RNA-Seq (1)

BCB 5250 Introduction to Bioinformatics II

Spring 2020

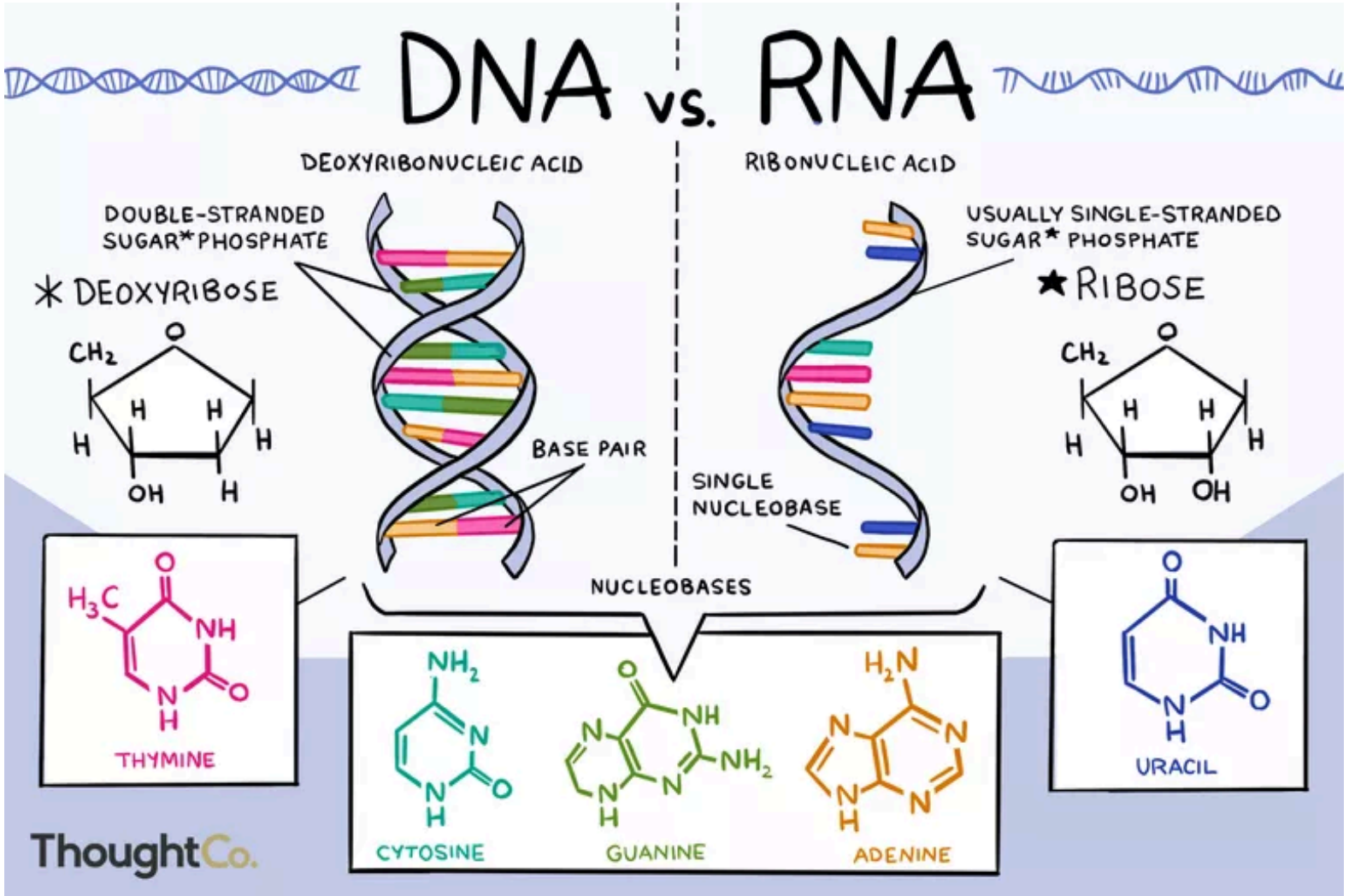
Tae-Hyuk (Ted) Ahn
Department of Computer Science
Program of Bioinformatics and Computational Biology
Saint Louis University

SAINT LOUIS
UNIVERSITY
EST. 1818

RNA-Seq

- RNA + Seq

The Differences Between DNA and RNA



https://www.thoughtco.com/dna-versus-rna-608191

The Differences Between DNA and RNA

- DNA contains the sugar deoxyribose, while RNA contains the sugar ribose. The only difference between ribose and deoxyribose is that ribose has one more -OH group than deoxyribose, which has -H attached to the second (2') carbon in the ring.
- DNA is a double-stranded molecule, while RNA is a single-stranded molecule.
- DNA is stable under alkaline conditions, while RNA is not stable.
- DNA and RNA perform different functions in humans. DNA is responsible for storing and transferring genetic information, while RNA directly codes for amino acids and acts as a messenger between DNA and ribosomes to make proteins.
- DNA and RNA base pairing is slightly different since DNA uses the bases adenine, thymine, cytosine, and guanine; RNA uses adenine, uracil, cytosine, and guanine. Uracil differs from thymine in that it lacks a methyl group on its ring.

https://www.thoughtco.com/dna-versus-rna-608191

What is RNA-Seq?

- RNA-seq (RNA-sequencing) is a technique that can examine the quantity and sequences of RNA in a sample using next generation sequencing (NGS).
- It analyzes the transcriptome of gene expression patterns encoded within our RNA.
- It is providing researchers with visibility into previously undetected changes occurring in disease states, in response to therapeutics, under different environmental conditions, and across a broad range of other study designs.
- RNA-Seq allows researchers to detect both known and novel features in a single assay, enabling the detection of transcript isoforms, gene fusions, single nucleotide variants, and other features without the limitation of prior knowledge.

Benefits of RNA Sequencing

RNA-Seq with next-generation sequencing (NGS) is increasingly the method of choice for researchers studying the transcriptome. It offers numerous advantages over gene expression arrays.

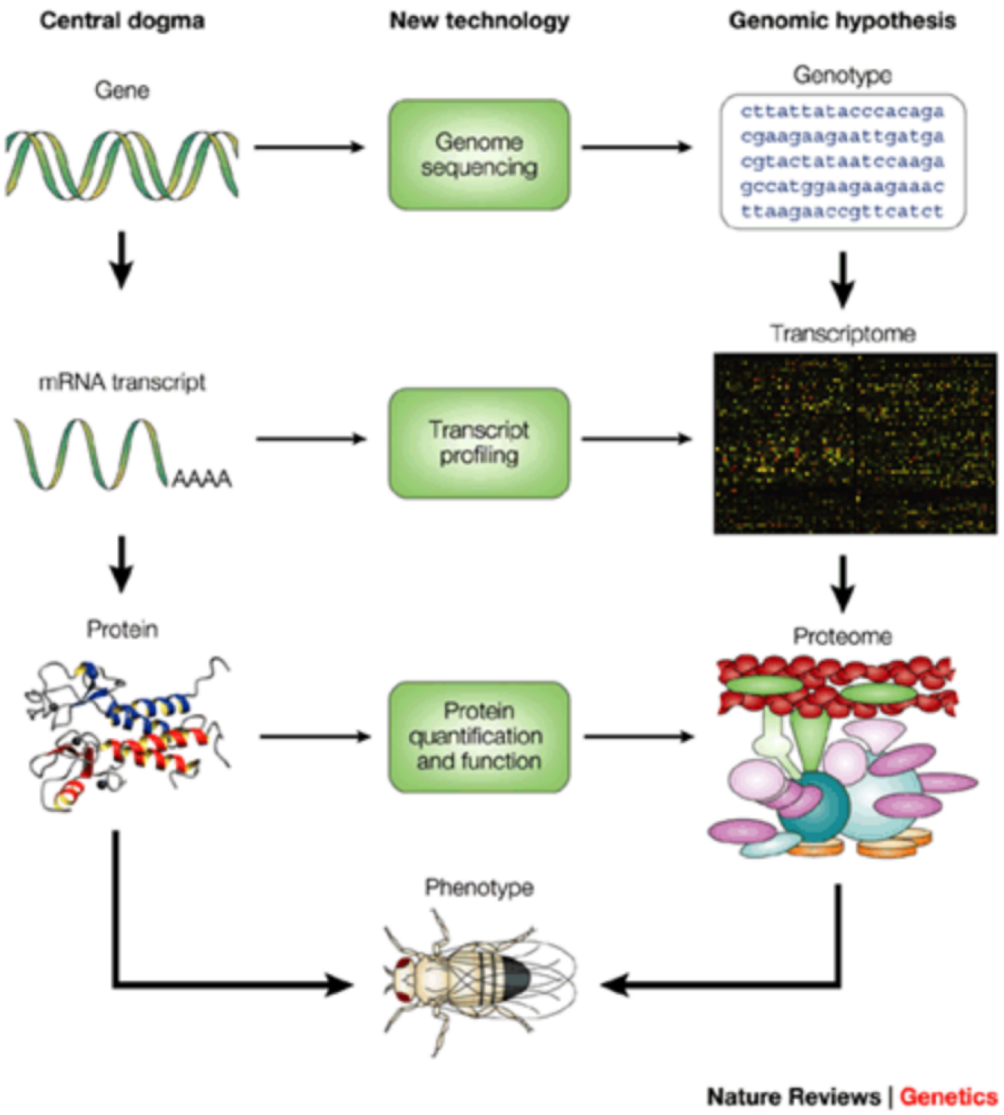
- Broader dynamic range enables more sensitive and accurate measurement of gene expression
- Not limited by prior knowledge - captures both known and novel features
- Can be applied to any species, even if reference sequencing is not available
- A better value, often delivering advantages at a comparable or lower price per sample than many arrays

https://www.illumina.com/techniques/sequencing/rna-sequencing.html

What are the applications of RNA-seq?

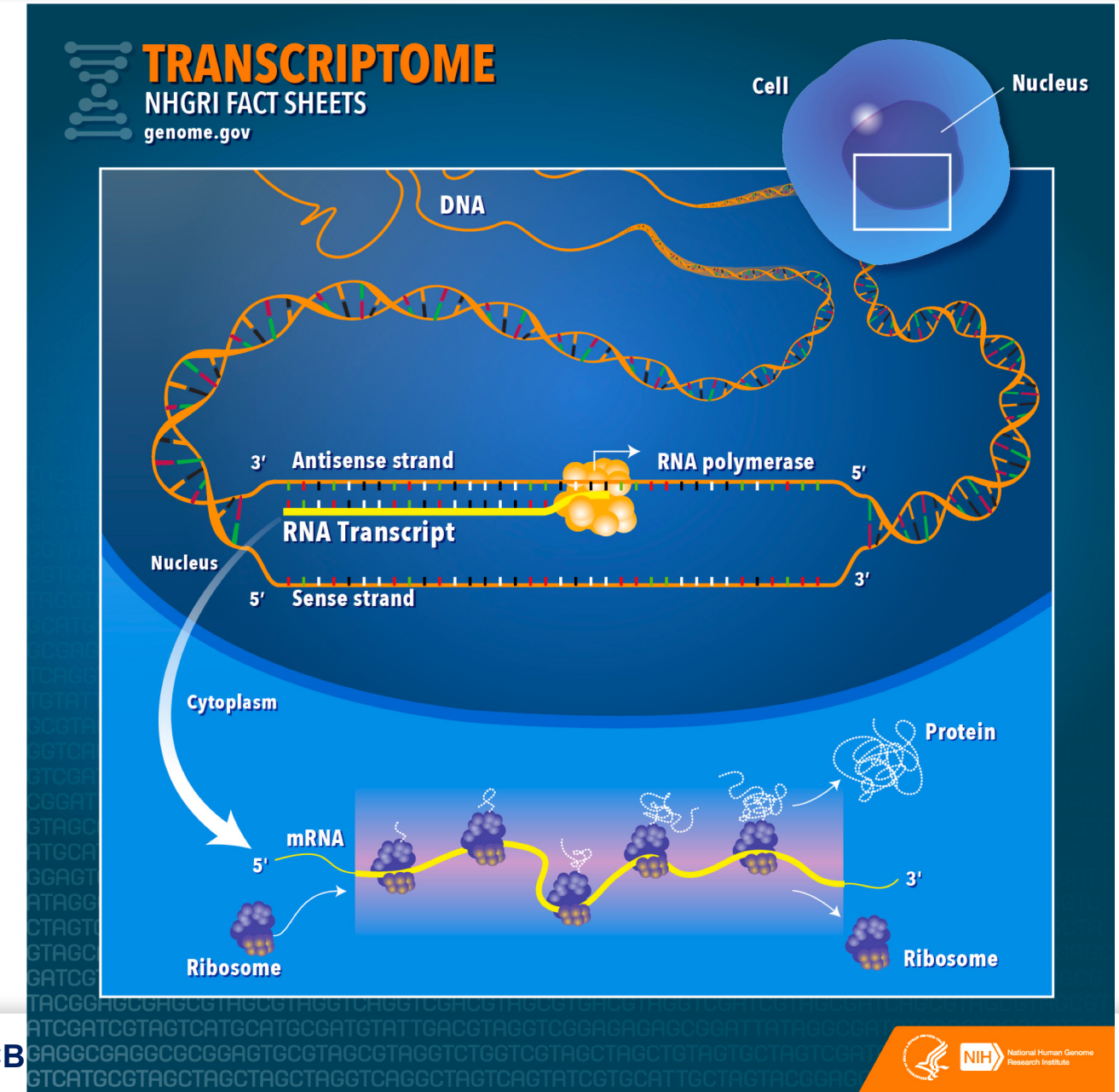
- RNA-seq lets us investigate and discover the transcriptome, the total cellular content of RNAs including mRNA, rRNA and tRNA.
- Understanding the transcriptome is key if we are to connect the information on our genome with its functional protein expression.
- RNA-seq can tell us which genes are turned on in a cell, what their level of expression is, and at what times they are activated or shut off2.
- This allows scientists to more deeply understand the biology of a cell and assess changes that may indicate disease.
- Some of the most popular techniques that use RNA-seq are transcriptional profiling, SNP identification, RNA editing and differential gene expression analysis3.

How Does a Genotype Turn Into a Phenotype?



- Genotypes is the genetic makeup of an organism, while the phenotype is the physical traits of that organism.
- For genotype to turn into phenotype, it starts with transcription, in which the DNA is copied into RNA. After transcription and before translation, RNA transcripts to produce mRNAs.

Transcriptome Fact Sheet



https://www.genome.gov/13014330/transcriptome-fact-sheet/

A transcriptome is a collection of all the gene readouts present in a cell.

How to obtain a transcriptome?

Review: RNA-Seq: a revolutionary tool for transcriptomics (*Nat Rev Genet*, 2010)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/

Various technologies have been developed to deduce and quantify the transcriptome, including hybridization-or sequence-based approaches.

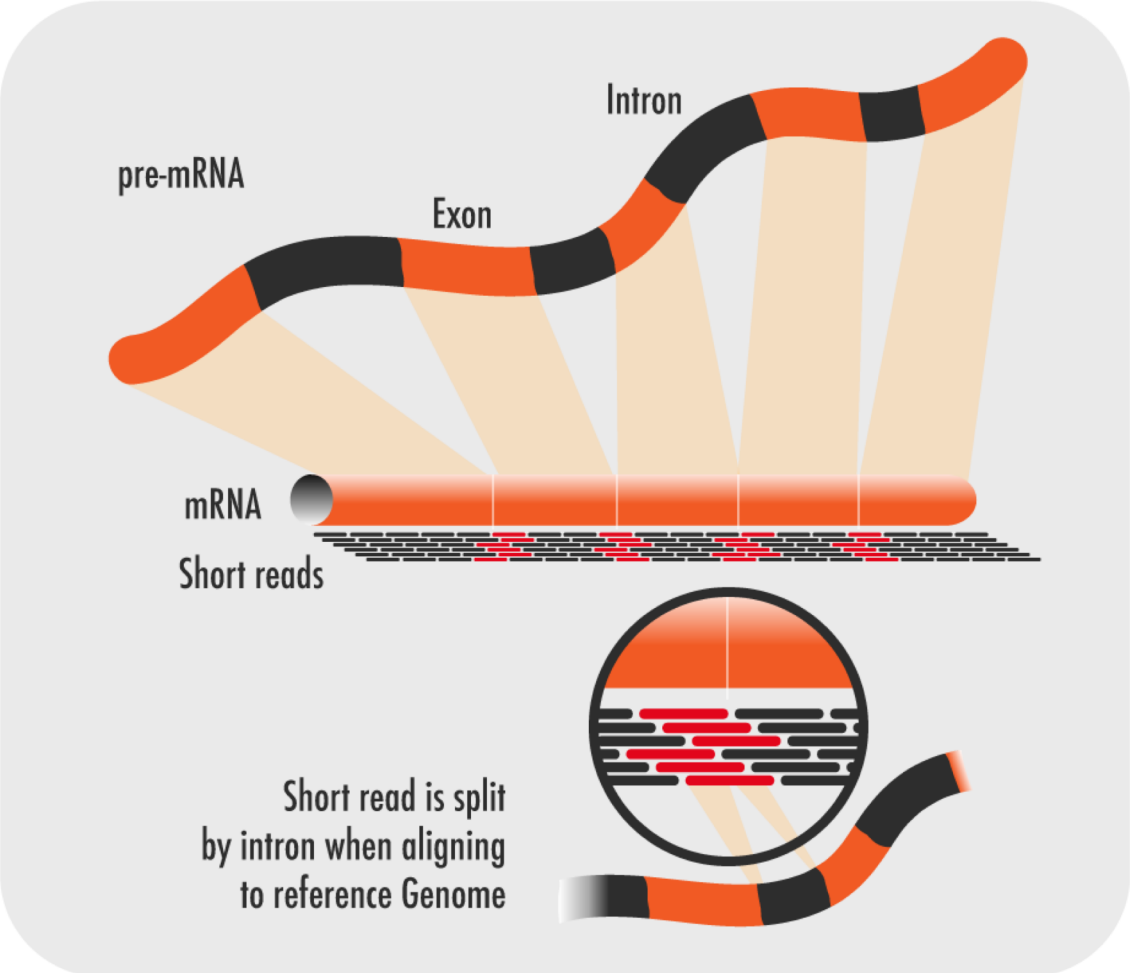
- Hybridization-based approaches typically involve incubating fluorescently labelled cDNA with custom-made microarrays or commercial high-density oligo microarrays.
- In contrast to microarray methods, sequence-based approaches directly determine the cDNA sequence.
  - Initially, Sanger sequencing of cDNA or EST libraries8,9 was used, but this approach is relatively low throughput, expensive and generally not quantitative.
  - Tag-based methods were developed to overcome these limitations, including serial analysis of gene expression (SAGE), cap analysis of gene expression (CAGE) and massively parallel signature sequencing (MPSS).
  - High-throughput DNA sequencing methods has provided a new method for both mapping and quantifying transcriptomes. This method, termed RNA-Seq (RNA sequencing).

How to obtain a transcriptome?

- RNA-sequencing
  - Traditional (Bulk) RNA-seq
  - Single-cell RNA-seq

Traditional RNA-Seq methods analyzed the RNA of an entire population of cells, but only yielded a bulk average of the measurement instead of representing each individual cell's transcriptome. By analyzing the transcriptome of a single cell at a time, the heterogeneity of a sample is captured and resolved to the fundamental unit of living organisms—the cell.

RNA-seq



RNA-seq data uses uses short reads of mRNA which is free of intronic non-coding DNA. These reads must then be aligned back to the reference genome.

An RNA-seq protocol

Experiment Planning

Preparation prior to starting your RNA-seq experiment is essential.

- What method of RNA purification are you using?
- How many reads will you need?
- Which platform will you use?
- What reference genome will you use?
- How are you assessing the quality of your RNA?
- Do you need to enrich your target RNA?
- Will you barcode your RNA?
- Have I got enough biological and technical replicates?
- Single-read or paired-end sequencing?

An RNA-seq protocol

cDNA Library Preparation

- After these points have been considered, you can start preparing your cDNA library.
- This will require adding the platform-specific “adapter sequences” and amplification of the DNA, but the exact procedure will be very specific to the platform used at this stage.
- The amplification of the DNA involves a reverse transcriptase mediated first strand synthesis followed by a DNA polymerase-mediated second strand synthesis.

An RNA-seq protocol

cDNA Sequencing

- Once the library is prepared, and adapters added, you can use your chosen sequencing platform to sequence your cDNA library to your desired depth.
- Once your transcript data has been produced, you can map the data to your reference genome.
- The alignment process can be complicated by the presence of splice variants and modifications, and the choice of reference genome used will also vary how difficult this stage is.

Only Short Reads in RNA-Seq? No

PacBio

(https://www.pacb.com/applications/rna-sequencing/)

DISCOVER FULL-LENGTH TRANSCRIPTS

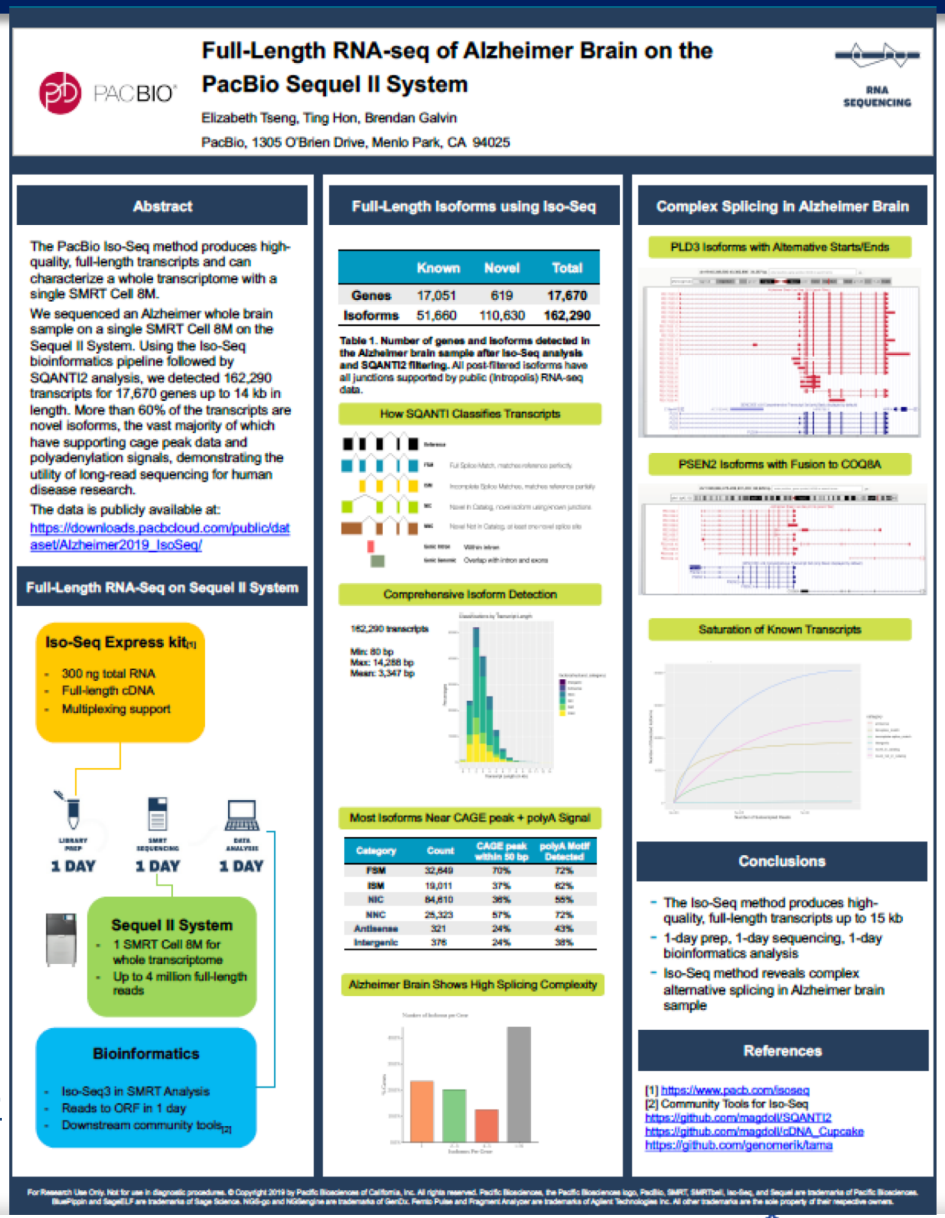
Get a complete view of transcript isoform diversity with PacBio long-read sequencing.

RNA Sequencing

Single Molecule, Real-Time (SMRT) Sequencing and Iso-Seq analysis allow you to generate full-length cDNA sequences — no assembly required — to characterize transcript isoforms within targeted genes or across an entire transcriptome so that you can easily and affordably:

- Discover new genes, transcripts and alternative splicing events
- Improve genome annotation to identify gene structure, regulatory elements, and coding regions
- Increase the accuracy of RNA-seq quantification with isoform-level resolution

https://www.pacb.com/wp-content/uploads/Tseng-ASHG-2019-Full-Length-RNA-seq-of-Alzheimer-brain-on-the-PacBio-Sequel-II-System.pdf



Full-Length RNA-seq of Alzheimer Brain on the PacBio Sequel II System

Elizabeth Tseng, Ting Hon, Brendan Galvin

PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

RNA SEQUENCING

Abstract

The PacBio Iso-Seq method produces high-quality, full-length transcripts and can characterize a whole transcriptome with a single SMRT Cell 8M.

We sequenced an Alzheimer whole brain sample on a single SMRT Cell 8M on the Sequel II System. Using the Iso-Seq bioinformatics pipeline followed by SQANTI2 analysis, we detected 162,290 transcripts for 17,670 genes up to 14 kb in length. More than 60% of the transcripts are novel isoforms, the vast majority of which have supporting cage peak data and polyadenylation signals, demonstrating the utility of long-read sequencing for human disease research.

The data is publicly available at: https://downloads.pacbcloud.com/public/dataset/Alzheimer2019_IsoSeq/

Full-Length RNA-Seq on Sequel II System

Iso-Seq Express kit[1]
- 300 ng total RNA
- Full-length cDNA
- Multiplexing support

1 DAY | 1 DAY | 1 DAY

Sequel II System
- 1 SMRT Cell 8M for whole transcriptome
- Up to 4 million full-length reads

Bioinformatics
- Iso-Seq3 in SMRT Analysis
- Reads to ORF in 1 day
- Downstream community tools[2]

Full-Length Isoforms using Iso-Seq

	Known	Novel	Total
Genes	17,051	619	17,670
Isoforms	51,660	110,630	162,290

Table 1. Number of genes and isoforms detected in the Alzheimer brain sample after Iso-Seq analysis and SQANTI2 filtering. All post-filtered isoforms have all junctions supported by public (Intropolis) RNA-seq data.

How SQANTI Classifies Transcripts

Comprehensive Isoform Detection

162,290 transcripts
Min: 80 bp
Max: 14,288 bp
Mean: 3,347 bp

Most Isoforms Near CAGE peak + polyA Signal

Category	Count	CAGE peak within 50 bp	polyA Motif Detected
FSM	32,649	70%	72%
ISM	19,011	37%	62%
NIC	84,610	36%	55%
NNC	25,323	57%	72%
Antisense	321	24%	43%
Intergenic	376	24%	38%

Alzheimer Brain Shows High Splicing Complexity

Complex Splicing in Alzheimer Brain

PLD3 Isoforms with Alternative Starts/Ends

PSEN2 Isoforms with Fusion to COQ8A

Saturation of Known Transcripts

Conclusions
- The Iso-Seq method produces high-quality, full-length transcripts up to 15 kb
- 1-day prep, 1-day sequencing, 1-day bioinformatics analysis
- Iso-Seq method reveals complex alternative splicing in Alzheimer brain sample

References

[1] https://www.pacb.com/isoseq
[2] Community Tools for Iso-Seq
https://github.com/magdoll/SQANTI2
https://github.com/magdoll/cDNA_Cupcake
https://github.com/genomerik/tama

RNA-seq data analysis

Two scenarios

- Reference genome sequence available
- No Reference genome sequence available

With reference genome or transcriptome

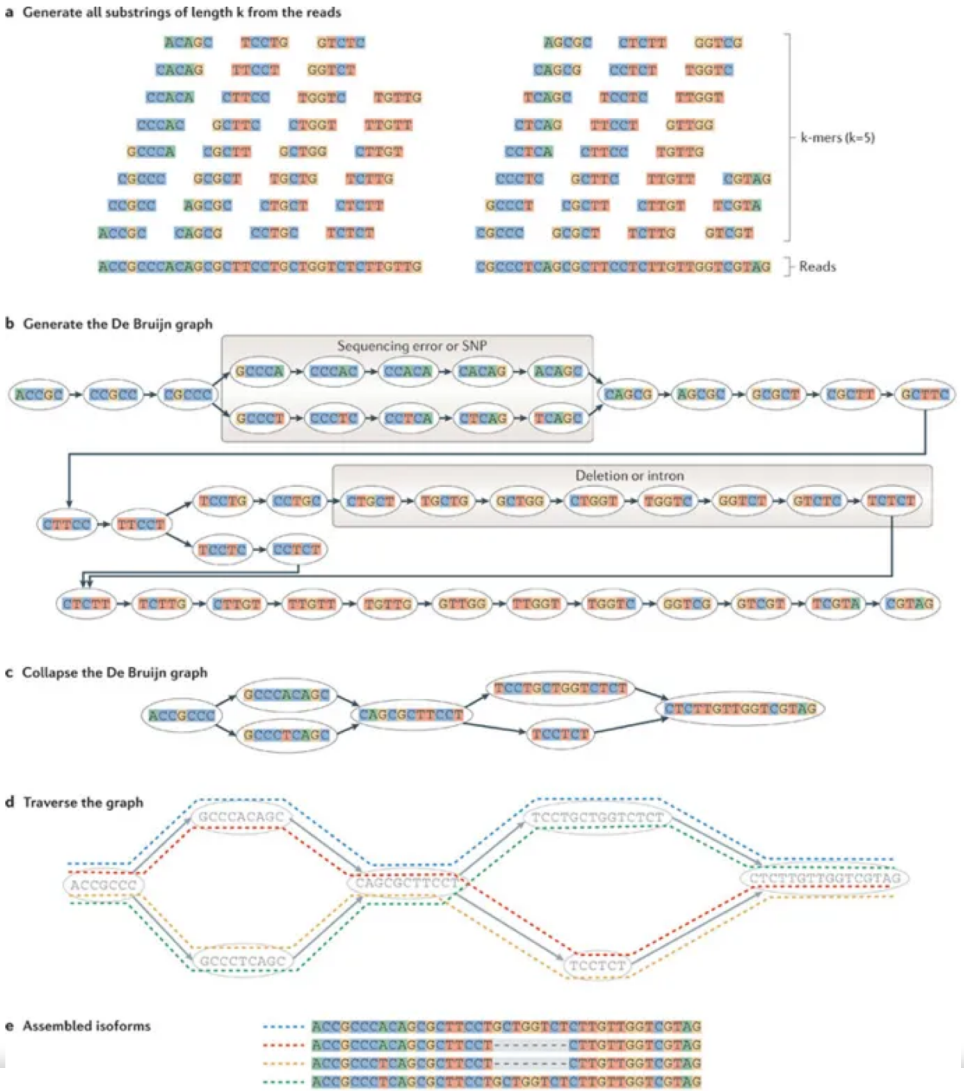
- Step 1: Read alignment:
  - align reads directly to a reference genome (spliced) or transcriptome (unspliced)
- Step 2: Transcriptome reconstruction
  - identify expressed genes and isoforms
- Step 3: Expression qualification and differentiation expression analysis
  - Expression quantification
    - Gene quantification
    - Isoform quantification
  - Differential expression
- Step 4: Functional Interpretation:

Without reference genome or transcriptome

- Step 1: *De novo* transcriptome assembly
- Step 2: Map reads back to assembled transcripts
- Step 3: Expression qualification and differentiation expression analysis
  - Expression quantification
    - Gene quantification
    - Isoform quantification
  - Differential expression
- Step 4: Functional Interpretation:

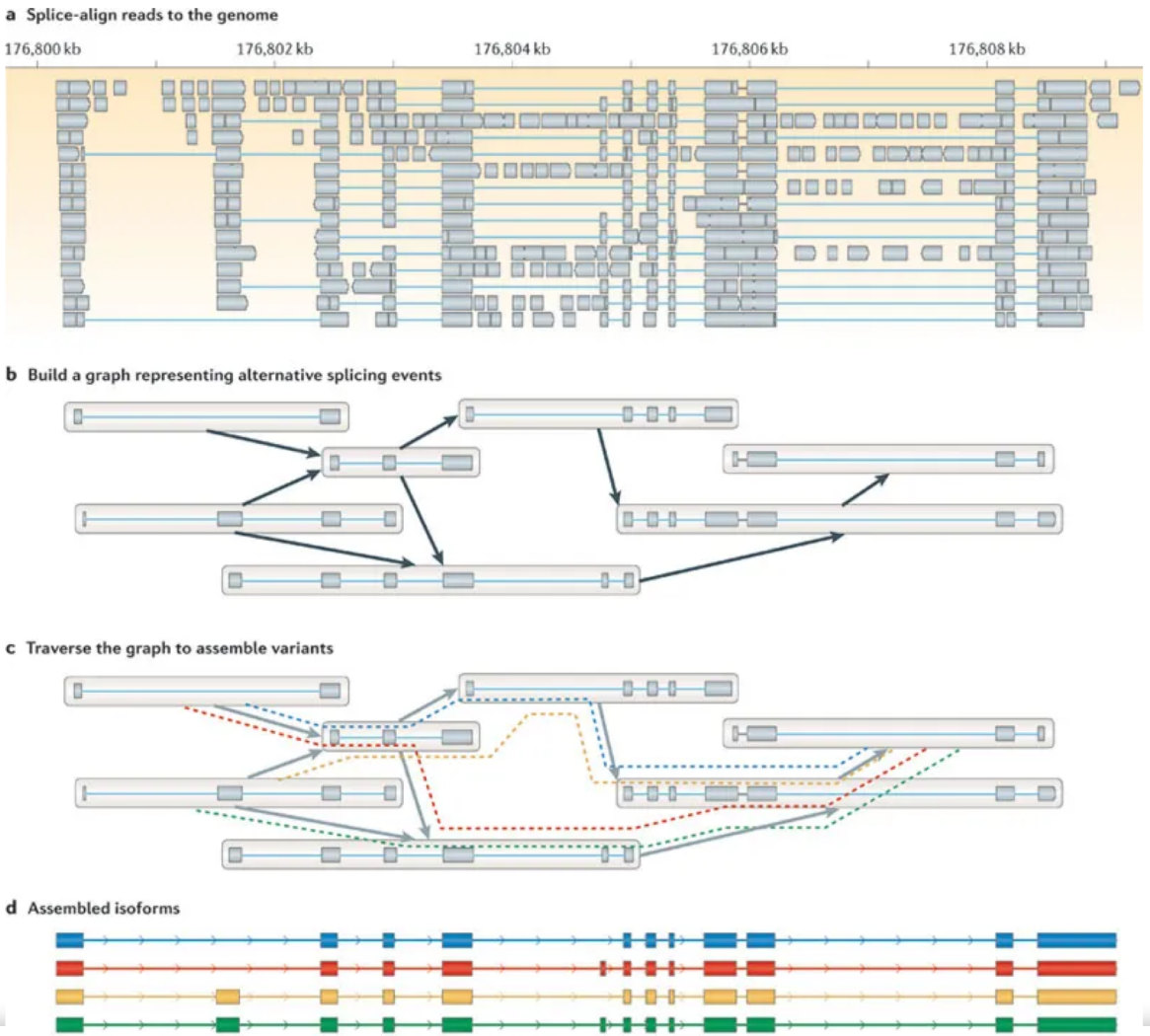
Overview of the *de novo* transcriptome assembly strategy.

- **Next-generation transcriptome assembly (https://www.nature.com/articles/nrg3068)**

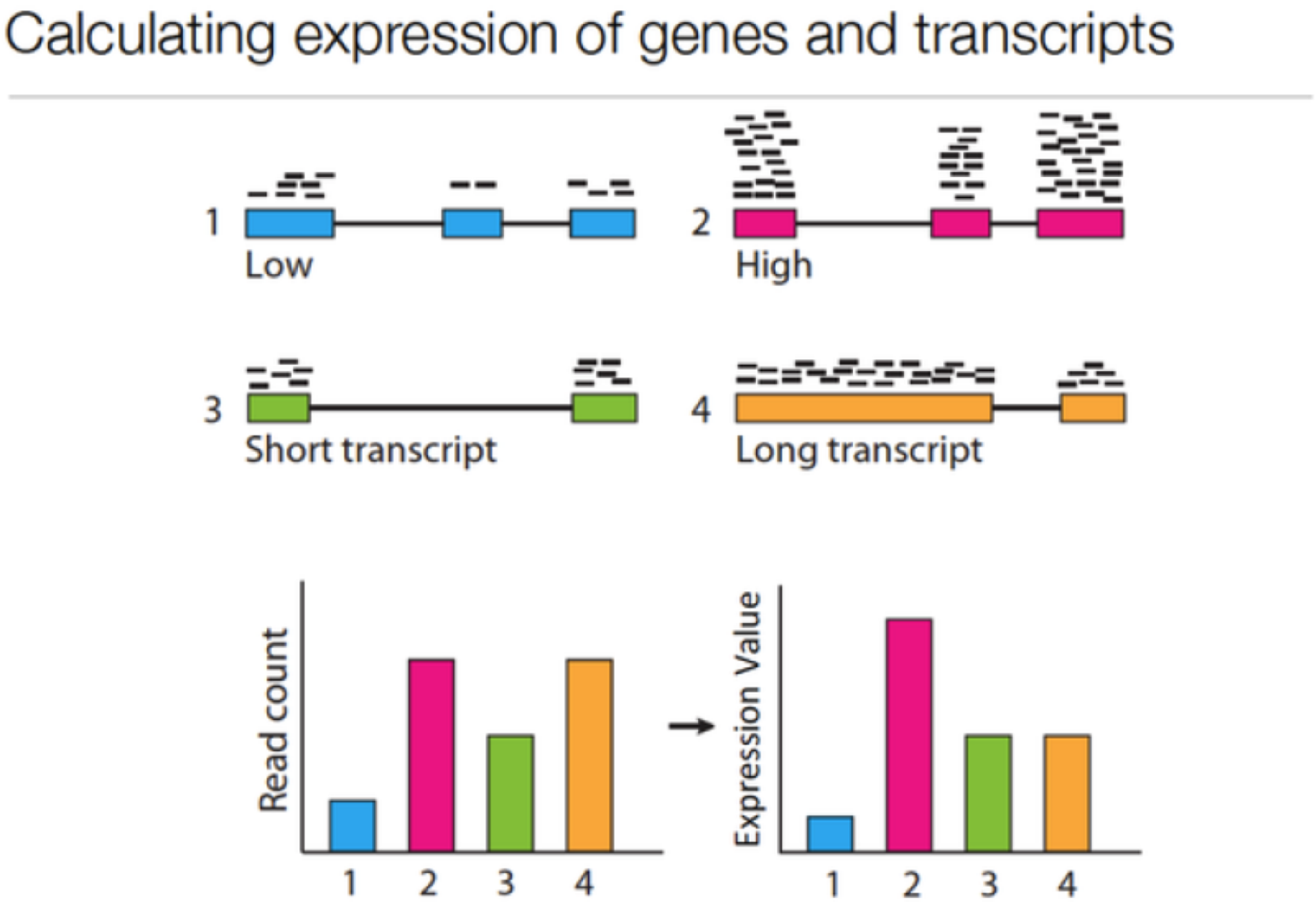


Overview of the reference-based transcriptome assembly

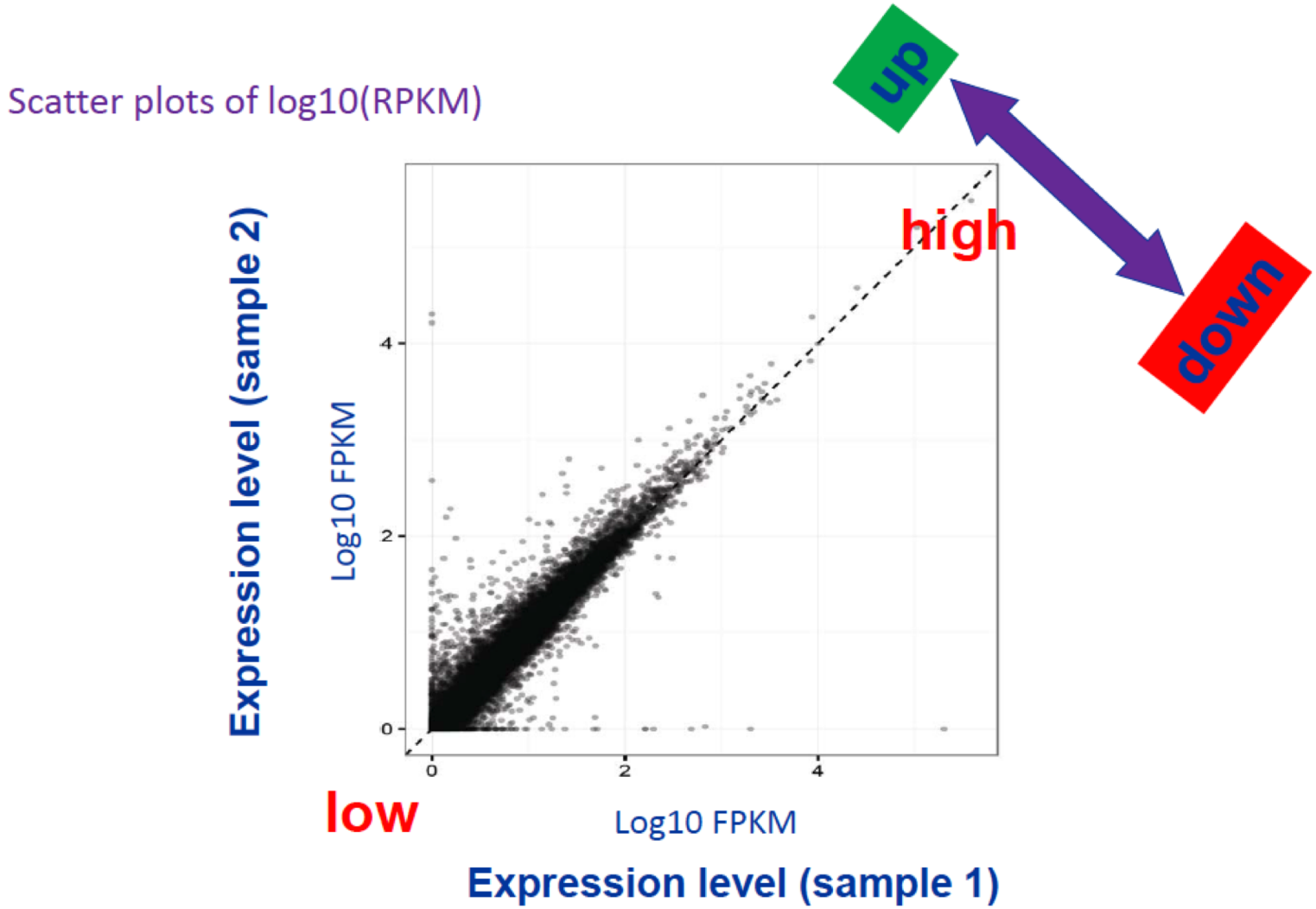
- **Next-generation transcriptome assembly (https://www.nature.com/articles/nrg3068)**



Normalize expression value



Differential expression analysis



Functional annotation Gene Ontology

Figure 3 : *SalvCKO* mice activate reparative molecular response to heart failure.

From: Hippo pathway deficiency reverses systolic heart failure after infarction