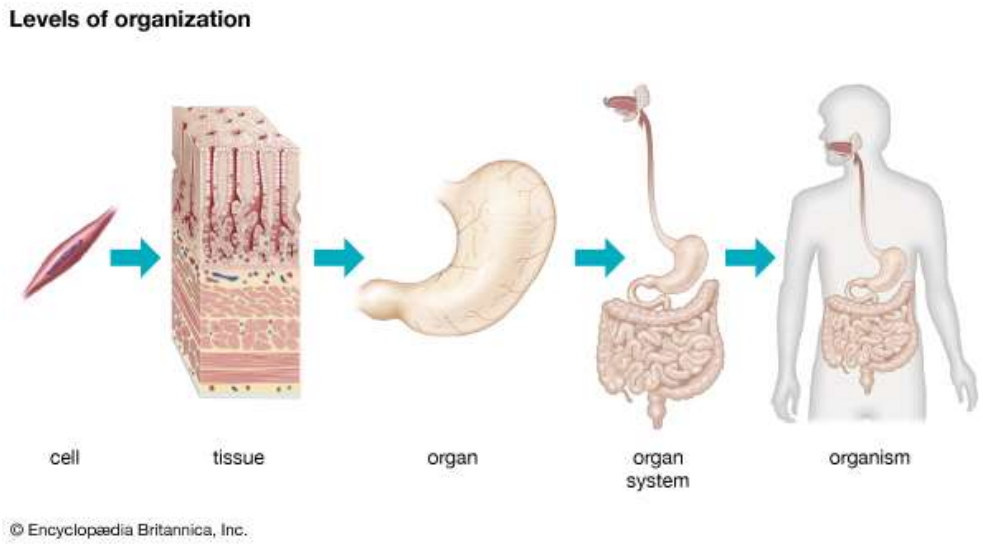
Introduction to 'Omics & RNA-Sequencing

Lauren Vanderlinden

BIOS 6660

Spring 2019

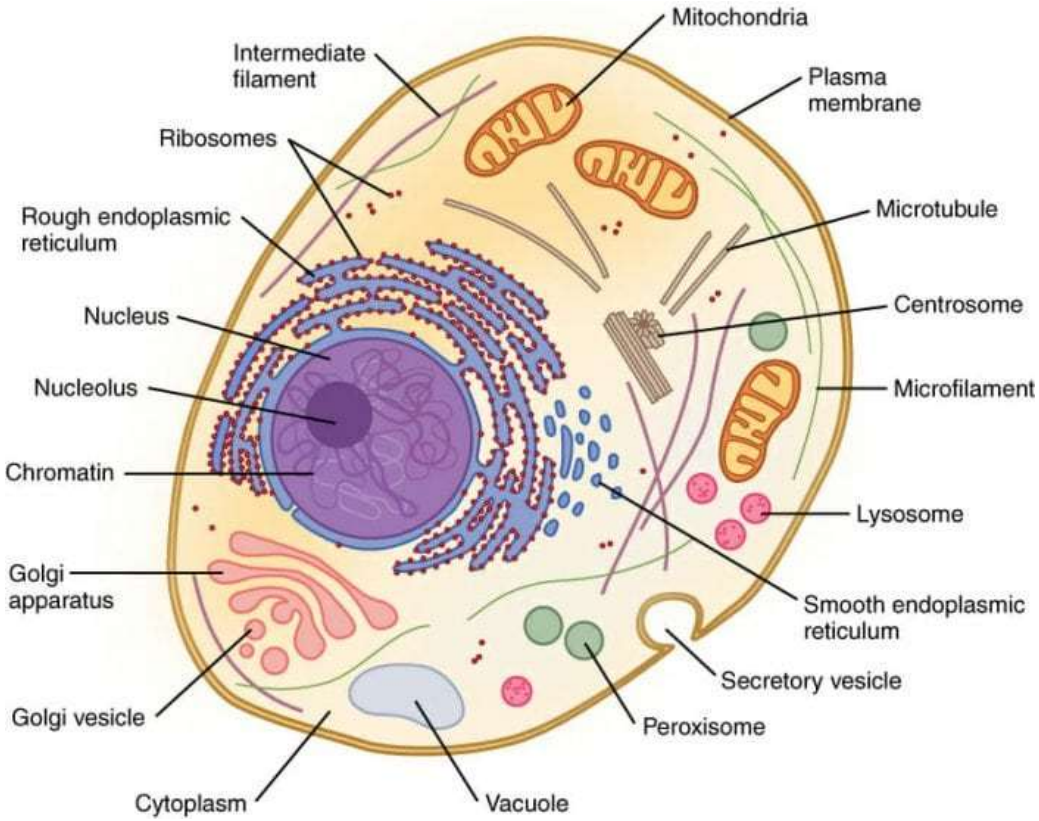
Levels of Biological Organization



source: Encyclopedia Britannica

Cellular Level

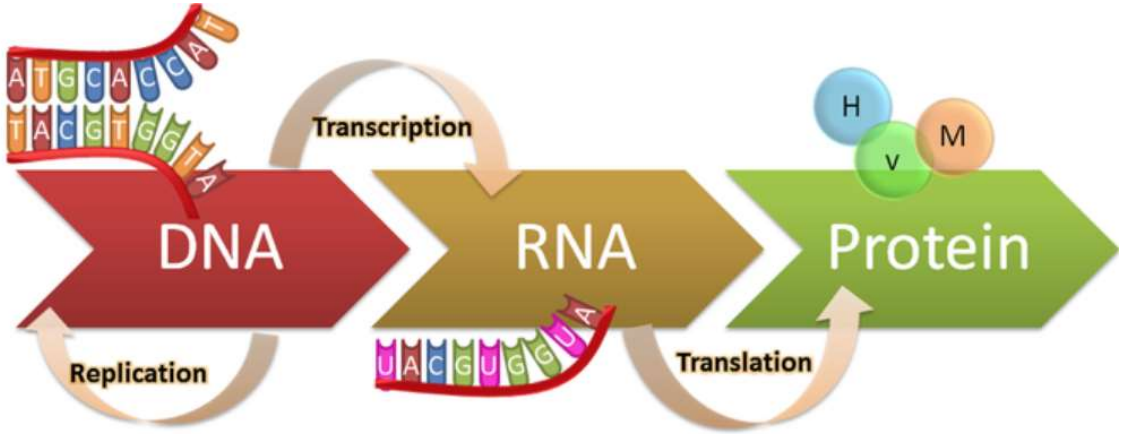
- Working unit of every living organism
- Made up of organelles
- Each has a specific function



source: ScienceTrends.com

Central Dogma of Microbiology

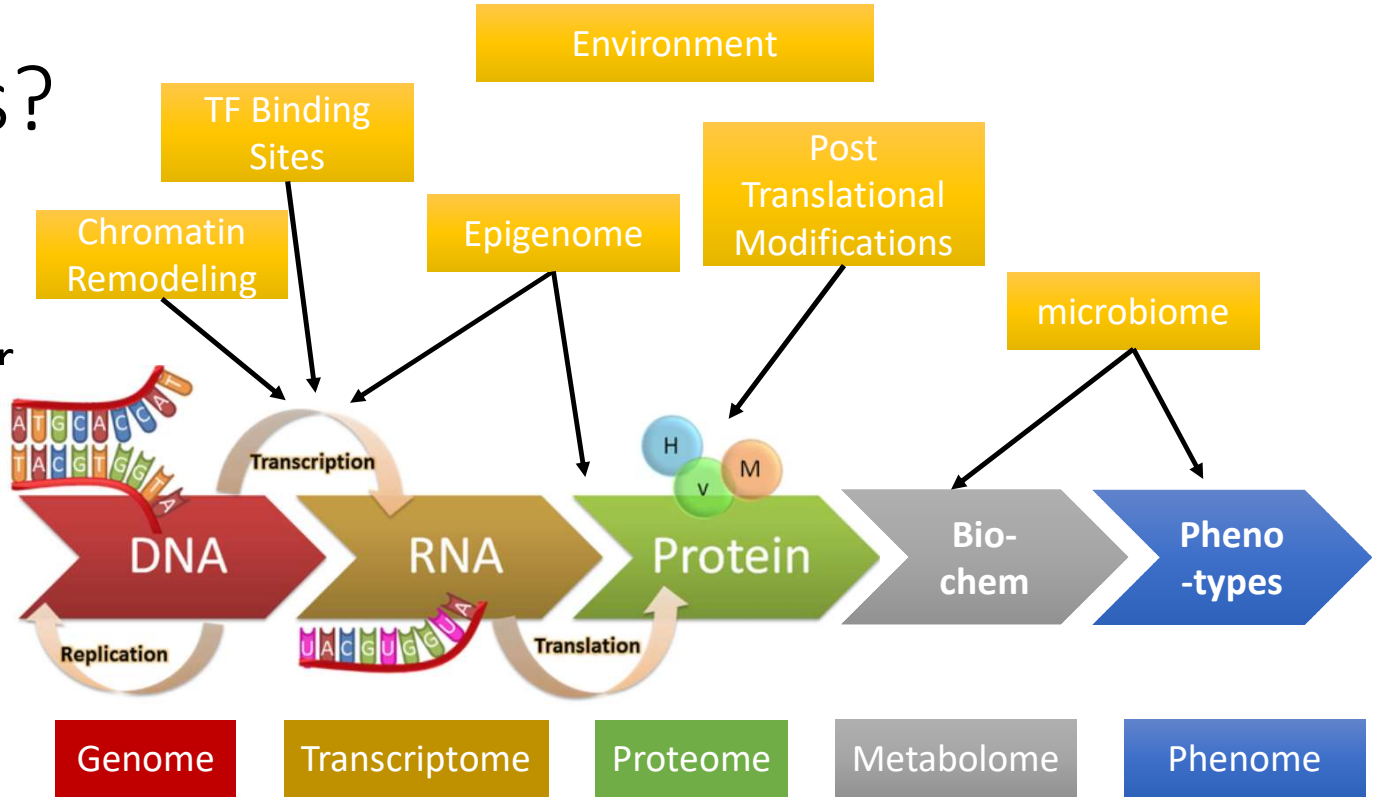
- Within each cell, this process is taking place
- Seems like a pretty simple linear progression
- 3 classes of information



source: rbpaonline.com

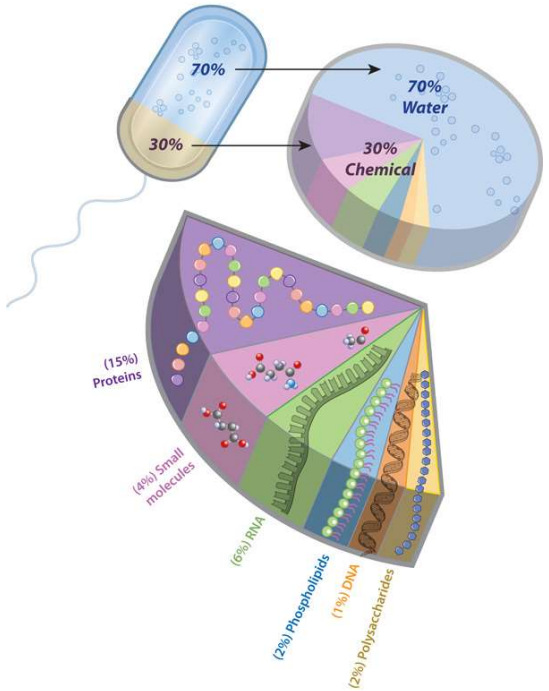
What is 'Omics?

- Goal is to understand the behavior of cells, tissues, organs, and the whole organism at the molecular level
- High throughput data relating to a biology "ome"
  - Genome
  - Transcriptome
  - Proteome
  - Metabolome
  - Epigenome
- Not as straight forward



Cellular Composition

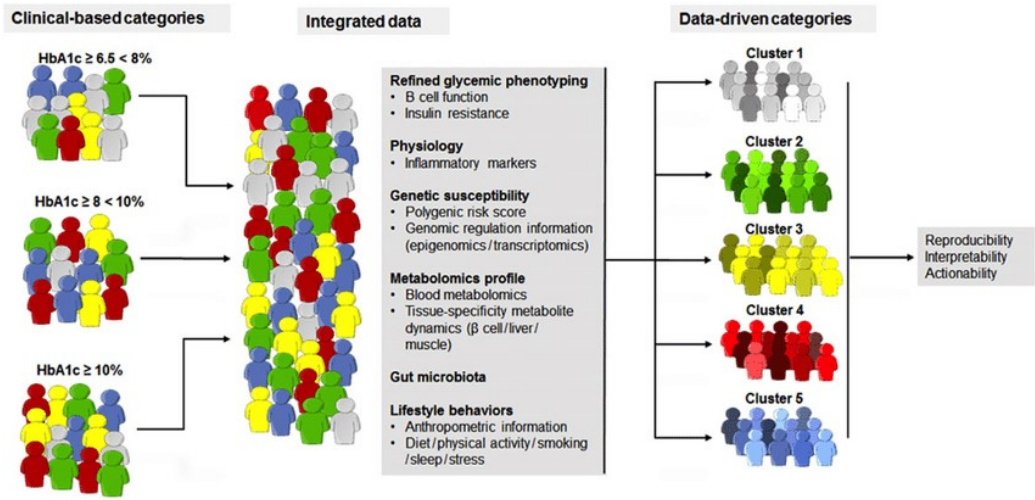
- Proteins make up the majority of chemical structure
- Various assays for each type of chemical structure



source: natureEducation

Goal: Personalized & Precision Medicine

- Tailor medical treatment
  - Gleevec to treat chronic myeloid leukemia (CML) in presence of a fusion gene (BCR-ABL)
- Disease prevention
  - Those with BRCA mutations can elect for mastectomy

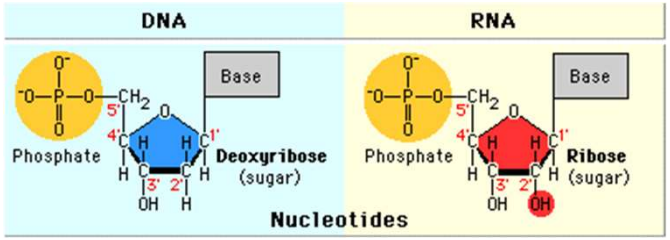


Source: Merino et al, 2018

Nucleotides

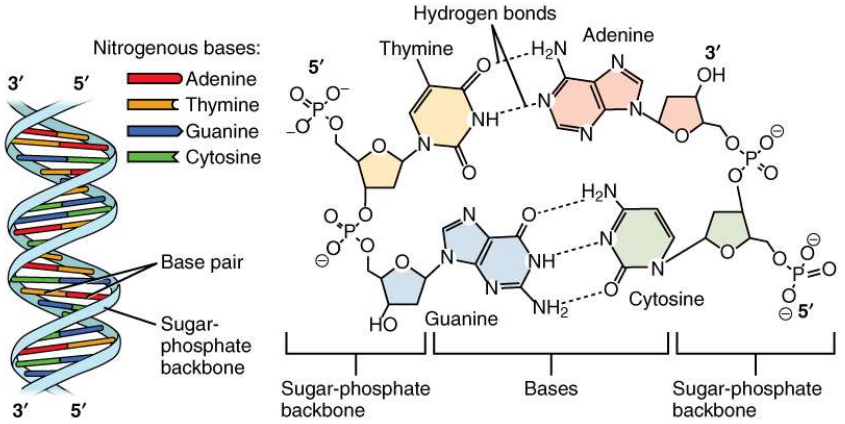
- 3 distinctive subunit

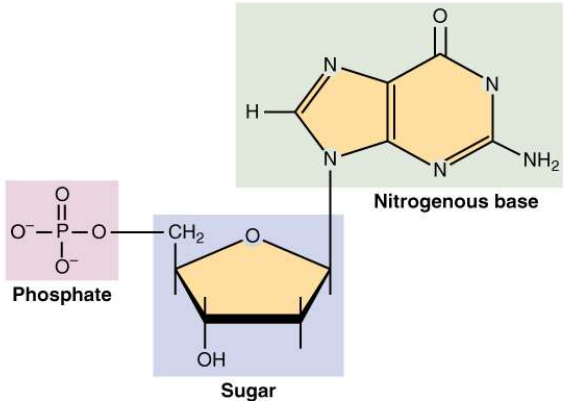
1. 5-carbon sugar
   - DNA: deoxyribose
   - RNA: ribose



Source: phschool

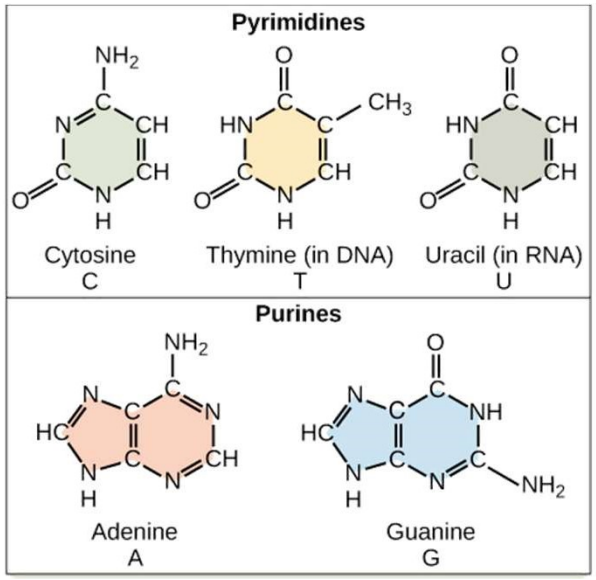
2. Nitrogenous base
3. Phosphate group



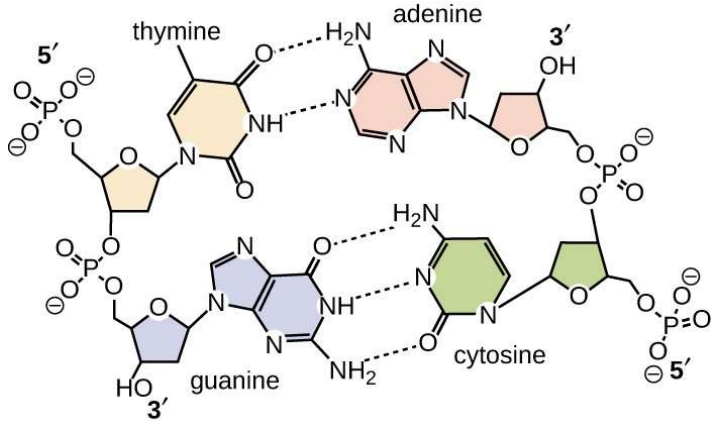


Source: wikipedia

Bases and Base-Pairing



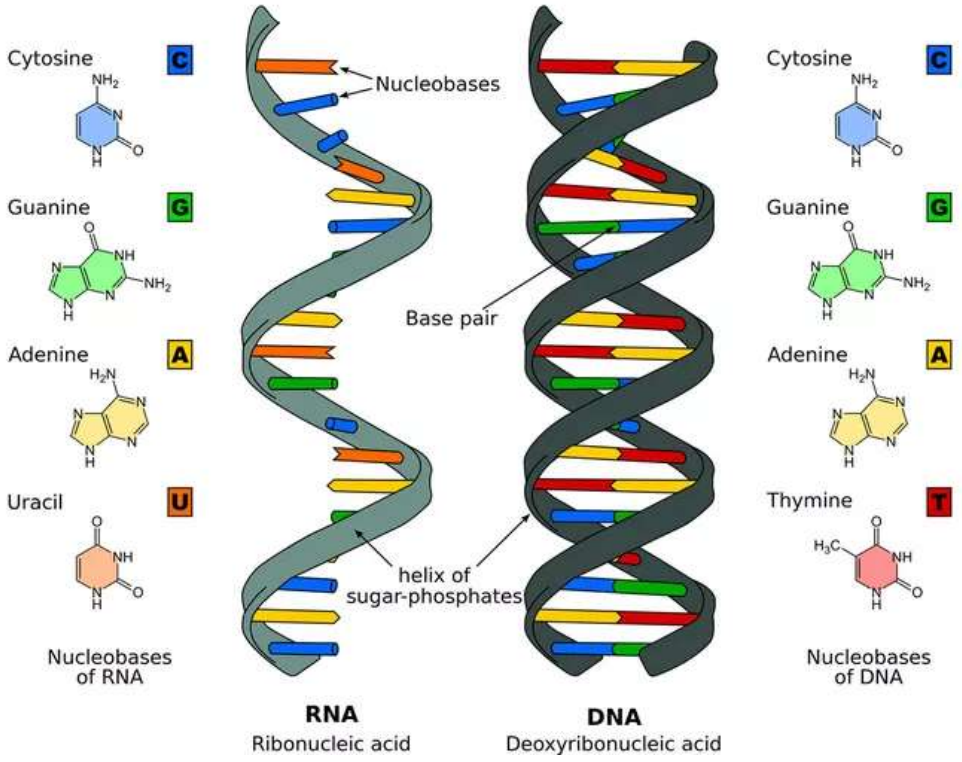
Source: Khan academy



- Hydrogen bonding
- Purine with pyrimidine
  - A-T
  - G-C

DNA vs RNA

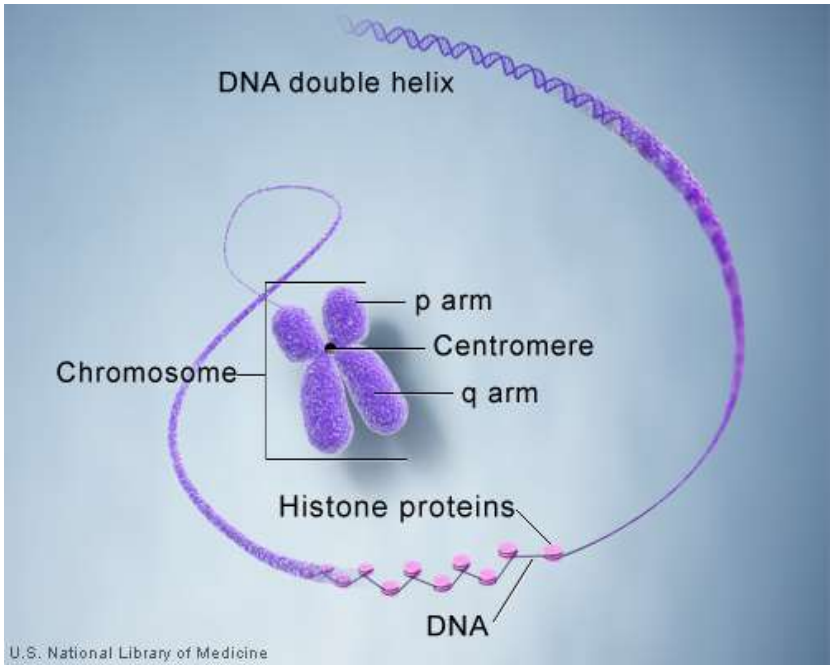
- Commonality: both consist of nucleotides
- DeoxyriboNucleic Acid (DNA)
  - Double stranded using base pairing
- RiboNucleic Acid (RNA)
  - Single-stranded
  - Uracil (U) instead of Thymine (T)



Source: wikipedia

Chromosomes

- The packing of DNA
- DNA tightly wound around proteins called histones
- Store lots of data in small area
- Multiple chromosomes/organism
  - E. coli only has 1



Source: US National Library of Medicine

Genomes Across Species

Species	# Chromosomes	# Nucleotides (Mb)	# Genes
E. coli	1	4,600,000	4,377
Fruit fly	4	180,000,000	17,000
Yeast	16	13,000,000	5,770
Rat	21	2,750,000,000	20,000
Mouse	20	2,800,000,000	23,000
Human	23	3,300,000,000	21,000

RNA (Transcriptomics)

- Messenger RNA (mRNA)
- Transfer RNA (tRNA)
- Ribsomal RNA (rRNA)
- Long Non-Coding RNA (lncRNA)
  - Circular RNA (circRNA)
- Small RNA (<200 nt)
  - microRNA (miRNA)
  - Piwi-interacting RNA (piRNA)
  - small interfering RNA (siRNA)
  - small nucleolar RNA (snoRNAs)
  - tRNA-derived small RNA (tsRNA)
  - small rDNA-derived RNA (srRNA)
  - small nuclear RNA (U-RNA)

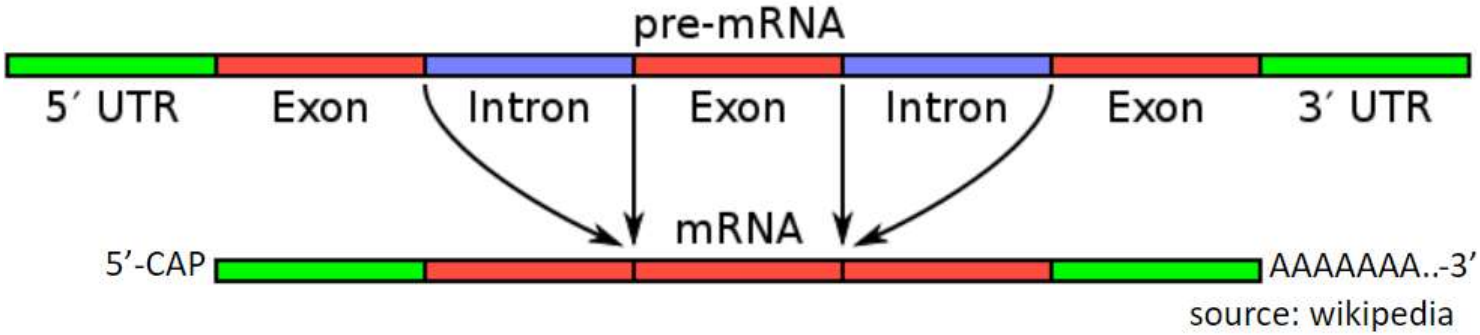
Most abundant & function known and defined

More regulatory functions & not clearly defined or known

Messenger RNA (mRNA)

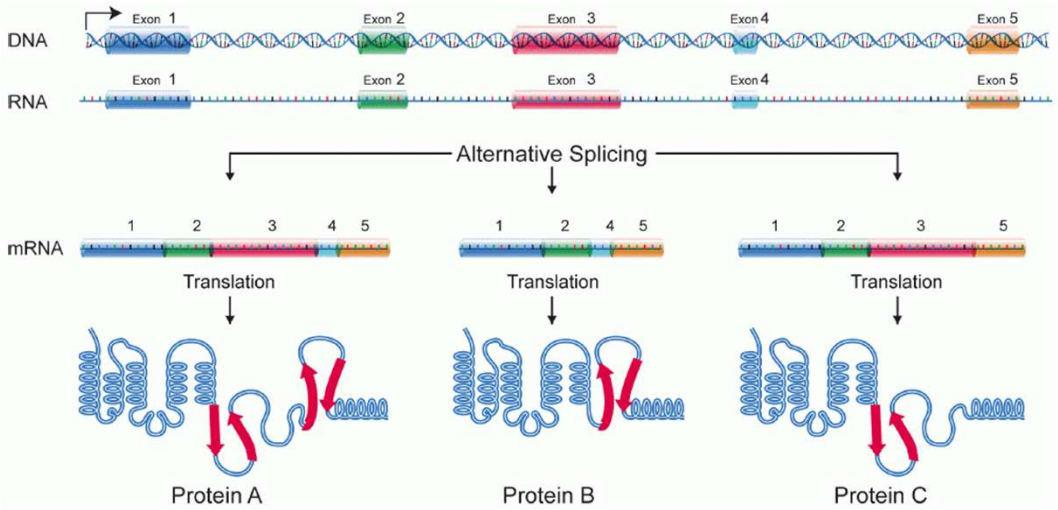
- Protein Coding
- Most commonly interrogated type of RNA

- Structure:
  - 5’ cap
  - 5’ untranslated region (UTR)
  - Coding sequence (CDS)
  - 3’ untranslated region (UTR)
  - Poly-A tail



Alternative Splicing and Isoforms

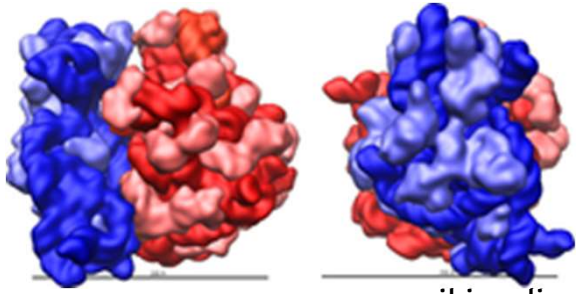
- Contributes to protein complexity
  - estimated 35-60% of human genes alternatively spliced
- ~15% disease-causing mutations cause errors in alternative splicing regulation
- Can alters protein structure & function
- Can changes ratios of protein isoforms



source: wikipedia

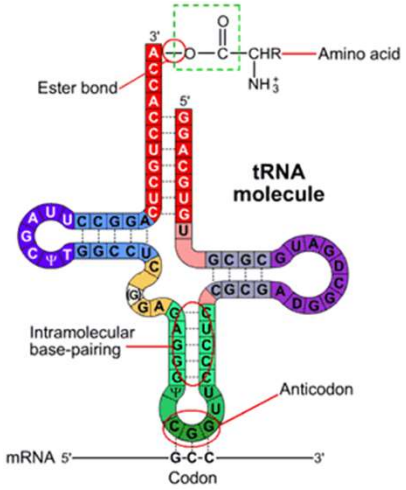
Ribosomal RNA (rRNA) & Transfer RNA (tRNA)

- rRNA is the most abundant type of RNA in cell (90-95% in human)
- 2 predominant rRNAs in human:
  - 18S (small component, blue)
  - 28S (large component, red)
- Removing rRNA can be difficult



source: wikipedia

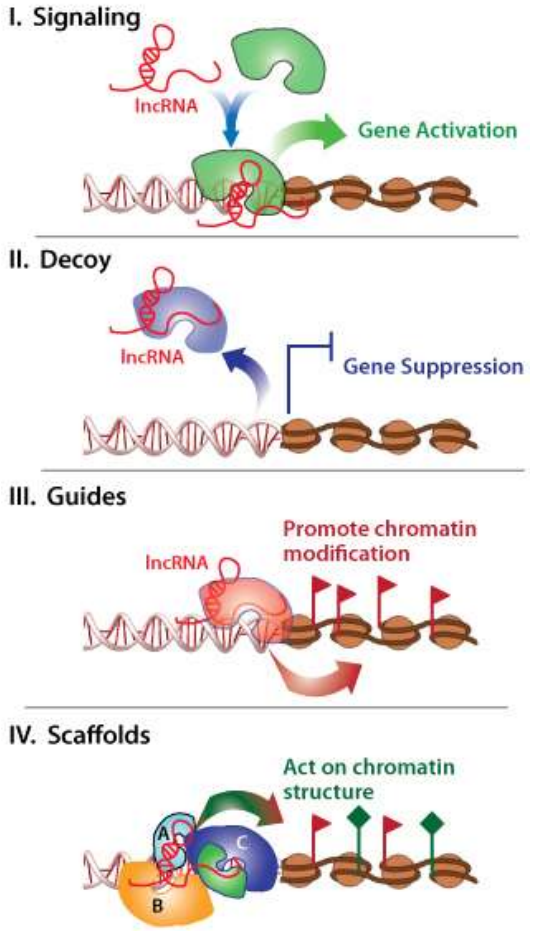
- tRNA is the physical link between mRNA and protein
- It transfers an amino acid to the ribosome



source: wiley

Long non-coding RNA (lncRNA)

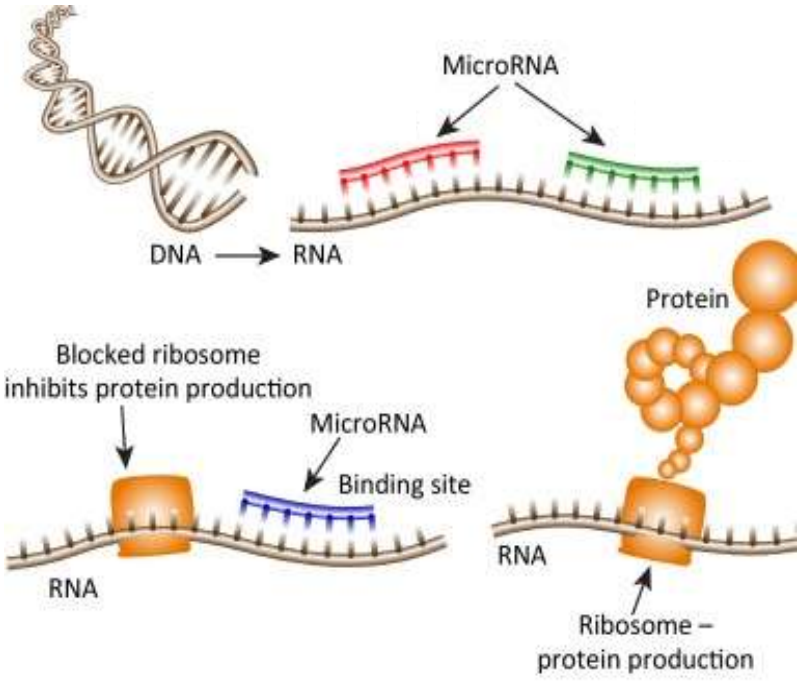
- Regulates transcription & translation process
- Many different mechanism
  - Inhibition
  - Promotion
  - Modifications
- Used to be considered “junk” DNA
- For example, circRNA is a class of lncRNA
  - Back-splicing to create a loop & can recruit miRNAs (another way to regulate mRNA translation)



source: Figure 1 Wang et. al, 2011

Micro RNA (miRNA)

- Regulates translation
- Thought to inhibit translation:
  - Inhibits binding capabilities of translational elements
  - Promotes degradation of targeted mRNA
- However, functionality is not as straight forward compared to other known RNA types
  - Wasn't recognized as it's own class of RNA till early 2000s.

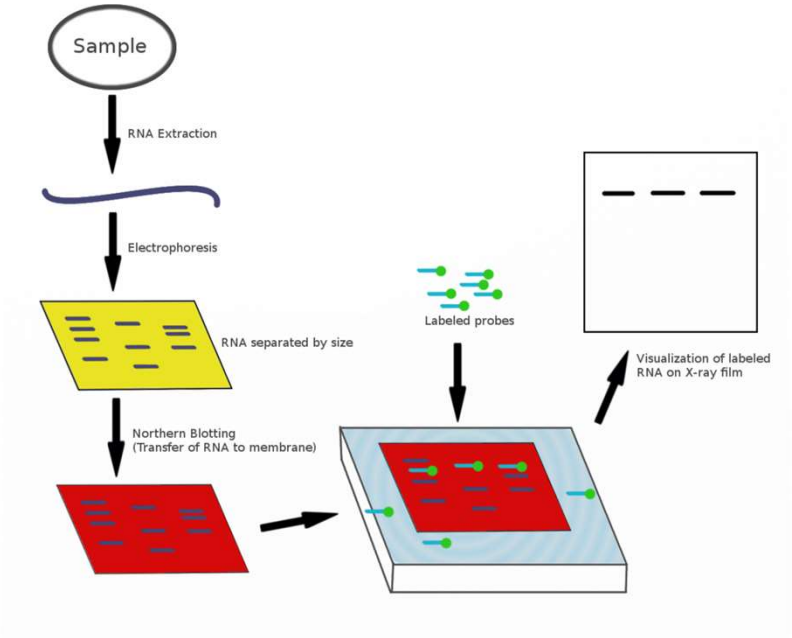


source: Figure 1 Meydan et. al, 2016

Major RNA expression technologies

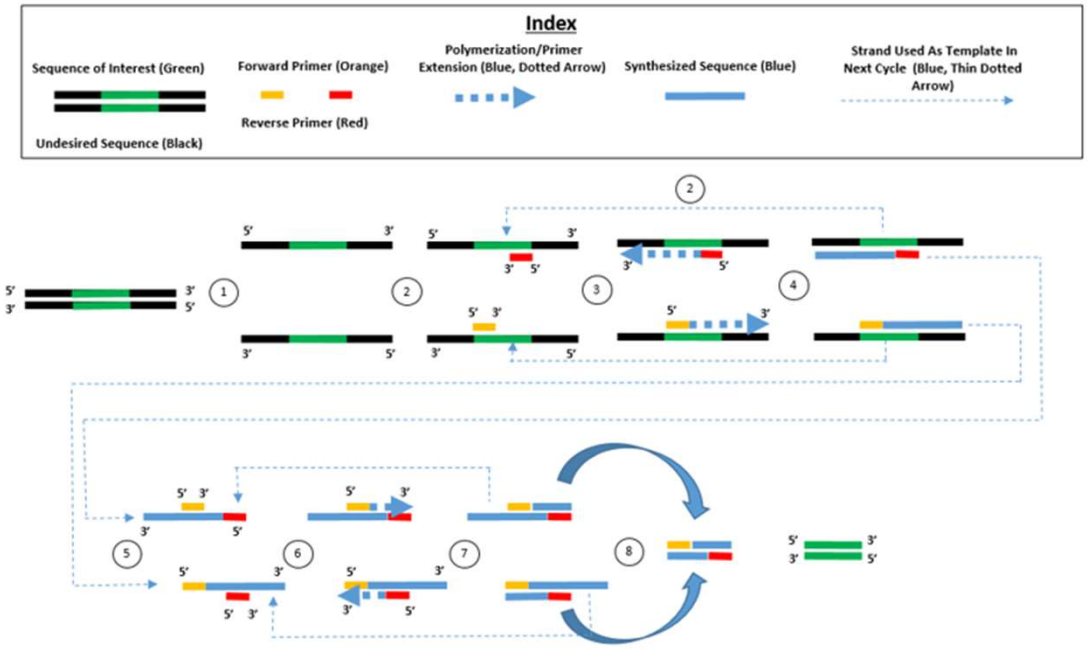
Technology	High Throughput	Quantitativeness	Single Cell Information	Ease of Use
Northern blot (1977)	----	++	---	+
RT-qPCR (1990)	+	+++++	---	+++
Microarrays (1995)	++	+++	---	++
RNA-Seq (2008)	+++	+++	---	---
scRNA-Seq (2014)	+++	+	++++	----

Northern Blot



source: wikipedia

RT-qPCR



source: wikipedia

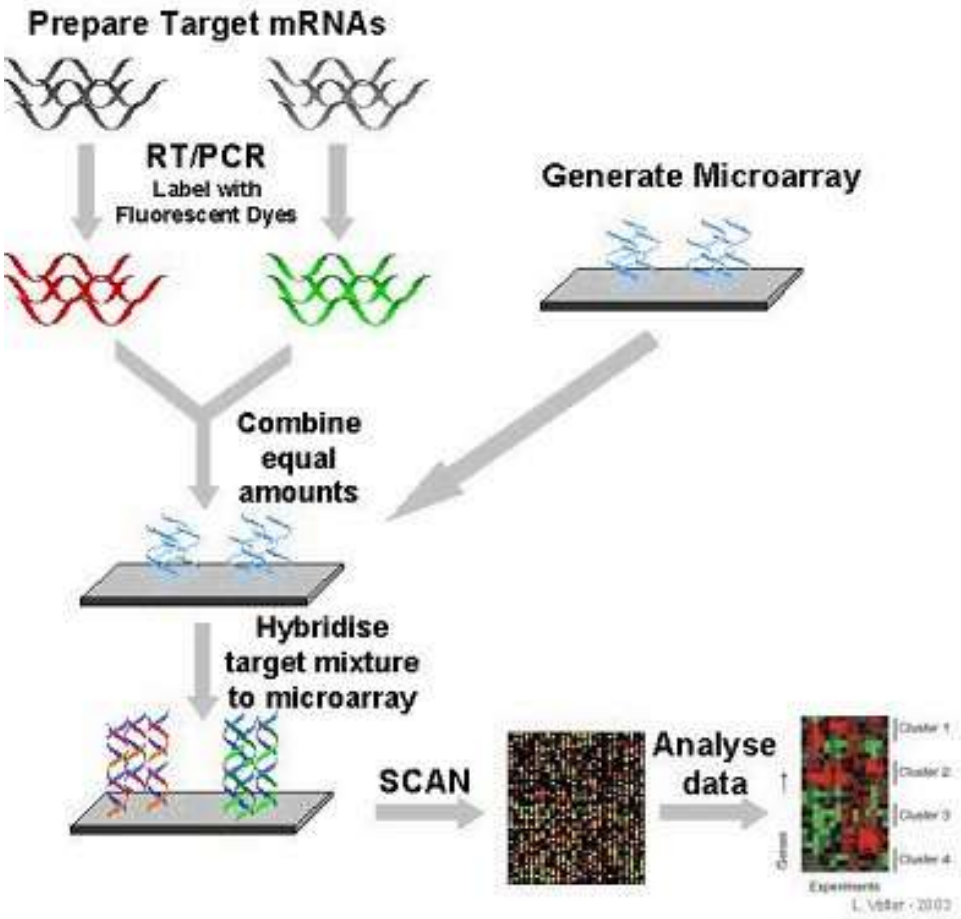
Limited to very small list of RNA

- membrane blot or primers

Microarray

First genome-wide application

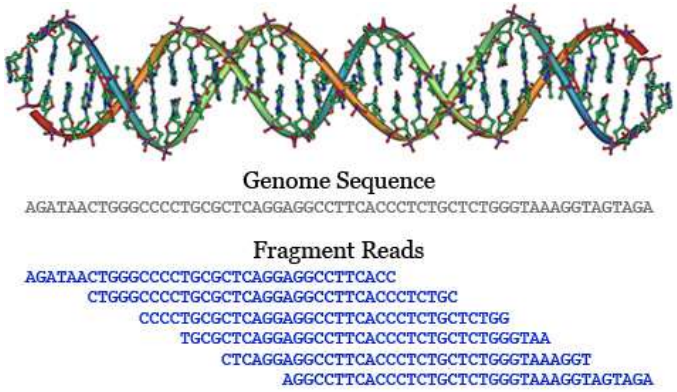
- Individual probes are 20-25 bp long
- Probeset is group of probes that probe same region which are summarized together
- Target specific areas of genes
  - Exons
  - 3’ end of gene
- Drawback: need to know sequence of RNA when designing the microarray



source: affymetrix

Sequencing Technology

- DNA or RNA
- Want the raw sequences
- This is a game changer as you don't need to know sequence



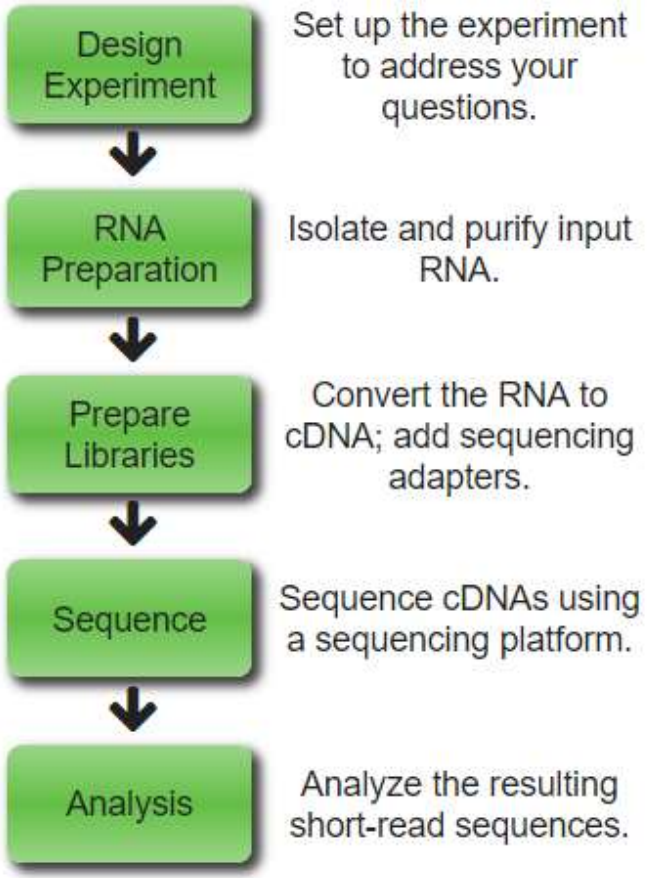
source: clarkesworld

RNA Sequencing Workflow

Step 1: Design Experiment

- Study Design:
  - Differential expression between case vs control
  - Longitudinal study of gene expression over time

- Do you just want mRNA?
- What about small RNA?
- Are you interested in isoform information?
- Are you interested in cell type proportions?
- How important is quantitation to you to get the most bang for your buck?



source: RNAseqlopedia

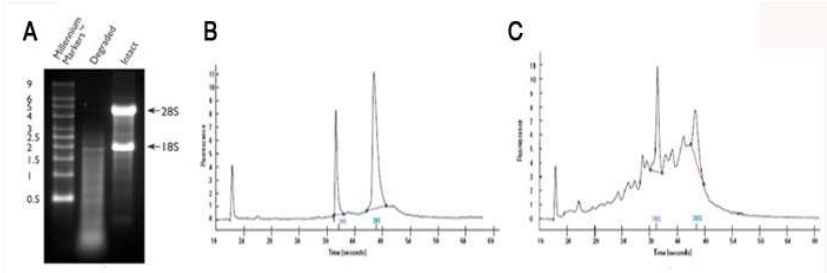
RNA-Seq: RNA Prep

- Stabilize RNA:
  - RNA degrades immediately once tissue is harvested
  - Flash freezing IS NOT ENOUGH
  - RNAstable® (Biomatrica)
  - RNAlater® (Qiagen)
- Isolate and purify RNA
  - Part of this step is a size filter
  - 200 bp filter
    - Many times <200 bp section is thrown out
    - NOT ideal for small RNA

- Assessing quality and quantity of RNA

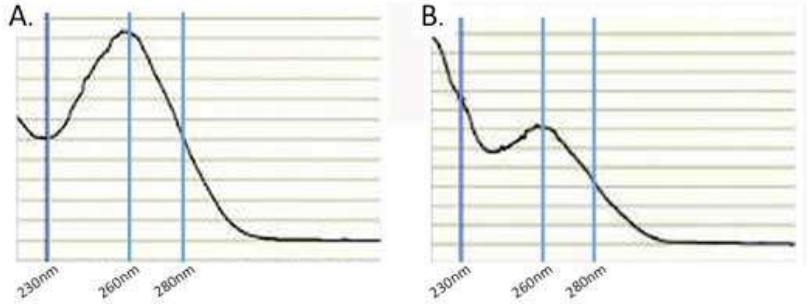
  Agarose gels and/or Bioanalyzer

  Intact Degraded



Quantitate RNA is by measuring the absorbance at 260 nm

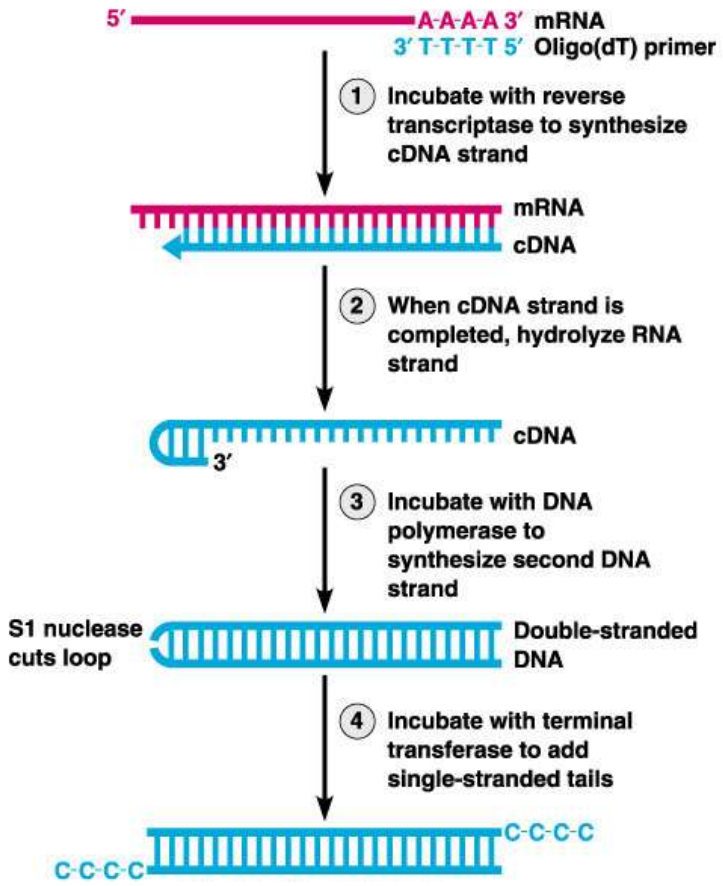
Clean Contaminated



source: RNAseqlopedia

Library Prep Part 1

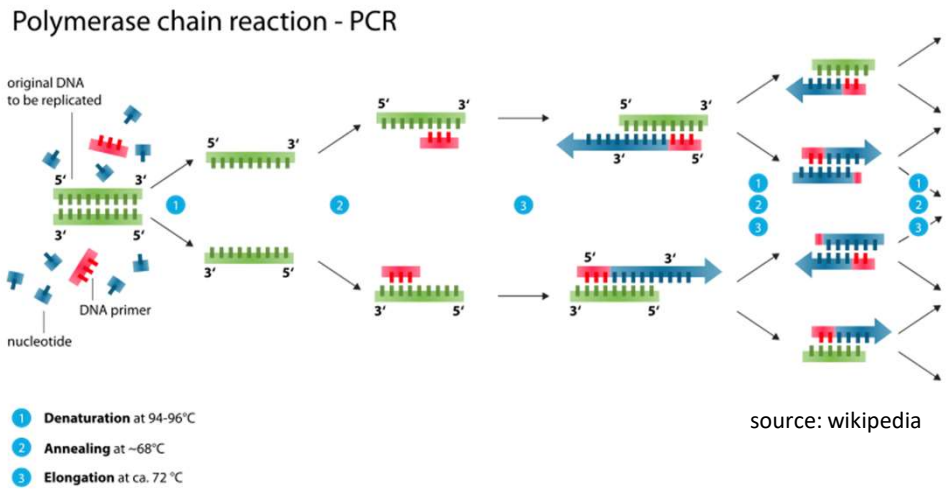
- Select/Clean RNA
  - polyA selection
  - rRNA depletion
- Fragment RNA
  - Optional to fragment here, most protocols do
- Convert RNA to DNA
  - Reverse transcriptase from retroviruses can synthesize DNA from RNA
- Or Fragment cDNA
- Repair ends add A overhang
  - In figure A would be where the C's are



Source: Addison Wesley Longman, Inc

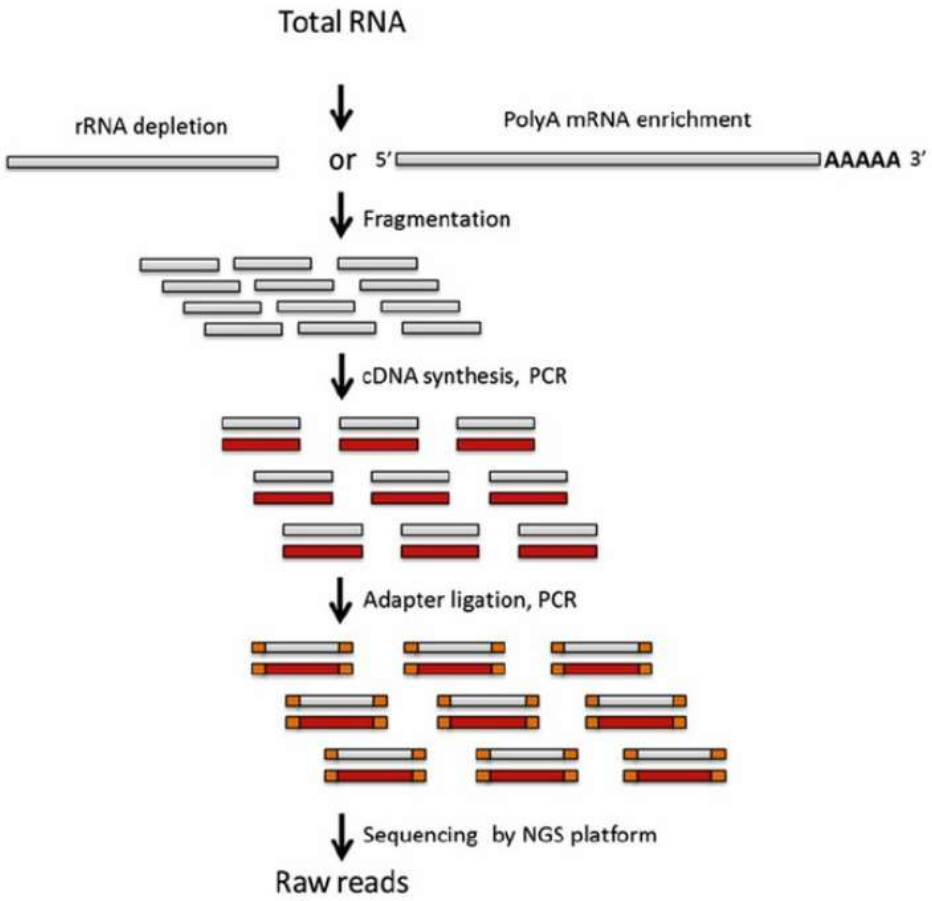
Library Prep Part 2

- Perform PCR
- Attach adapters
  - Illumina: clusters are attached to the flow cell
  - 454, IonTorrent and SOLiD: platforms are clustered on beads using emulsion PCR
- 2 types of sequence elements required:
  - Terminal platform-dependent sequences required for clonal amplification
  - Sequences for priming sequencing reaction
- Additional optional element are multiplexing (barcode) or 2nd priming site (paired-end sequencing)
- More PCR



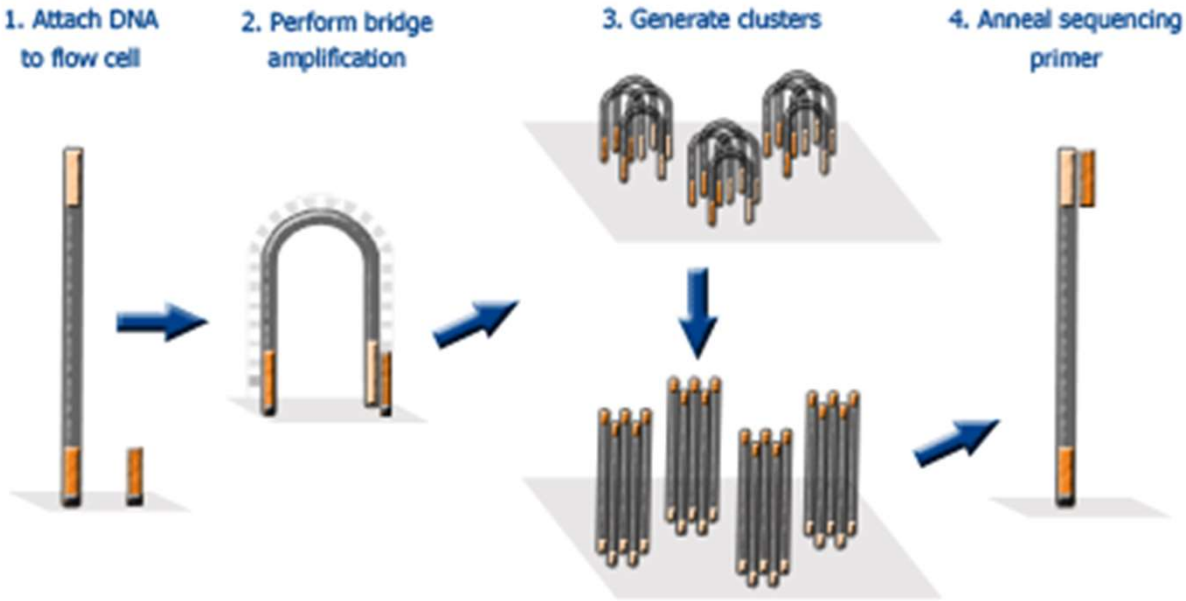
source: wikipedia

Library Prep Overview



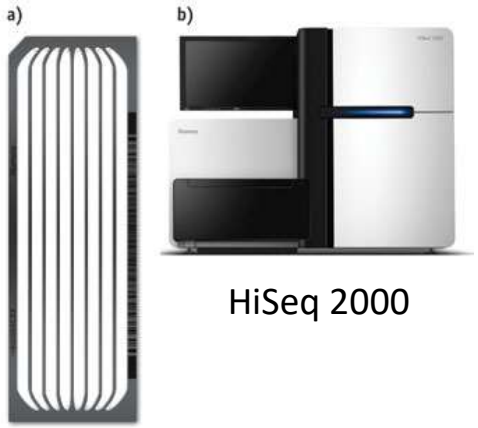
source: RNAseqlopedia

Illumina Sequencing Part 1



Source: www.eurofinsgenomics.co.in

Flow cell is coated with single stranded oligos that correspond to the adapters

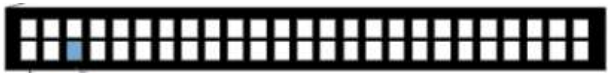


HiSeq 2000

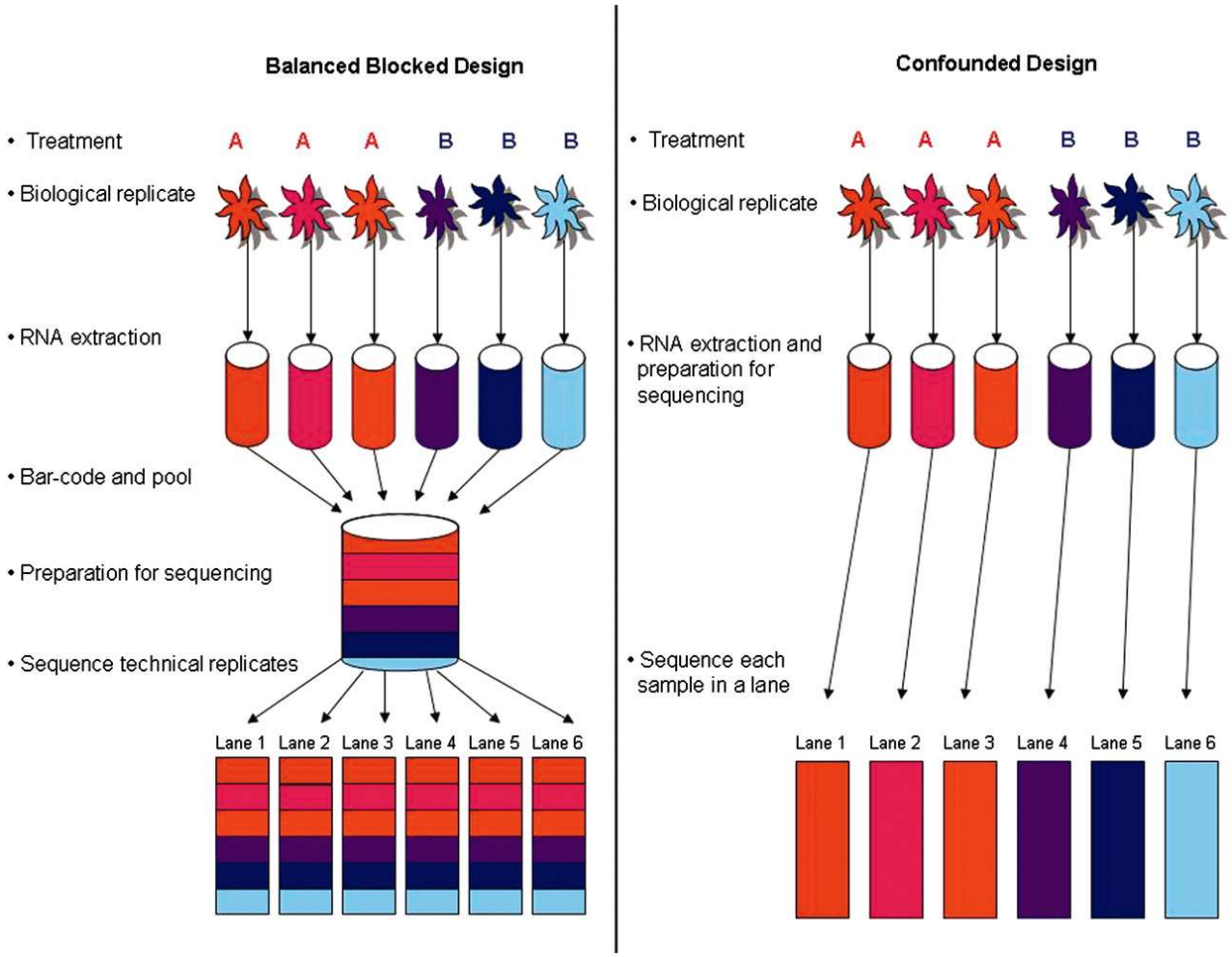
Source: bioopticsworld

Multiplexing

- Make sure you have a balanced design
- Each lane looks like:

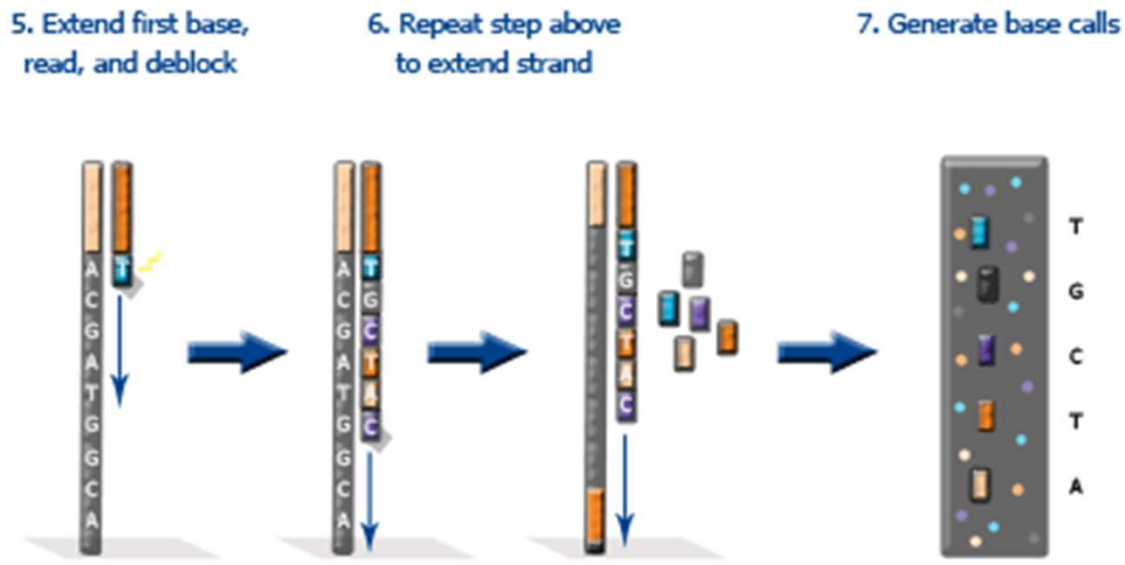


- Has 2 columns
- Each column up to 50 tiles
- Talk to lab and set up design



Source: Auer & Doerge 2010

Illumina Sequencing Part 2

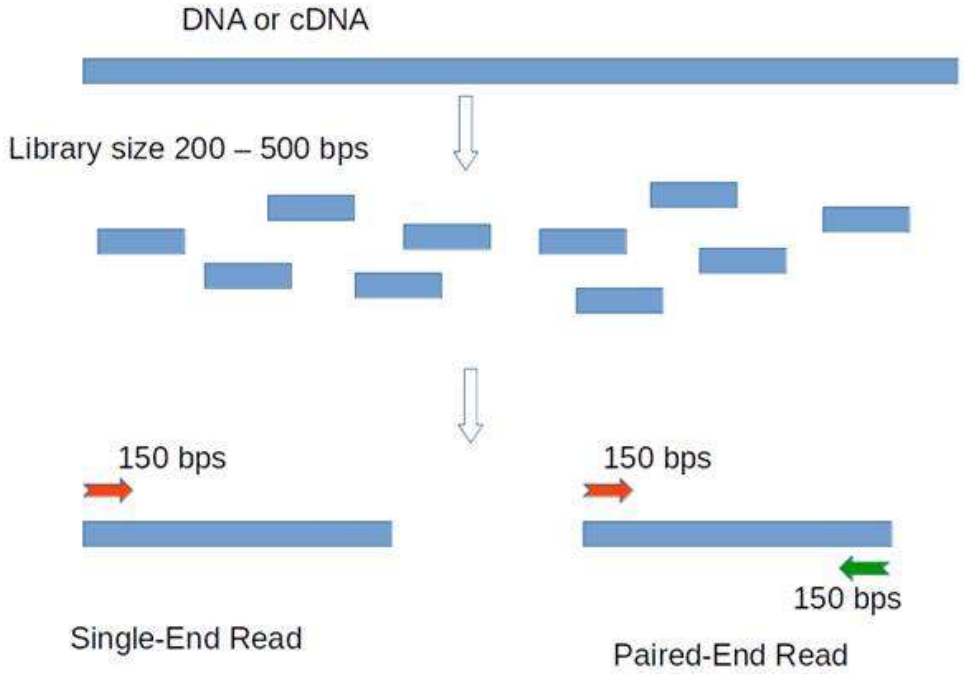


Source: www.eurofinsgenomics.co.in

- Cycles of extension and imaging
- Incorporates a single florescent nucleotide
  - each base has a unique color for each base
- Each base call has a quality score corresponding with it
  - Longer you go along, the quality diminishes

Single-End vs Paired-End Reads

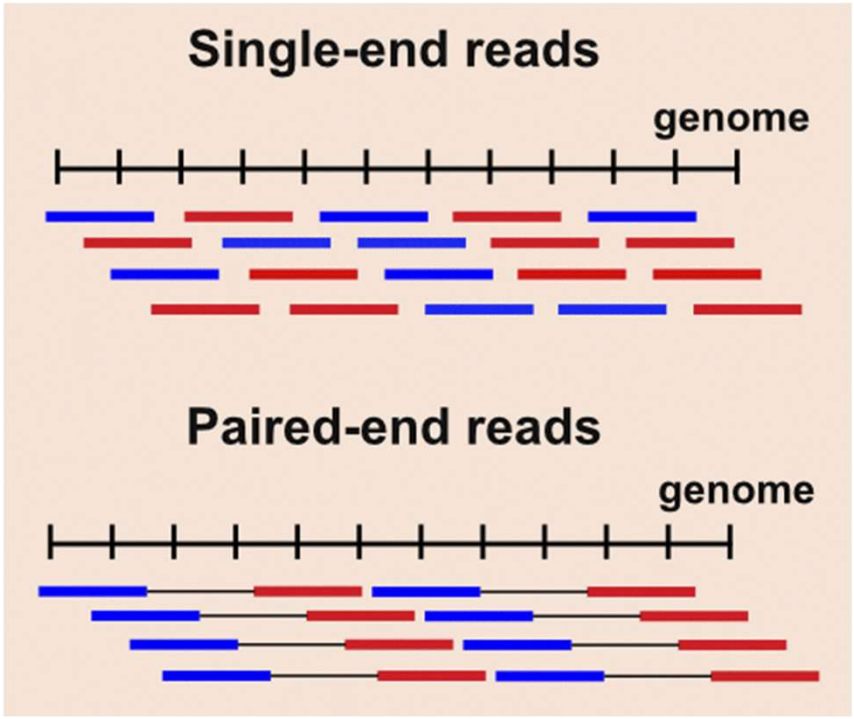
- Single only sequencing from only 1 end of the read
- Paired end sequences both ends



Source: 1010genome.com

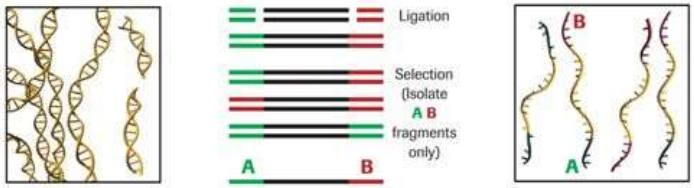
Single-End vs Paired-End Pro's and Con's

- Paired-end can detect quality alignments as it can detect splice junctions, genomic rearrangements, repetitive sequencing elements, gene fusions
- Single-end is more affordable and easier to process
- When it comes to quantitation and differential expression, do not gain substantially more in paired-end reads

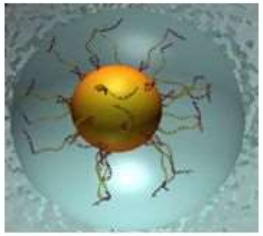


Source: genetics.org

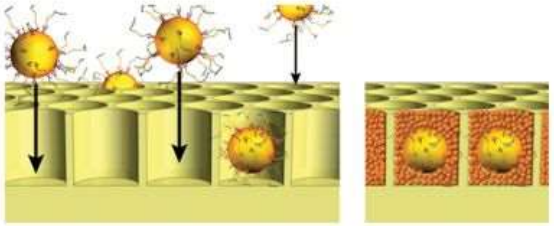
Overview of The 454 Sequencing System



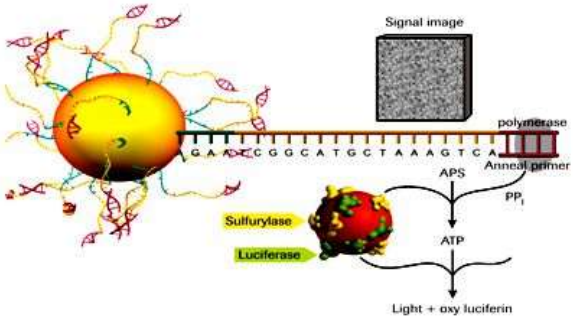
1) Prepare Adapter Ligated ssDNA Library (A-[insert]-B)



2) EmPCR: Clonal Amplification on 28 µ beads followed by enrichment



3) Load beads and enzymes in PicoTiter Plate™



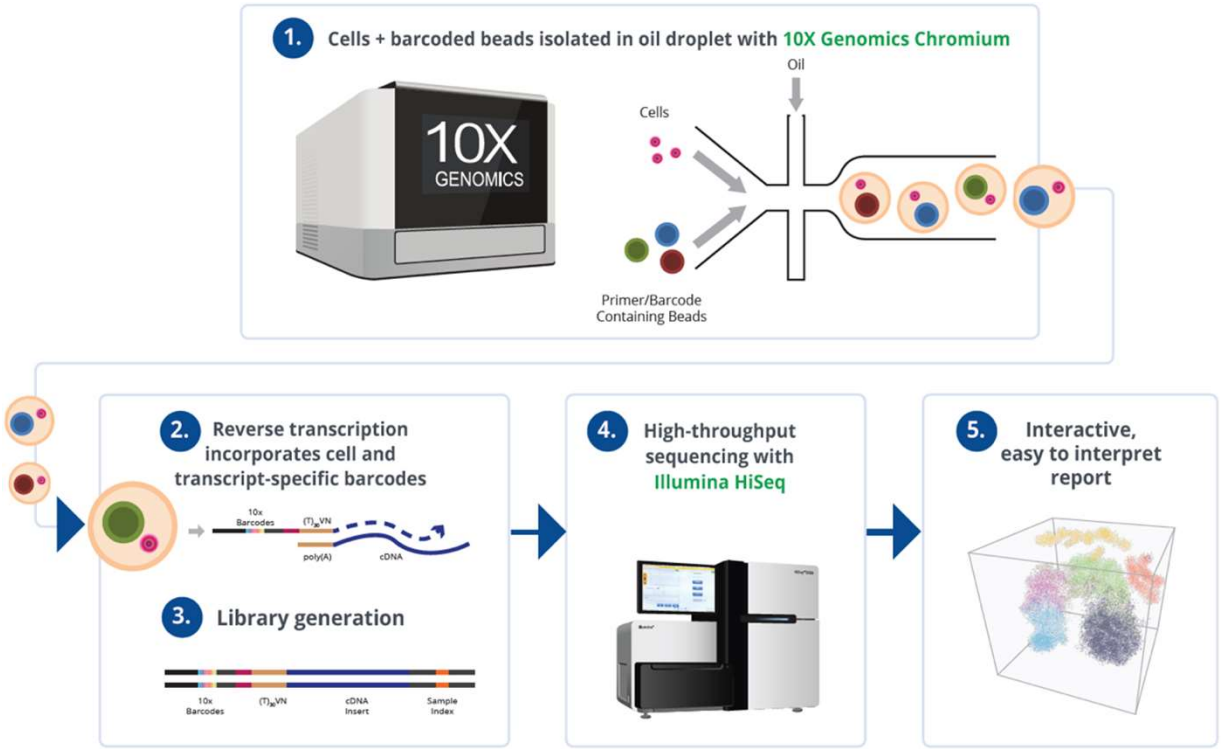
4) Perform sequencing-by-synthesis on the 454 Sequencer

CSB2008 August 2008

UCSC Sequencing Center

Single-cell Sequencing: 10x Genomics

- Separating out cells prior to sequencing
- Adding on specific barcode to each cDNA to tell which cell it came from
- Pro: single cell level
- In practice I've only seen analyses where they determine different proportions of cell types (or clusters) between groups.

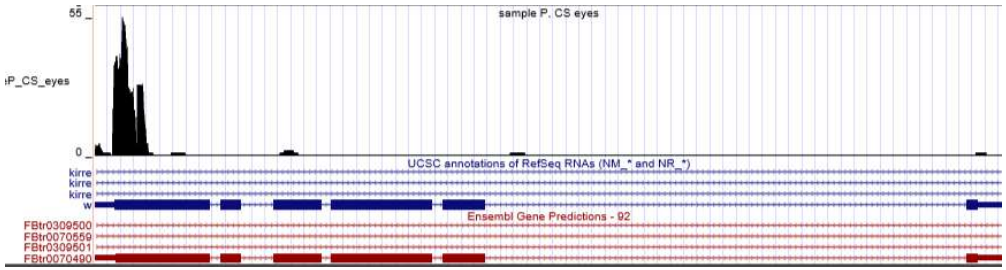


Source: 10x Genomics

Tag-Seq

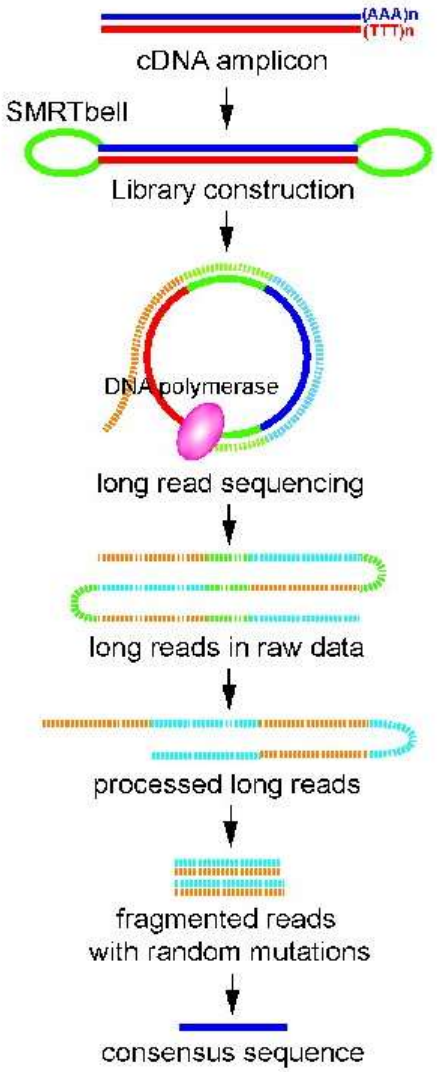
- Also known as: 3'Tag-RNA-Seq, Digital RNA-seq, Quant-Seq
- Only generates only a single library molecule per transcript, complementary to 3' sequence
- Low cost and low noise for gene expression profiling
  - 1/10 cost of traditional RNA-seq
- NOT good for isoform analysis or identifying splice junctions

- Example data viewed in UCSC genome browser



Isoform-Seq

- Also known as: Iso-Seq, single molecular real time (SMRT)-seq
- PacBio
- This gives you full length transcript information
- Biased toward shorter transcripts as your enzyme needs to stay and keep looping around
- Fantastic if you are interested in isoform information
- Very expensive



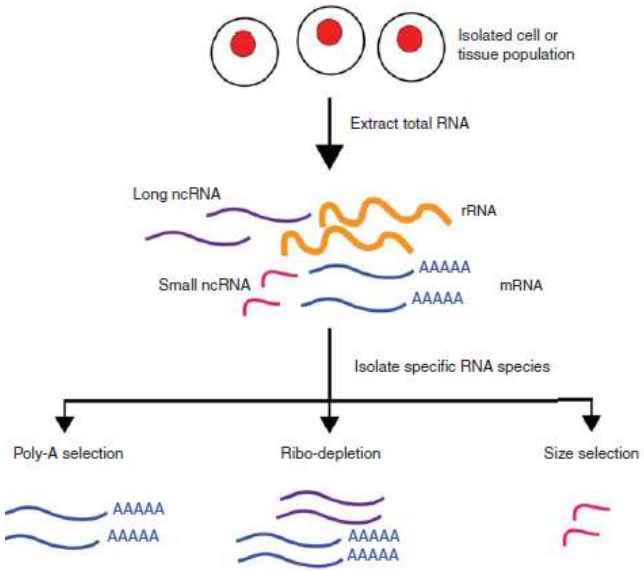
Source: Yeh et al, 2017

Read Depth/Coverage

- Coverage = # Bases Generated / Total Genome
- Reads are not distributed evenly across a transcript (or genome)
- Differential expression: 25-30M reads/sample
- Rare transcript, splice variants, de novo transcriptome: 100-200M reads/sample

Main Questions to Ask

What type of RNA do you want to look at?



Source: RNA-Seq Blog

What sequencing technology most benefits you and your hypothesis?

- Do you want isoform information?
  - Potentially reconstruct a transcriptome?
  - Want to look at gene fusions?
  - Is there a well annotated transcriptome for your species?
- Do you need individual cell information?
- Do you need high sample sizes for modeling?
  - Longitudinal models
  - Data-driven network analysis (WGCNA)

References

Merino J, Florez JC, (2018) *Precision medicine in diabetes: an opportunity for clinical translation*. Ann. N.Y. Acad. Sci. 1411:140–152.

García-Sevillano MÁ, García-Barrera T, Gómez-Ariza JL. (2015) *Environmental metabolomics: Biological markers for metal toxicity*. Electrophoresis. 2015 Sep;36(18):2348-2365.

Meydan C, Shenhar-Tsarfaty S, Soreq H. (2016) *MicroRNA Regulators of Anxiety and Metabolic Disorders*. Trends Mol Med. 2016 Sep;22(9):798-812.

Wang, KC and Chang HY. (2011) *Molecular Mechanisms of Long Noncoding RNAs*. Mol Cell. 2011 Sep 16;43(6):904-14.

Yeh, Hsin-Sung and Zhang, Wei and Yong, Jeongsik. (2017) Analyses of alternative polyadenylation: from old school biochemistry to high-throughput technologies. BMB Rep. 50(4): 201–207.

Auer & Doerge (2010) *Statistical Design and Analysis of RNA Sequencing Data.* 185:405-416.