization tools

- GBrowse: GBrowse is a combination of database and interactive web pages for manipulating and displaying annotations on genomes.
- IGB: The Integrated Genome Browser (IGB, pronounced ig-bee) is an application intended for visualization and exploration of genomes and corresponding annotations from multiple data sources.
- IGV: The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets.
- SeqMonk: SeqMonk is a program to enable the visualisation and analysis of mapped sequence data. It was written for use with mapped next generation sequence data but can in theory be used for any dataset which can be expressed as a series of genomic positions.
- Tablet: Tablet is a lightweight, high-performance graphical viewer for next generation sequence assemblies and alignments.

RNA-Seq Databases

- ENCODE: Encyclopedia of DNA Elements
- RNA-Seq Atlas: a reference database for gene expression profiling in normal tissue by next-generation sequencing.
- SRA: The Sequence Read Archive (SRA) stores raw sequence data from “next-generation” sequencing technologies including 454, IonTorrent, Illumina, SOLiD, Helicos and Complete Genomics. In addition to raw sequence data, SRA now stores alignment information in the form of read placements on a reference sequence.

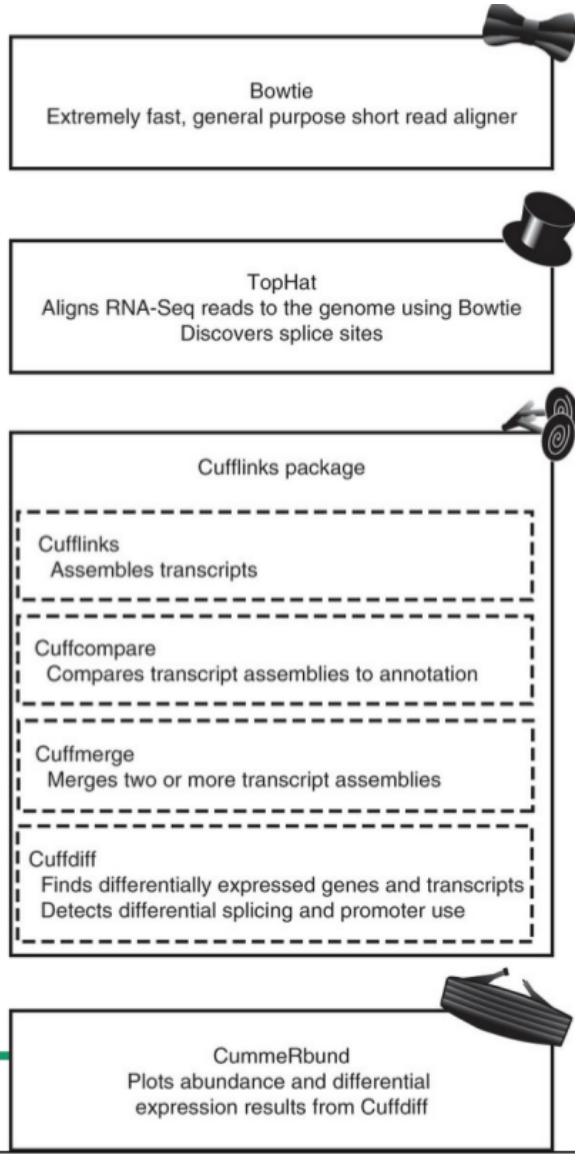
Tuxedo

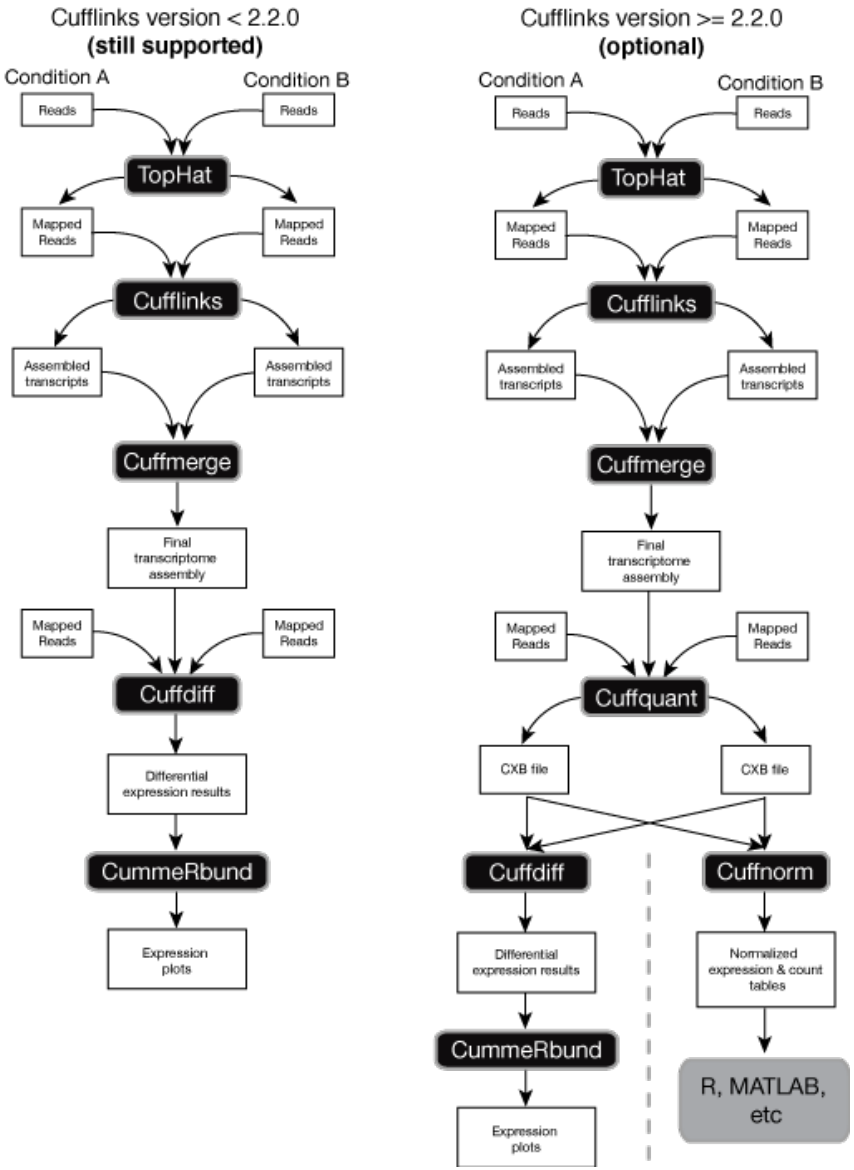
The RNA-seq pipeline “Tuxedo” consists of the TopHat spliced read mapper, that internally uses Bowtie/Bowtie2 short read aligners, and several Cufflinks tools that allows one to assemble transcripts, estimate their abundances, and tests for differential expression and regulation in RNA-Seq samples.

Tool		Tool description
Bowtie		Ultrafast short read aligner
Tophat		Aligns RNA-seq reads to the genome using Bowtie. Discovers splice sites
Cufflinks package	Cufflinks	Assembles transcripts
	Cuffcompare	Compares transcript assemblies to annotation
	Cuffmerge	Merges two or more transcript assemblies
	Cuffdiff	Finds differentially expressed genes and transcripts. Detects differential splicing and promoter use

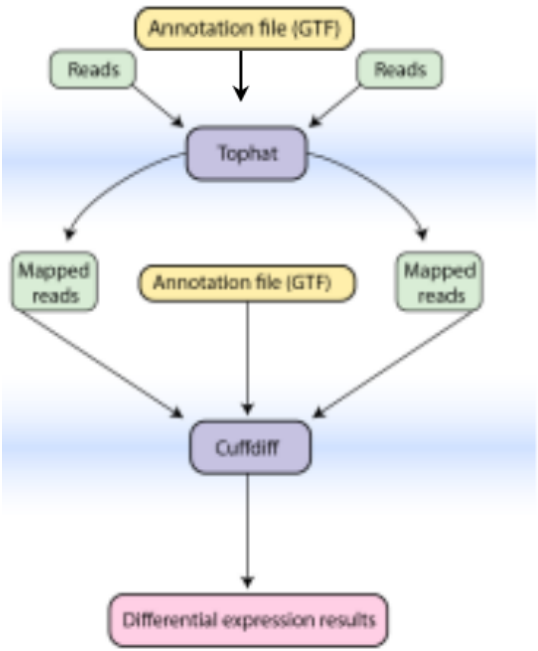
The Tuxedo suite
a complete solution (not the best anymore)

- Bowtie
- TopHat
- Cufflinks
  - **Cuffmerge**
  - **Cuffdiff**
- CummRbund

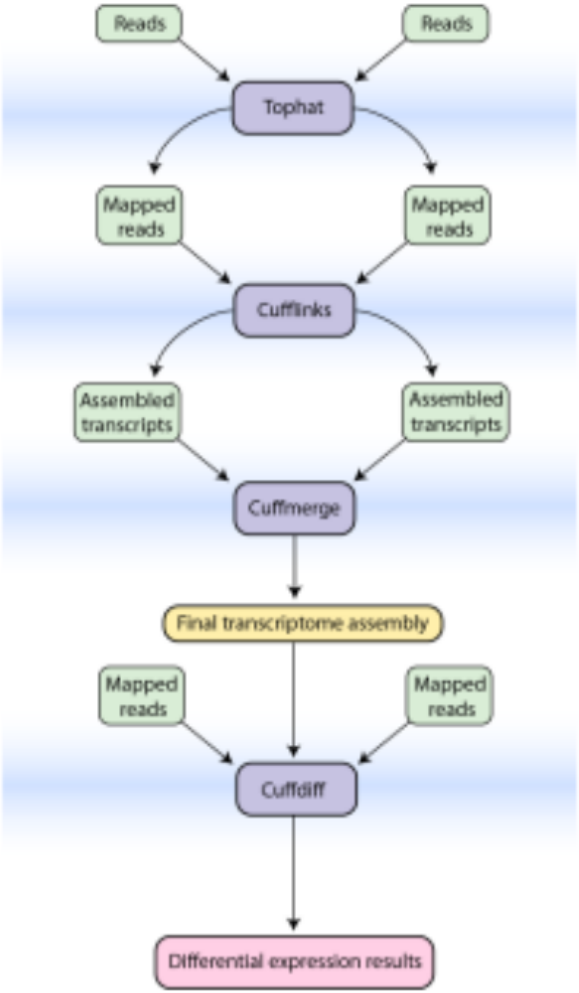



Cufflinks version < 2.2.0
(still supported)
Condition A
Condition B
Reads
Reads
TopHat
Mapped Reads
Mapped Reads
Cufflinks
Assembled transcripts
Assembled transcripts
Cuffmerge
Final transcriptome assembly
Mapped Reads
Mapped Reads
Cuffdiff
Differential expression results
CummeRbund
Expression plots
Cufflinks version >= 2.2.0
(optional)
Condition A
Condition B
Reads
Reads
TopHat
Mapped Reads
Mapped Reads
Cufflinks
Assembled transcripts
Assembled transcripts
Cuffmerge
Final transcriptome assembly
Mapped Reads
Mapped Reads
Cuffquant
CXB file
CXB file
Cuffdiff
Cuffnorm
Differential expression results
Normalized expression & count tables
CummeRbund
R, MATLAB, etc
Expression plots

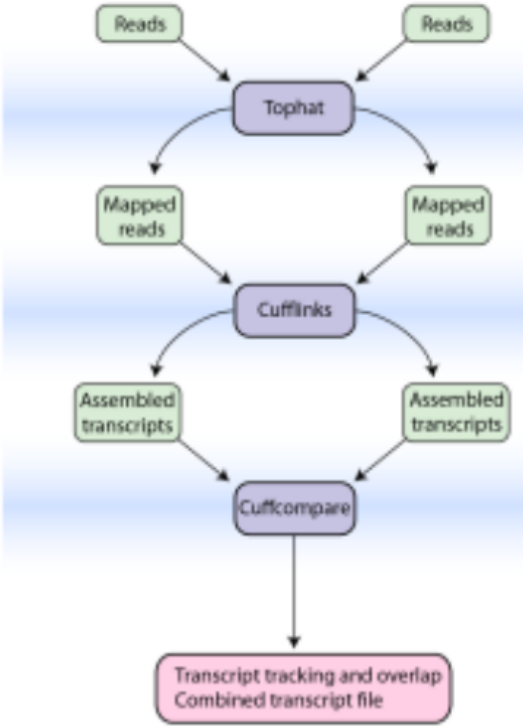
Hands-on 1
Differential expression analysis

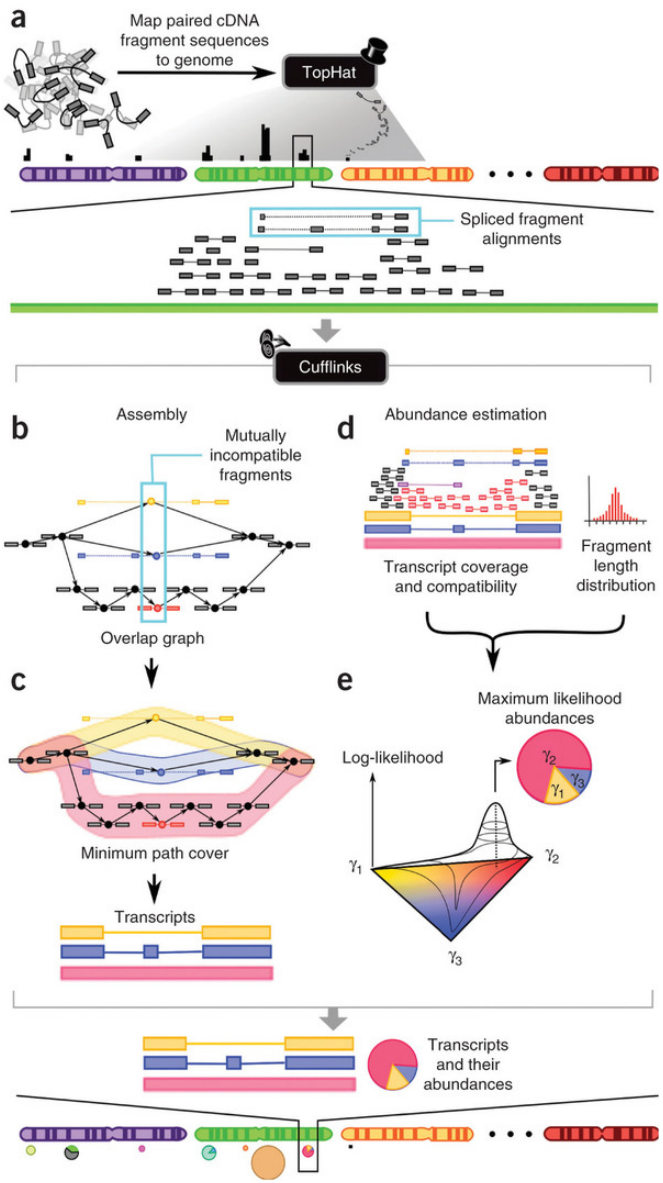


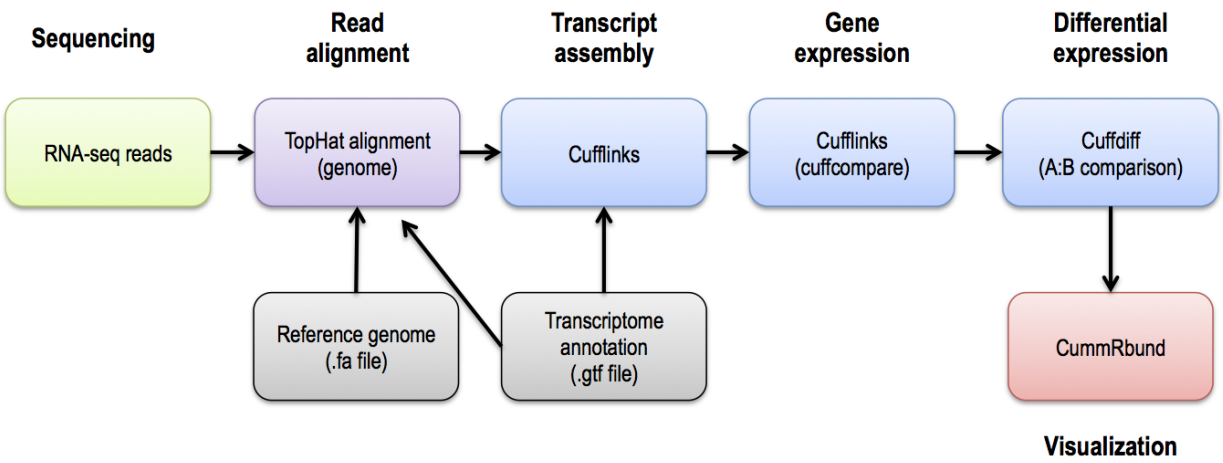
Hands-on 2
Transcript assembly and differential expression analysis

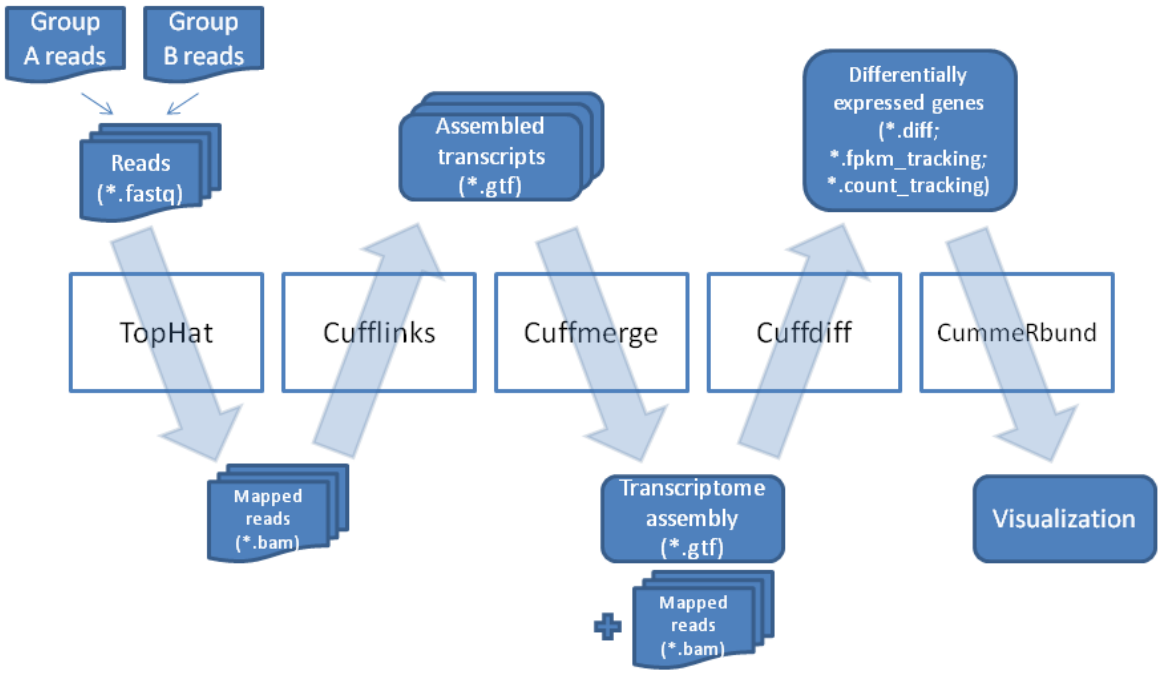


Hands-on 3
Transcript assembly and transcript comparison




a
Map paired cDNA fragment sequences to genome
TopHat
Spliced fragment alignments
Cufflinks
b
Assembly
Mutually incompatible fragments
Overlap graph
d
Abundance estimation
Transcript coverage and compatibility
Fragment length distribution
c
Minimum path cover
Transcripts
e
Maximum likelihood abundances
Log-likelihood
Transcripts and their abundances


Sequencing
Read alignment
Transcript assembly
Gene expression
Differential expression
RNA-seq reads
TopHat alignment (genome)
Cufflinks
Cufflinks (cuffcompare)
Cuffdiff (A:B comparison)
Reference genome (.fa file)
Transcriptome annotation (.gtf file)
CummRbund
Visualization


Group A reads
Group B reads
Reads (*.fastq)
Assembled transcripts (*.gtf)
Differentially expressed genes (*.diff; *.fpkm_tracking; *.count_tracking)
TopHat
Cufflinks
Cuffmerge
Cuffdiff
CummeRbund
Mapped reads (*.bam)
Transcriptome assembly (*.gtf)
Mapped reads (*.bam)
Visualization

Bowtie: ultrafast short read alignment

Bowtie is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.

TopHat: alignment of short RNA-Seq reads

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.

Bowtie: ultrafast short read alignment

Bowtie is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.

TopHat: alignment of short RNA-Seq reads

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.

Cufflinks: transcriptome assembly and differential expression analysis for RNA-Seq

Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols.

CummeRbund: visualization of RNA-Seq differential analysis

CummeRbund is an R package that is designed to aid and simplify the task of analyzing Cufflinks RNA-Seq output.

Cufflinks: transcriptome assembly and differential expression analysis for RNA-Seq

Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols.

CummeRbund: visualization of RNA-Seq differential analysis

CummeRbund is an R package that is designed to aid and simplify the task of analyzing Cufflinks RNA-Seq output.

Tophat

http://ccb.jhu.edu/software/tophat/

Two-step approach:

- (optional) Align to transcriptome first.
- Use bowtie to align whole reads, identify potential exon.
- Split left-over reads into small segments, align independently.
  - Make a database of splice junctions.
  - Map the reads to confirm the splice junctions.

Some considerations:

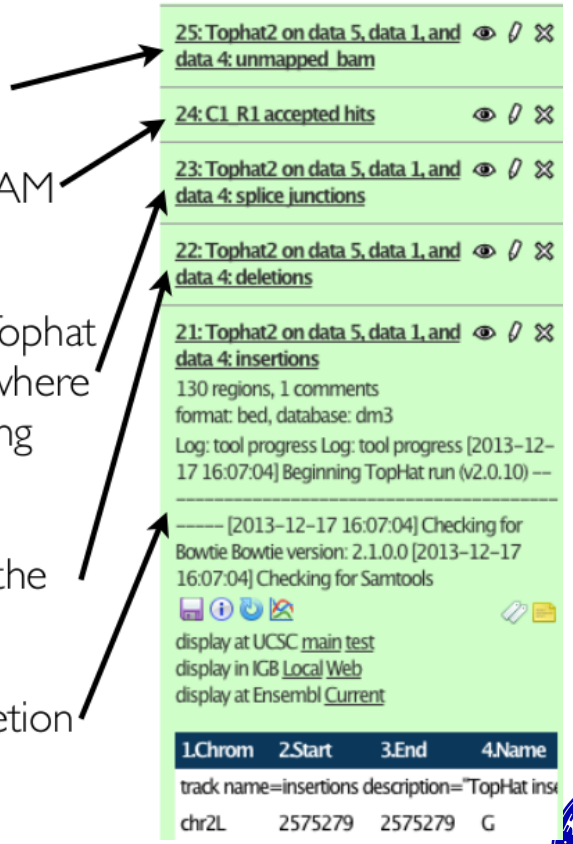
- Does not support soft-clipping

Mapping with Tophat

- Initial Tophat run
- Determine insert size
- Rerun Tophat with correct insert size
- Review mapping statistics

Tophat Output

- unmapped.bam (BAM)
- accepted_hits.bam (BAM): a list of read alignments in BAM/SAM format
- junctions.bed (BED): list BED track of junctions reported by Tophat where each junction consists of two connected BED blocks where each block is as long as the max overhang of nay read spanning juction
- deletions.bed (BED): mentions the last genomic base before the deletion
- insertions.bed (BED): mentions the first genomic base of deletion



Cufflinks

Input:

- Aligned reads.
  - Gmap / Gsnap.
  - Tophat.

What it can do:

- Assemble transcripts.
- Estimate transcript abundance.

Cufflinks

Modes of operation:

- Use predefined transcripts.
- Assemble transcripts assisted by known transcripts.
- Assemble transcripts with no prior knowledge.

When to use:

- Only interested in expression.
- Alternative splicing.

Discover new transcripts (Cuffcompare).

Cuffdiff

Find significant changes in transcript expression, splicing, and promoter use.

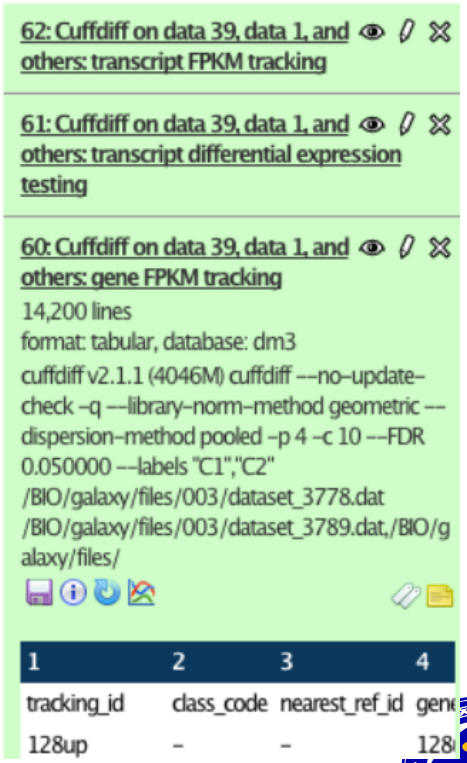
- Support multiple samples and replicates
- Models to estimate the distribution
  - pooled, per-condition, blind
- Output a number of statistic tests
  - fold change, p values, q values

Differential Expression

- Cuffdiff (Cufflinks package)
  - Pairwise comparisons
  - Differnetial gene, transcript, and primary transcript expression
  - Easy to use, well documented
  - Input: transcriptome, SAM/BAM read alignments

Cuffdiff output

- Genes: gene differential FPKM
- Isoforms: Transcript differential FPKM
- CDS: Coding sequence differential FPKM


62: Cuffdiff on data 39, data 1, and others: transcript FPKM tracking

61: Cuffdiff on data 39, data 1, and others: transcript differential expression testing

60: Cuffdiff on data 39, data 1, and others: gene FPKM tracking

14,200 lines

format: tabular, database: dm3

cuffdiff v2.1.1 (4046M) cuffdiff --no-update-check -q --library-norm-method geometric --dispersion-method pooled -p 4 -c 10 --FDR 0.050000 --labels "C1","C2" /BIO/galaxy/files/003/dataset_3778.dat /BIO/galaxy/files/003/dataset_3789.dat,/BIO/galaxy/files/

1	2	3	4
tracking_id	class_code	nearest_ref_id	gene
128up	-	-	128

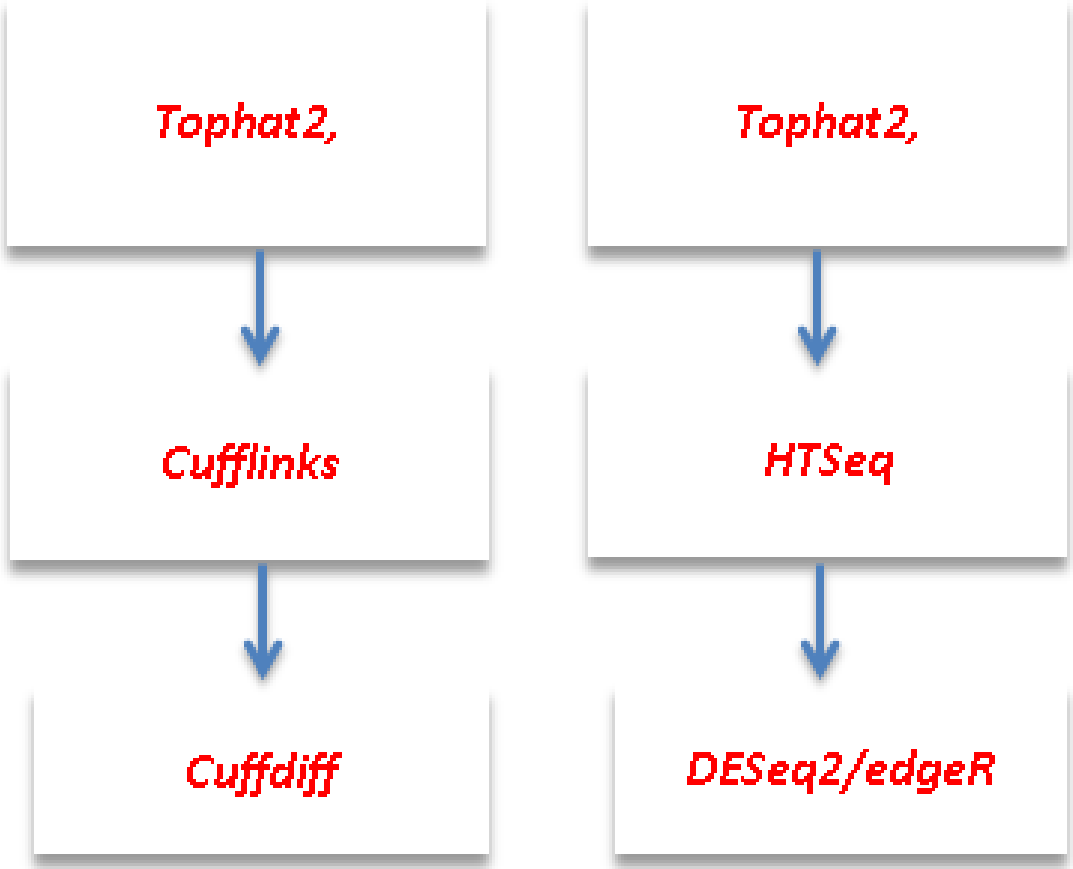
Cuffdiff: differentially expressed genes

Column	Contents
test_stat	value of the test statistic used to compute significance of the observed change in FPKM
p_value	Uncorrected P value for test statistic
q_value	FDR-adjusted p-value for the test statistic
status	Was there enough data to run the test?
significant	and, was the gene differentially expressed?

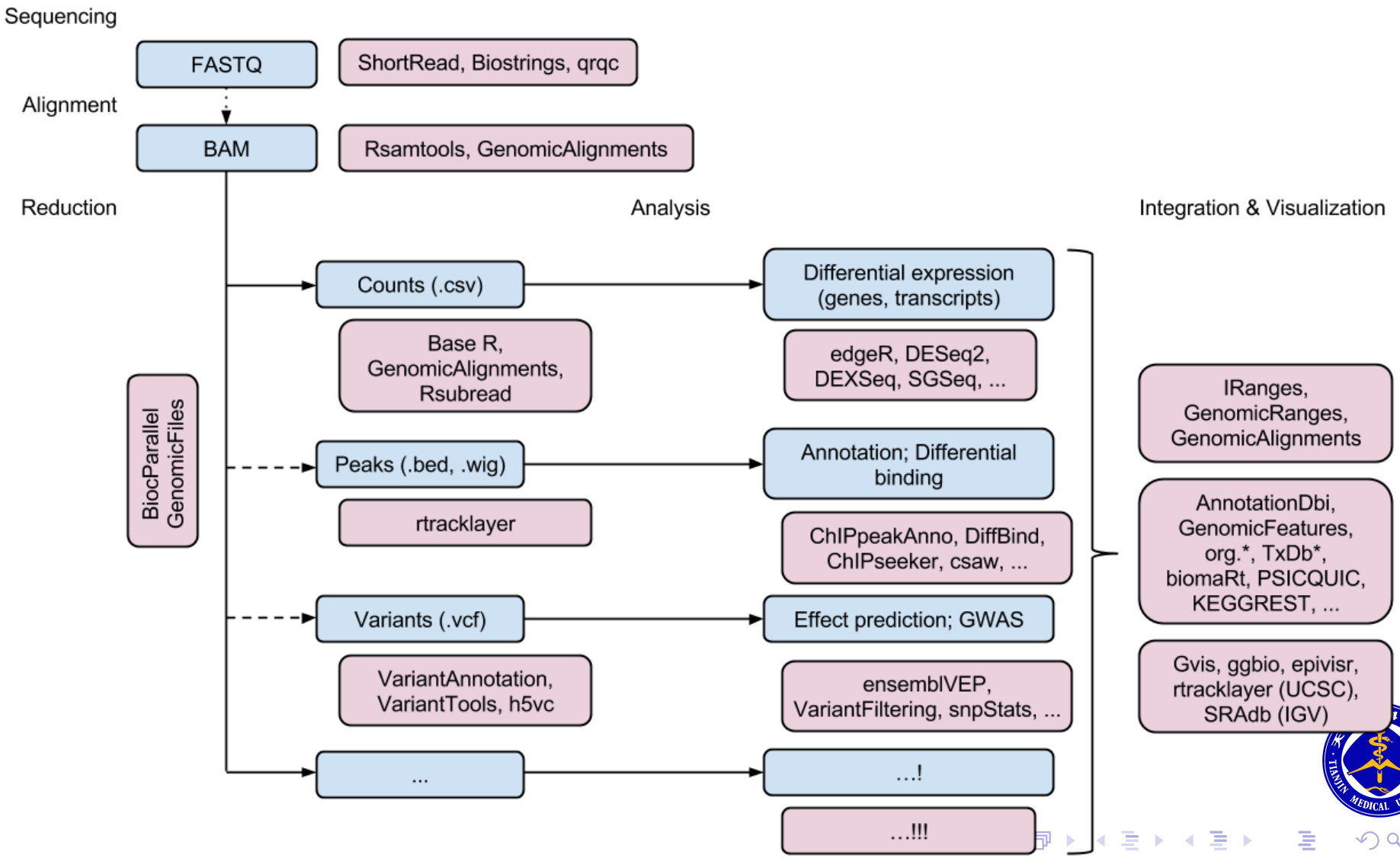
Cuffdiff

- Column 7 (“status”) can be FAIL, NOTEST, LOWDATA or OK
  - Filter and Sort → Filter
    - c7 == 'OK'
- Column 14 (“significant”) can be yes or no
  - Filter and Sort → Filter
    - c14 == 'yes'

Returns the list of genes with
1) enough data to make a call, and
2) that are called as differentially expressed.


Tophat2,
Cufflinks
Cuffdiff
Tophat2,
HTSeq
DESeq2/edgeR

转录组学 | RNA-Seq | 分析 | 工具 | Pipeline | R/Bioconductor



Multiple testing problem

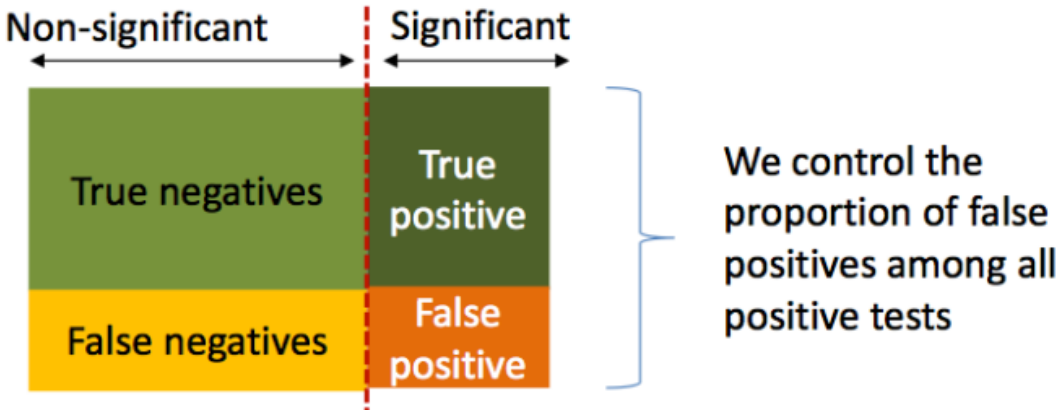
p-value: probability of observing a test statistic that is at least as extreme as the observed one if the null hypothesis were true.

In RNA-seq, we perform 1 test per gene

For example: ~20'000 human genes means ~1'000 significant tests (at P<5%) expected by chance even if there is NO differential expression!

False discovery rate (FDR) correction

For example Benjamini & Hochberg (1995) J. Roy. Stat. Soc. B



UNIVERSITÉ DE FRIBO

Correcting for multiple test

By default DESeq2 uses a FDR of 10%

This means that max 10% of the significant genes could be false positives

	baseMean	log2FC	lfcSE	stat	pvalue	padj
WBGene00000001	366.24625	0.73464	0.27865	2.63646	0.00838	0.02992
WBGene00000002	289.48852	-0.58337	0.32379	-1.80172	0.07159	0.16243
WBGene00000003	93.21683	0.11464	0.27268	0.42044	0.67417	0.79280
WBGene00000004	165.72585	-0.69165	0.32537	-2.12575	0.03352	0.09026
WBGene00000005	439.78883	-0.74071	0.29184	-2.53804	0.01115	0.03771
WBGene00000006	244.67827	-1.16500	0.40588	-2.87031	0.00410	0.01658
WBGene00000007	367.81227	-0.00548	0.38734	-0.01416	0.98870	0.99445
WBGene00000008	19.25137	-0.09673	0.56984	-0.16975	0.86521	0.92061

Adjusted p-values using Benjamini-Hochberg procedure

After DGE RNA-seq?

You get your "gene list", finished?

- Validate
- Typically expect some false-positives
- Genes not in your list may be differentially expressed

Important to always remember

- Your list of genes is produced with an arbitrary significance threshold!

Next?

- Gene-set enrichment tests
- Novel transcripts, novel splice-variants, ...

Downstream analysis

You get your favorite DE gene list, what do you do next?

We want to annotate them functionally!

Which biological functions (or pathways) are enriched among the highly differentially expressed genes?

Are there any common TF/enhancers controlling the expression of the most highly differentially expressed genes?

Shift focus from single genes to collections of genes (gene sets)

Why gene set analysis?

Genes are believed to work together

Many small, but concordant, effects may be detectable together even if they are not detectable individually!

Fewer gene sets than individual genes - less severe multiple testing problem.

A gene set, what's that?

A collection of genes that have something in common, e.g.,

genes that are part of the same pathway co-expressed genes

genes with similar chromosomal positions

genes that are known to be involved in some cancer type

The gene sets have to be defined in advance (i.e., they are not generated in the gene set analysis).

Gene sets and software tools

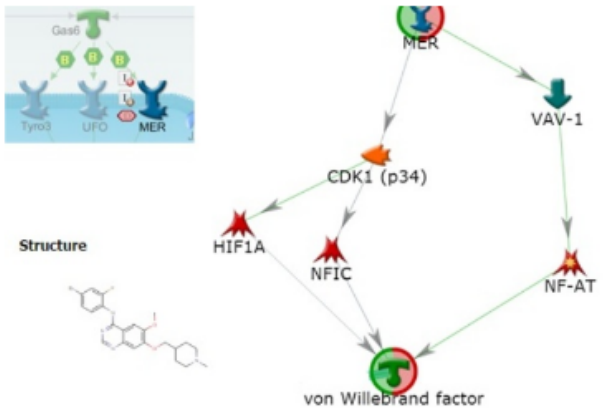
MSigDB (http://www.broadinstitute.org/gsea/msigdb/index.jsp)
GSEA (http://www.broadinstitute.org/gsea/index.jsp)
DAVID (http://david.abcc.ncifcrf.gov)
GO gene ontology (http://www.geneontology.org)
KEGG (http://www.genome.jp/kegg/)
Reactome (http://www.reactome.org)
Publications

Integrated in many software suits and commercial packages like GeneGo MetaCore, Pathway Studio, Ingenuity

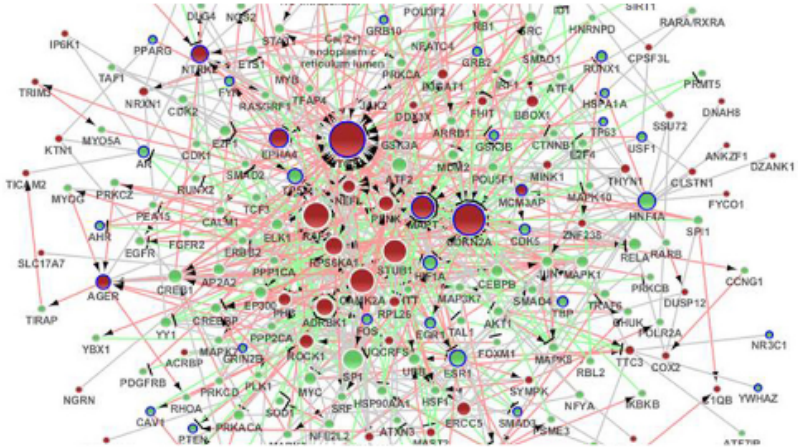
Where to find gene sets?
Some are available in R database packages (KEGG.db, GO.db, reactome.db)

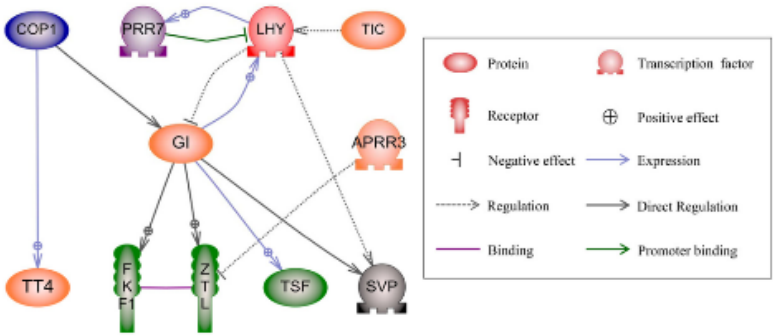
Pathway analysis

Hoping for this...



You often get this...



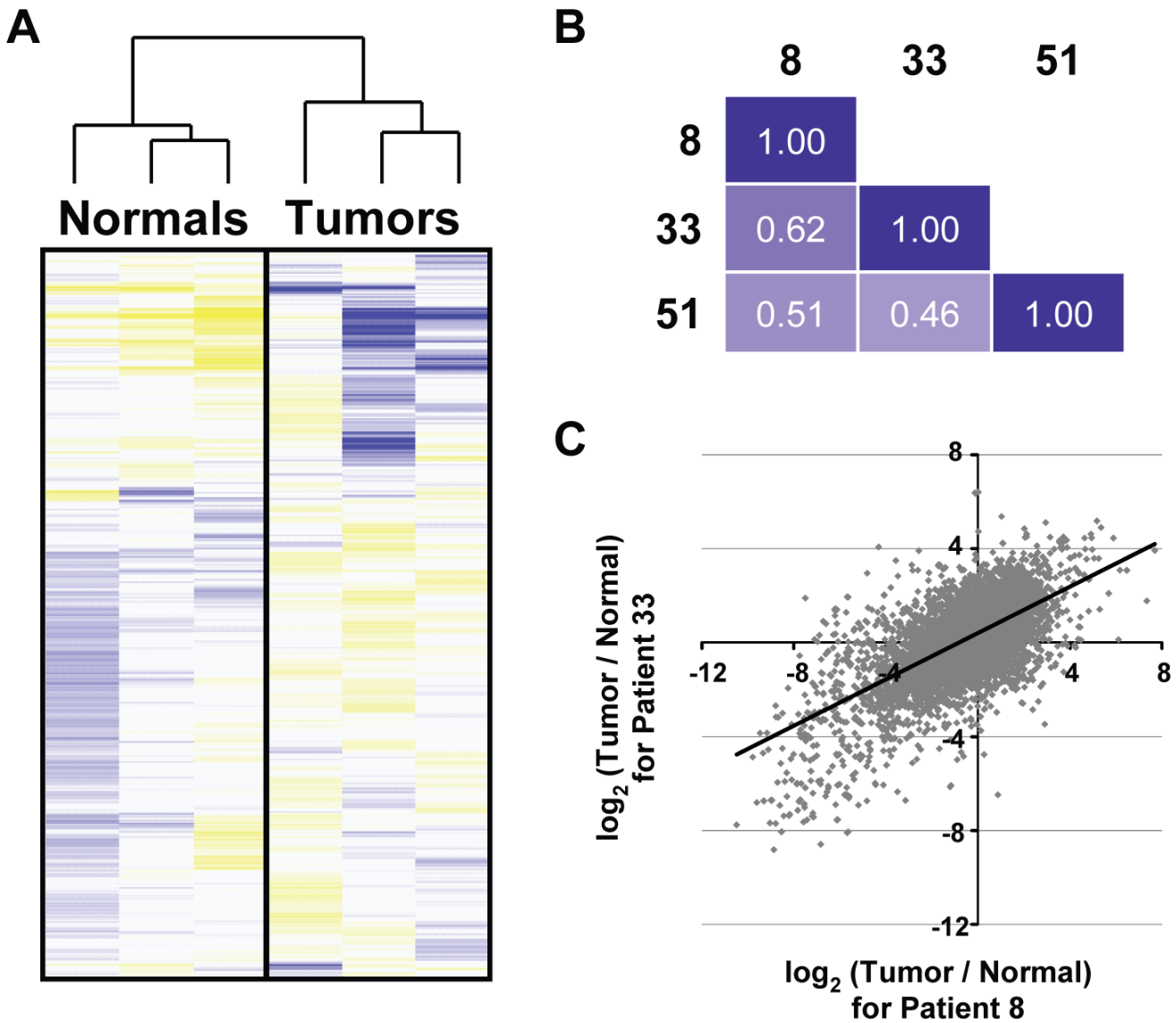


教学提纲

PLoS ONE, 2010

As an example of clinical applications, researchers at the Mayo Clinic used an RNA-Seq approach to identify differentially expressed transcripts between oral cancer and normal tissue samples. They also accurately evaluated the allelic imbalance (AI), ratio of the transcripts produced by the single alleles, within a subgroup of genes involved in cell differentiation, adhesion, cell motility and muscle contraction identifying a unique transcriptomic and genomic signature in oral cancer patients.

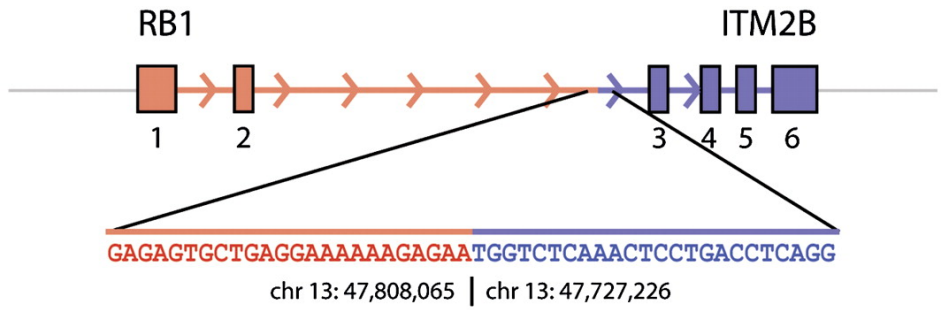
Tuch BB, Laborde RR, Xu X, et al. (2010). “Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations”. PLoS ONE. 5 (2): e9317.

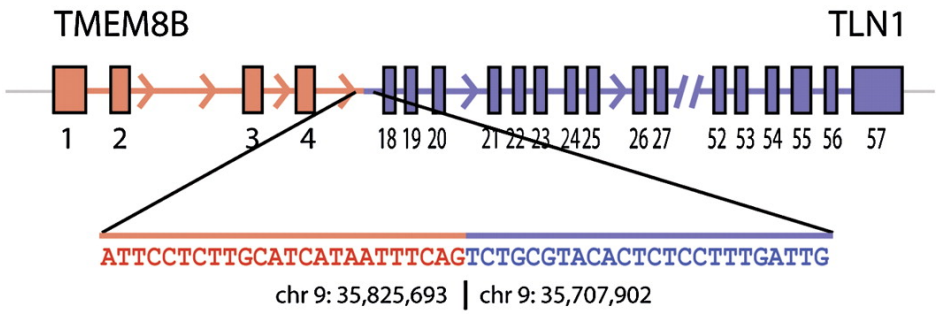

A
Normals
Tumors
B
8
33
51
8
1.00
33
0.62
1.00
51
0.51
0.46
1.00
C
8
4
-4
-8
-12
-12
-8
-4
4
8
log2 (Tumor / Normal) for Patient 33
log2 (Tumor / Normal) for Patient 8

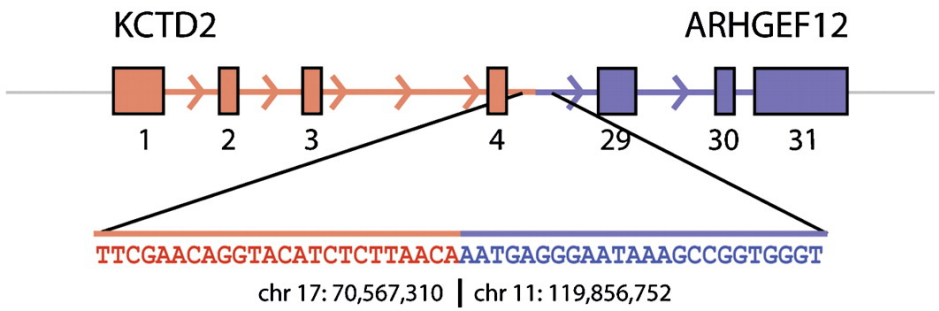
Genome Research, 2010

Novel insight on skin cancer (melanoma) also come from RNA-Seq of melanoma patients. This approach led to the identification of eleven novel gene fusion transcripts originated from previously unknown chromosomal rearrangements. Twelve novel chimeric transcripts were also reported, including seven of those that confirmed previously identified data in multiple melanoma samples.

Berger MF, Levin JZ, Vijayendran K, et al. (April 2010). “Integrative analysis of the melanoma transcriptome”. Genome Res. 20 (4): 413–27.


RB1
ITM2B
1
2
3
4
5
6
GAGAGTGCTGAGGAAAAAAGAGAATGGTCTCAAACTCCTGACCTCAGG
chr 13: 47,808,065 | chr 13: 47,727,226


TMEM8B
TLN1
1
2
3
4
18 19 20
21 22 23 24 25
26 27
52 53 54 55 56
57
ATTCCTCTTGCATCATAATTTCAGTCTGCGTACACTCTCCTTTGATTG
chr 9: 35,825,693 | chr 9: 35,707,902

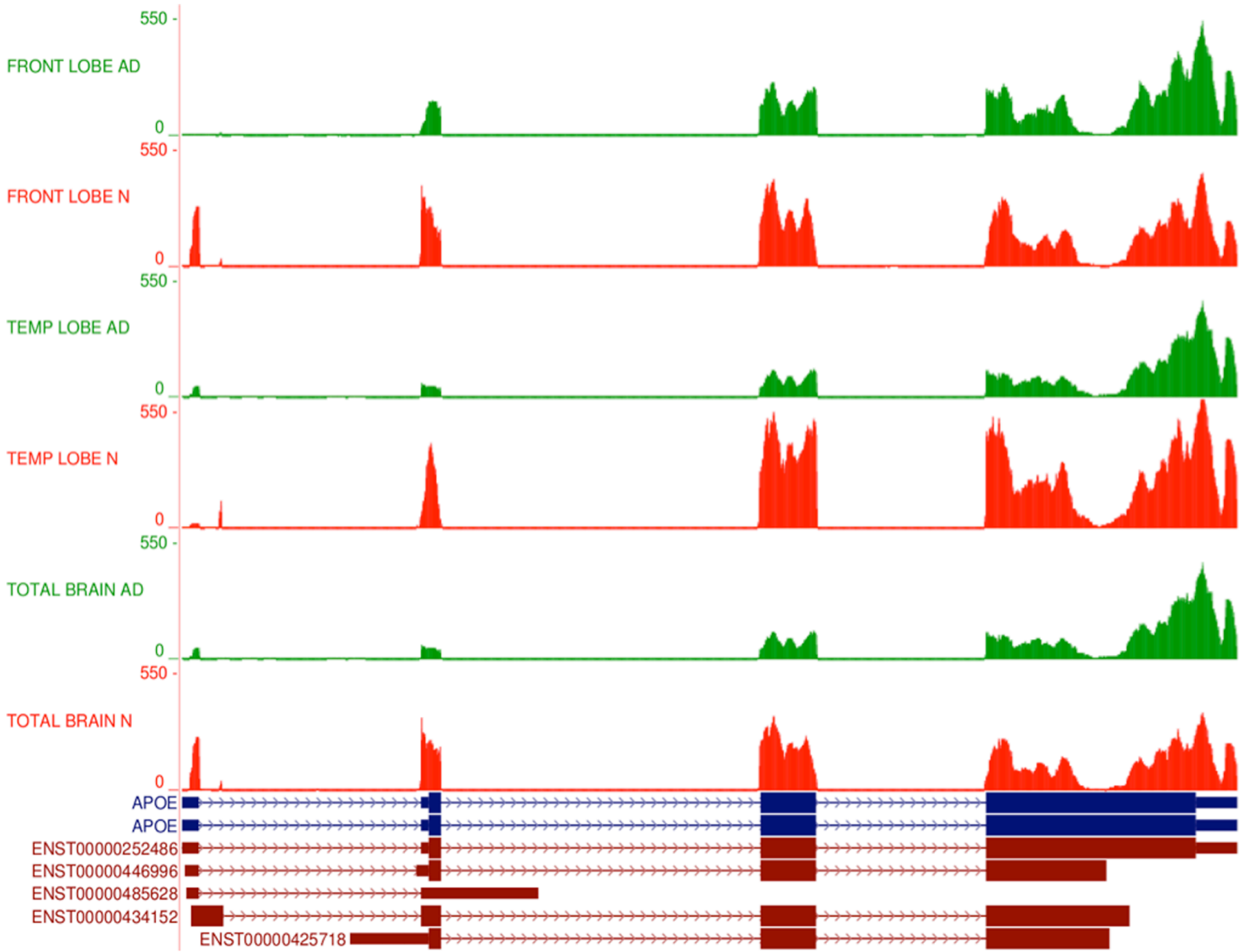

KCTD2
ARHGEF12
1
2
3
4
29
30
31
TTCGAACAGGTACATCTCTTAACAAATGAGGGAATAAAGCCGGTGGGT
chr 17: 70,567,310 | chr 11: 119,856,752

PLoS ONE, 2011

RNA-Seq has been used to study other important chronic diseases such as Alzheimer (AD) and diabetes. In the former case, Twine and colleagues compared the transcriptome of different lobes of deceased AD's patient's brain with the brain of healthy individuals identifying a lower number of splice variants in AD's patients and differential promoter usage of the APOE-001 and -002 isoforms in AD's brains.

Twine NA, Janitz K, Wilkins MR, Janitz M (2011). "Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease". PLoS ONE. 6 (1): e16266.

转录组学 | RNA-Seq | 实例 | splice variant



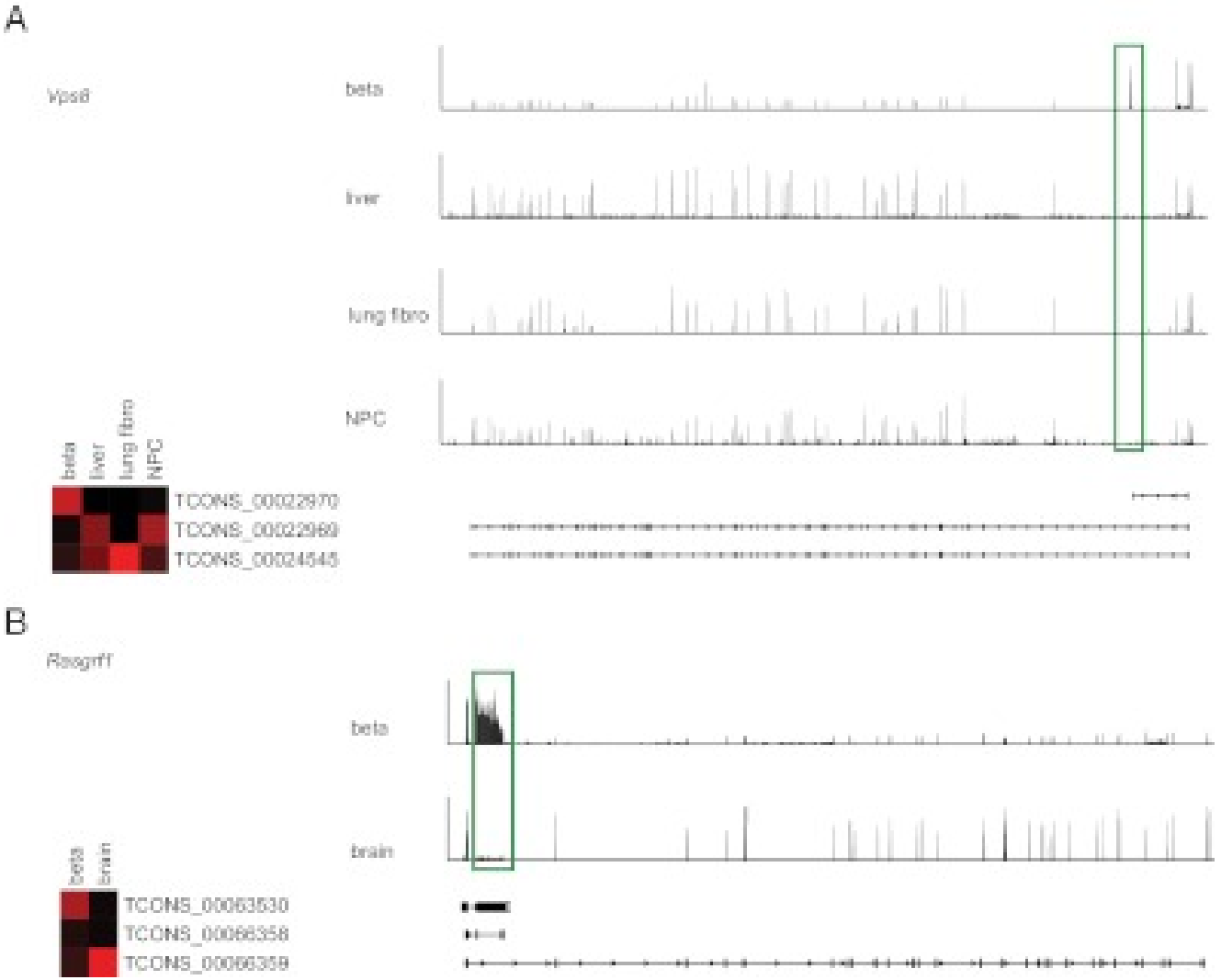
Mol. Endocrinol, Cell Metab, 2012

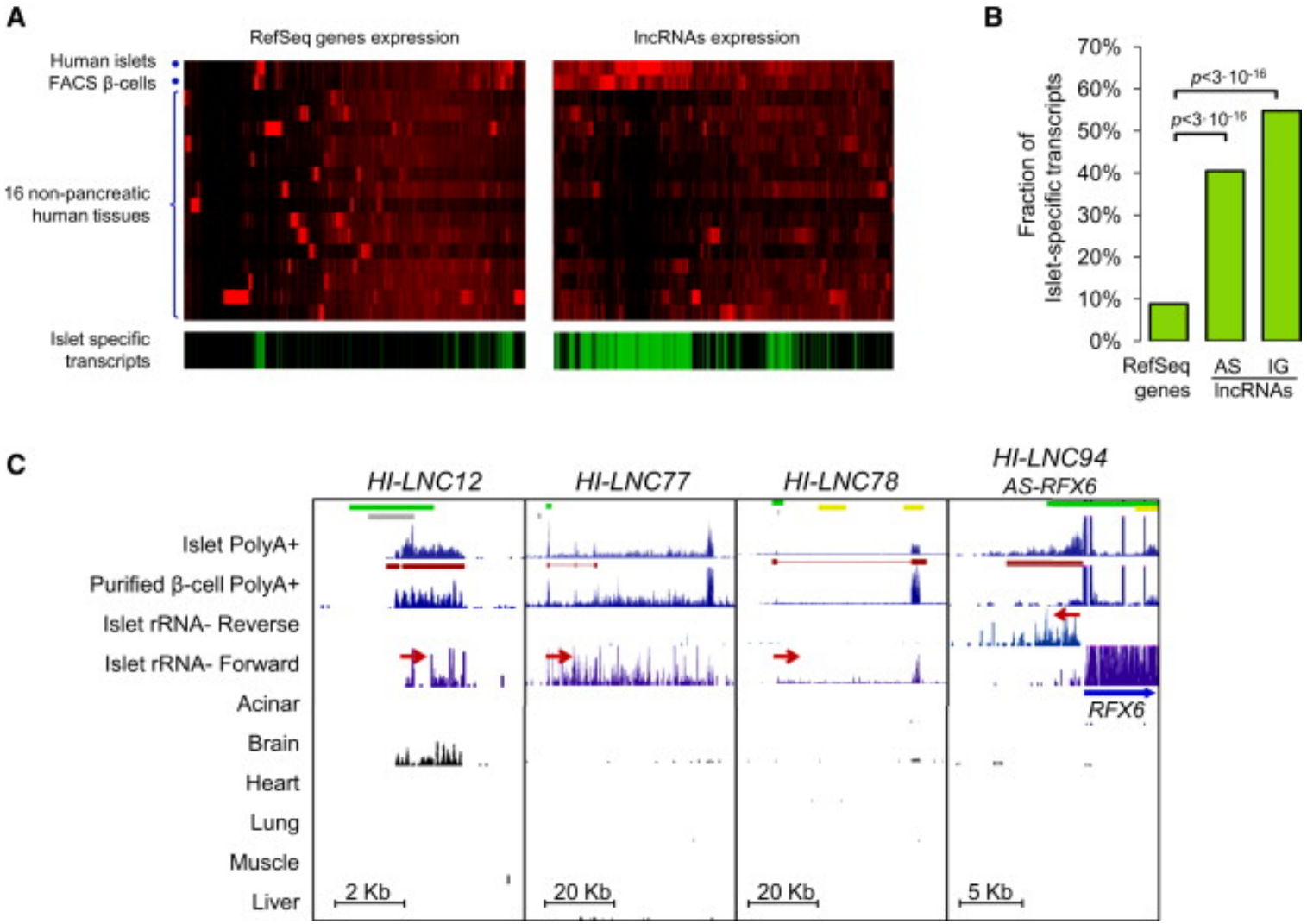
In the latter case, different groups showed the unicity of the beta-cells transcriptome in diabetic patients in terms of transcripts accumulation and differential promoter usage and long non coding RNAs (lncRNAs) signature.

Ku GM, Kim H, Vaughn IW, et al. (October 2012). "Research resource: RNA-Seq reveals unique features of the pancreatic β-cell transcriptome". Mol. Endocrinol. 26 (10): 1783–92.

Morán I, Akerman I, van de Bunt M, et al. (October 2012). "Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes". Cell Metab. 16 (4): 435–48.

转录组学 | RNA-Seq | 实例 | splicing events and alternative promoter use



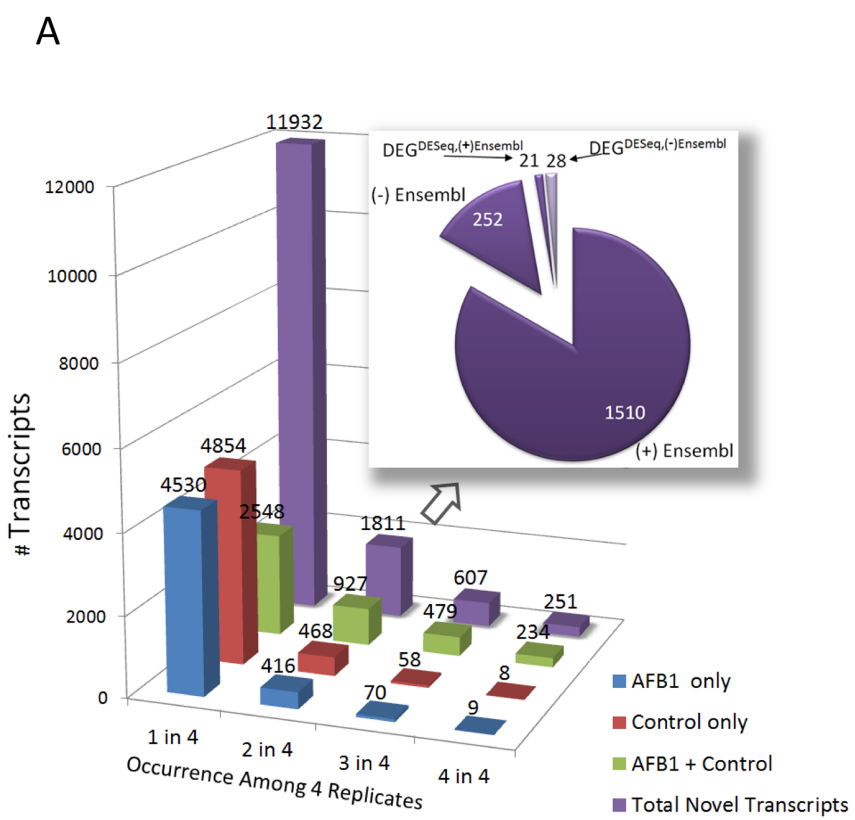

A
RefSeq genes expression
lncRNAs expression
Human islets
FACS β-cells
16 non-pancreatic human tissues
Islet specific transcripts
B
Fraction of Islet-specific transcripts
70%
60%
50%
40%
30%
20%
10%
0%
p<3·10-16
p<3·10-16
RefSeq genes
AS lncRNAs
IG
C
HI-LNC12
HI-LNC77
HI-LNC78
HI-LNC94
AS-RFX6
Islet PolyA+
Purified β-cell PolyA+
Islet rRNA- Reverse
Islet rRNA- Forward
Acinar
Brain
Heart
Lung
Muscle
Liver
RFX6
2 Kb
20 Kb
20 Kb
5 Kb

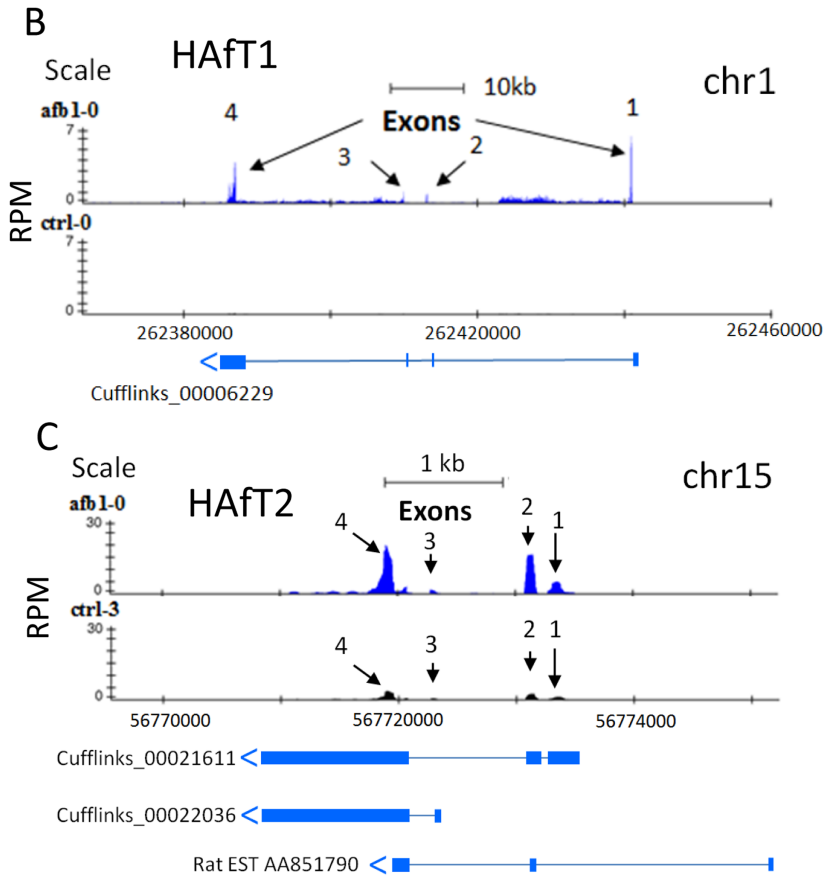
PLoS ONE, 2013

NGS technology identified several previously undocumented differentially-expressed transcripts in rats treated with AFB1, a potent hepatocarcinogen. Nearly 50 new differentially-expressed transcriptions were identified between the controls and AFB1-treated rats. Additionally potential new exons were identified, including some that are responsive to AFB1. The next-generation sequencing pipeline identified more differential gene expressions compared with microarrays, particularly when DESeq software was utilized. Cufflinks identified two novel transcripts that were not previously annotated in the Ensembl database; these transcripts were confirmed using cloning PCR.

Merrick B. A.; Phadke D. P.; Auerbach S. S.; Mav D.; Stiegelmeyer S. M.; Shah R. R.; Tice R. R. (2013). “RNA-seq profiling reveals novel hepatic gene expression pattern in Aflatoxin B1 treated rats”. PLoS ONE. 8: e61768.

转录组学 | RNA-Seq | 实例 | novel transcript

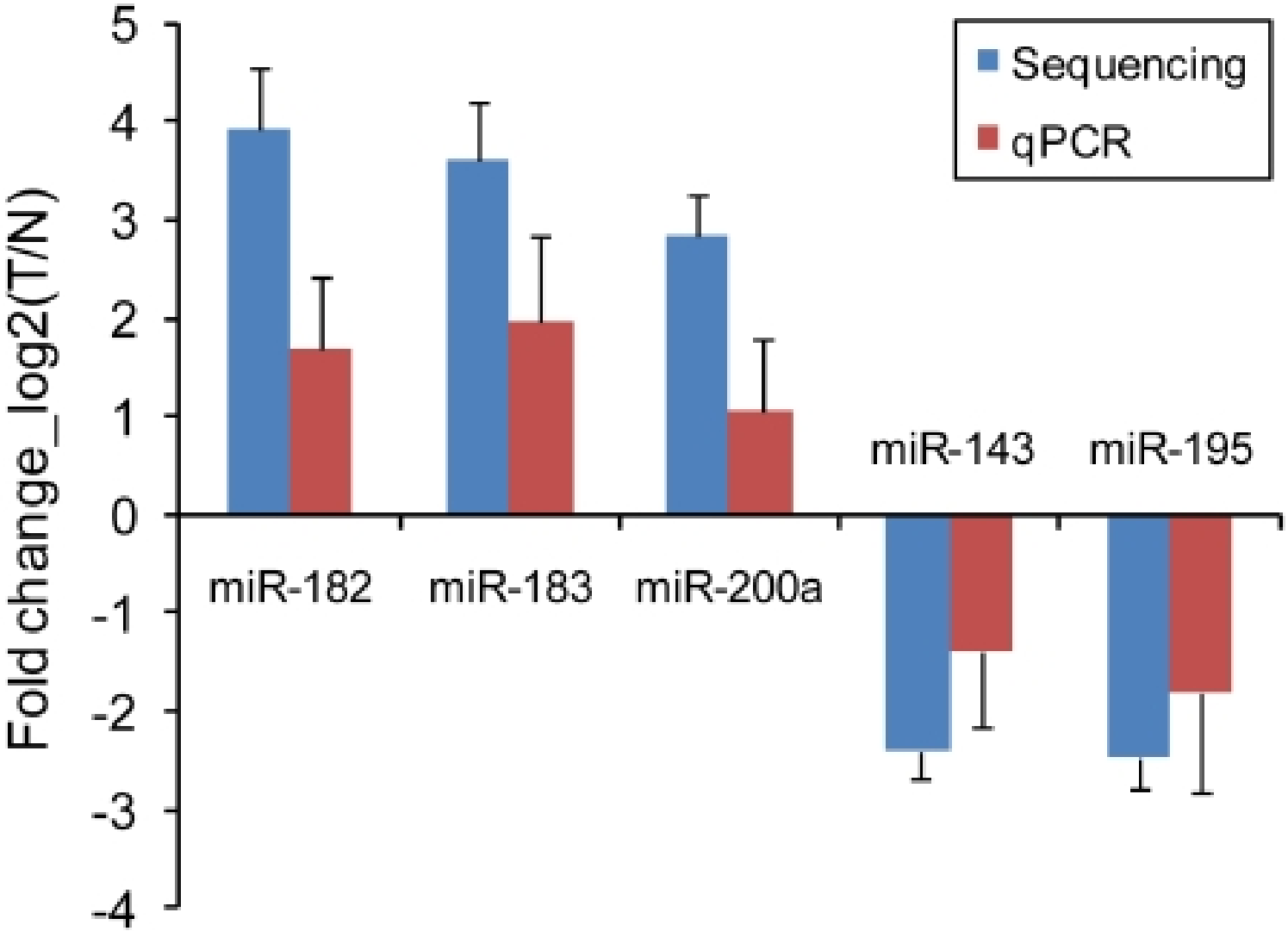




PLoS ONE, 2011

Han et al. (2011) examined microRNA expression differences in bladder cancer patients in order to understand how changes and dysregulation in microRNA can influence mRNA expression and function. Several microRNAs were differentially expressed in the bladder cancer patients. Upregulation in the aberrant microRNAs was more common than downregulation in the cancer patients. One of the upregulated microRNAs, has-miR-96, has been associated with carcinogenesis, and several of the overexpressed microRNAs have also been observed in other cancers, including ovarian and cervical. Some of the downregulated microRNAs in cancer samples were hypothesized to have inhibitory roles.

Han Y.; Chen J.; Zhao X.; Liang C.; Wang Y.; Sun L.; Jiang Z.; Zhang Z.; Yang R.; Chen J.; Li Z.; Tang A.; Li X.; Ye J.; Guan Z.; Gui Y.; Cai Z. (2011). “MicroRNA expression signatures of bladder cancer revealed by deep sequencing”. PLoS ONE. 6: e18286.


Sequencing
qPCR
Fold change_log2(T/N)
5
4
3
2
1
0
-1
-2
-3
-4
miR-182
miR-183
miR-200a
miR-143
miR-195

教学提纲

教学提纲

知识点

-
-

技能

-
-

教学提纲

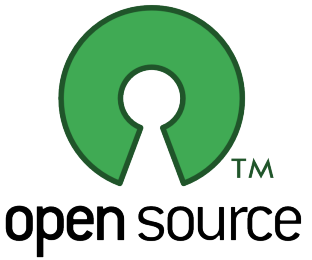
1.

2.

下节预告

-
-

Powered by



TEX LATEX XETEX Beamer