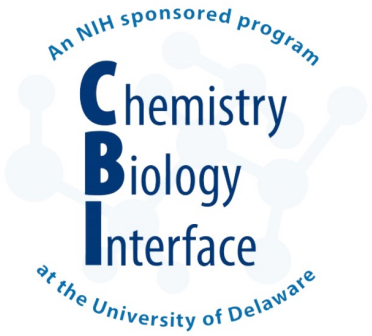
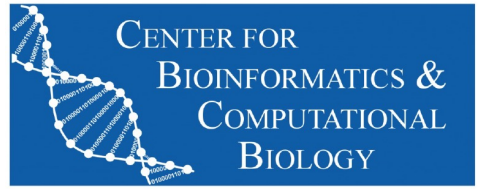
Omics Overview

FAIR Data Practices for Omics Analysis Workshop

University of Delaware

April 19 (Day 2)





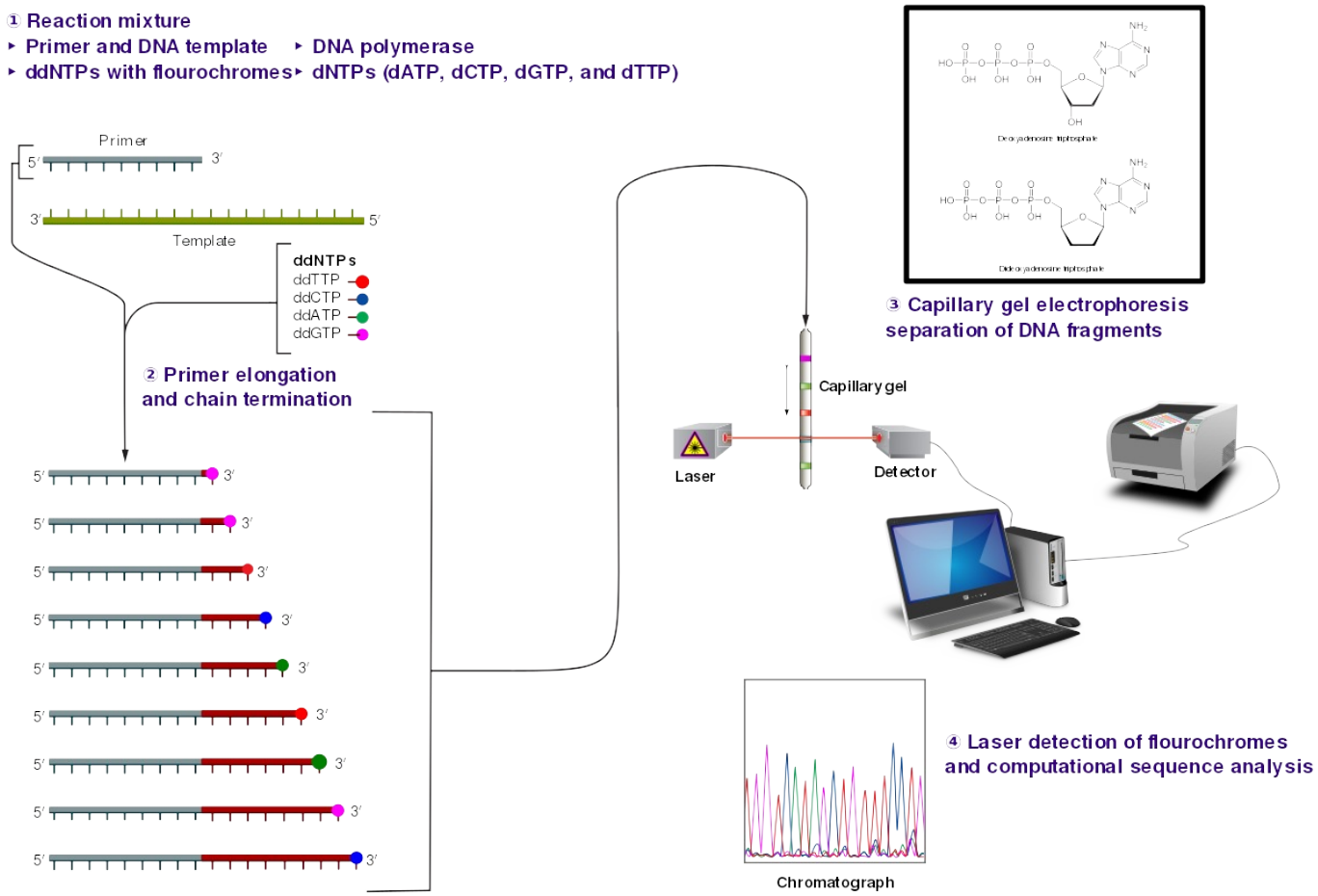
First Generation DNA Sequencing

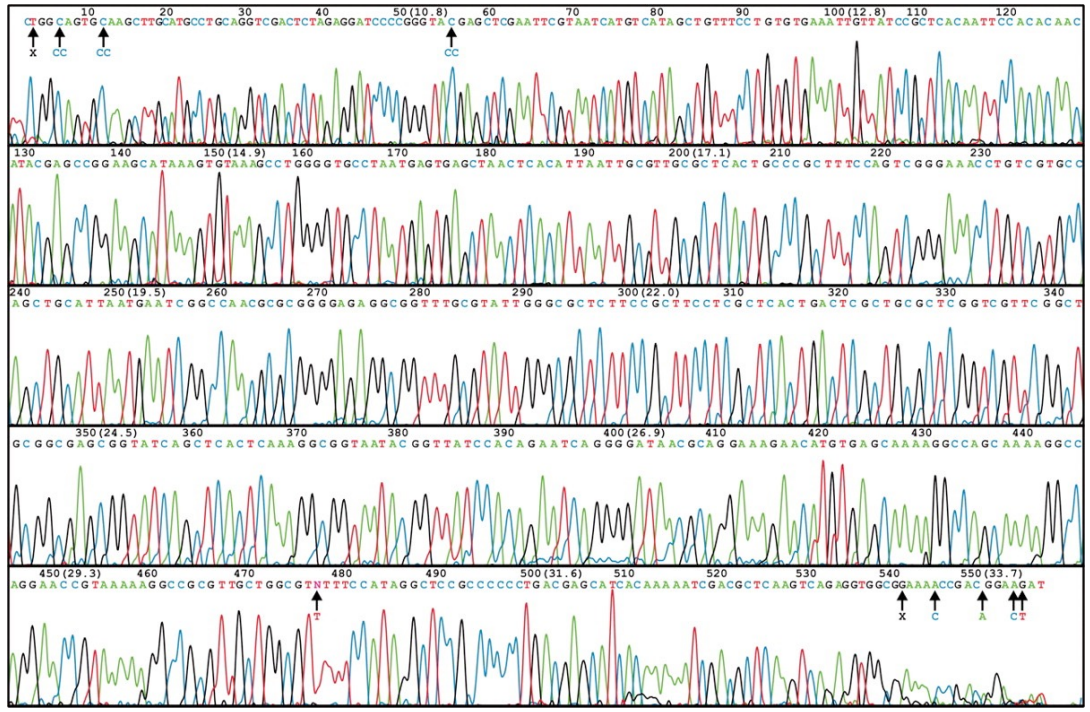
Key Concept:

- **Input**: Many copies of same DNA template
- **Output**: One sequence

Automated “Sanger”

1986 – Slab gel reader
1996 – Capillary electrophoresis





Characteristics of First Generation Sequencing

- A single template is being sequenced
- Many copies of the fragments are needed for detection
- Size separation (typically by gel electrophoresis)
- Read lengths up to 800-1000bp (much shorter earlier)

Applied Biosystems (ABI) 3100

First Gen Sequencing Today (Sanger)

- Technology of choice when a single template sequence needs to be determined
- **Advantages:** Targeted, Accurate, Rapid
- **Disadvantages:** Cloning or PCR dependent (can cause biases)
- Common Applications
  - Clone sequencing
  - Amplicon sequencing
  - Diagnostics
  - Confirming experiments / other technologies

Second Generation DNA Sequencing

Key Concept:

- **Input**: Many copies of **many** DNA templates
- **Output**: Many sequences

Second (Next) Gen Sequencing

- Massively Parallel
- PCR-dependent
- Short-ish: 25-700bp
- Technology Varies
  - Pyrosequencing
  - **Sequencing by synthesis**
  - Ion Sequencing
  - Nanoball
  - Polony
  - Sequencing by Ligation
- Usually better known by commercial platform names →

POLONATOR

illumina®

Complete genomics A BGI Company

ion torrent by life technologies™

SOLiD™ applied biosystems part of life technologies

Roche 454 SEQUENCING

Third Generation Sequencing

Key Concept:

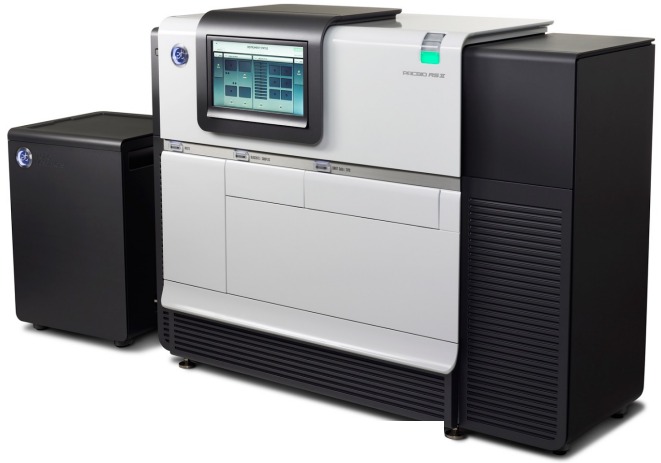
- **Input**: Single molecule of **many** templates
- **Output**: Many sequences

Third Generation Sequencing

- AKA: Single molecule, "Next, Next Gen"
- Advance: a single molecule of DNA is sequenced (no amplification needed)
- Presently, two commercially viable approaches:
  - Single Molecule Real Time Sequencing (PacBio)
  - Nanopore Sequencing (Oxford Nanopore)
- **Similar advantages:** long sequencing reads (reads >10,000bp, sometimes much longer)
- **Similar disadvantages:** relatively high error rates

Single Molecule Real Time (SMRT) Sequencing

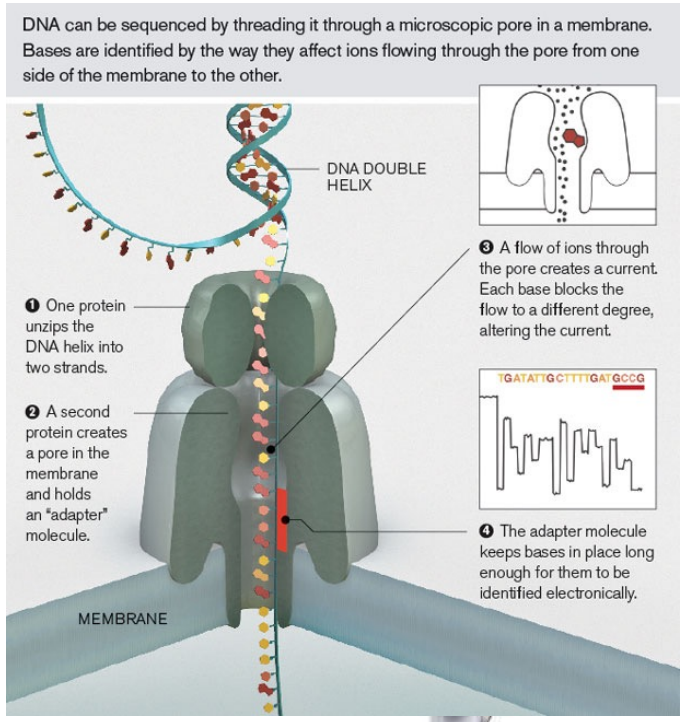
- PacBio Sequencing Platforms: RS2, Sequel, Sequel IIe
- Similar technology
- Advantages:
  - long read length; can be >>150kb
  - No amplification bias
  - Detection of base modifications
  - PacBio works with community on software development
- Disadvantages:
  - Single Pass Accuracy rate is 84-87%
  - BUT, errors are random and almost always inserts (~11%)…so with multi-pass accuracy >99% is possible (HiFi)
  - Sequencers are large and expensive
- Applications
  - Genome assembly, isoform detection, epigenetics, long amplicons



Oxford Nanopore MinION

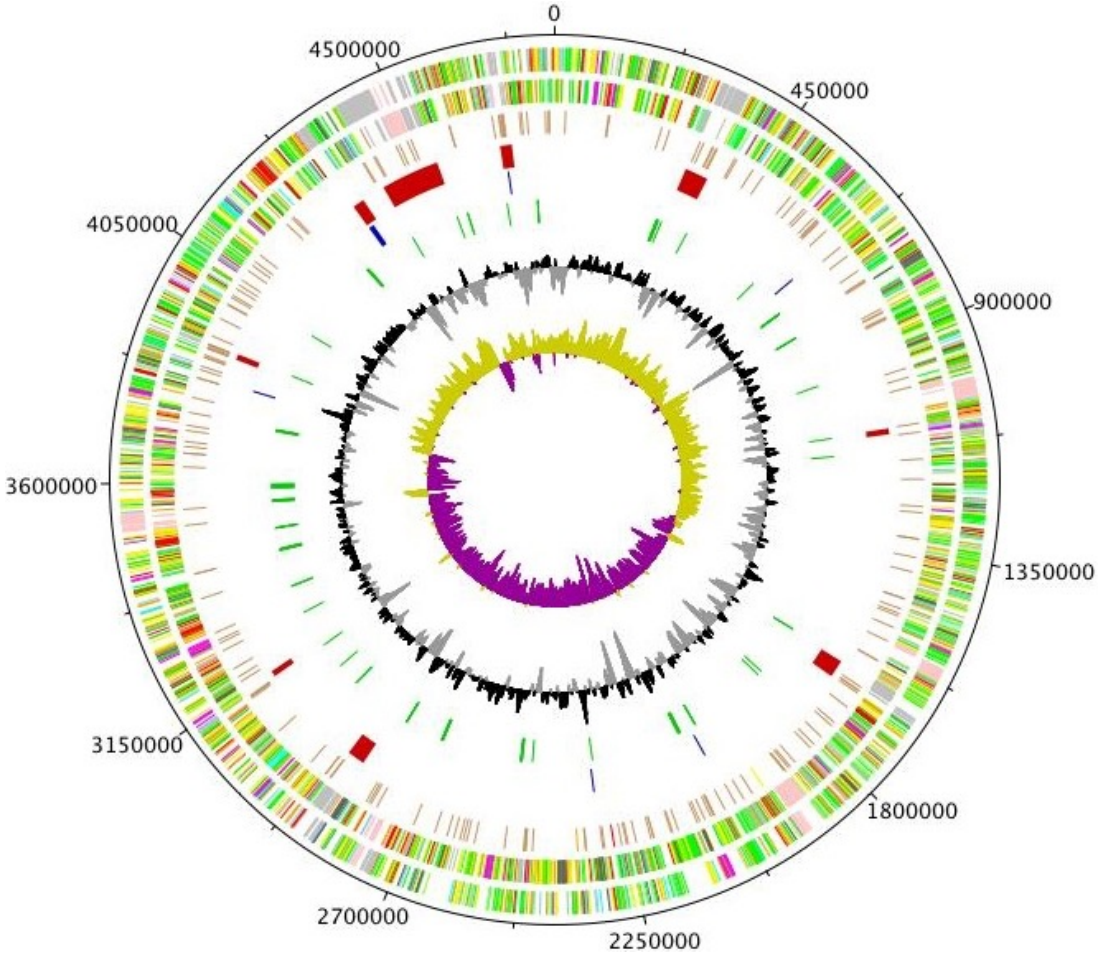
- Most common Nanopore sequencer
- Very small!
- Data is analyzed immediately, opportunity for real time analysis
- No Amplification
- Sample prep very simple
- Error rate
  - High and variable (some reads may have 5%, others 30%)
  - Avg similar to PacBio, but not as consistent
  - InDel and substitution errors common
- Potential to detect base modifications
- Can sequence direct from RNA
- Applications: Still Evolving
- https://youtu.be/Wq35ZXyayuU

MIT Technology Review



3rd Gen Sequencing Impact

- PacBio has changed genome assembly
  - Long reads with errors are superior to short accurate reads for genome assembly
  - Most bacterial genomes completely closed for ~$1000
  - Also very useful at improving Eukaryotic genome assemblies
- Transcript sequencing: Enables isoform detection
- Promise of epigenetics (base modification) looms . . . Need more throughput



Second Generation DNA Sequencing

Key Concept:

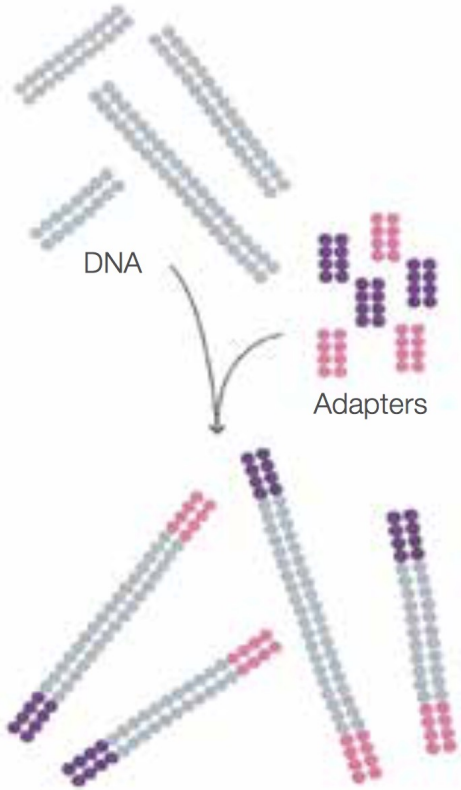
- **Input**: Many copies of **many** DNA templates
- **Output**: Many sequences

Illumina – The Basics

- Step 1. Nucleic Acids Extracted and Purified
  - Only DNA can be sequenced
  - If RNA additional steps will reverse transcribe to DNA during Step 2

- Step 2. Library Preparation
  - Quality is checked
  - Nucleic Acids fragmented to a target size by chemical or physical (sonication) means
  - Adapter sequences are appended to both ends
    - Illumina Adapters (called P5 and P7)
    - Barcodes (Indexes) – multiple samples on one lane
    - Sequencing Adapters

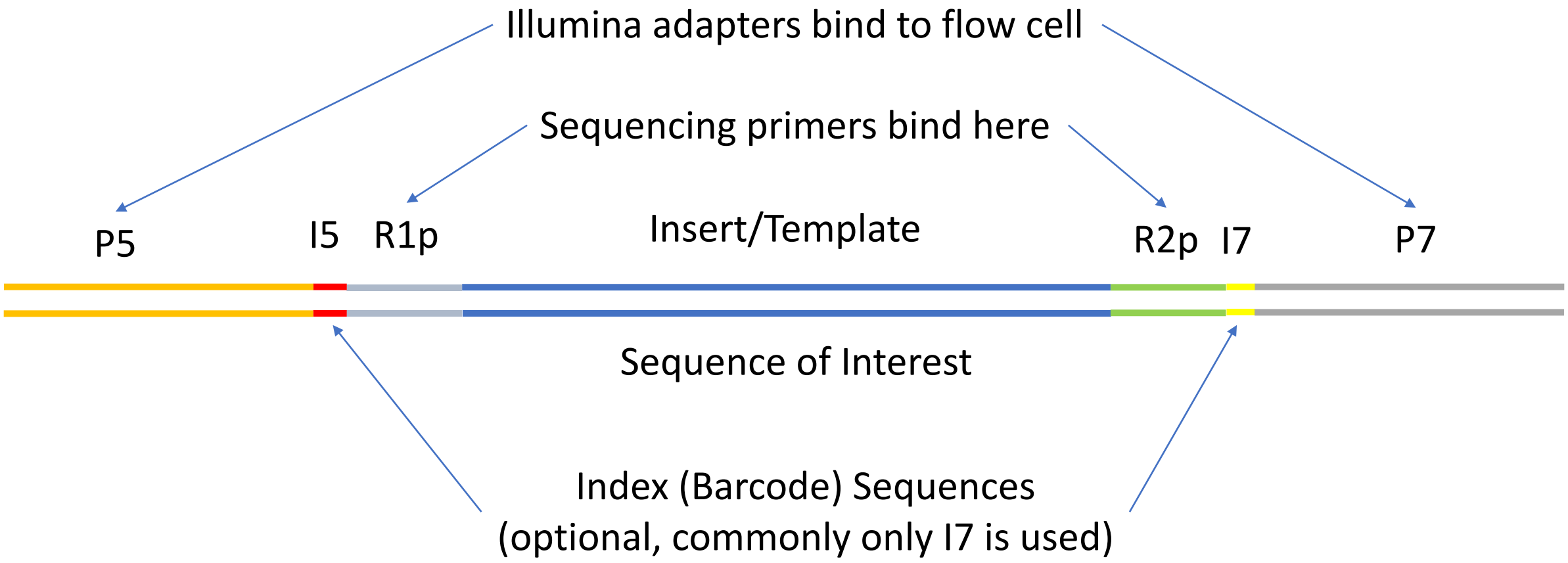
2



Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Illumina – The Basics

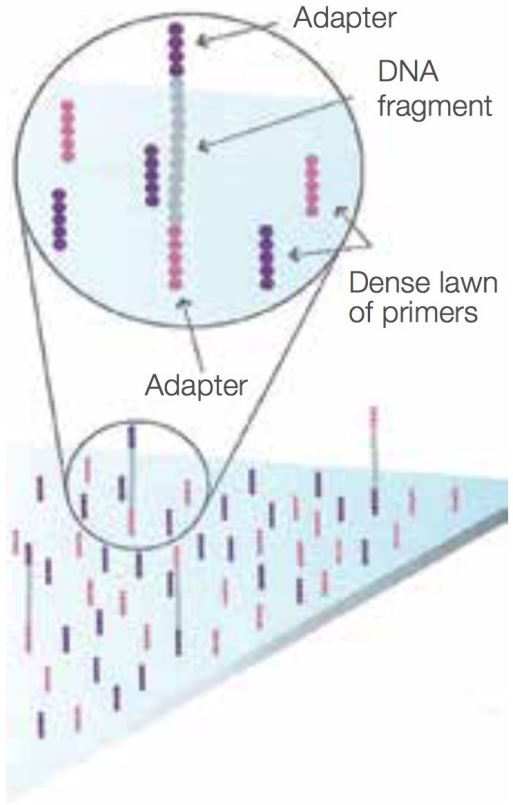
Fragment After Library Preparation



Illumina – The Basics

- Step 3. Bind to Flow Cell
  - Flow Cell Lanes are covered in a lawn of oligonucleotides that match both "Illumina adapters" from Step 2.
  - Thus when the library is denatured (strands separated) and flowed across a lane, both strands can bind to the lane
  - Polymerase and nucleotides are added, replicating the DNA once
  - This replicated strand is now anchored to the flow cell and when the flow cell is heated the original strand flows away
  - Figure at right shows the result

3

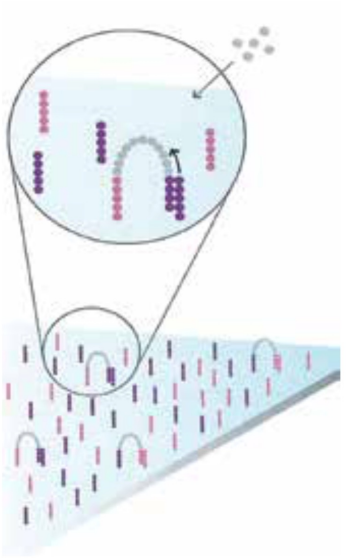


Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Illumina – The Basics

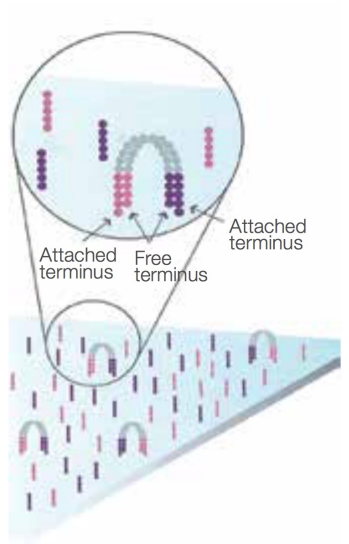
- Step 4-6. Bridge Amplification
  - Each ssDNA molecule has adapter on both ends, so it can bend over and bind to the primer for its opposite end on the flow cell
  - DNA Polymerase and nucleotides flooded across the lane convert the bridged ssDNA sequences into dsDNA
  - Heat is applied to denature these strands such that the number of molecules has doubled
  - This cycle is repeated many times (essentially PCR on a slide)

4



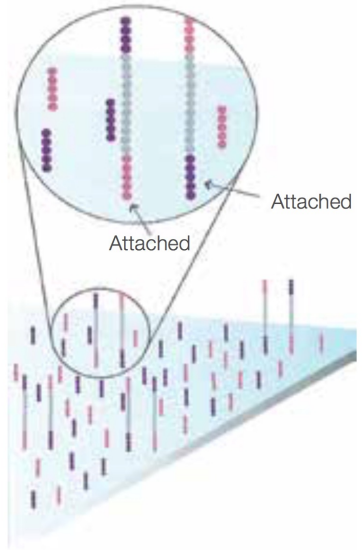
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

5



The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

6

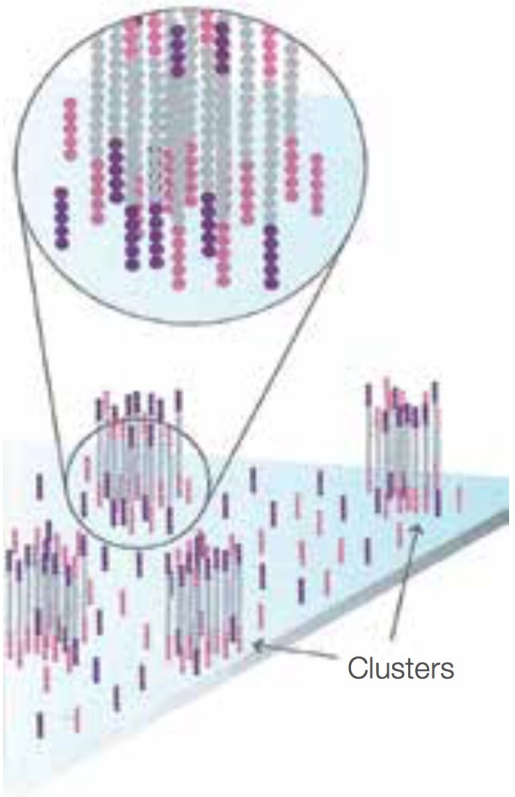


Denaturation leaves single-stranded templates anchored to the substrate.

Illumina – The Basics

- Step 7. Clustering
  - After many rounds of bridge amplification dense "Clusters" are formed
  - Each cluster represents many copies of the same sequence (attached in both directions)
  - The sequencer will map where the clusters are located and image that location in subsequent steps
  - Each cluster will produce one sequence in the output
  - Depending on the platform ~10 million to 10 billion clusters can be simultaneously sequenced in one flow cell
  - Reverse sequences bound to the flow cell are clipped and removed, such that only forward sequences remain

7

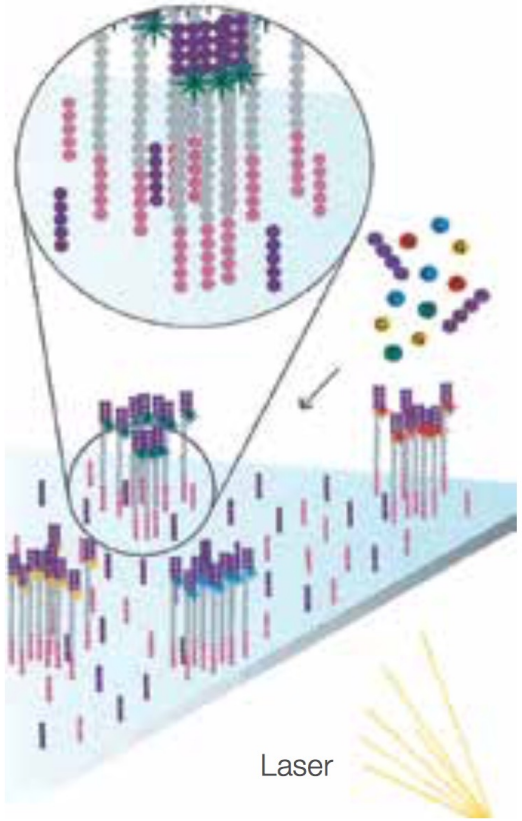


Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Illumina – The Basics

- Step 8. Begin forward sequencing reaction
  - Sequencing primer, polymerase and all four fluorescently labeled NTP's are added
  - NTP's have a reversible terminator which prevents more than one NTP from being added to each molecule

8

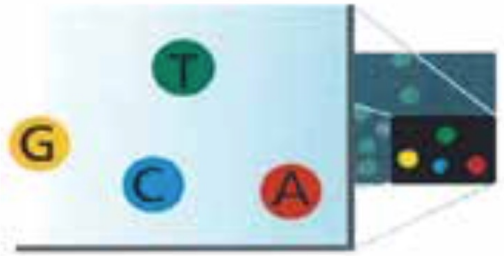


The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

Illumina – The Basics

- Step 9. Read the first bp
  - Lasers are used to excite the fluorescently labeled NTPs
  - Each cluster should have had the same dNTP added to all molecules, so each cluster should illuminate a single color
  - Optics record the color for each cluster
- Step 10. Reset
  - Chemicals/enzymes are added that remove the fluorophore and reversible terminator from each molecule
- Step 11. Repeat 8-10 for each bp to add

9



After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

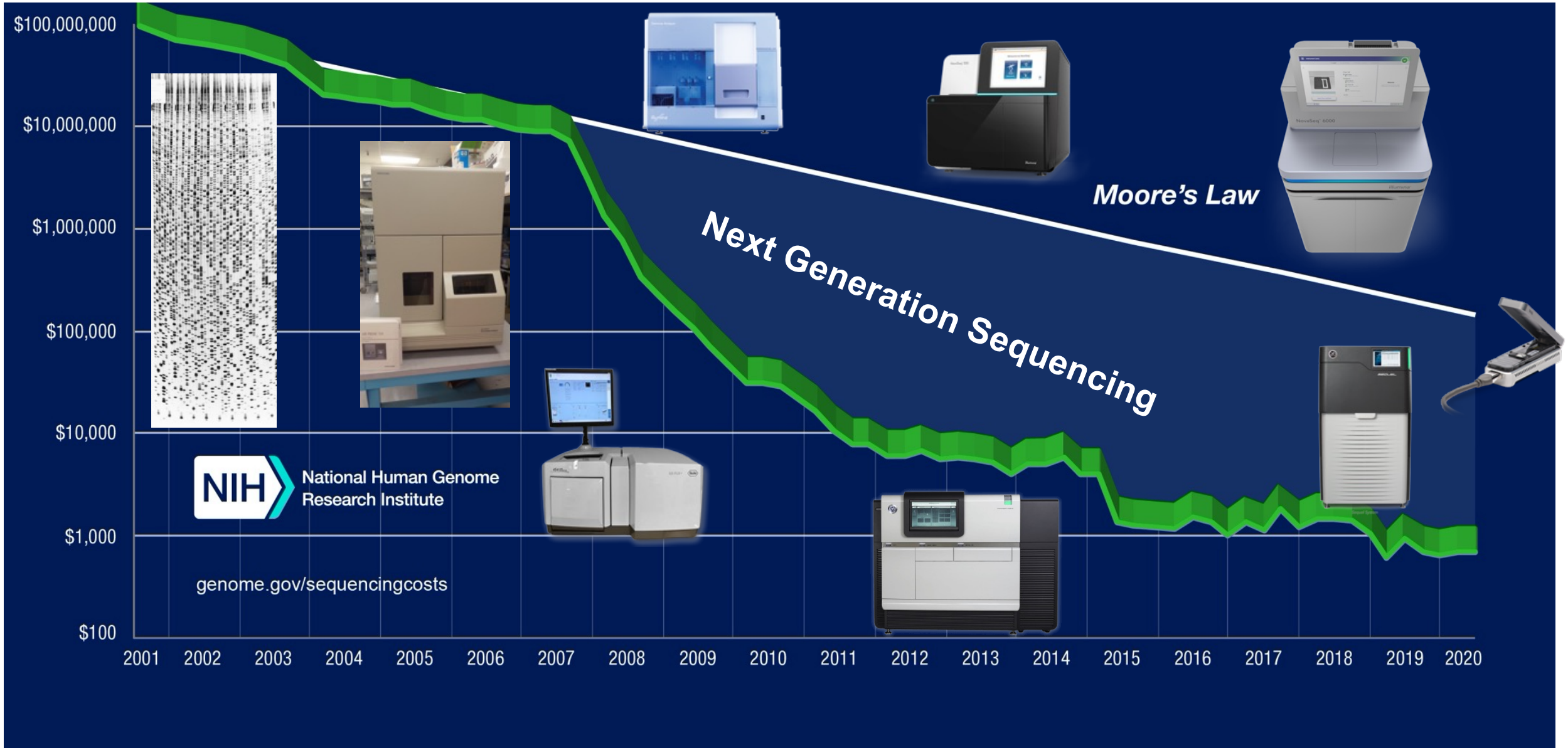
Why do we need to know all of this to do bioinformatics?

Experimental Design & Troubleshooting

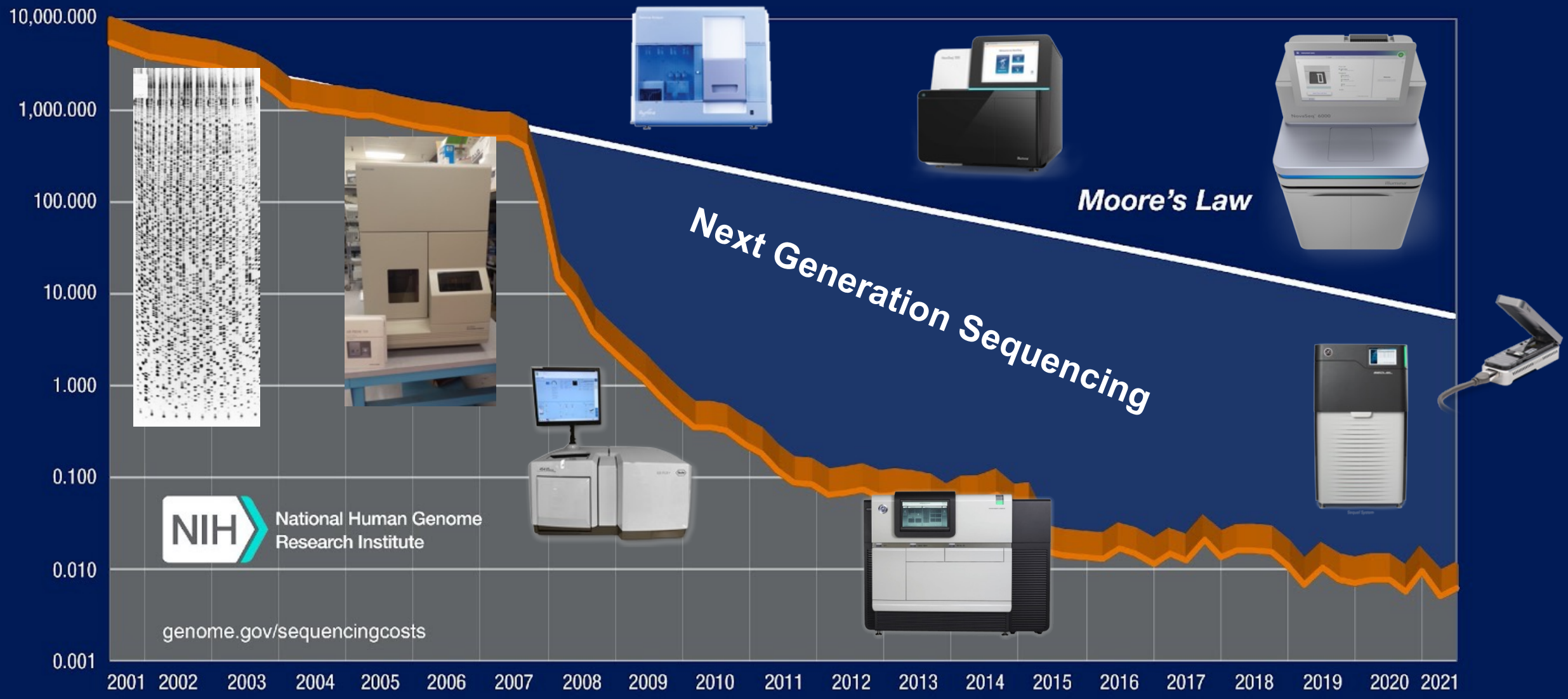
What did second generation sequencing do for biology?

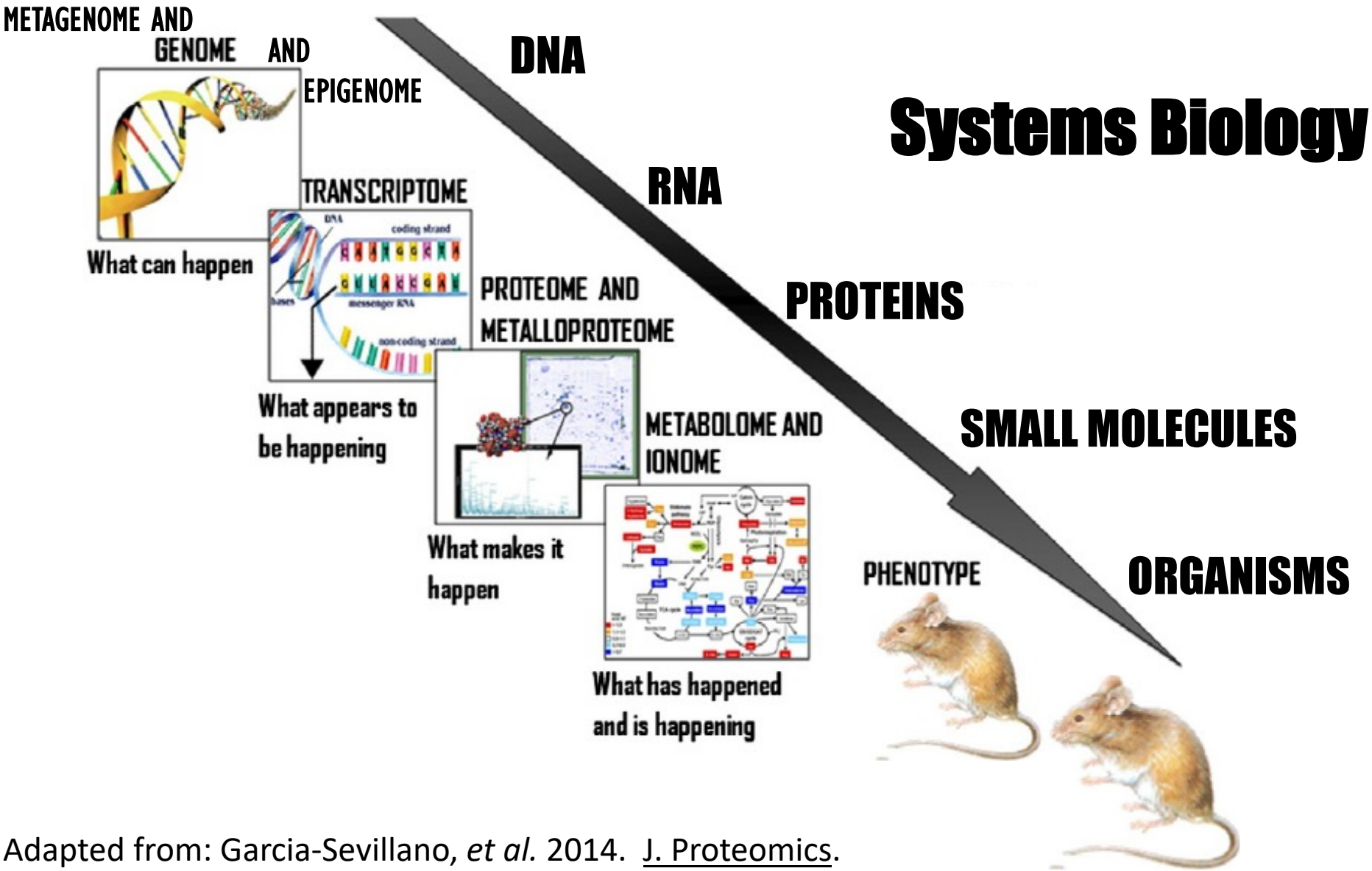
Changed How Research is Done

Cost per Human Genome



Cost per Raw Megabase of DNA Sequence





Adapted from: Garcia-Sevillano, *et al.* 2014. J. Proteomics.

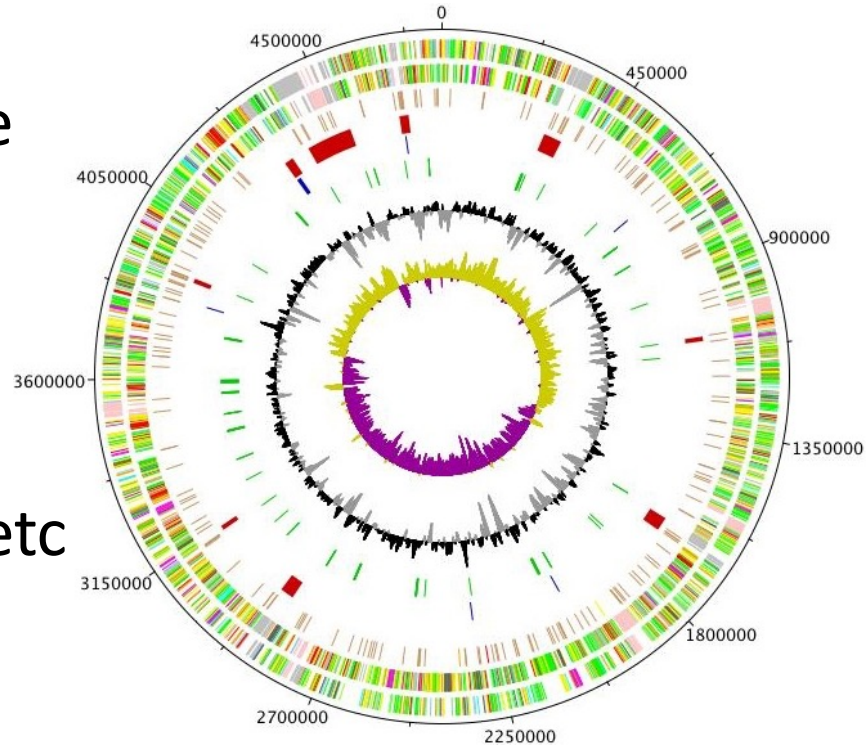
De novo Genome Assembly

Objective: Understand the genomic potential of an organism.

Approach: Shotgun (random) sequencing. Assemble the pieces. Annotate gene features.

Advantages: Provides the knowledge base for cell and molecular analysis of an organisms. Simplifies genetics, transcriptomics, proteomics, metabolomics, epigenetics, etc

Disadvantages: Rarely captures the full genetic potential. Takes many, many iterations to refine.



Omics for:
Understanding Populations and Genetic Diversity

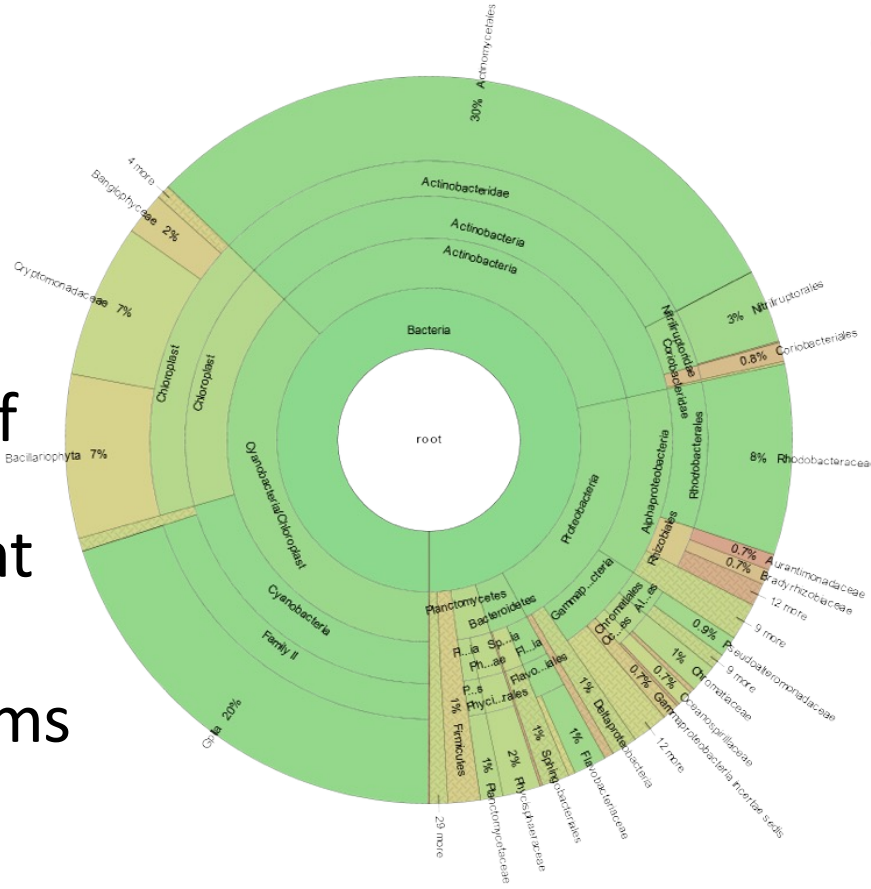
Amplicon Metagenomics

Objective: Understand population-level differences between communities of organisms

Approach: Amplify marker gene (16S rRNA) by PCR. Sequence many amplicons from each sample.

Advantages: Inexpensive, high-resolution snapshots of organisms are in a community and their relative population sizes. Analysis straight forward.

Disadvantages: No functional data (what are the organisms doing?), primer/gene choice can limit analysis



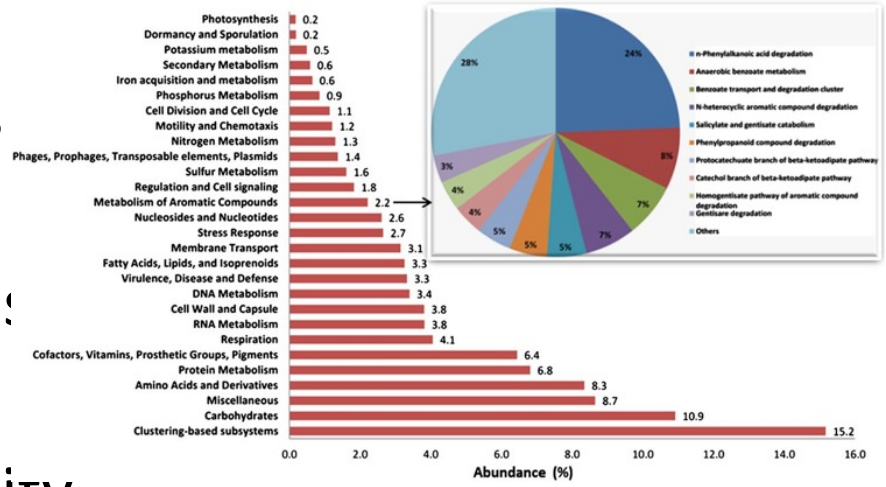
Shotgun Metagenomics (MetaG)

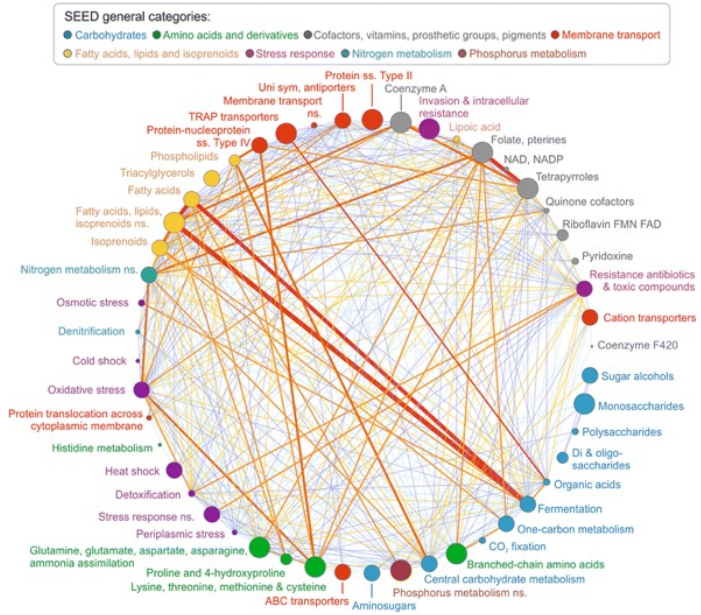
Objective: Understand population-level differences between communities of organisms

Approach: Isolate DNA from a community of organis Shotgun (random) sequence.

Advantages: Comprehensive view of what a community of organisms is <u>functionally capable</u> of doing.

Disadvantages: Expensive. May not actually performing those functions. Lower resolution than amplicon. Analysis can be tricky.





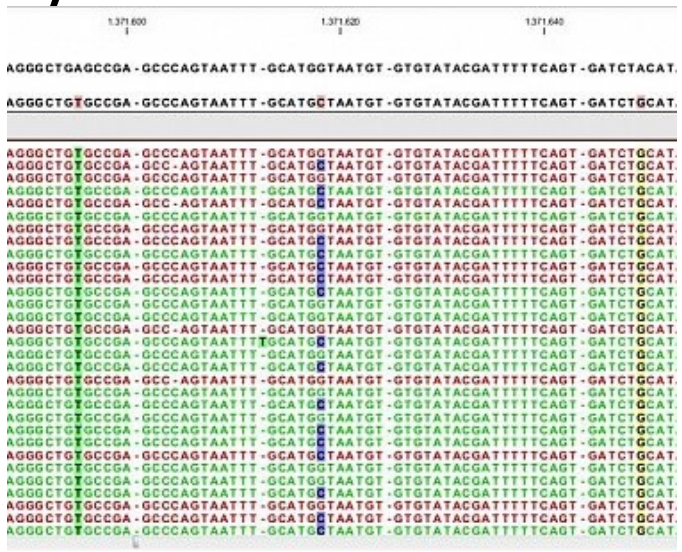
DNA Resequencing (Variant Analysis)

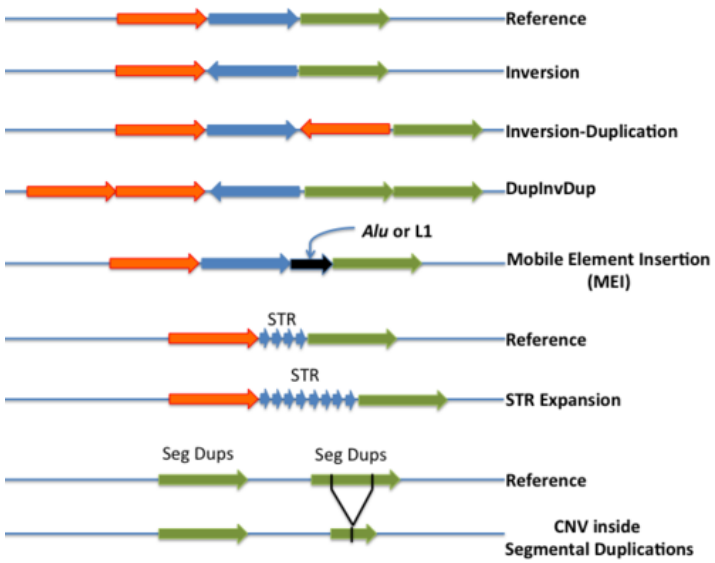
Objective: Understand the genetic variability of organisms.

Approach: DNA sequencing of whole genome or exome. Map to reference genome. Identify sequence variants: small (SNV, InDel) or structural (rearrangements, duplications, etc)

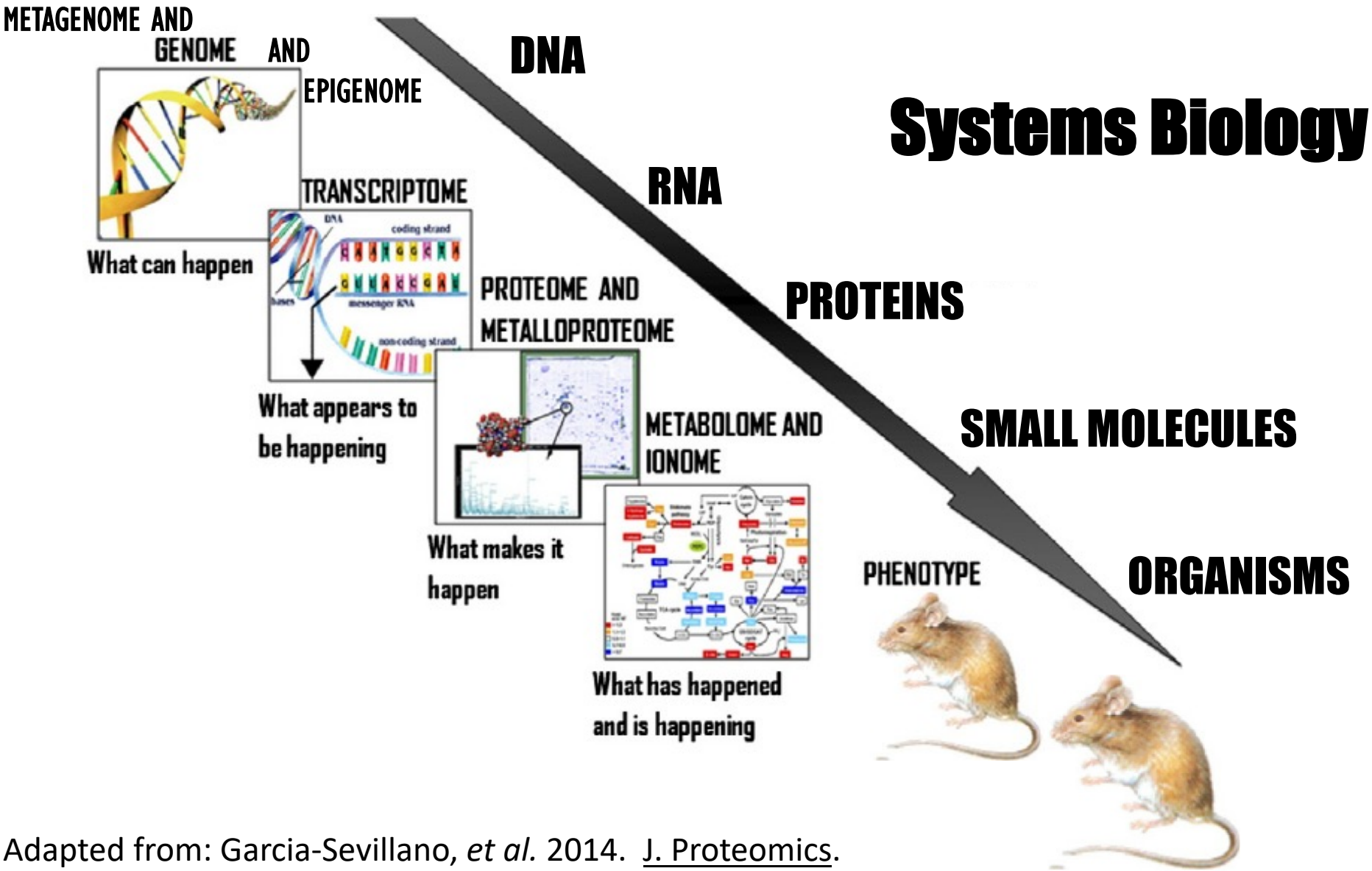
Advantages: Allows the genetic variations of groups and individuals to be compared.

Disadvantages: Zygosity, InDels, sequencing error can cause misidentification of variants.





Epigenetics – Variation Beyond Genome Sequence



Adapted from: Garcia-Sevillano, *et al.* 2014. J. Proteomics.

Epigenetics

- Interactions that affect the genome without changing its sequence
- Nucleotide modifications
  - Methylated Nucleotides (5-methyl Cytosine: 5mC)
  - Others . . . Often DNA damage indicators
- Proteins that bind DNA
  - Histones – chromatin structure may be changed by modifications of histones (e.g. protein methylations)
  - Transcription Factors – Gene regulation
- Often seen as an intersection of genetics and environmental response.

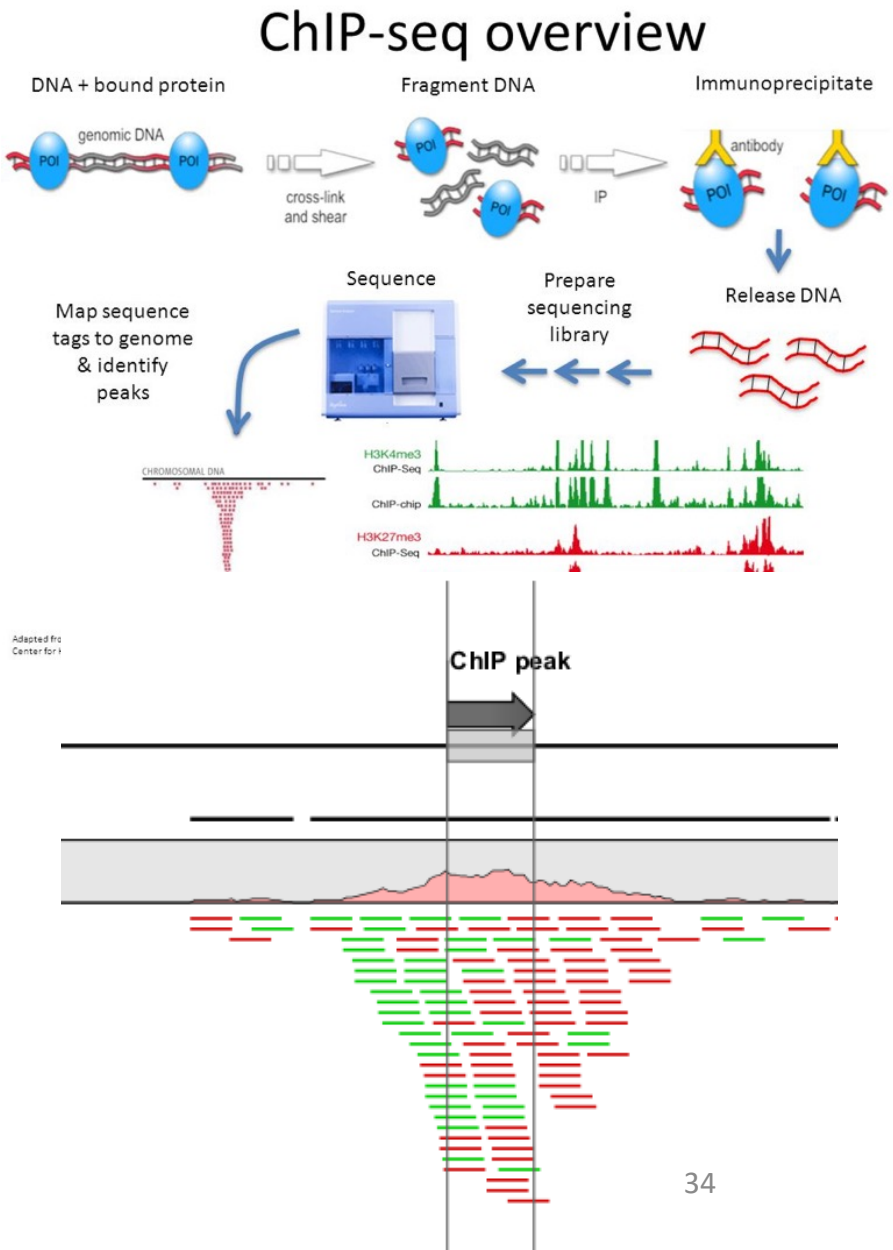
Epigenetics (ChIP-Seq)

Objective: Understand how proteins interact with genomic DNA

Approach: Use antibody to "pull down" a specific protein (transcription factor, histone). Sequence DNA attached to protein. Map to reference genome . . . look for peaks.

Advantages: DNA protein-interactions can be key for regulation of gene expression and other processes.

Disadvantages: Analysis not always clear cut. Non-target DNA is sometimes pulled down. Appropriate controls very important (input DNA, IgG). May be difficult to determine exact boundaries.



Epigenetics (ATAC-Seq)

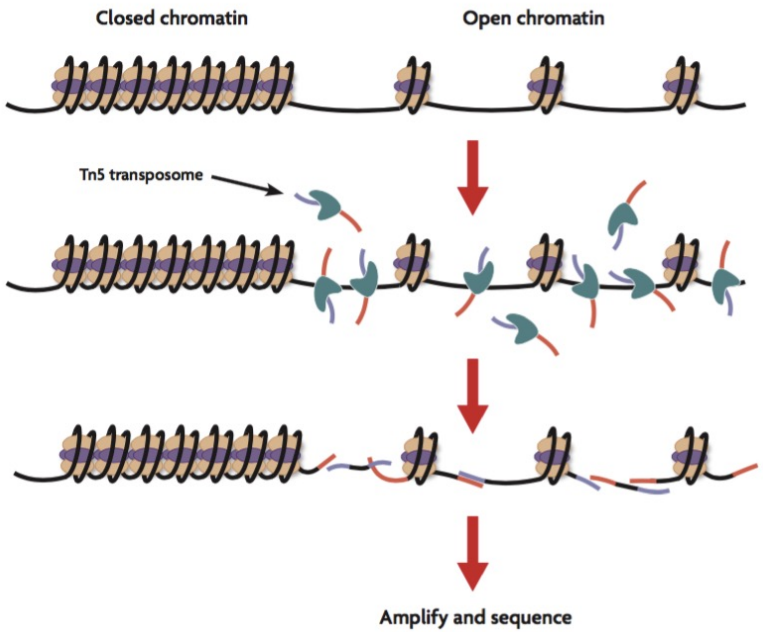
Assay for Transposase-Accessible Chromatin

Objective: Understand which regions of chromatin are accessible (not bound to proteins)

Approach: Use transposase (Tn5) to and insert adapters (priming sites) in accessible regions of the genome, then amplify and sequence.

Advantages: Can provide insights into what genes are accessible for active transcription given chromatin structure.

Disadvantages: When performed on bulk samples (i.e. tissue samples) it can be difficult to interpret as different cells may have different accessible DNA.



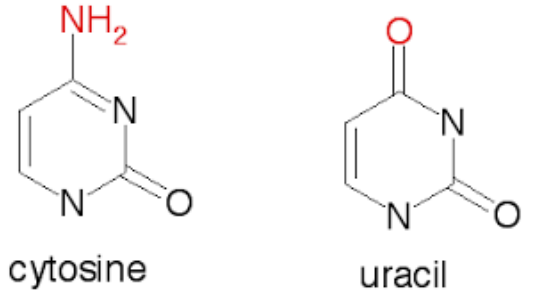
Epigenetics (BiS-Seq)

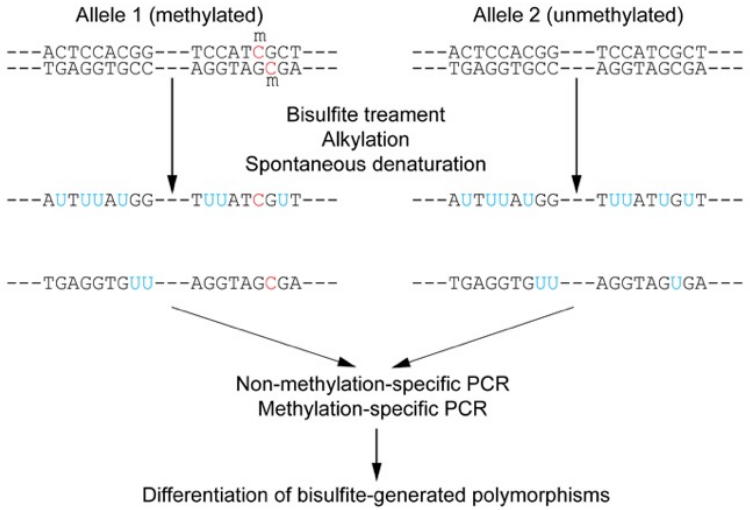
Objective: Identify modified nucleotides in the genome (5-methyl-Cytosine)

Approach: Bisulfite treatment will convert unmethylated C to Uracil. Sequence converted & unconverted libraries. Compare to find C -> T shifts (or lack of).

Advantages: Most sensitive way to identify global 5mC patterns.

Disadvantages: Bisulfite conversion is messy . . . Doesn't always work. 5-hydroxymethyl-Cytosine also converted. Only works for Cytosine modifications. Requires a lot of sequence coverage (expensive).





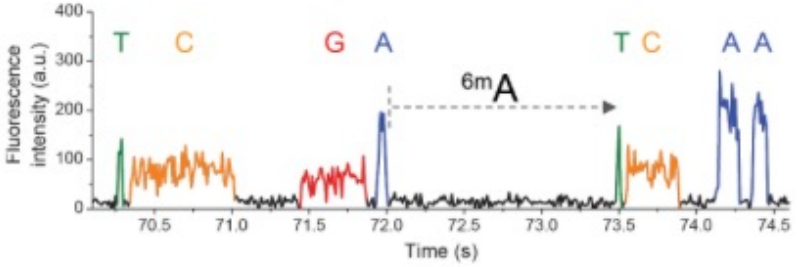
Epigenetics (PacBio/ONT)

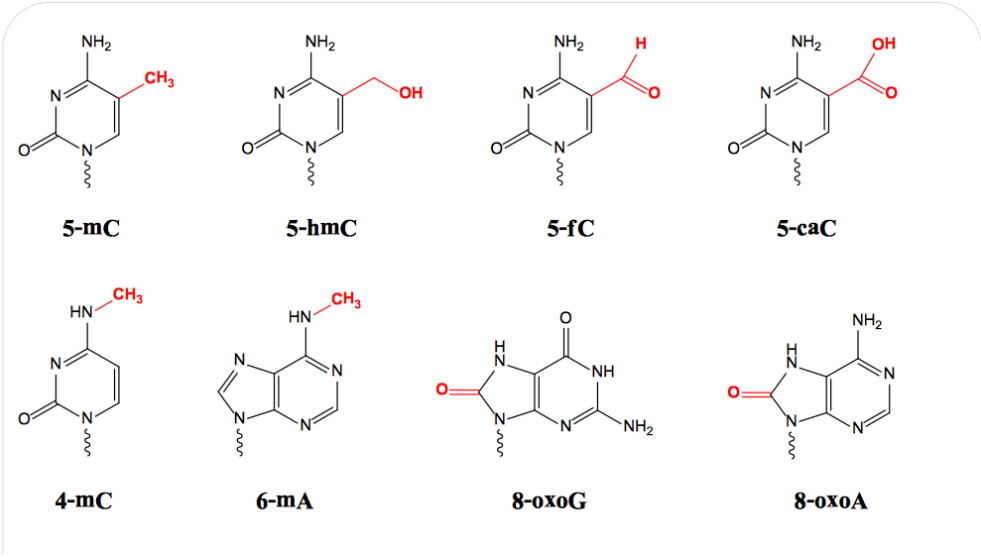
Objective: Identify modified nucleotides in the genome

Approach: PacBio sequencing can detect modified bases from the base incorporation rates

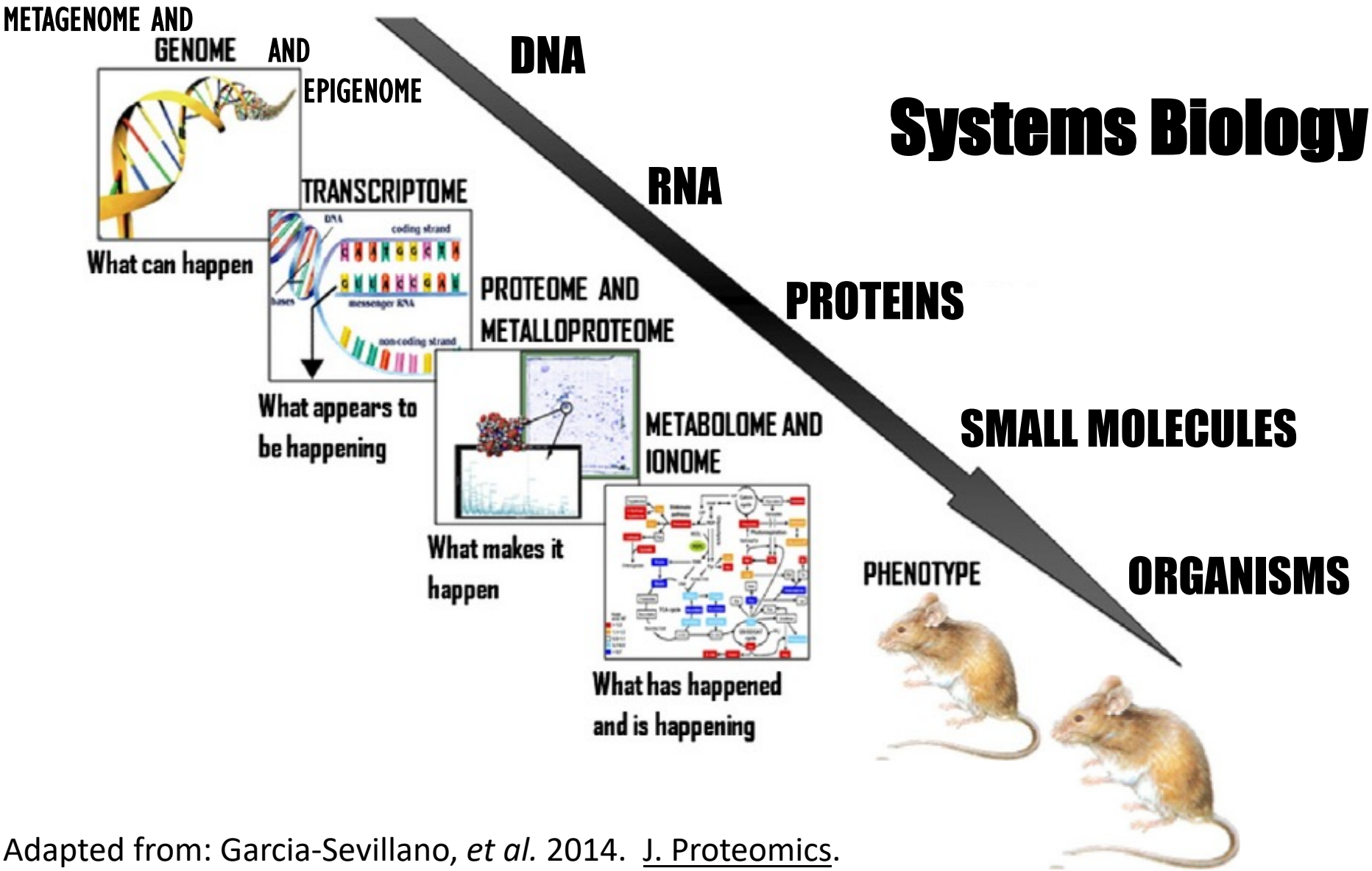
Advantages: Capable of detecting several base modifications without conversion

Disadvantages: High coverage needed for some: 5mC requires ~500X – expensive. Low sensitivity. Algorithms evolving.





Gene Expression



Adapted from: Garcia-Sevillano, *et al.* 2014. J. Proteomics.

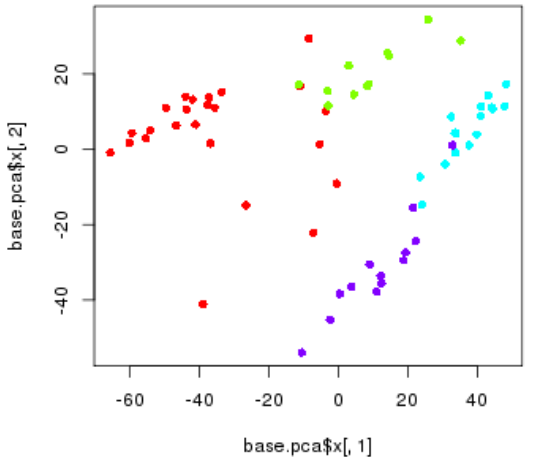
Transcriptomics (mRNA-Seq)

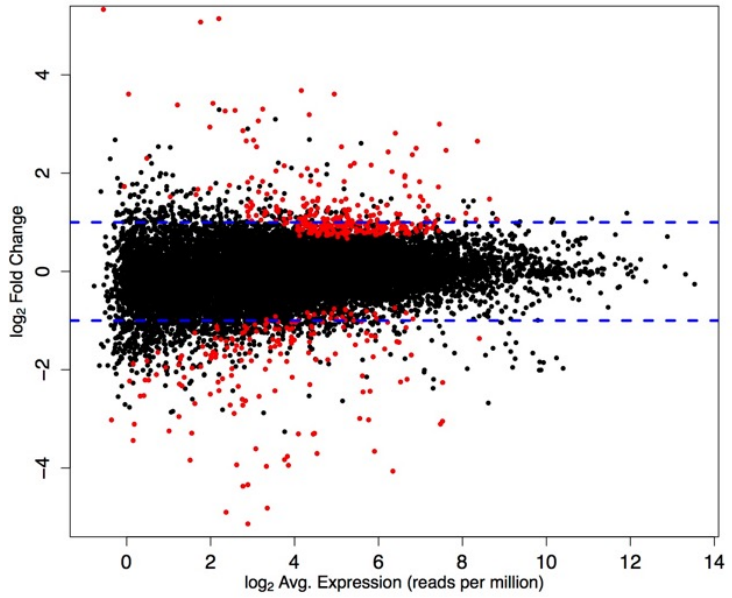
Objective: Determine differentially expressed genes

Approach: Select mRNA. Sequence fragments. Map to reference. Determine differentially expressed genes.

Advantages: Most sensitive global transcriptome profiling.

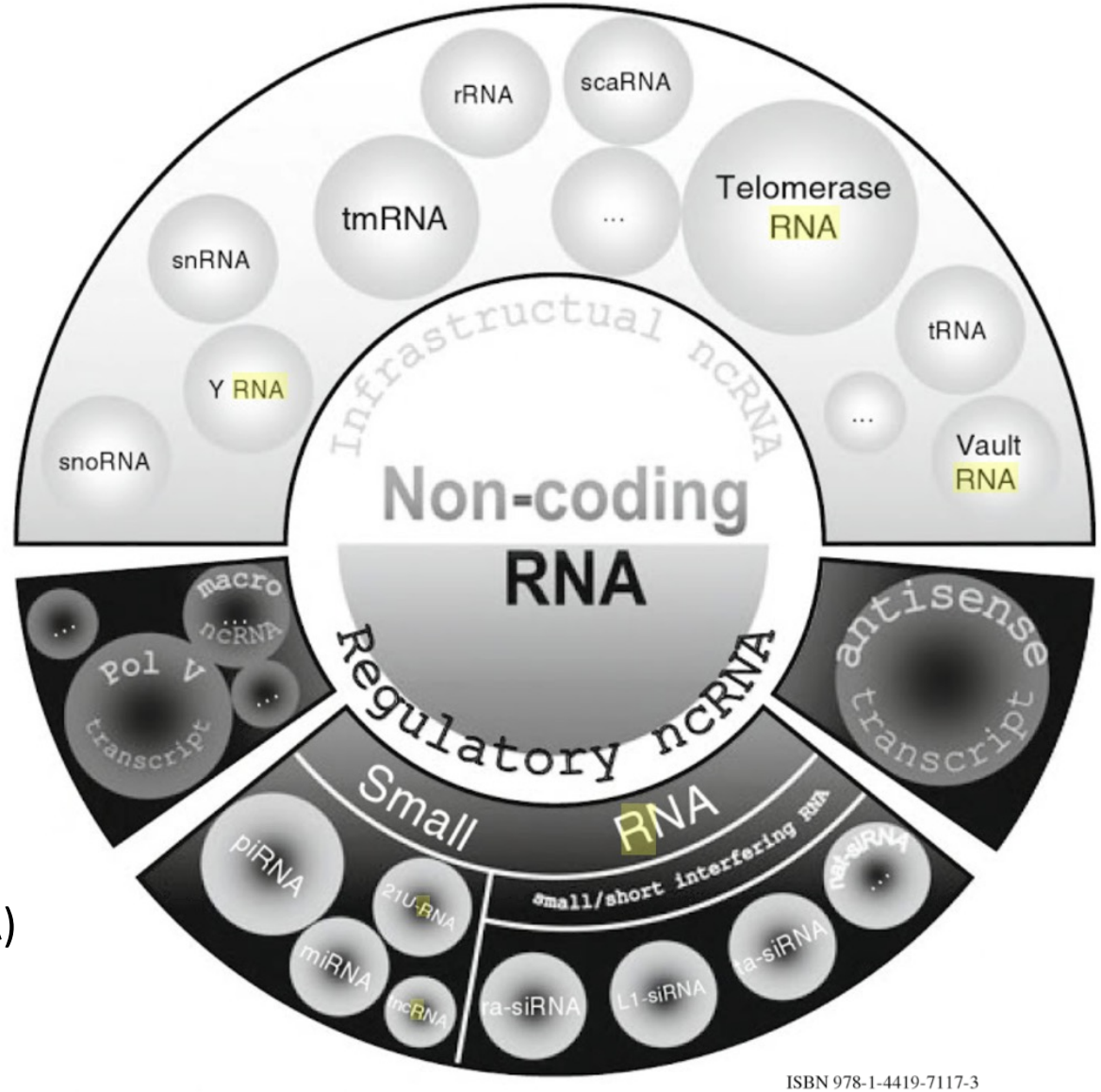
Disadvantages: Isoforms difficult to differentiate. More sensitive than approaches which could confirm results.





Non-coding RNAs

- Long non-coding RNAs (lncRNA)
  - Many sub-types
  - Functions vary to unknown
  - Some may even be translated
    . . . but not seemingly functional???
- Small RNAs (smRNA)
  - Silencing of transcripts
    - microRNA (miRNA)
    - Piwi-interacting RNA (piRNA)
    - Small/short interfering RNA (siRNA)
    - Trans-acting small interfering RNA (tasiRNA)
    - Phased secondary small interfering RNA (phasiRNA)
  - Splicing and post-transcriptional modification
    - Small nuclear RNA (snRNA)
    - Small nucleolar RNA (snoRNA)



ISBN 978-1-4419-7117-3
DOI 10.1007/978-1-4419-7118-0

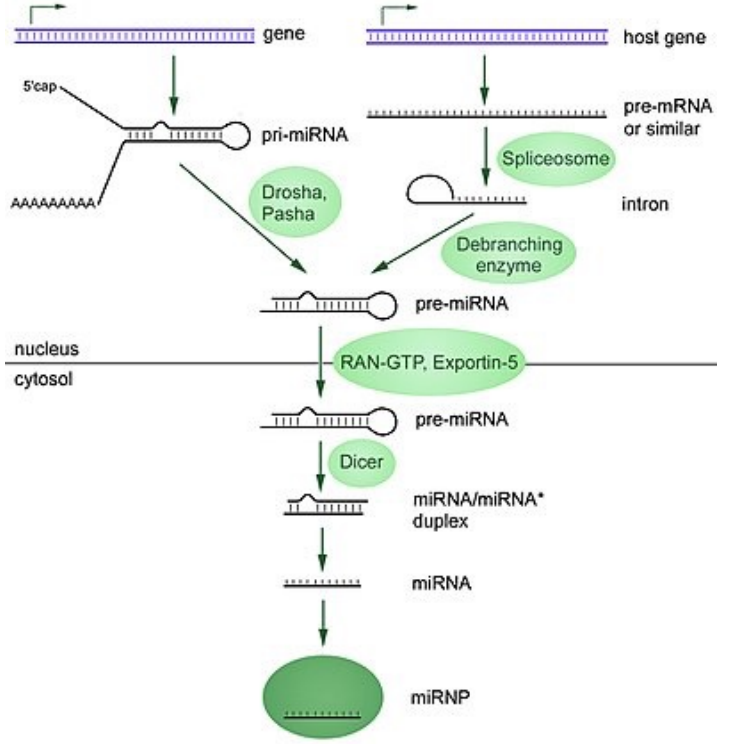
Transcriptomics (smRNA-Seq)

Objective: Determine differentially expressed smRNA.

Approach: Select for small RNA. Sequence fragments. Keep only fragments that are short (reverse adapter is identified). Identify unique sequences. Determine differentially expressed smRNA.

Advantages: Most sensitive global profiling for smRNA (miRNA, piRNA, etc.)

Disadvantages: More challenging (but possible) to characterize novel smRNA.



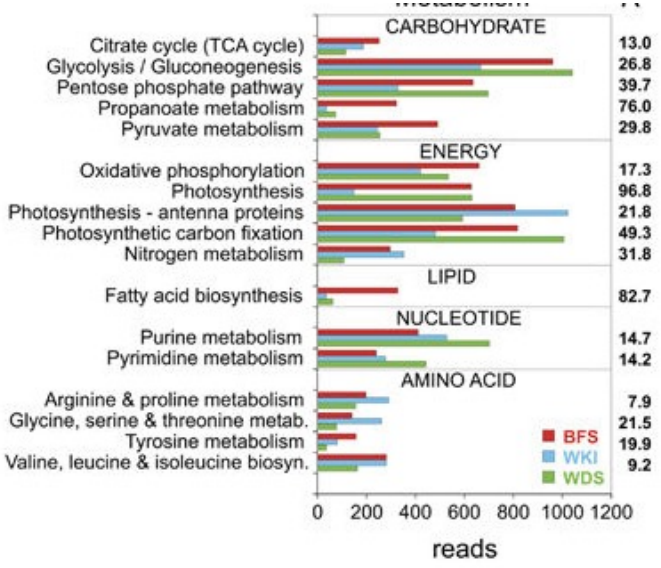
Metatranscriptomes (MetaT)

Objective: Identify active genes in microbial communities

Approach: Isolate RNA. Deplete rRNA, so you have mostly mRNA. Sequence. Map to reference (MetaT assembly, MetaG assembly, reference genomes)

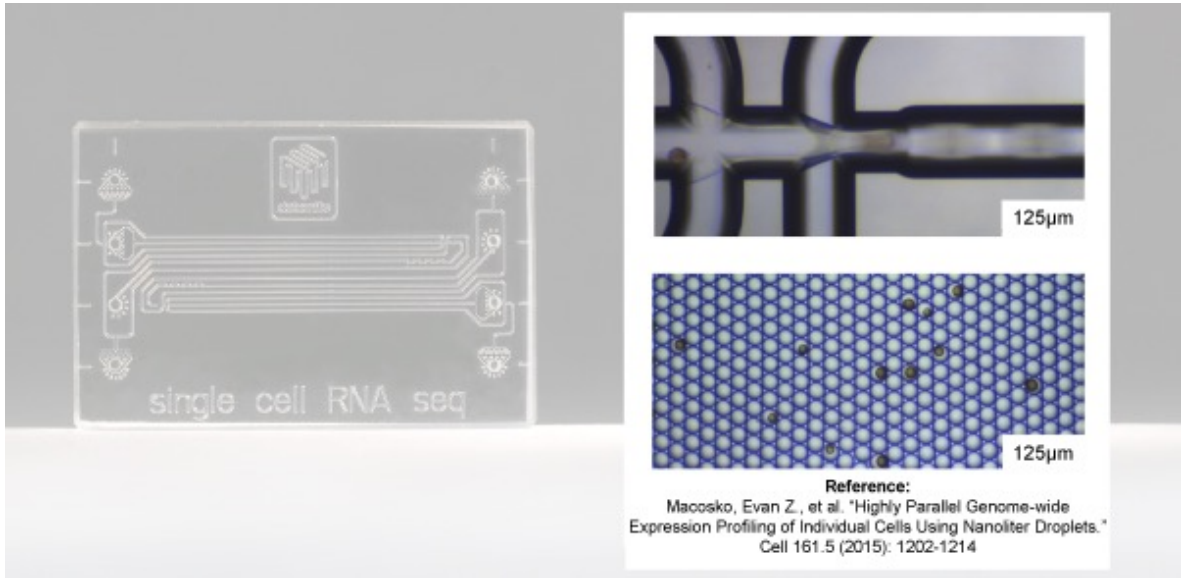
Advantages: Let's you see what microbial genes are active in a community.

Disadvantages: Very challenging to analyze. No standardized workflows. Differential expression inexact.

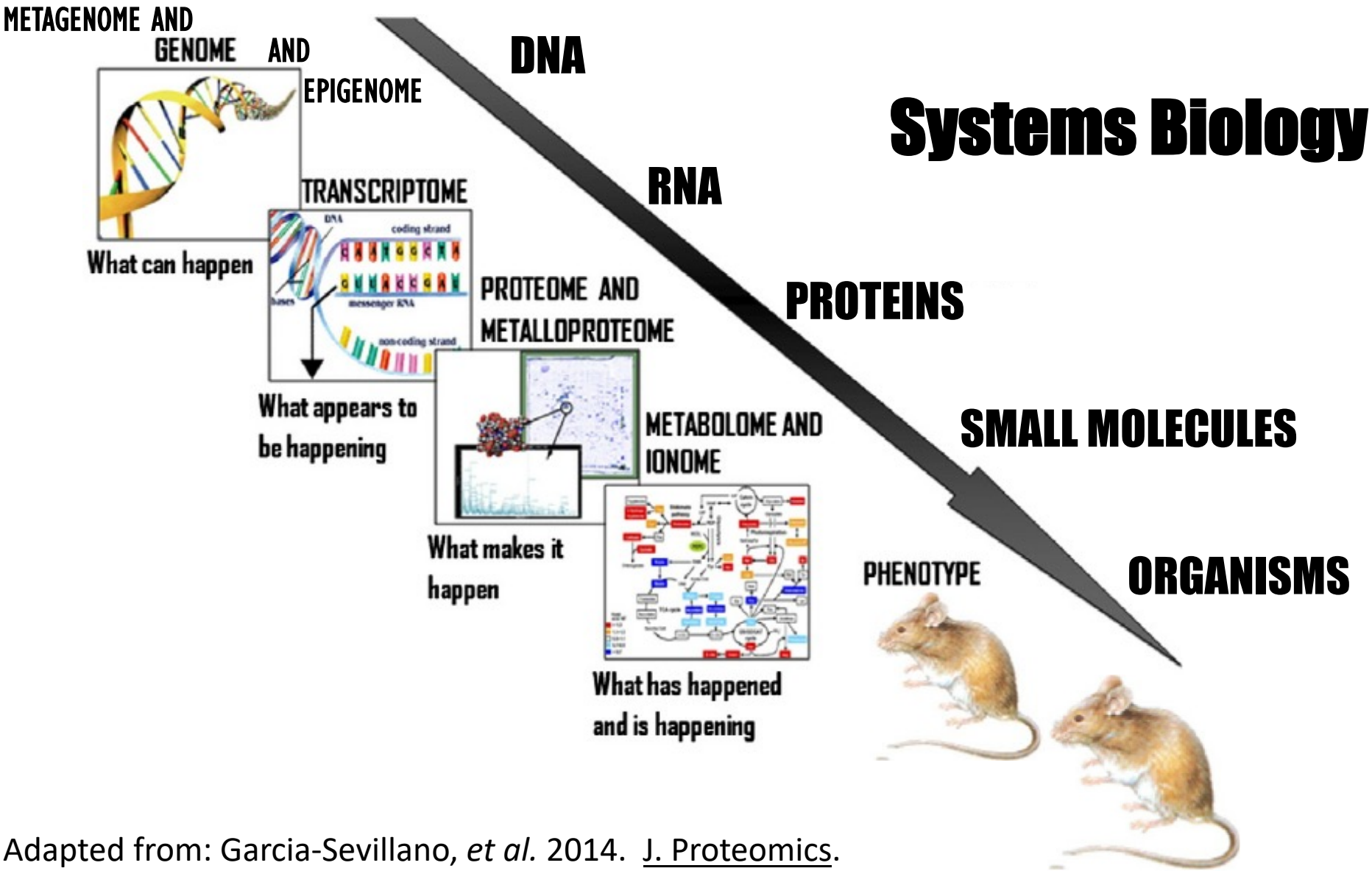


(Single Cell) scRNA-Seq

- Emerging technologies are using nano-fluidics to sort single cells into droplets
- Each droplet can then be barcoded and prepped for RNA-Seq
- Allows individual cells to be characterized…gets around bulk tissue sample issues
- Can also be used for ATAC-Seq…sometimes done simultaneously with RNA-Seq



Beyond Nucleic Acids



Adapted from: Garcia-Sevillano, *et al.* 2014. J. Proteomics.

Proteomics (Mass Spectrometry)

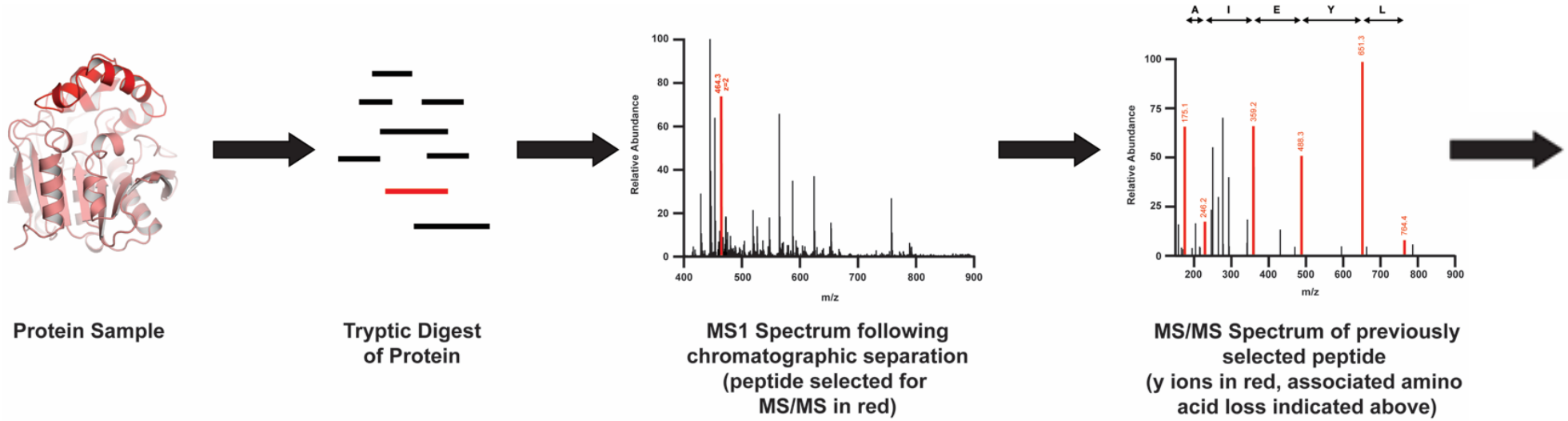
Objective: Identify proteins that are present in a sample.

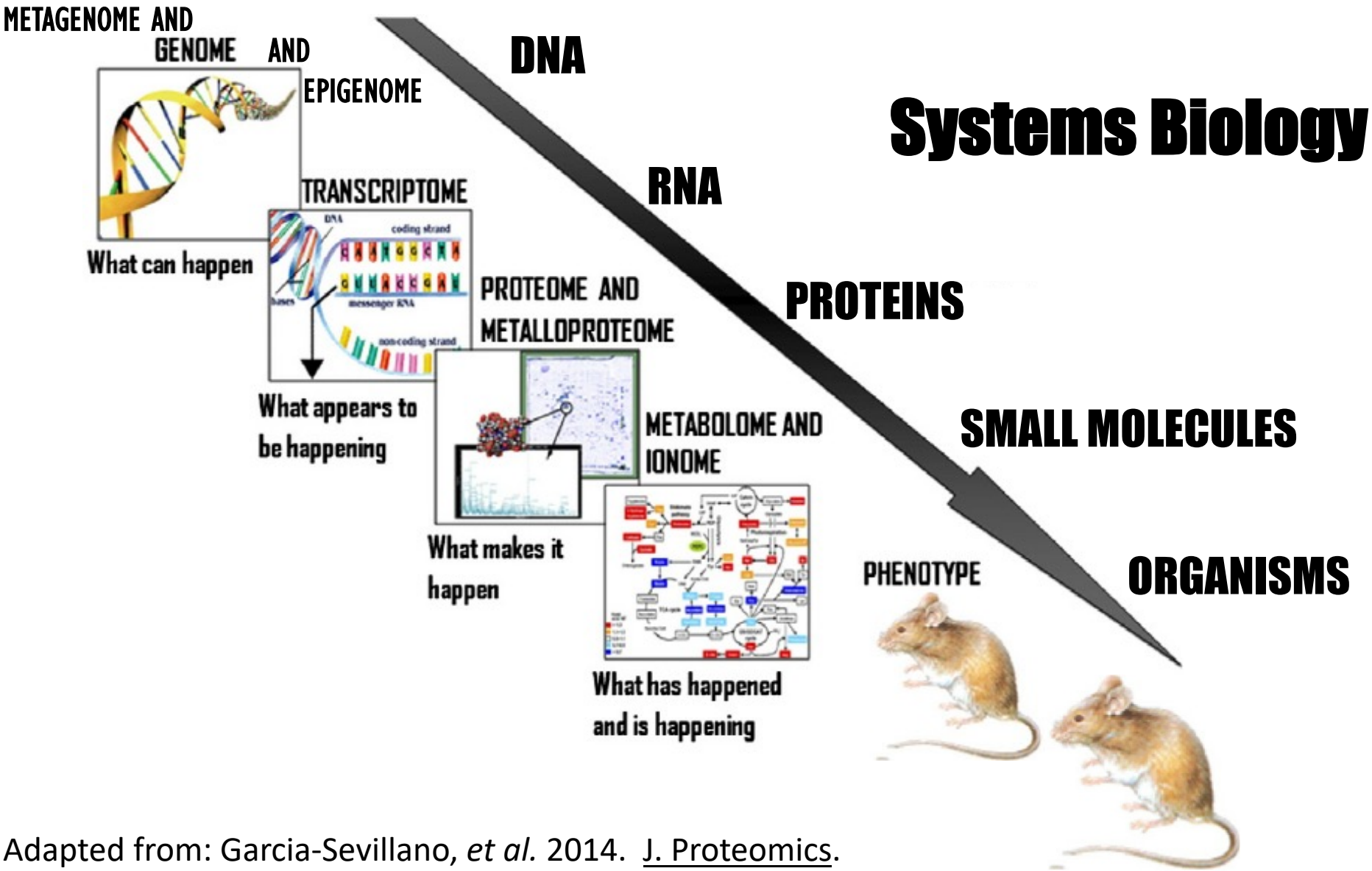
Approach: Fragment proteins. Use mass spectrometric techniques to identify protein fragments by mass.

Advantages: Provides amino acid sequence data for protein fragments.

Disadvantages: Does not sequence full proteins (only fragments). Often only the most common proteins in a sample may be differentiated. Database-dependent.

Derivatives: label-free vs tagged, Post translational modification (e.g. phosphoproteomics), metalloproteomics, metaproteomics





Adapted from: Garcia-Sevillano, *et al.* 2014. J. Proteomics.

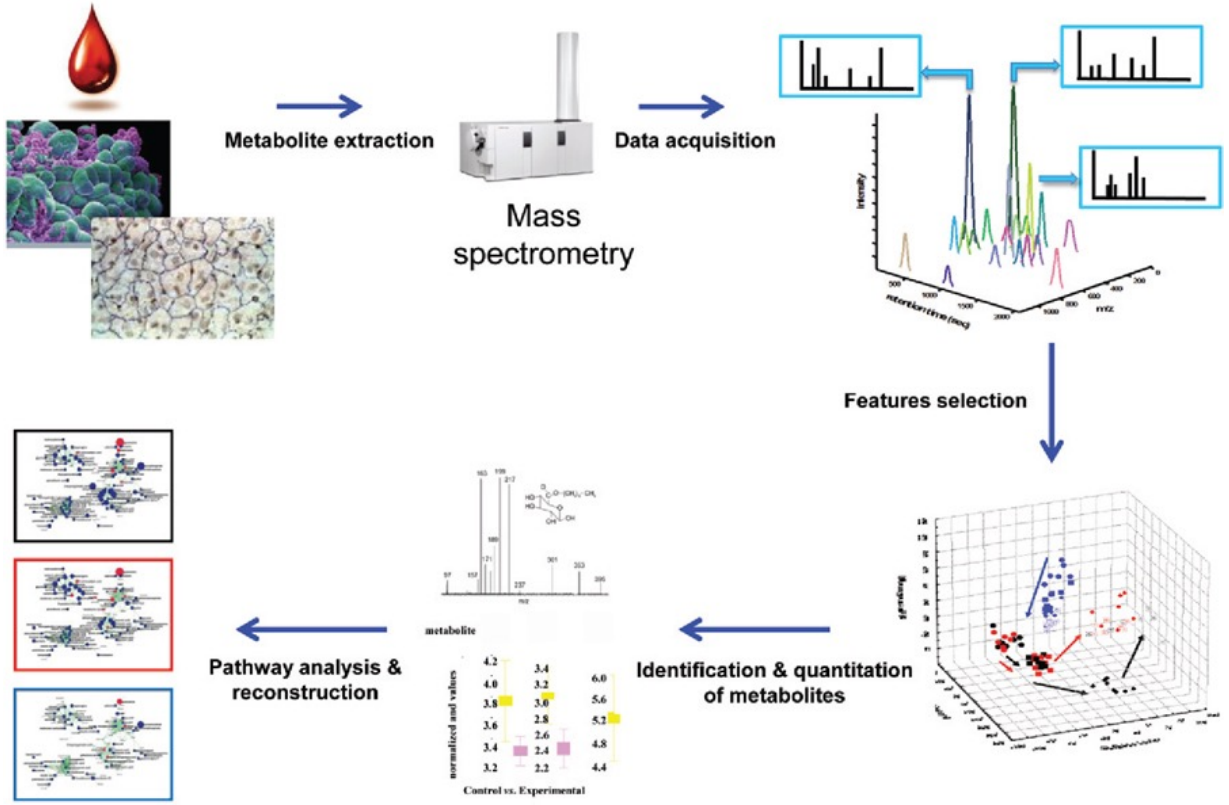
Metabolomics/Lipodomics

Objective: Identify small molecules (metabolites) in a sample.

Approach: Use chromatography/mass spectrometric techniques to separate and identify small molecules by mass. Combine with genomic, transcriptomic data to model metabolic flux.

Advantages: Ability to characterize what a cell/tissue is actually doing.

Disadvantages: Sensitivity can be an issue. Metabolic flux modeling can be tricky in organisms without well characterized genomic annotations.



How do we tie this all together?

Multi-omics Data Integration is the Frontier of Bioinformatics…

Fluxomics

Network Analysis, AI/Machine Learning