CPBS 7711 Course Review

Presented by Evan Stene

Topics Covered

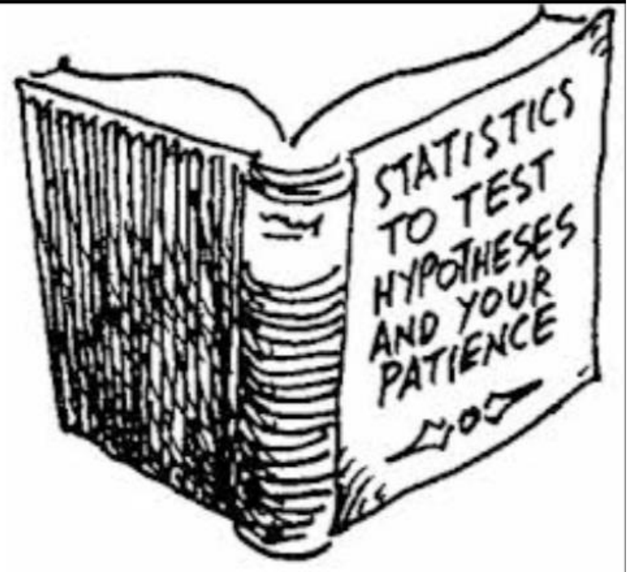
- Databases
- Genetics
- Machine Learning
- Protein Folding
- Phylogenetic Trees
- Visualization
- And More…

MolBio Databases

- Biological data is very heterogeneous
- Data is scattered among thousands of databases
- Much of the data stored in redundant

- Slides in this section from: Dr. Robin Dowell

Some statistics



- More than 1000 different bio-databases
- Generally accessible through the web

(useful link: www.expasy.ch/alinks.html)

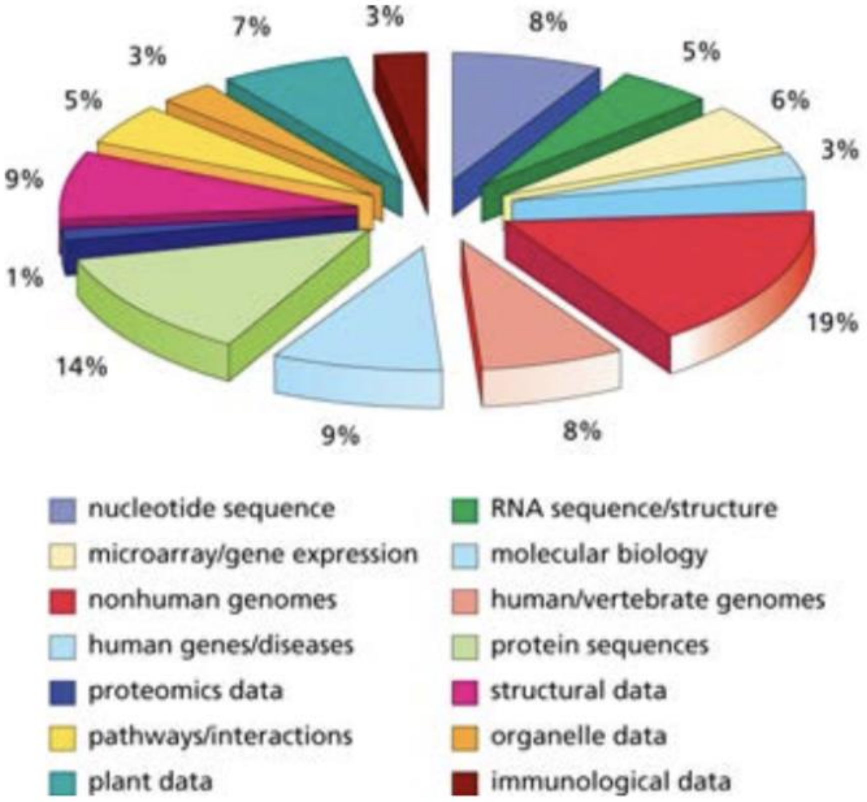
- Variable size: <100Kb to >10Gb
  - DNA: > 10 Gb
  - Protein: 1 Gb
  - 3D structure: 5 Gb
  - Other: smaller

- Update frequency: daily to annually

NAR Database Issue

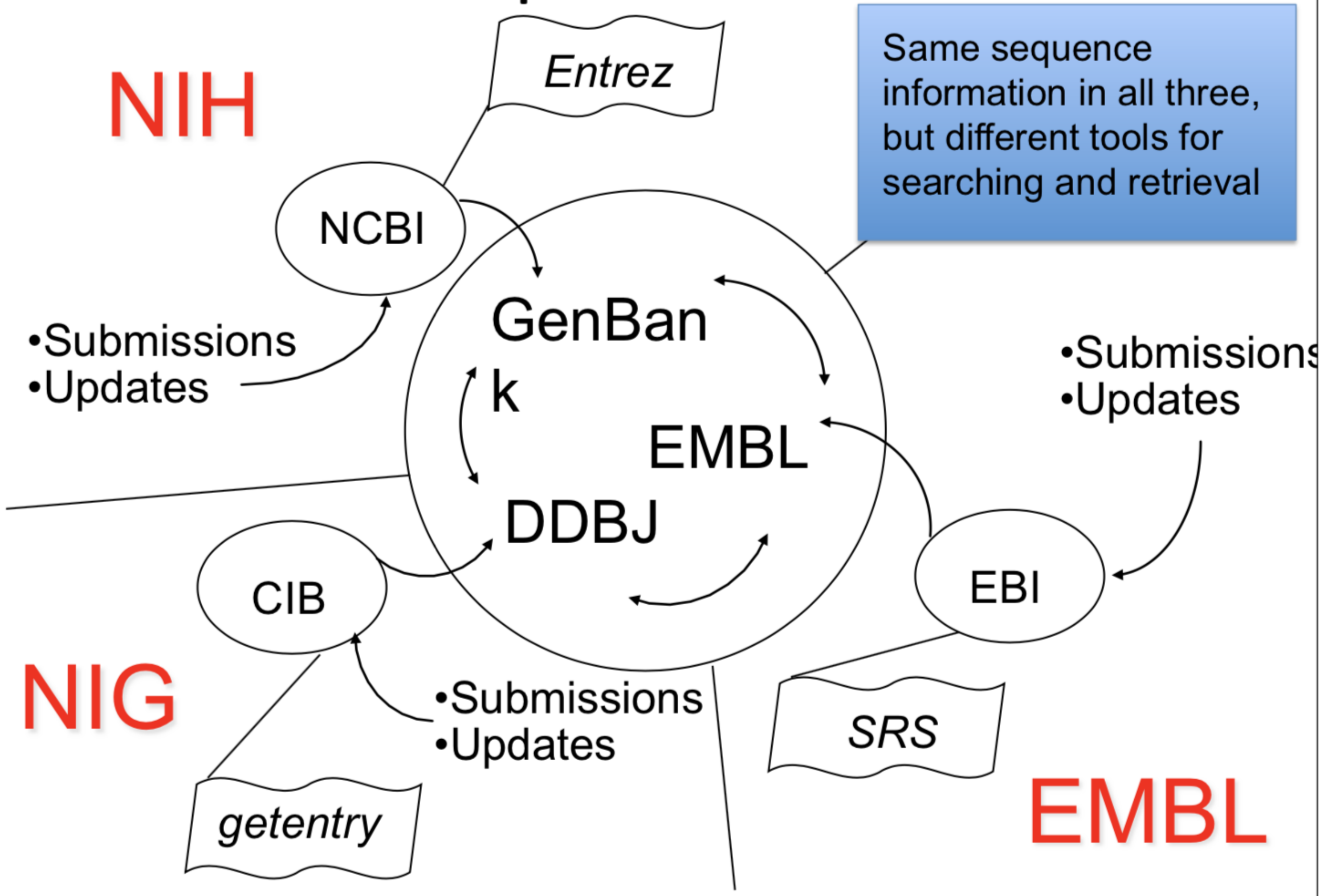
- Online collection of biological databases:

http://www.oxfordjournals.org/nar/database/c/



Public Sequence Databases

Same sequence information in all three, but different tools for searching and retrieval

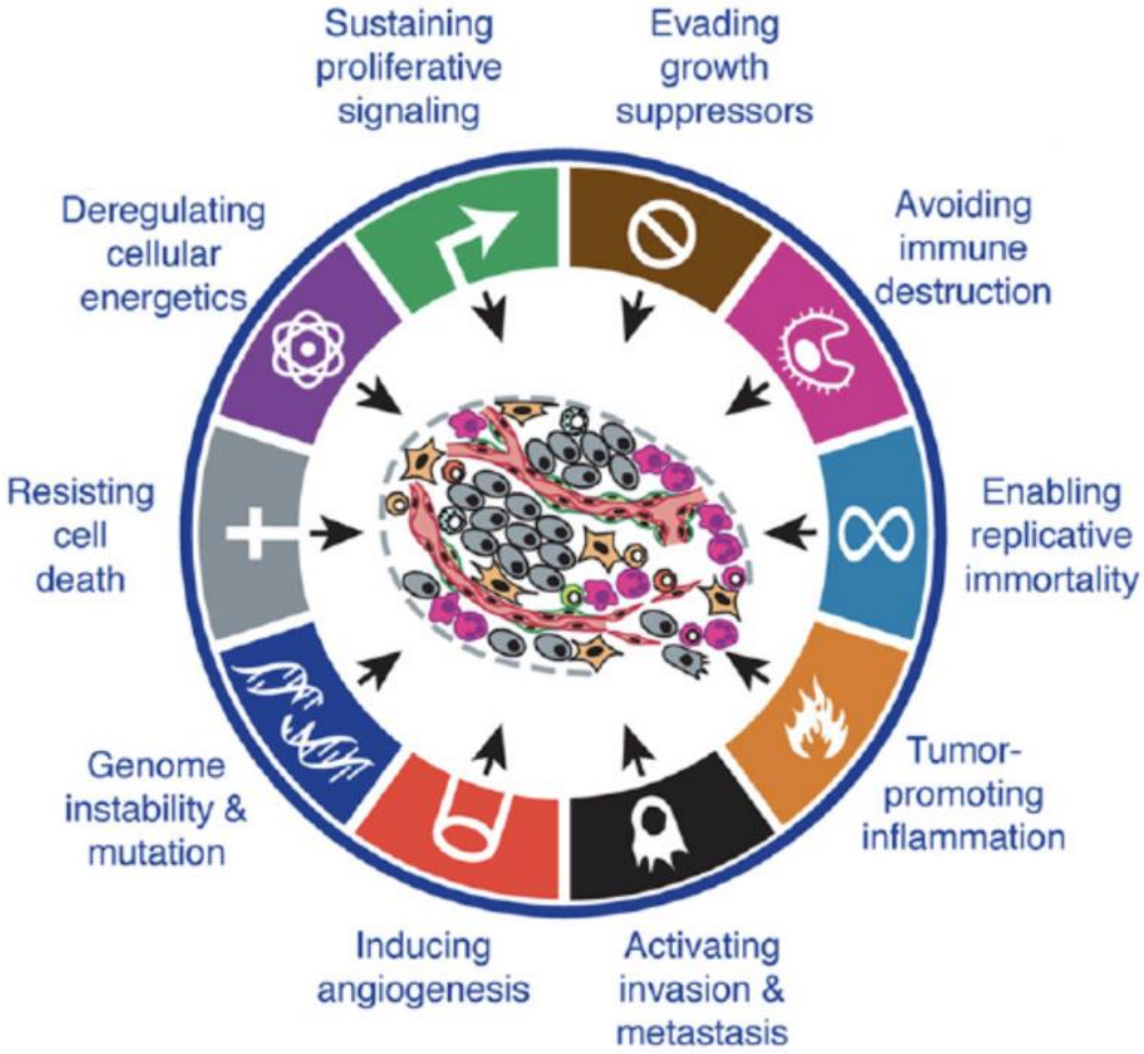


Cancer Genomics

- Humans are adept at survival
    - Errors (mutations) in DNA are repaired (usually)
    - Redundant systems take over when errors occur
    - Cancerous cells are usually killed before becoming damaging
- Cancer is a complex disease
    - Many mutations must accumulate to cause cancer
    - There is no single way to treat cancer

- Slides in this section from Dr. James Costello

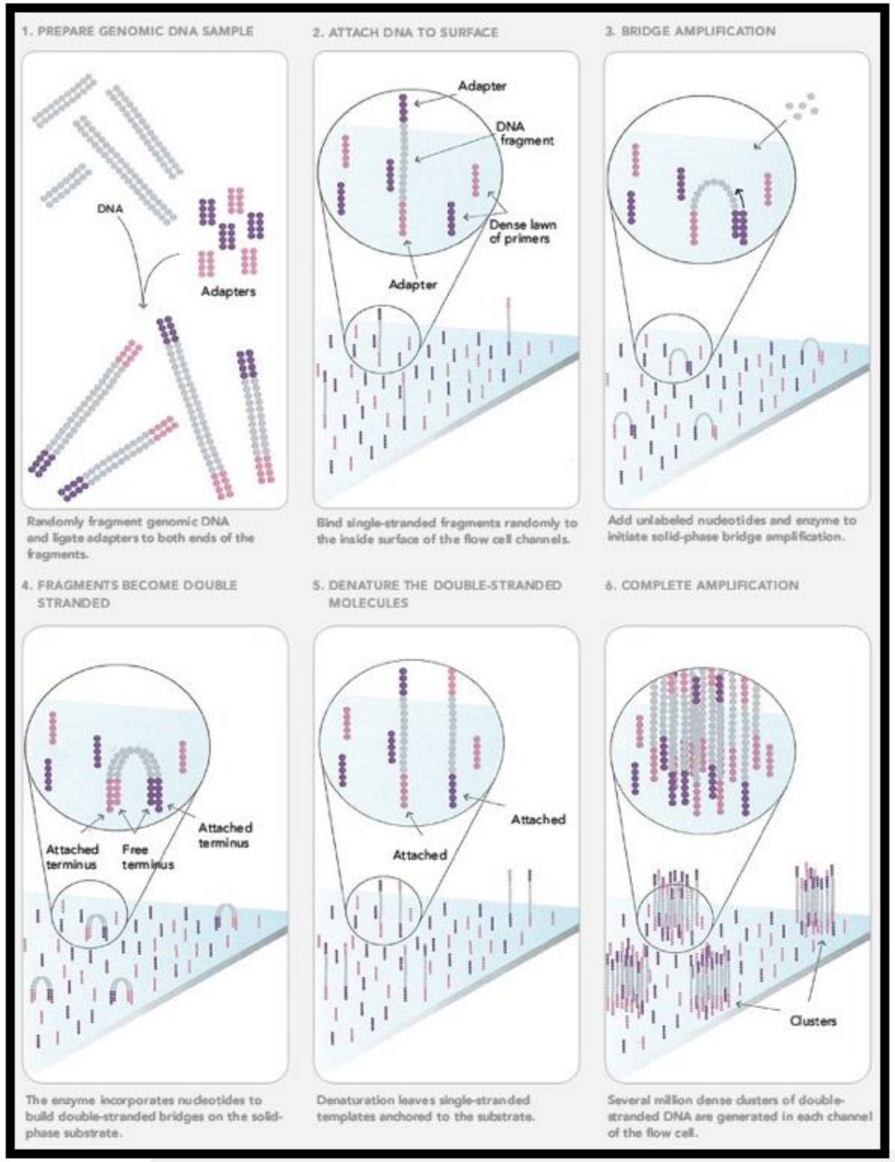
Hallmarks of Cancer



Sustaining proliferative signaling
Evading growth suppressors
Avoiding immune destruction
Enabling replicative immortality
Tumor-promoting inflammation
Activating invasion & metastasis
Inducing angiogenesis
Genome instability & mutation
Resisting cell death
Deregulating cellular energetics

Hanahan and Weinberg, Cell. 2011

Illumina Sequencing



1. PREPARE GENOMIC DNA SAMPLE

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

2. ATTACH DNA TO SURFACE

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

3. BRIDGE AMPLIFICATION

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

4. FRAGMENTS BECOME DOUBLE STRANDED

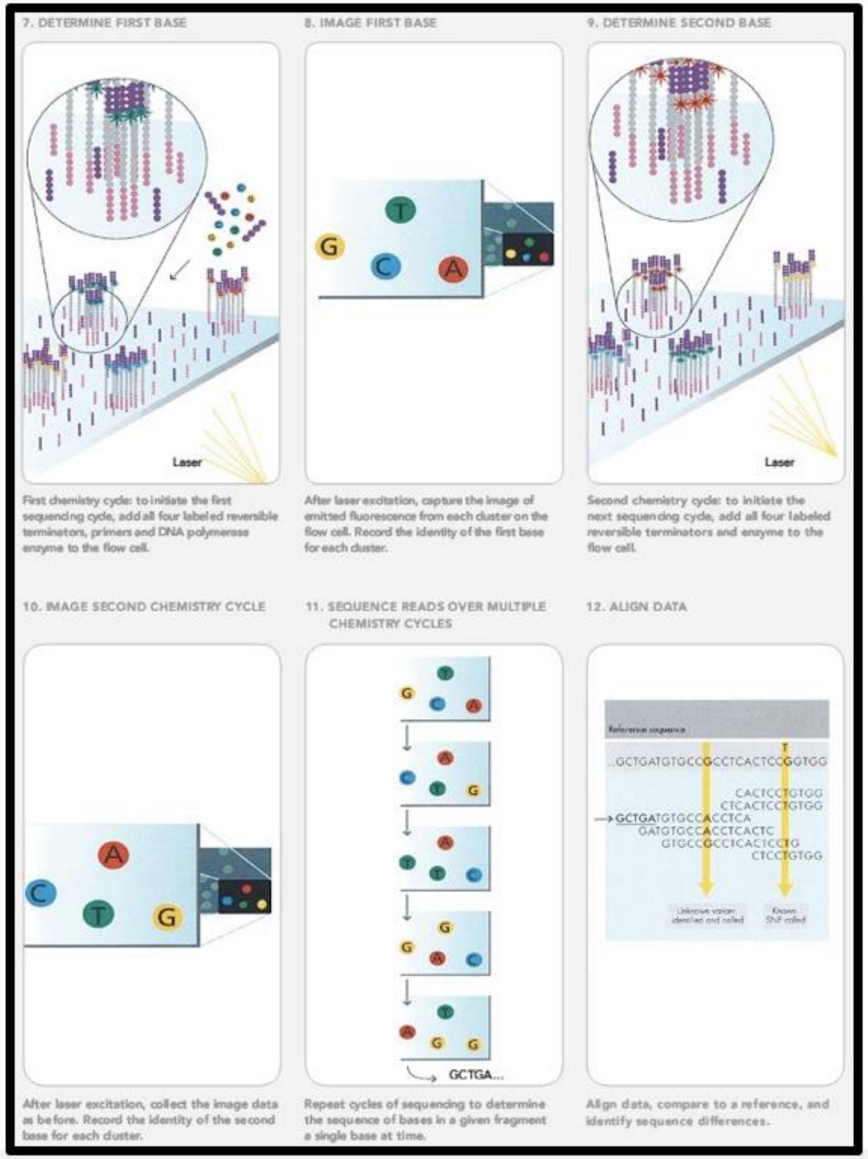
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES

Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION

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



7. DETERMINE FIRST BASE

First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE

Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

10. IMAGE SECOND CHEMISTRY CYCLE

After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

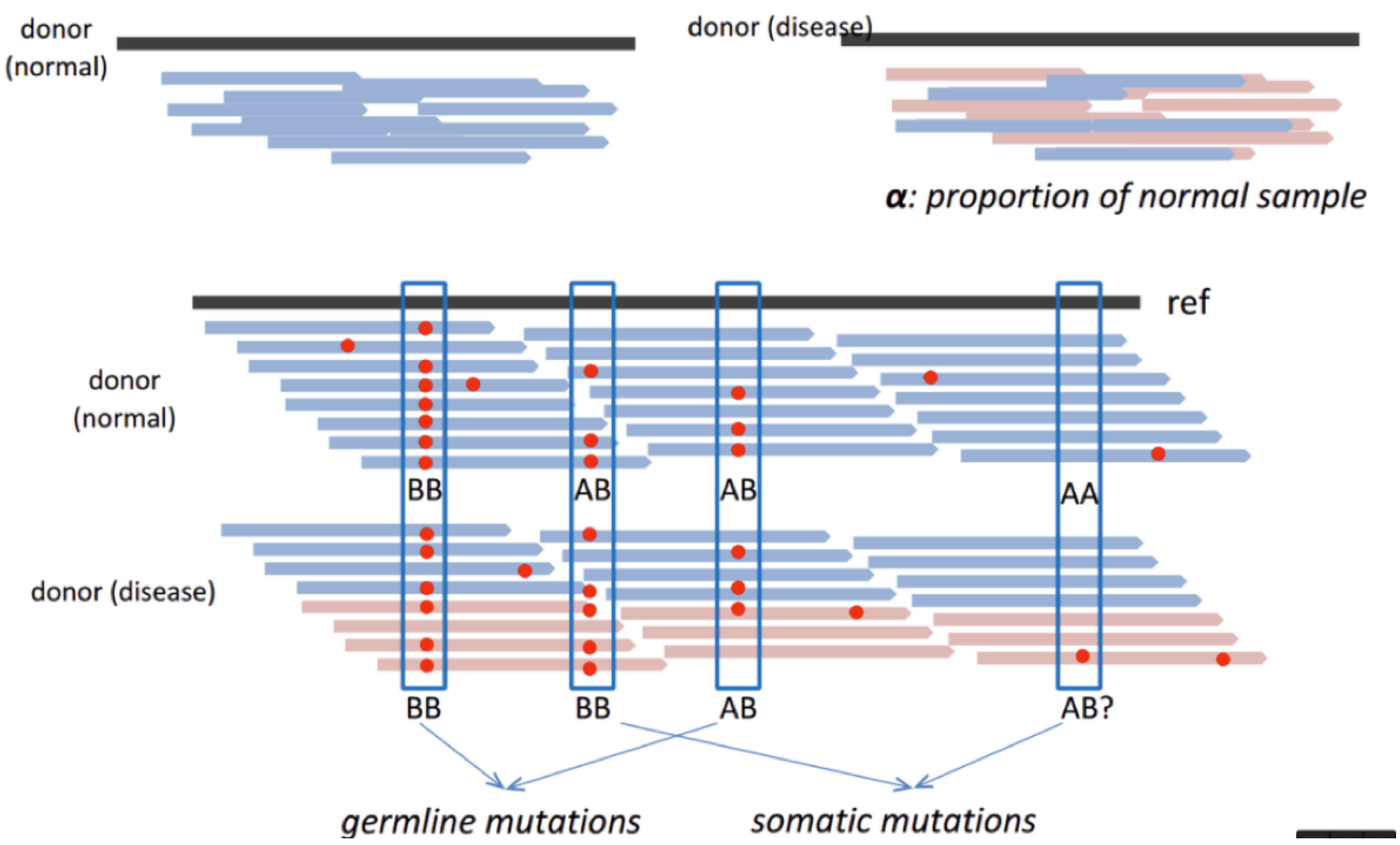
11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

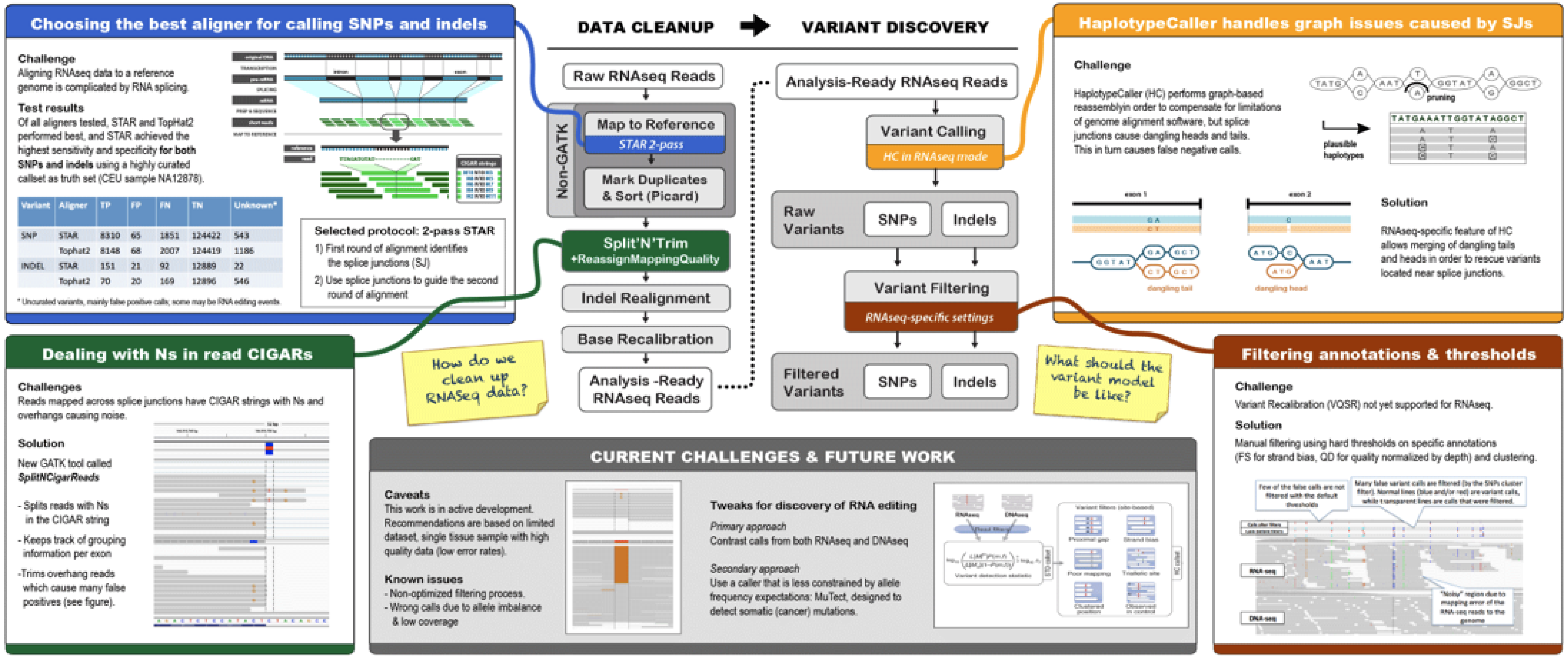
12. ALIGN DATA

Align data, compare to a reference, and identify sequence differences.

Mutation Calling



Mutation Calling Pipeline



Choosing the best aligner for calling SNPs and indels

Challenge
Aligning RNAseq data to a reference genome is complicated by RNA splicing.

Test results
Of all aligners tested, STAR and TopHat2 performed best, and STAR achieved the highest sensitivity and specificity **for both SNPs and indels** using a highly curated callset as truth set (CEU sample NA12878).

Variant	Aligner	TP	FP	FN	TN	Unknown*
SNP	STAR	8310	65	1851	124422	543
	Tophat2	8148	68	2007	124419	1186
INDEL	STAR	151	21	92	12889	22
	Tophat2	70	20	169	12896	546

* Uncurated variants, mainly false positive calls; some may be RNA editing events.

Selected protocol: 2-pass STAR
1) First round of alignment identifies the splice junctions (SJ)
2) Use splice junctions to guide the second round of alignment

Dealing with Ns in read CIGARs

Challenges
Reads mapped across splice junctions have CIGAR strings with Ns and overhangs causing noise.

Solution
New GATK tool called *SplitNCigarReads*

- Splits reads with Ns in the CIGAR string
- Keeps track of grouping information per exon
- Trims overhang reads which cause many false positives (see figure).

DATA CLEANUP → VARIANT DISCOVERY

Raw RNAseq Reads
Non-GATK: Map to Reference (*STAR 2-pass*); Mark Duplicates & Sort (Picard)
Split'N'Trim +ReassignMappingQuality
Indel Realignment
Base Recalibration
Analysis -Ready RNAseq Reads

Analysis-Ready RNAseq Reads
Variant Calling (*HC in RNAseq mode*)
Raw Variants: SNPs, Indels
Variant Filtering (*RNAseq-specific settings*)
Filtered Variants: SNPs, Indels

How do we clean up RNAseq data?

What should the variant model be like?

HaplotypeCaller handles graph issues caused by SJs

Challenge

HaplotypeCaller (HC) performs graph-based reassemblyin order to compensate for limitations of genome alignment software, but splice junctions cause dangling heads and tails. This in turn causes false negative calls.

Solution

RNAseq-specific feature of HC allows merging of dangling tails and heads in order to rescue variants located near splice junctions.

Filtering annotations & thresholds

Challenge
Variant Recalibration (VQSR) not yet supported for RNAseq.

Solution
Manual filtering using hard thresholds on specific annotations (FS for strand bias, QD for quality normalized by depth) and clustering.

Few of the false calls are not filtered with the default thresholds
Many false variant calls are filtered (by the SNPs cluster filter). Normal lines (blue and/or red) are variant calls, while transparent lines are calls that were filtered.
"Noisy" region due to mapping error of the RNA-seq reads to the genome
RNA-seq
DNA-seq

CURRENT CHALLENGES & FUTURE WORK

Caveats
This work is in active development. Recommendations are based on limited dataset, single tissue sample with high quality data (low error rates).

Known issues
- Non-optimized filtering process.
- Wrong calls due to allele imbalance & low coverage

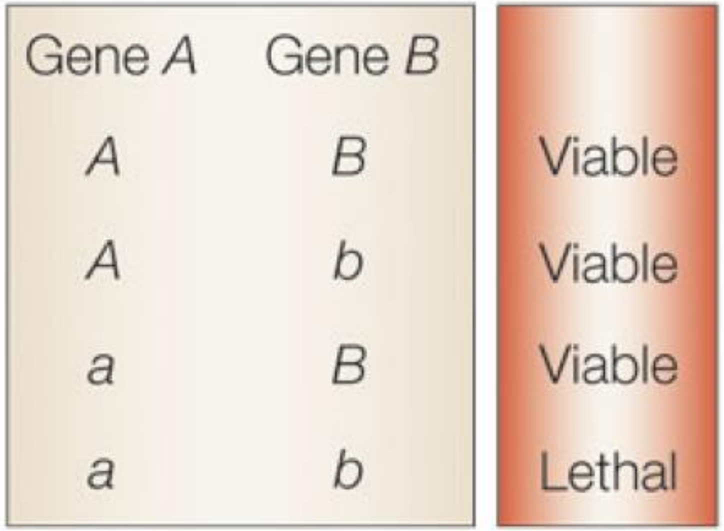
Tweaks for discovery of RNA editing

Primary approach
Contrast calls from both RNAseq and DNAseq

Secondary approach
Use a caller that is less constrained by allele frequency expectations: MuTect, designed to detect somatic (cancer) mutations.

GATK Best Practices

Synthetic Lethality



Gene *A*	Gene *B*	
A	*B*	Viable
A	*b*	Viable
a	*B*	Viable
a	*b*	Lethal

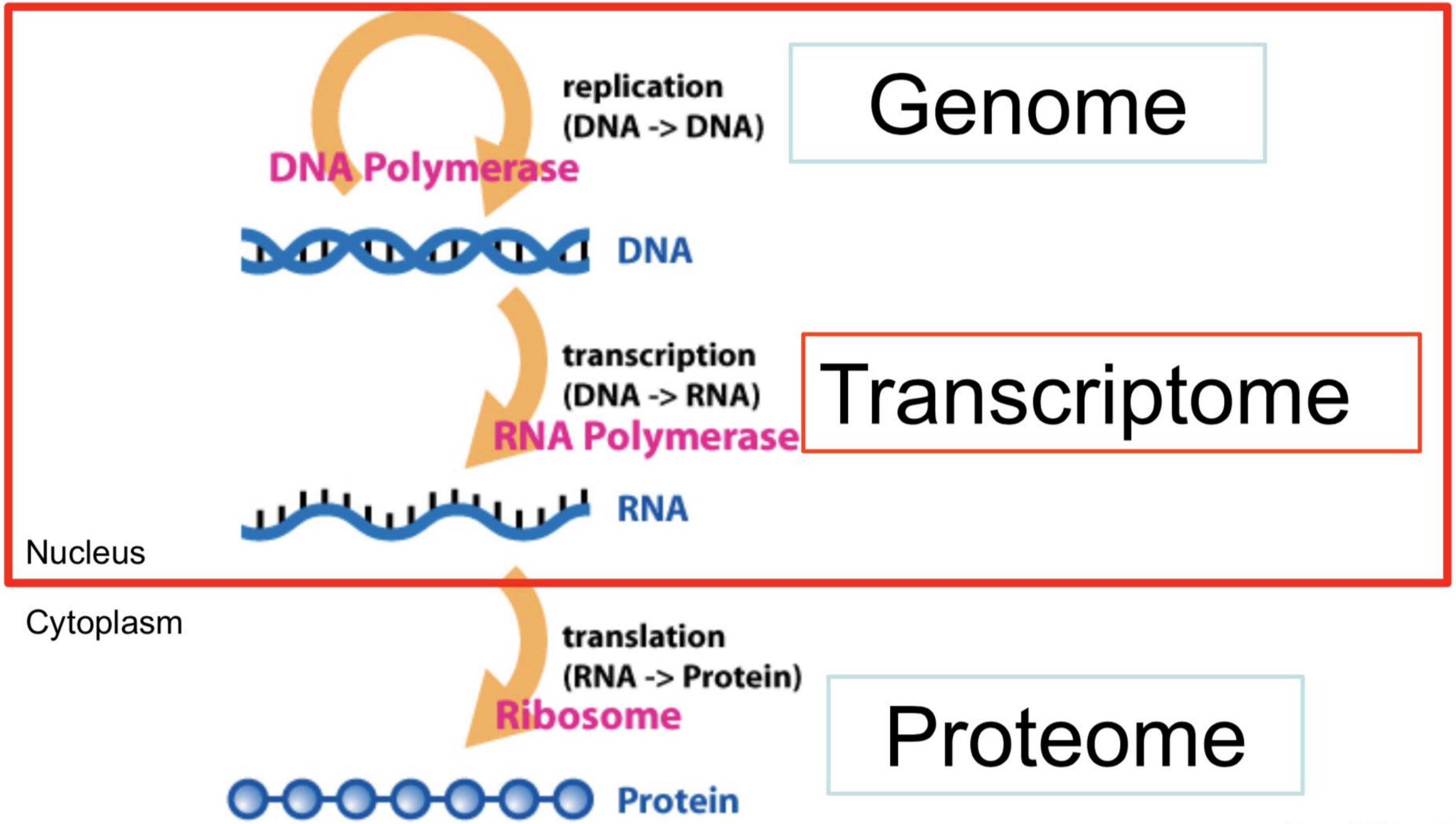
Kaelin (2005) Nature Reviews Cancer. 5:689-98.

Gene Expression Analysis

- Studying the transcriptome
  - Parts of the DNA that act like blueprints
  - Why cells with the same DNA can be entirely different

- Slides in this section from Dr. Aik Choon Tan

The Central Dogma of Molecular Biology



replication (DNA -> DNA)

DNA Polymerase

Genome

DNA

transcription (DNA -> RNA)

RNA Polymerase

Transcriptome

RNA

Nucleus

Cytoplasm

translation (RNA -> Protein)

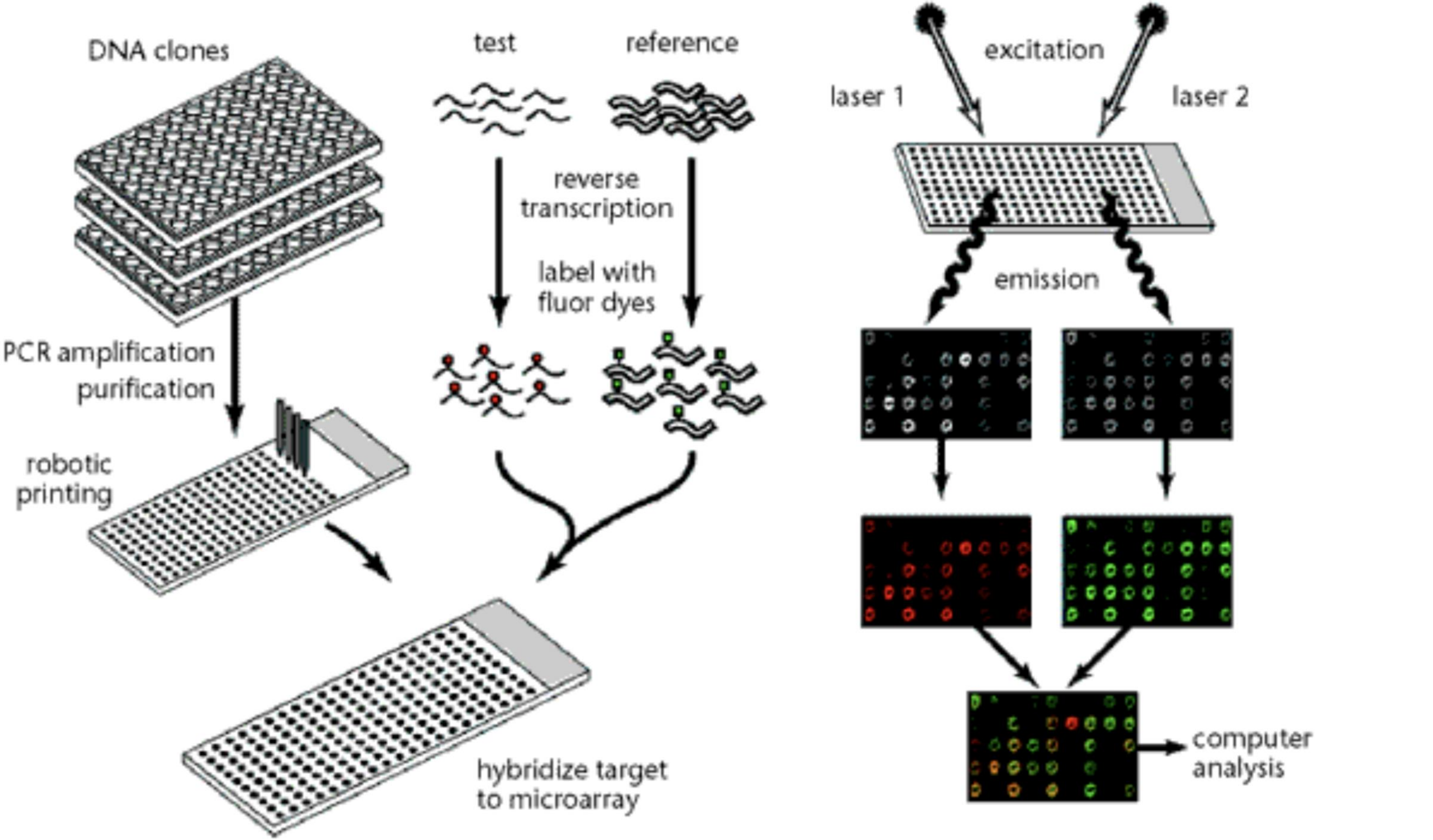
Ribosome

Proteome

Protein

(From Wikipedia)

cDNA microarray schema



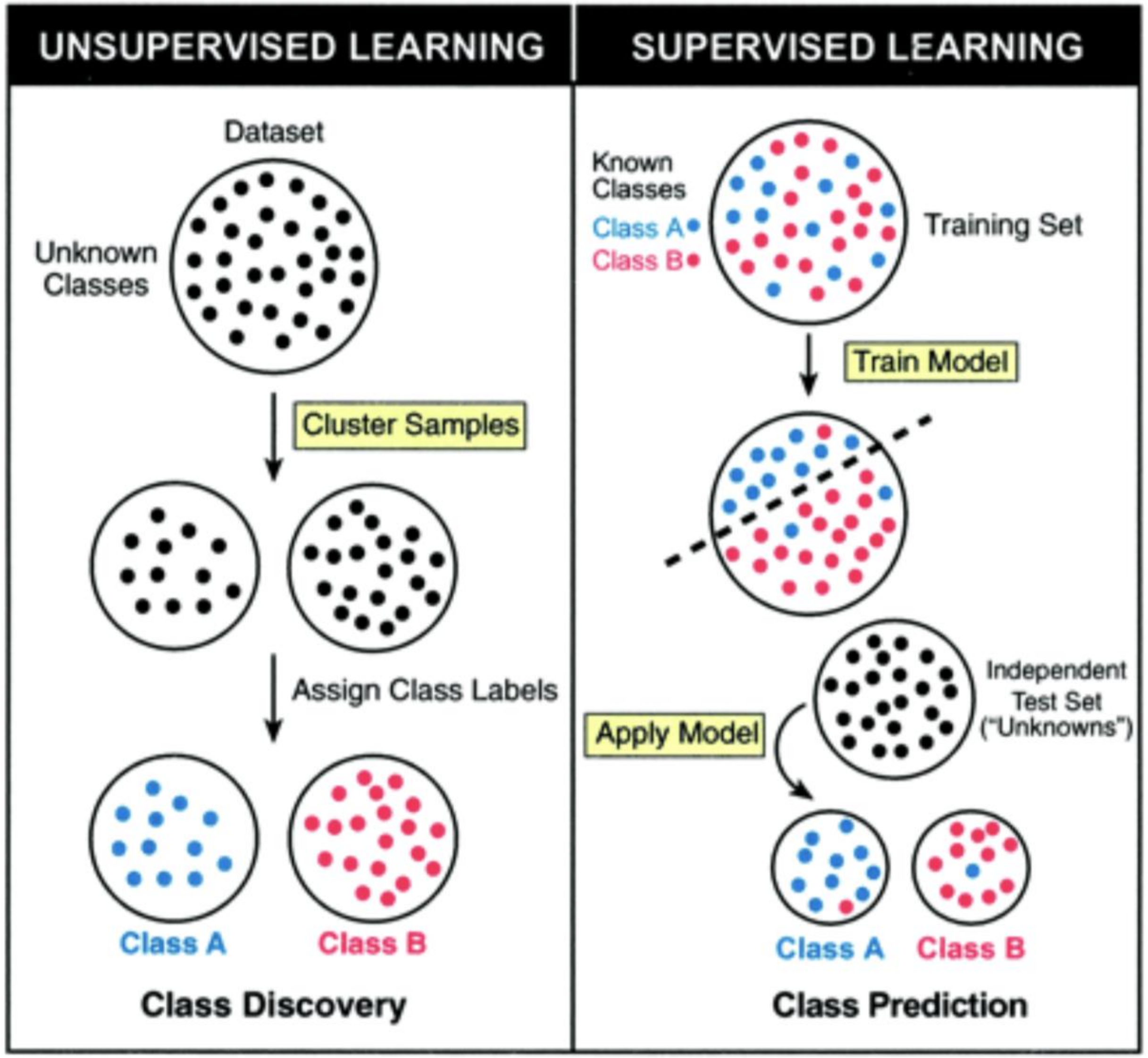
Gene Expression Profile

m samples

n genes

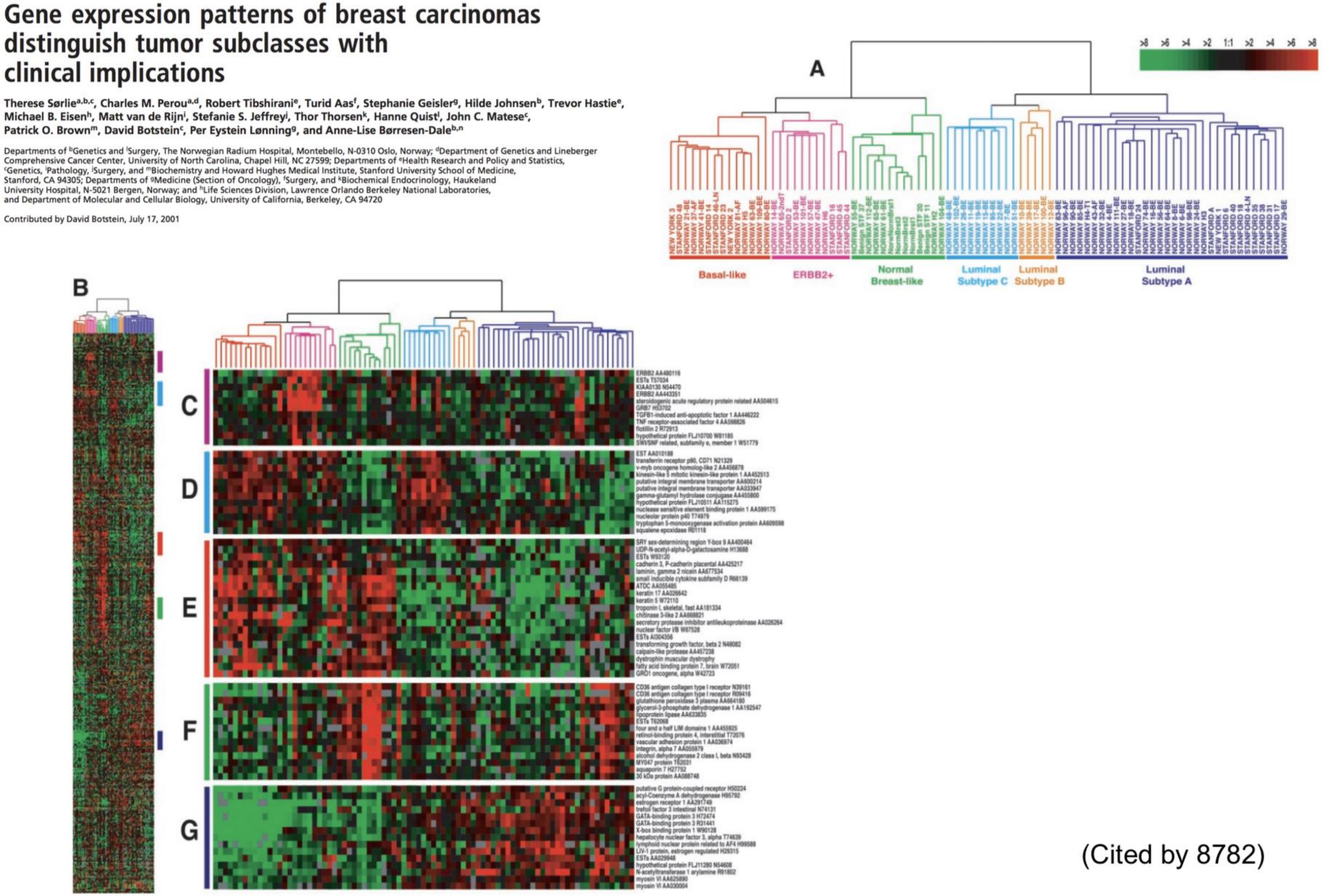
Geneid	Condition 1	Condition 2	…	Condition *m*
Gene1	103.02	58.79	…	101.54
Gene2	40.55	1246.87	…	1432.12
…	…	…	…	…
Gene *n*	78.13	66.25	…	823.09

Gene expression data analysis



(Ramaswamy and Golub 2002)

Gene Expression Clustering



Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications

Therese Sørlie[a,b,c], Charles M. Perou[a,d], Robert Tibshirani[e], Turid Aas[f], Stephanie Geisler[g], Hilde Johnsen[b], Trevor Hastie[e], Michael B. Eisen[h], Matt van de Rijn[i], Stefanie S. Jeffrey[j], Thor Thorsen[k], Hanne Quist[l], John C. Matese[c], Patrick O. Brown[m], David Botstein[c], Per Eystein Lønning[g], and Anne-Lise Børresen-Dale[b,n]

Departments of [b]Genetics and [l]Surgery, The Norwegian Radium Hospital, Montebello, N-0310 Oslo, Norway; [d]Department of Genetics and Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599; Departments of [e]Health Research and Policy and Statistics, [c]Genetics, [i]Pathology, [j]Surgery, and [m]Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305; Departments of [g]Medicine (Section of Oncology), [f]Surgery, and [k]Biochemical Endocrinology, Haukeland University Hospital, N-5021 Bergen, Norway; and [h]Life Sciences Division, Lawrence Orlando Berkeley National Laboratories, and Department of Molecular and Cellular Biology, University of California, Berkeley, CA 94720

Contributed by David Botstein, July 17, 2001

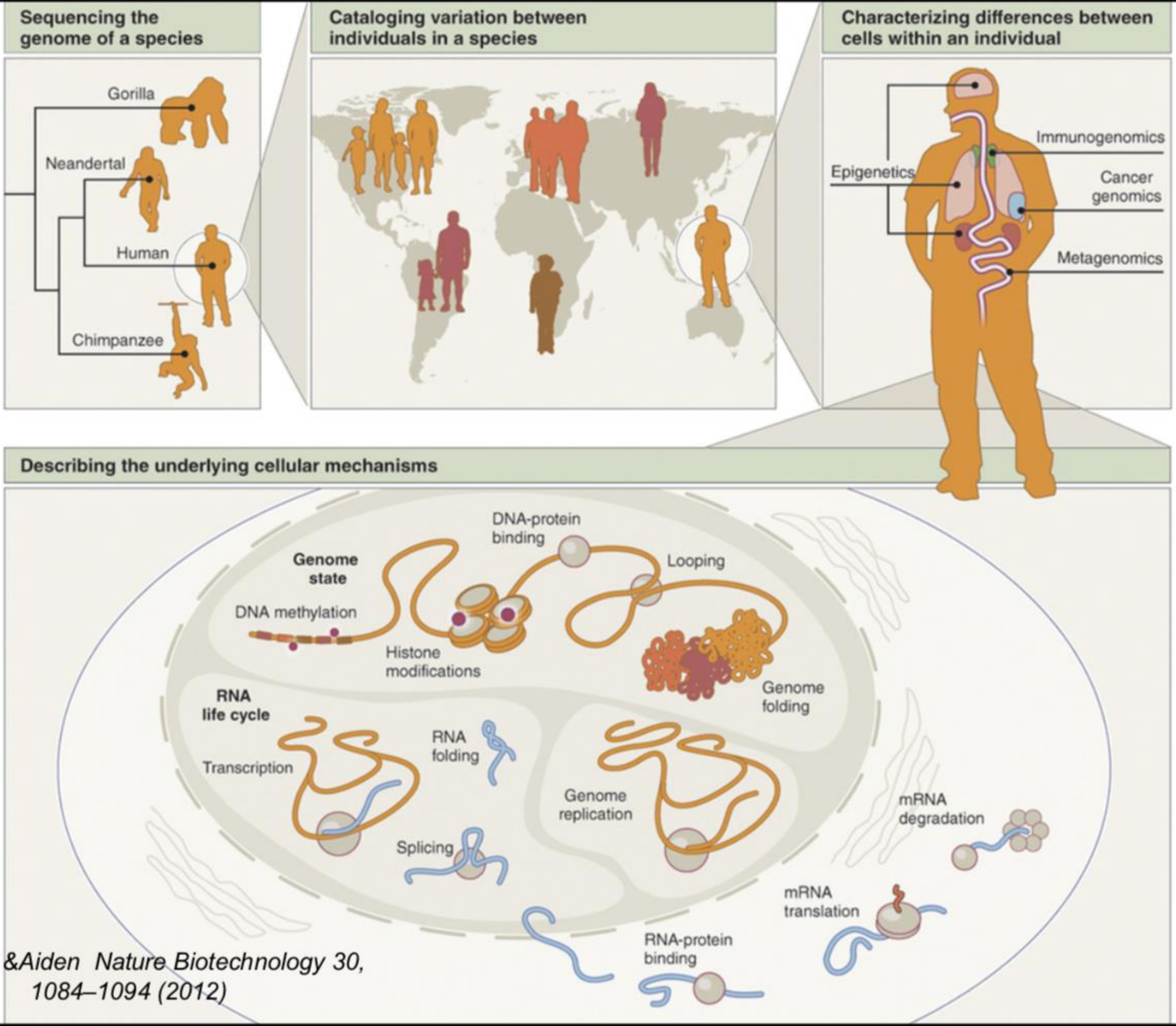
A
>8 >6 >4 >2 1:1 >2 >4 >6 >8
NEW YORK 3
STANFORD 48
NORWAY 21-BE
NORWAY 37-AF
NORWAY 41-BE
STANFORD 14
STANFORD 46-LN
STANFORD 23
NEW YORK 2
NORWAY 81-AF
NORWAY H5
NORWAY 63-BE
NORWAY 109-BE
NORWAY 80-BE
NORWAY 14-BE
NORWAY 65-2ndT
STANFORD 2
NORWAY 53-BE
NORWAY 101-BE
NORWAY 57-BE
NORWAY 47-BE
NORWAY H6
STANFORD 16
STANFORD 45
STANFORD 44
NORWAY 55-BE
Benign STF 37
NORWAY 112-BE
NORWAY 65-BE
NORWAY 61-BE
NormNormBrst1
NormBrst3
NormBrst2
NormBrst1
Benign STF 20
Benign STF 11
NORWAY H2
NORWAY 104-BE
NORWAY 48-BE
NORWAY 102-BE
NORWAY 26-BE
NORWAY 11-BE
NORWAY 19-BE
NORWAY 15-BE
NORWAY 95-BE
NORWAY 22-BE
NORWAY 7-BE
NORWAY 51-BE
NORWAY 10-BE
NORWAY 39-BE
NORWAY 17-BE
NORWAY 100-BE
NORWAY 12-BE
NORWAY 83-BE
NORWAY 96-AF
NORWAY 90-BE
NORWAY 85-BE
NORWAY H4-T1
NORWAY 43-AF
NORWAY 32-BE
NORWAY 4-BE
NORWAY 111-BE
NORWAY 27-BE
NORWAY 18-BE
STANFORD 24
NORWAY 74-BE
NORWAY 16-BE
NORWAY 56-BE
NORWAY 64-BE
NORWAY 8-BE
NORWAY 6-BE
NORWAY 98-BE
NORWAY 24-BE
NORWAY H3
STANFORD A
NEW YORK 1
STANFORD 6
STANFORD 40
STANFORD 18
STANFORD 4-LN
STANFORD 35
STANFORD 38
STANFORD 31
STANFORD 17
NORWAY 29-BE
Basal-like
ERBB2+
Normal Breast-like
Luminal Subtype C
Luminal Subtype B
Luminal Subtype A
B
C
ERBB2 AA480116
ESTs T57034
KIAA0130 N54470
ERBB2 AA443351
steroidogenic acute regulatory protein related AA504615
GRB7 H53702
TGFB1-induced anti-apoptotic factor 1 AA446222
TNF receptor-associated factor 4 AA598826
flotillin 2 R72913
hypothetical protein FLJ10700 W81185
SWI/SNF related, subfamily e, member 1 W51779
D
EST AA010188
transferrin receptor p90, CD71 N21329
v-myb oncogene homolog-like 2 AA456878
kinesin-like 5 mitotic kinesin-like protein 1 AA452513
putative integral membrane transporter AA600214
putative integral membrane transporter AA033947
gamma-glutamyl hydrolase conjugase AA455800
hypothetical protein FLJ10511 AA115275
nuclease sensitive element binding protein 1 AA599175
nucleolar protein p40 T74979
tryptophan 5-monooxygenase activation protein AA609598
squalene epoxidase R01118
E
SRY sex-determining region Y-box 9 AA400464
UDP-N-acetyl-alpha-D-galactosamine H13688
ESTs W93120
cadherin 3, P-cadherin placental AA425217
laminin, gamma 2 nicein AA677534
small inducible cytokine subfamily D R66139
ATDC AA055485
keratin 17 AA026642
keratin 5 W72110
troponin I, skeletal, fast AA181334
chitinase 3-like 2 AA668821
secretory protease inhibitor antileukoproteinase AA026264
nuclear factor VB W87528
ESTs AI304356
transforming growth factor, beta 2 N48082
calpain-like protease AA457238
dystrophin muscular dystrophy
fatty acid binding protein 7, brain W72051
GRO1 oncogene, alpha W42723
F
CD36 antigen collagen type I receptor N39161
CD36 antigen collagen type I receptor R09416
glutathione peroxidase 3 plasma AA664180
glycerol-3-phosphate dehydrogenase 1 AA192547
lipoprotein lipase AA633835
ESTs T62068
four and a half LIM domains 1 AA455925
retinol-binding protein 4, interstitial T72076
vascular adhesion protein 1 AA036974
integrin, alpha 7 AA055979
alcohol dehydrogenase 2 class I, beta N93428
MY047 protein T62031
aquaporin 7 H27752
30 kDa protein AA088748
G
putative G protein-coupled receptor H50224
acyl-Coenzyme A dehydrogenase H95792
estrogen receptor 1 AA291749
trefoil factor 3 intestinal N74131
GATA-binding protein 3 H72474
GATA-binding protein 3 R31441
X-box binding protein 1 W90128
hepatocyte nuclear factor 3, alpha T74639
lymphoid nuclear protein related to AF4 H99588
LIV-1 protein, estrogen regulated H29315
ESTs AA029948
hypothetical protein FLJ11280 N54608
N-acetyltransferase 1 arylamine R91802
myosin VI AA625890
myosin VI AA030004

(Cited by 8782)

Next Generation Sequencing (NGS)

- High throughput sequencers that produce many short sequences
- Lower accuracy (than single sequence Sanger sequencing)
- Requires processing data before use in analysis pipelines

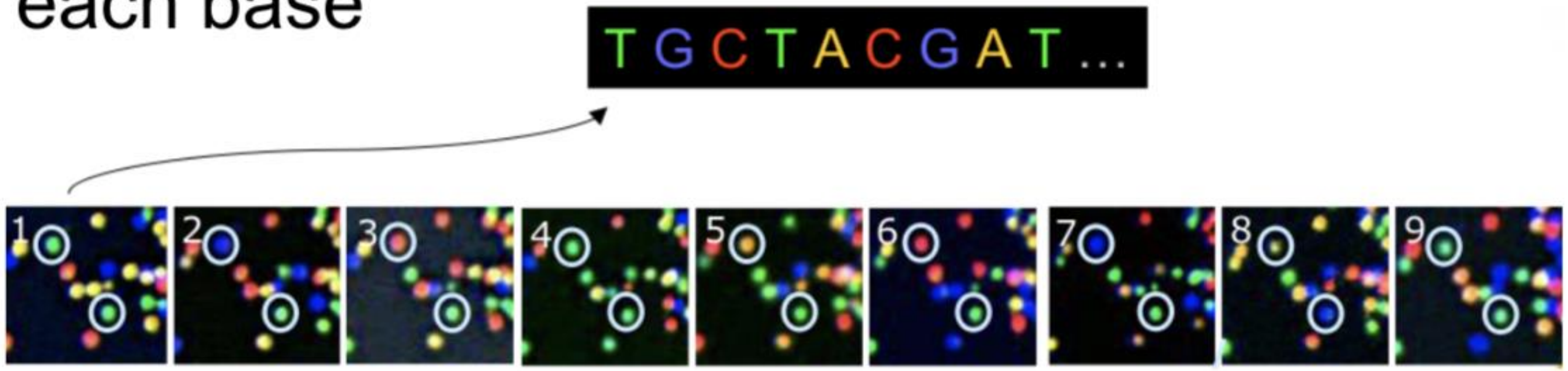
- Slides in this section from Dr. Katerina Kechris



Shendure &Aiden Nature Biotechnology 30, 1084–1094 (2012)

Base-Calling

- Converts the fluorescence signals of four nucleotides for each cycle into sequence data
- Methods differ on correcting for cross-talk, phasing/prephasing, signal decay
- Return sequence read and quality score for each base



T G C T A C G A T ...

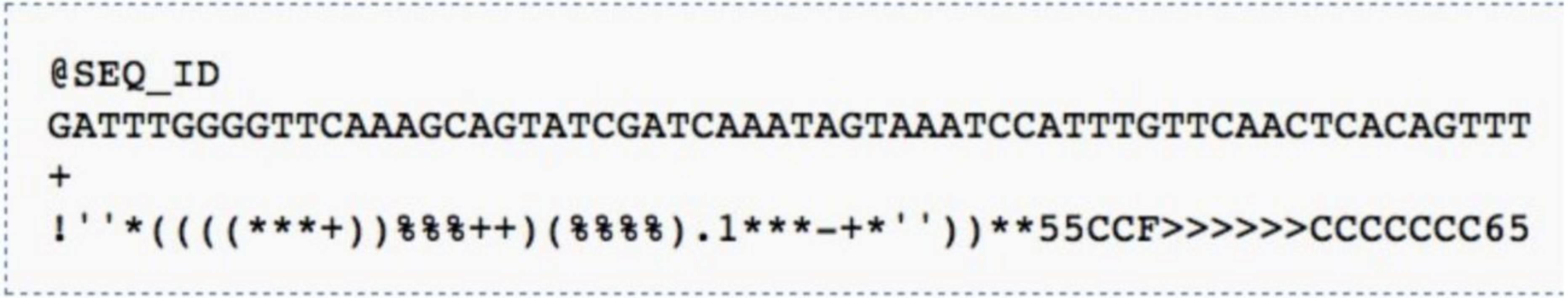
Quality Score Per Base

- Assess reliability of base call
- For each base, converts error probability (p) to integer score (rounded)

$$q = -10\log_{10}(p)$$

q	p	**Probability called base is correct**
10	0.1	0.9
20	0.01	0.99
30	0.001	0.999
40	0.0001	0.9999

Output: FASTQ Format

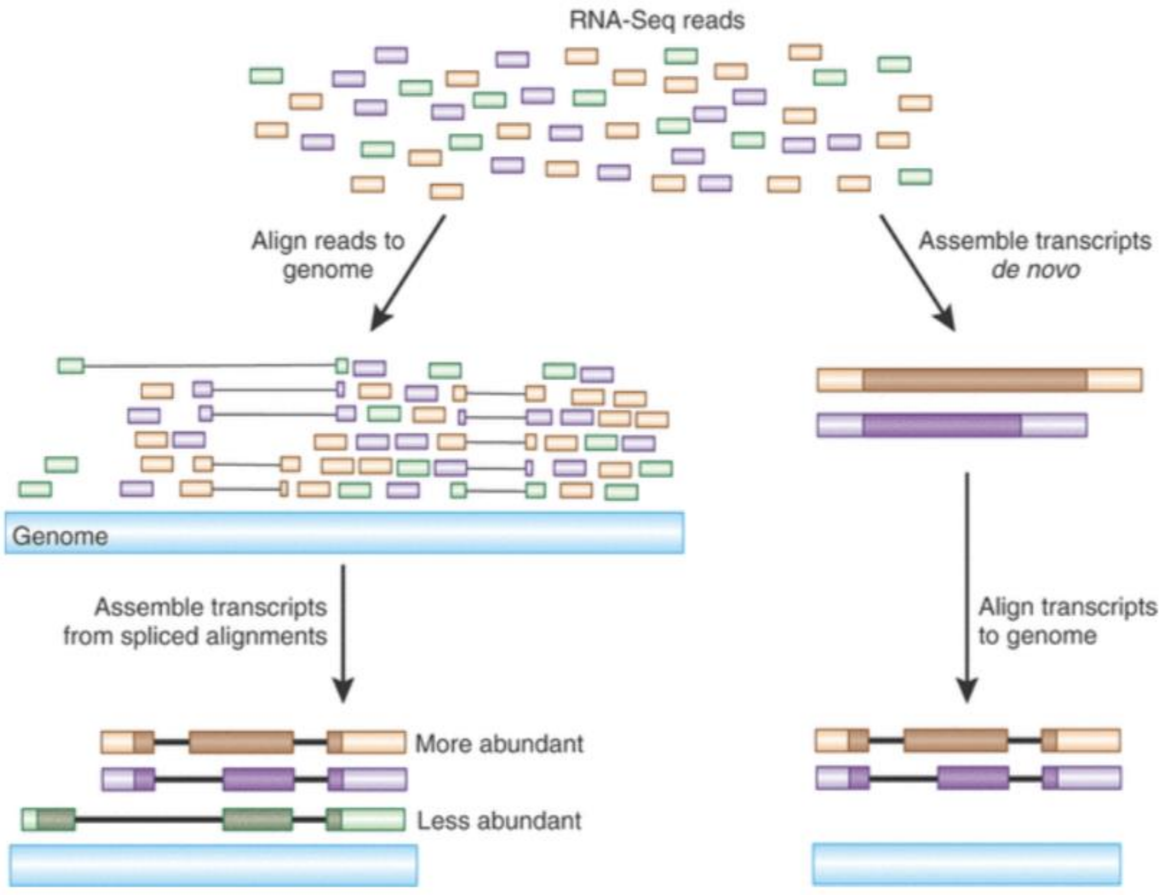


```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- ASCII text
- 4 lines per sequence
- Line 1 begins with the @ character, a sequence ID, and an optional description
- Line 2 is the sequence for the read
- Line 3 begins with the + character, followed by the same sequence ID, and another optional description
- Line 4 encodes quality values in hexadecimal format for the sequence letters in line 2
  - Must contain the same number of characters as the sequence in line 2
  - Hexadecimal – use single ASCII to represent up to 92 numeric value (save space)

Mapping & Reconstruction

1. Mapping: Align reads to reference genome (read mapping)
2. Reconstruct transcriptome (using reference or *de novo*)


RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Mapping Options

Align and then assemble

vs

Assemble and then align

Align to genome

vs

Transcriptome

Non-spliced vs Spliced

Challenges for Mapping

- Multiple hits
- Allowing mismatches
  - High error rate
- Paired-end reads
- Span exon-exon junctions
  - Align with gaps
- Very large number of reads
  - Storage & processing power

Pharmacogenomics & Personalized Medicine

- Use patient data (DNA) to determine treatments on an individual level
- Requires understanding of the disease and drug(s)
- Benefits from tracking if the patient took the prescribed drug (adherance)

- Slides in this section from: Dr. Debashis Ghosh

What is Personalized Medicine?

- Also known as “precision medicine.”
- Provide "the right patient with the right drug at the right dose at the right time."
- Tailor medical treatment to the individual characteristics, needs, and preferences of a patient during all stages of care.
- Identify genetic or protein biomarkers to predict drug response and side effects.

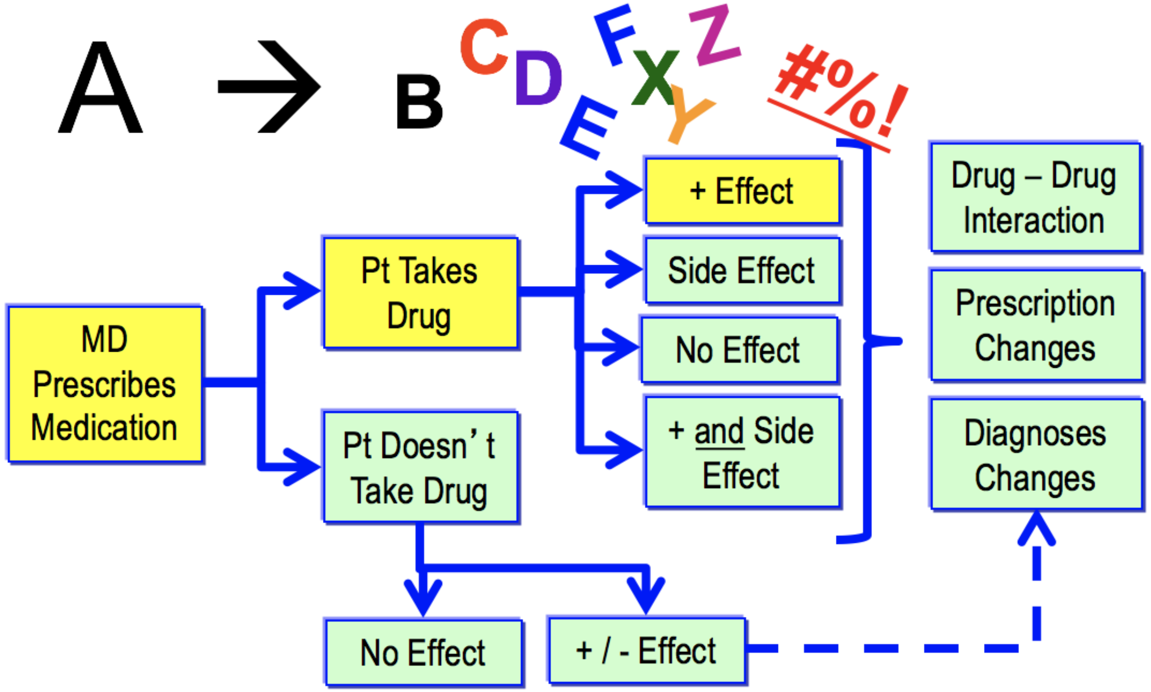
Pharmacogenetics Terminology

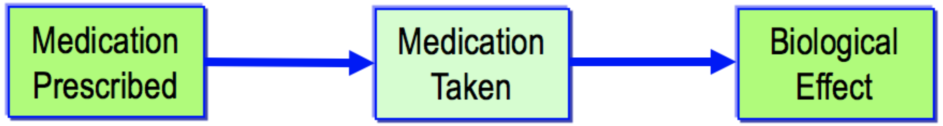
- A genetic variant is a difference in the DNA sequence compared with a reference sequence.
  - Polymorphism: A genetic variant that is common, often defined as ≥ 1% in the population.
  - Mutation: A genetic variant that is rare, often defined as <1% in the population.

Examples of Types of Genetic Variants	Definition
Single nucleotide polymorphism (SNP)	Difference in one nucleotide (base pair)
Insertion or deletion (indel)	Insertion or deletion of multiple consecutive nucleotides
Repeat polymorphism	Variable number of nucleotides that are repeated
Copy number variation (CNV)	Abnormal number of copies of one or more DNA regions (e.g., gene duplication or deletion)

Realities of Drug Therapy

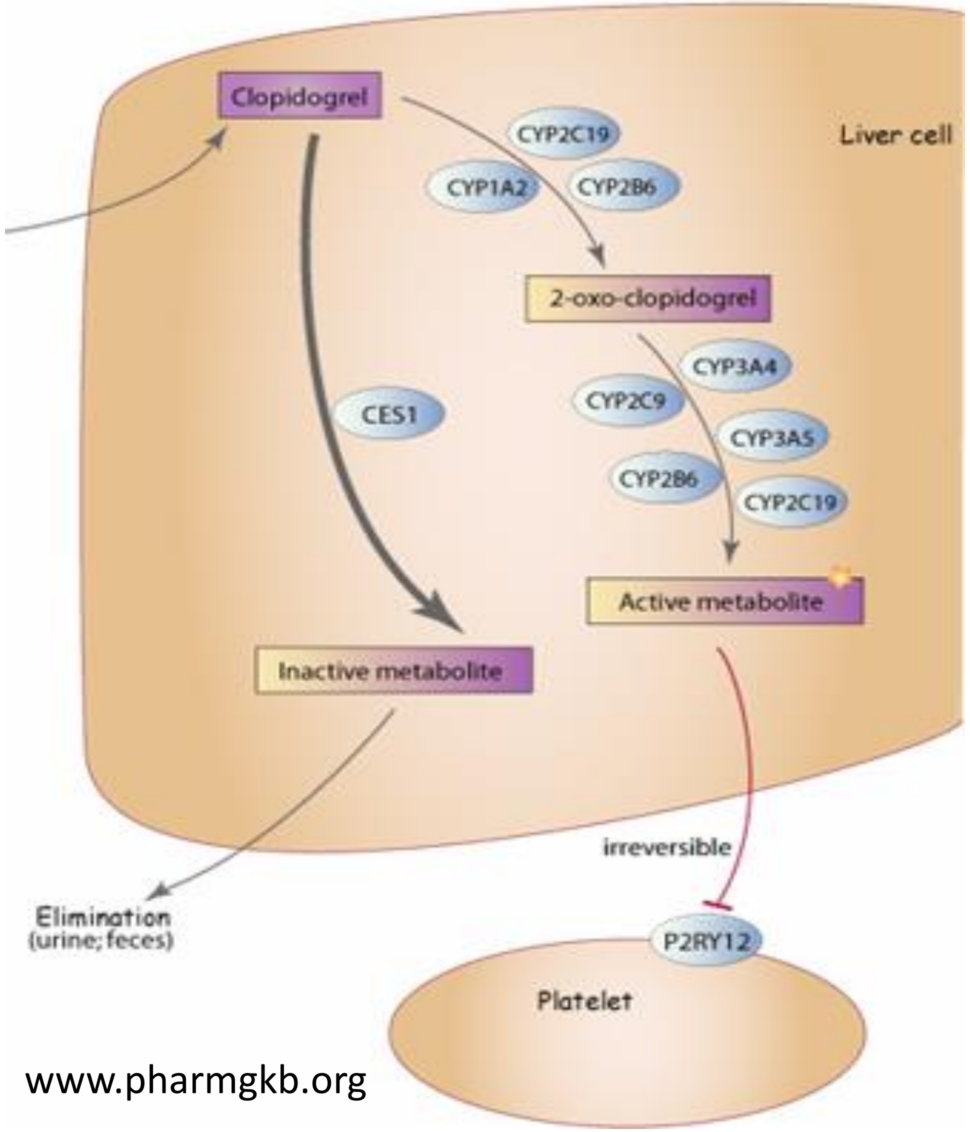
Physician Prescribes a Medication → Anything goes





Clopidogrel (Plavix)

- Antiplatelet agent used to reduce the risk of atherosclerotic events.
- Metabolized to 2-oxo-clopidogrel and an active thiol metabolite by numerous CYP enzymes.
- CYP2C19 is the key enzyme involved in the formation of the active metabolite.
- The active metabolite binds to and inhibits P2Y12 receptors on platelets, thereby inhibiting platelet activation and aggregation.



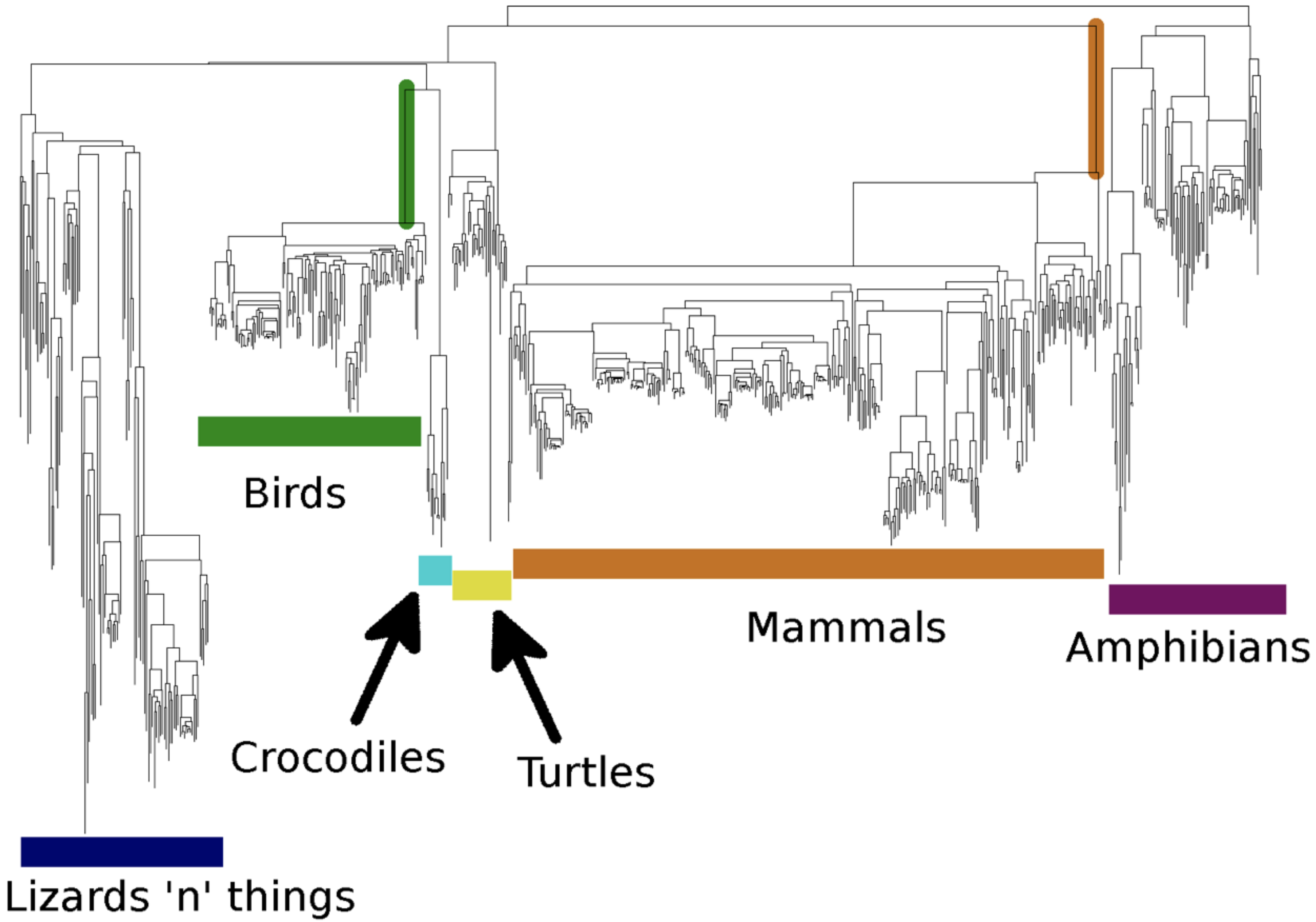
www.pharmgkb.org

Computational Phylogeny

- Determine evolutionary relationships using genetic information
- Similar sequence => similar function

- Slides in this section from: Dr. David Pollock

Tetrapods

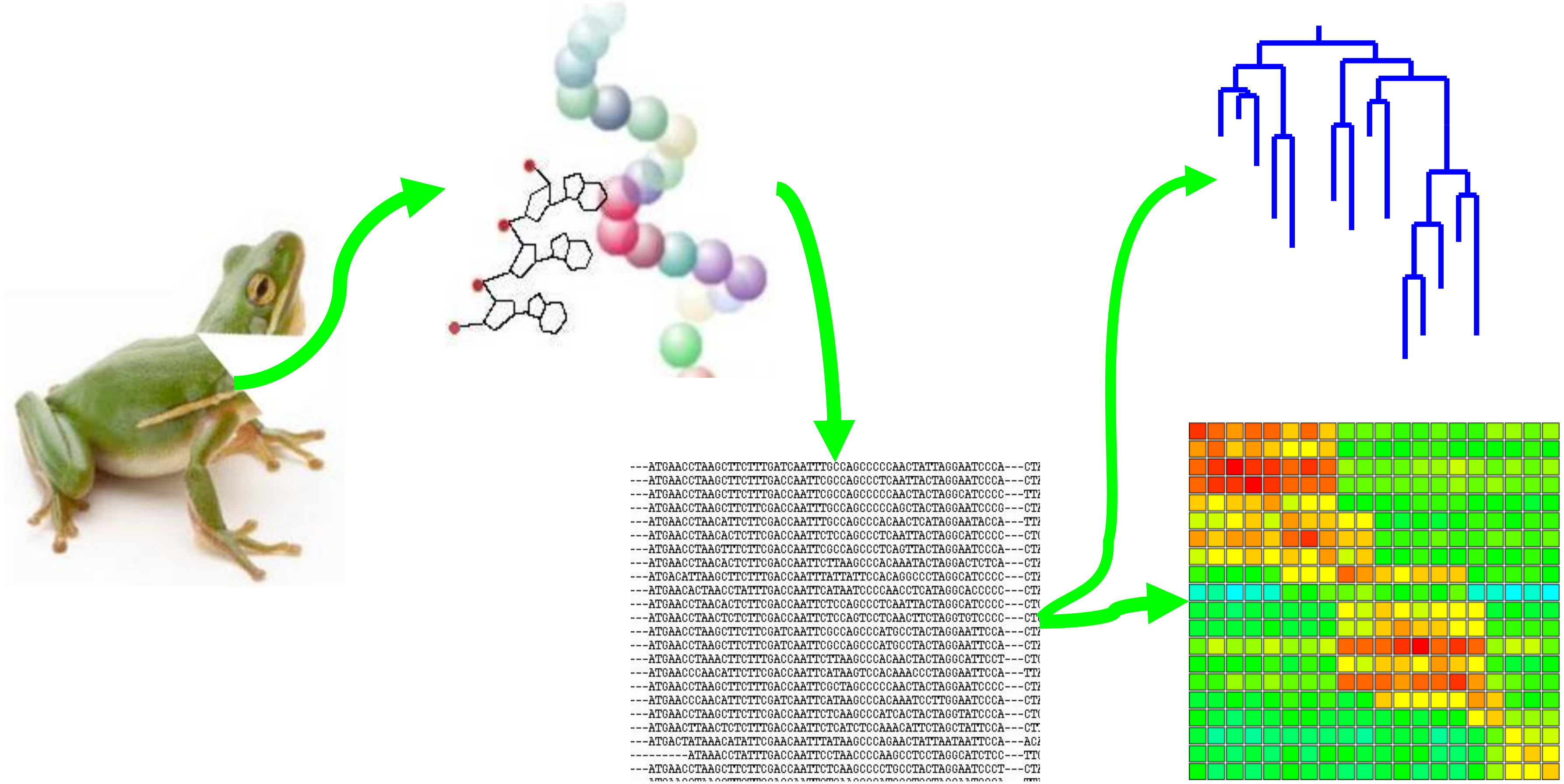


Cytochrome C Oxidase I Bayesian consensus tree

Why Phylogenetics?

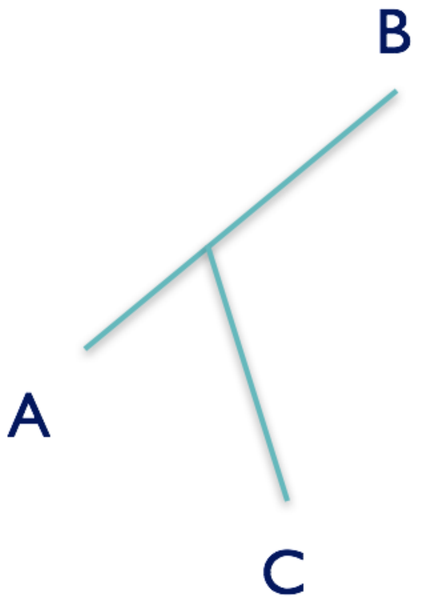
- Resolve evolutionary history
  - Important for comparative analysis to account for correlations due to relatedness
- Disease origins, paths of infection
  - Influenza, HIV
- Origin of genes, systems, functions

Standard Empirical Models of Sequence Evolution



```
---ATGAACCTAAGCTTCTTTGATCAATTTGCCAGCCCCCAACTATTAGGAATCCCA---CTA
---ATGAACCTAAGCTTCTTTGACCAATTCGCCAGCCCTCAATTACTAGGAATCCCA---CTA
---ATGAACCTAAGCTTCTTTGACCAATTCGCCAGCCCCCAACTACTAGGCATCCCC---TTA
---ATGAACCTAAGCTTCTTCGACCAATTTGCCAGCCCCCAGCTACTAGGAATCCCG---CTA
---ATGAACCTAACATTCTTCGACCAATTTGCCAGCCCACAACTCATAGGAATACCA---TTA
---ATGAACCTAACACTCTTCGACCAATTCTCCAGCCCTCAATTACTAGGCATCCCC---CTC
---ATGAACCTAAGTTTCTTCGACCAATTCGCCAGCCCTCAGTTACTAGGAATCCCA---CTA
---ATGAACCTAACACTCTTCGACCAATTCTTAAGCCCACAAATACTAGGACTCTCA---CTA
---ATGACATTAAGCTTCTTTGACCAATTTATTATTCCACAGGCCCTAGGCATCCCC---CTA
---ATGAACACTAACCTATTTGACCAATTCATAATCCCCAACCTCATAGGCACCCCC---CTA
---ATGAACCTAACACTCTTCGACCAATTCTCCAGCCCTCAATTACTAGGCATCCCC---CTC
---ATGAACCTAACTCTCTTCGACCAATTCTCCAGTCCTCAACTTCTAGGTGTCCCC---CTC
---ATGAACCTAAGCTTCTTCGATCAATTCGCCAGCCCATGCCTACTAGGAATTCCA---CTA
---ATGAACCTAAGCTTCTTCGATCAATTCGCCAGCCCATGCCTACTAGGAATTCCA---CTA
---ATGAACCTAAACTTCTTTGACCAATTCTTAAGCCCACAACTACTAGGCATTCCT---CTC
---ATGAACCCAACATTCTTCGACCAATTCATAAGTCCACAAACCCTAGGAATTCCA---TTA
---ATGAACCTAAGCTTCTTTGACCAATTCGCTAGCCCCCAACTACTAGGAATCCCC---CTA
---ATGAACCCAACATTCTTCGATCAATTCATAAGCCCACAAATCCTTGGAATCCCA---CTA
---ATGAACCTAAGCTTCTTCGACCAATTCTCAAGCCCATCACTACTAGGTATCCCA---CTC
---ATGAACTTAACTCTCTTTGACCAATTCTCATCTCCAAACATTCTAGCTATTCCA---CT
---ATGACTATAAACATATTCGAACAATTTATAAGCCCAGAACTATTAATAATTCCA---ACA
---------ATAAACCTATTTGACCAATTCCTAACCCCAAGCCTCCTAGGCATCTCC---TTC
---ATGAACCTAAGCTTCTTCGACCAATTCTCAAGCCCCTGCCTACTAGGAATCCCT---CTA
```

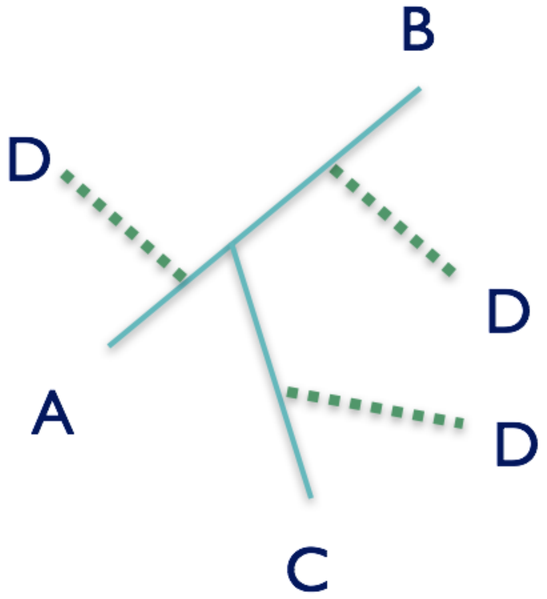
Topology Space

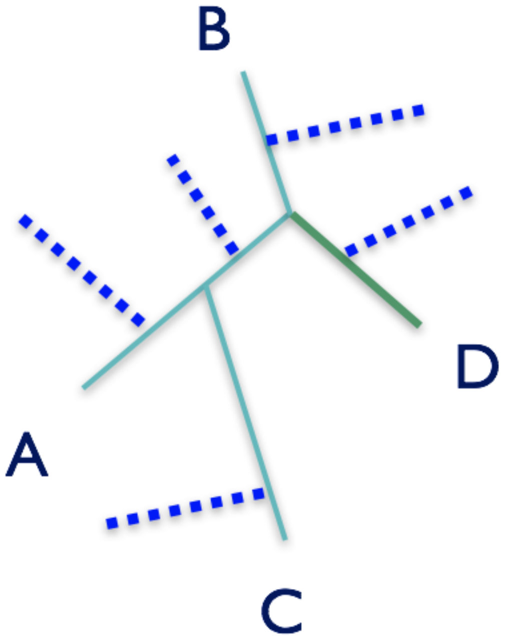


$$\prod_{i=3}^{T}(2i-5)$$

Times branch lengths

Need tree search algorithm





Heuristic, Optimal, Posterior

- NP hard, so need tricks
- Distance
  - Estimate of amount of change separating two sequences (species)
  - Calculate analytically (limited), or ML
  - Requires a reversible model of evolution
- Parsimony
  - Minimal number of changes
  - Poorly specified model (but there is one)
  - Easy to calculate
- ML, Bayes
  - Model based, don’t toss the data

Multiple Sequence Alignment

- Compare multiple sequences to identify similarities
- Parts of a sequence that are often preserved are likely important
- Combines fields of Biochemistry, Statistics, & Comp Sci

- Slides in this section from: Dr. Scott Walmsley

.

Why MSA?

“Whether the ultimate aim is a ***phylogenetic*** analysis of several orthologues, the identification of a ***pattern*** for particular feature or motif, or the basis for ***structural modelling***, multiple sequence alignments allow the researcher to gather more biological information than a single sequence can offer”

“The importance of a residue for maintaining the structure and function of a protein can usually be inferred from how conserved it appears in a multiple sequence alignment of that protein and its homologues”

Valdar WS. Scoring residue conservation. Proteins. 2002 Aug 1;48(2):227-41. Review

Pre-requisite knowledge

Computational / Math / Statistics & Biochemistry

- Alphabets

DNA (n= 4)

RNA (n = 4)

Amino Acids (n = 20)

CODON (n=64)

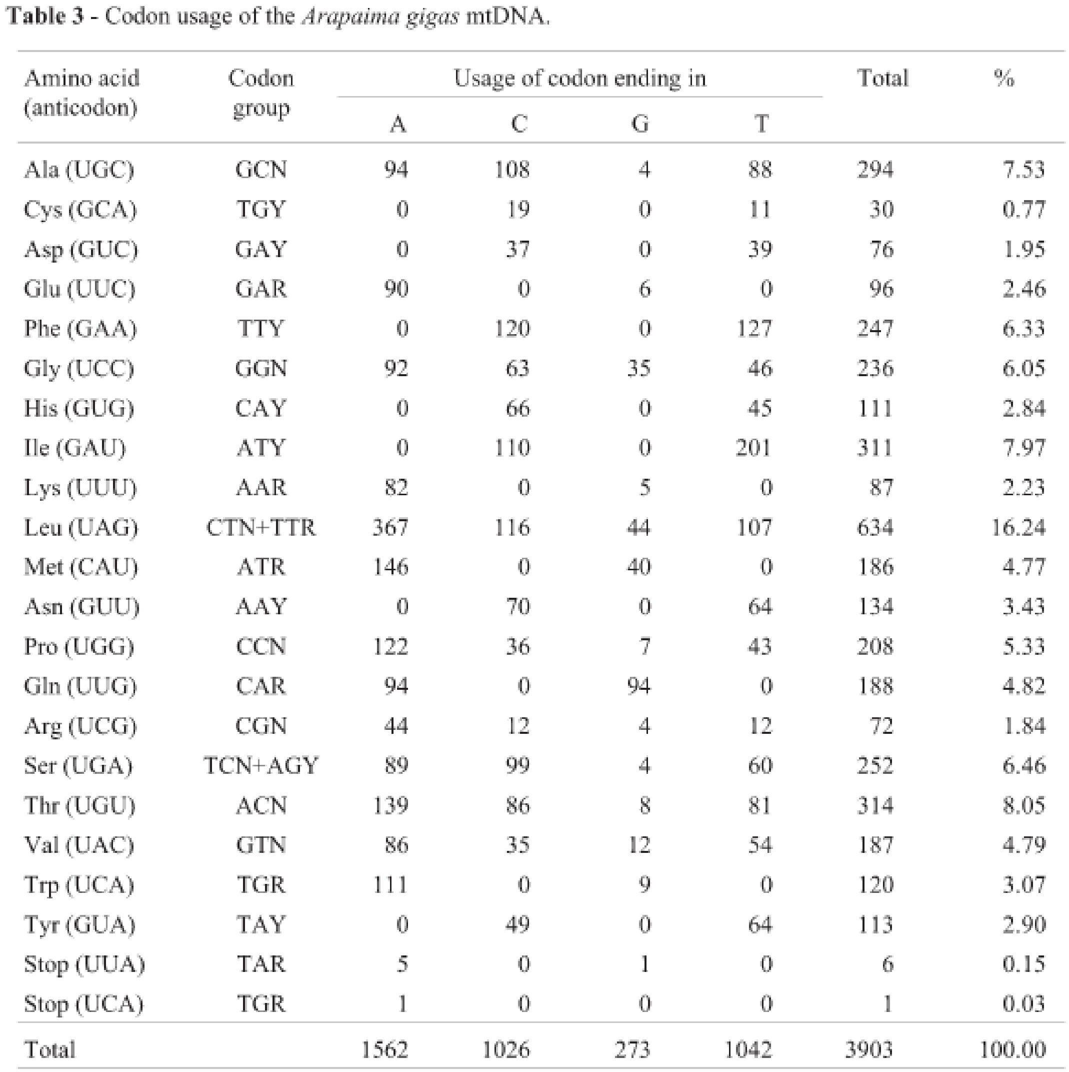


Table 3 - Codon usage of the *Arapaima gigas* mtDNA.

Amino acid (anticodon)	Codon group	Usage of codon ending in				Total	%
		A	C	G	T		
Ala (UGC)	GCN	94	108	4	88	294	7.53
Cys (GCA)	TGY	0	19	0	11	30	0.77
Asp (GUC)	GAY	0	37	0	39	76	1.95
Glu (UUC)	GAR	90	0	6	0	96	2.46
Phe (GAA)	TTY	0	120	0	127	247	6.33
Gly (UCC)	GGN	92	63	35	46	236	6.05
His (GUG)	CAY	0	66	0	45	111	2.84
Ile (GAU)	ATY	0	110	0	201	311	7.97
Lys (UUU)	AAR	82	0	5	0	87	2.23
Leu (UAG)	CTN+TTR	367	116	44	107	634	16.24
Met (CAU)	ATR	146	0	40	0	186	4.77
Asn (GUU)	AAY	0	70	0	64	134	3.43
Pro (UGG)	CCN	122	36	7	43	208	5.33
Gln (UUG)	CAR	94	0	94	0	188	4.82
Arg (UCG)	CGN	44	12	4	12	72	1.84
Ser (UGA)	TCN+AGY	89	99	4	60	252	6.46
Thr (UGU)	ACN	139	86	8	81	314	8.05
Val (UAC)	GTN	86	35	12	54	187	4.79
Trp (UCA)	TGR	111	0	9	0	120	3.07
Tyr (GUA)	TAY	0	49	0	64	113	2.90
Stop (UUA)	TAR	5	0	1	0	6	0.15
Stop (UCA)	TGR	1	0	0	0	1	0.03
Total		1562	1026	273	1042	3903	100.00

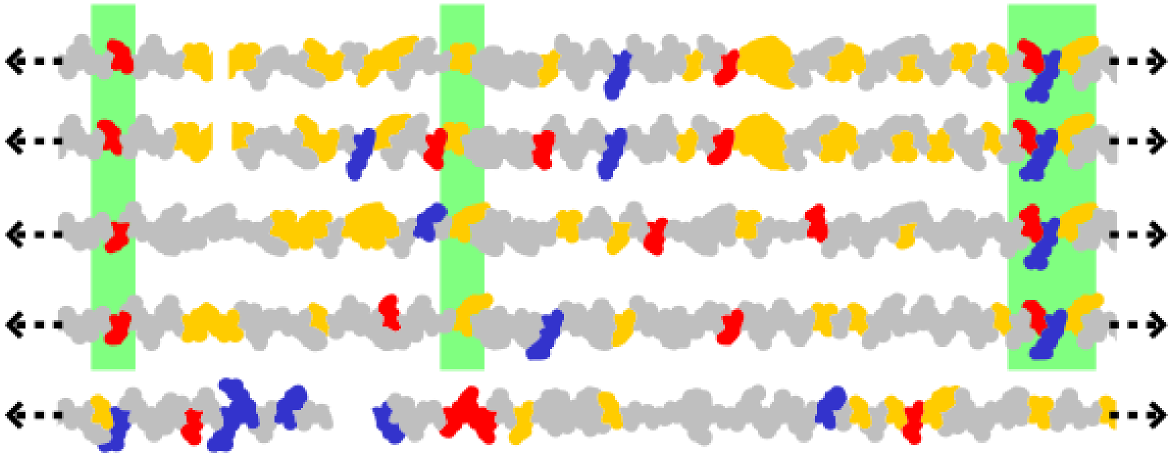
Pre-requisite knowledge

Biochemistry / Molecular Biology

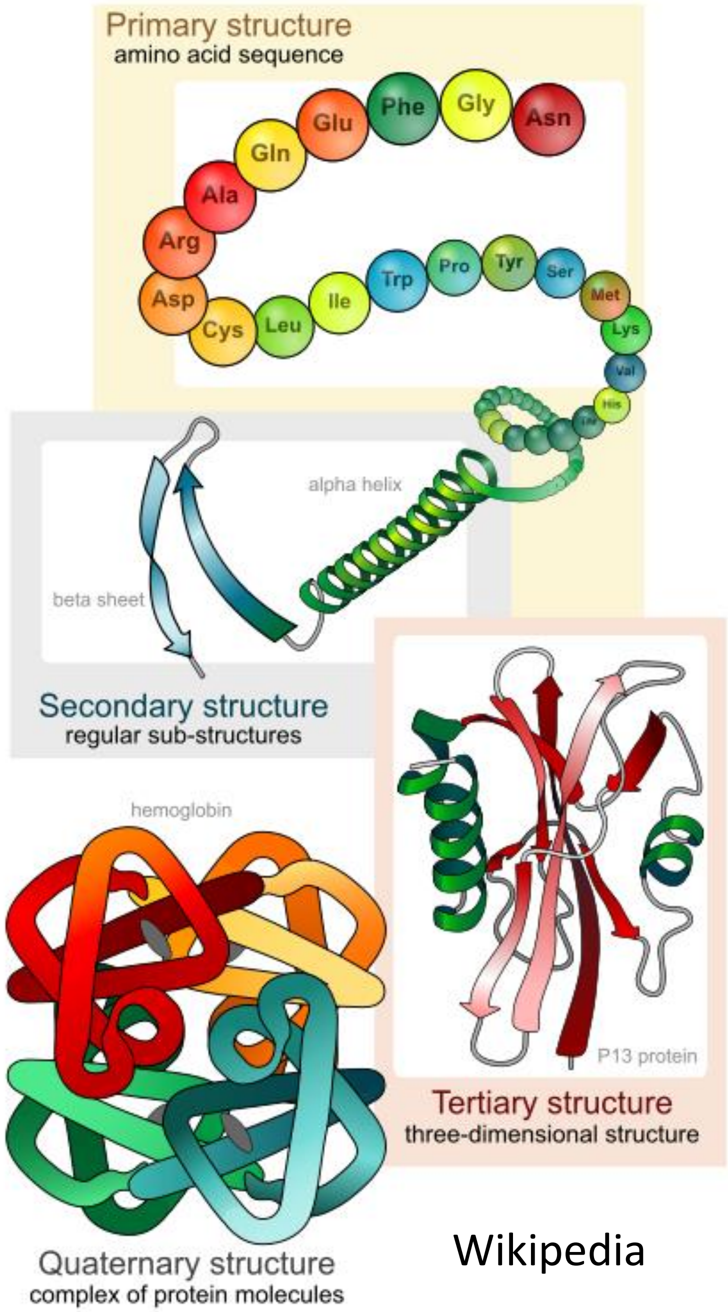
- Mutation rates drive evolution
- Biophysical mechanisms produce mutation rates:

  DNA / RNA Polymerase

- Insertion /Deletion : frameshift → altered CODON

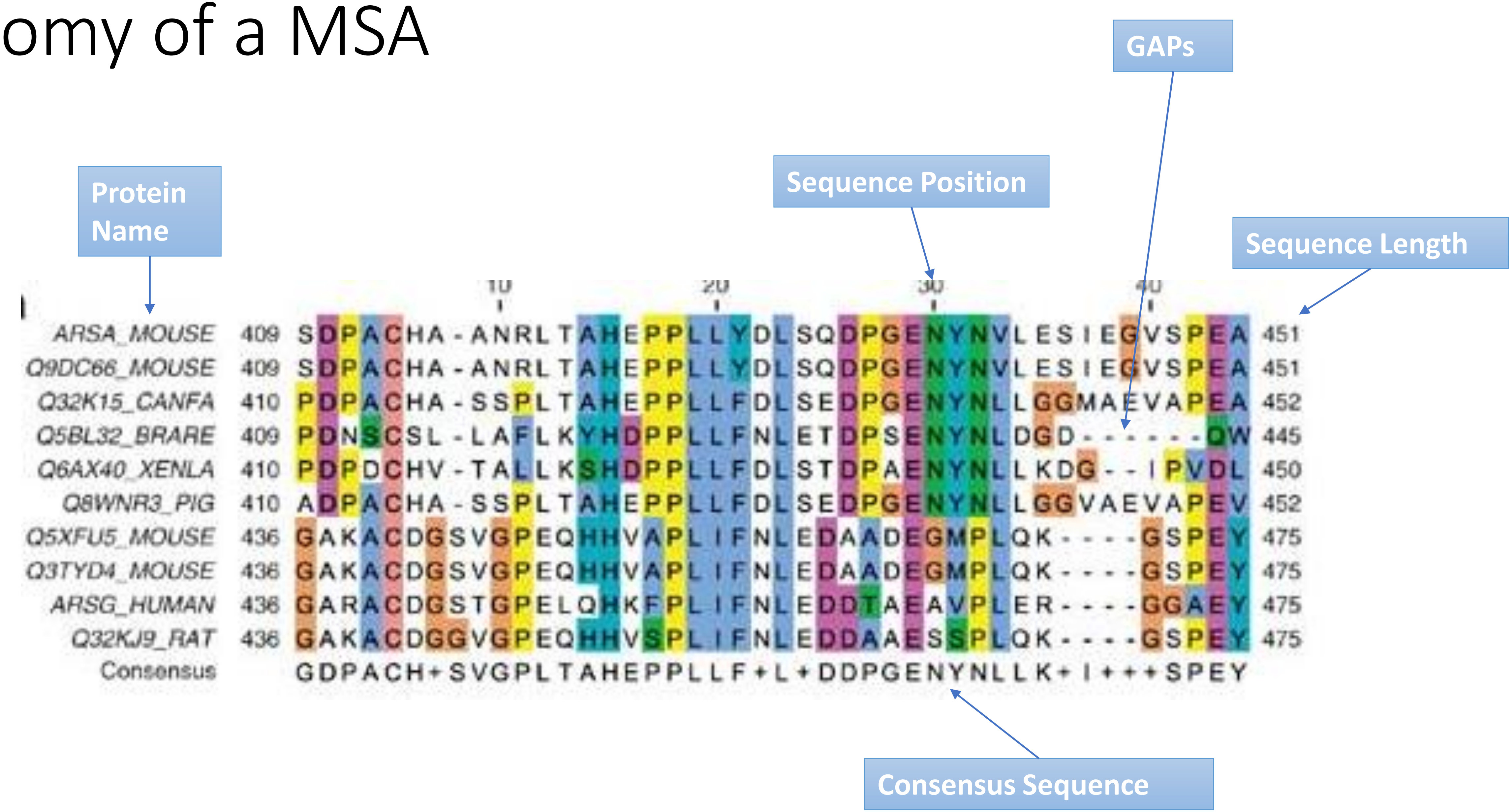


Wikipedia



Wikipedia

Anatomy of a MSA



```
                                  10        20        30        40
                                   |         |         |         |
ARSA_MOUSE    409  SDPACHA-ANRLTAHEPPLLYDLSQDPGENYNVLESIEGVSPEA  451
Q9DC66_MOUSE  409  SDPACHA-ANRLTAHEPPLLYDLSQDPGENYNVLESIEGVSPEA  451
Q32K15_CANFA  410  PDPACHA-SSPLTAHEPPLLFDLSEDPGENYNLLGGMAEVAPEA  452
Q5BL32_BRARE  409  PDNSCSL-LAFLKYHDPPLLFNLETDPSENYNLDGD------QW  445
Q6AX40_XENLA  410  PDPDCHV-TALLKSHDPPLLFDLSTDPAENYNLLKDG--IPVDL  450
Q8WNR3_PIG    410  ADPACHA-SSPLTAHEPPLLFDLSEDPGENYNLLGGVAEVAPEV  452
Q5XFU5_MOUSE  436  GAKACDGSVGPEQHHVAPLIFNLEDAADEGMPLQK----GSPEY  475
Q3TYD4_MOUSE  436  GAKACDGSVGPEQHHVAPLIFNLEDAADEGMPLQK----GSPEY  475
ARSG_HUMAN    436  GARACDGSTGPELQHKFPLIFNLEDDTAEAVPLER----GGAEY  475
Q32KJ9_RAT    436  GAKACDGGVGPEQHHVSPLIFNLEDDAAESSPLQK----GSPEY  475
Consensus          GDPACH+SVGPLTAHEPPLLF+L+DDPGENYNLLK+I+++SPEY
```

Protein Name

Sequence Position

GAPs

Sequence Length

Consensus Sequence

Procter JB, Thompson J, Letunic I, Creevey C, Jossinet F, Barton GJ. Visualization of multiple alignments, phylogenies and gene family evolution. Nat Methods. 2010 Mar;7(3 Suppl):S16-25.

Protein Structure and Prediction

- Given a sequence of DNA can we compute:
  - The translated protein sequence?
  - The protein's shape?
  - The protein's functionality?

- Slides in this section from: Dr. Michael Strong

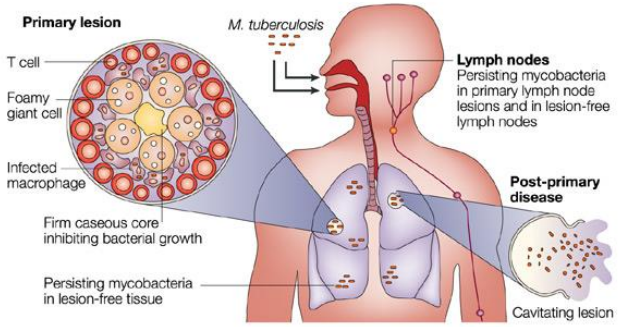
Why do we care about protein structures

Combining Structure and Genomic Information

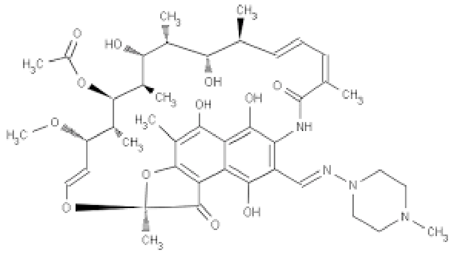
- Help us understand implications of mutations

Drug Resistance

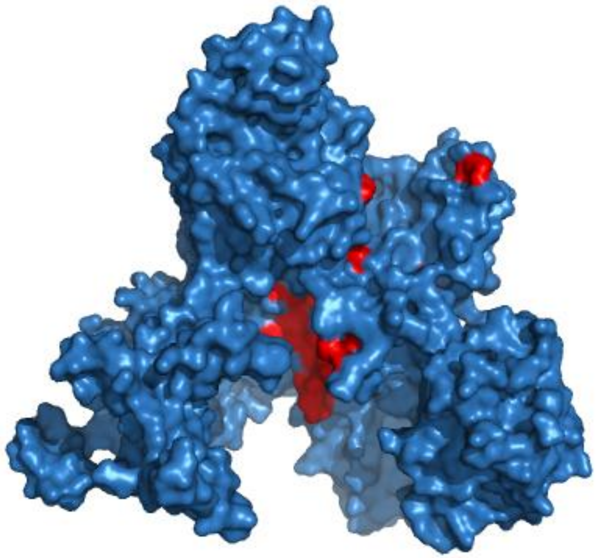
Tuberculosis


Primary lesion
M. tuberculosis
T cell
Foamy giant cell
Infected macrophage
Firm caseous core inhibiting bacterial growth
Persisting mycobacteria in lesion-free tissue
Lymph nodes
Persisting mycobacteria in primary lymph node lesions and in lesion-free lymph nodes
Post-primary disease
Cavitating lesion

Nature Reviews | Microbiology



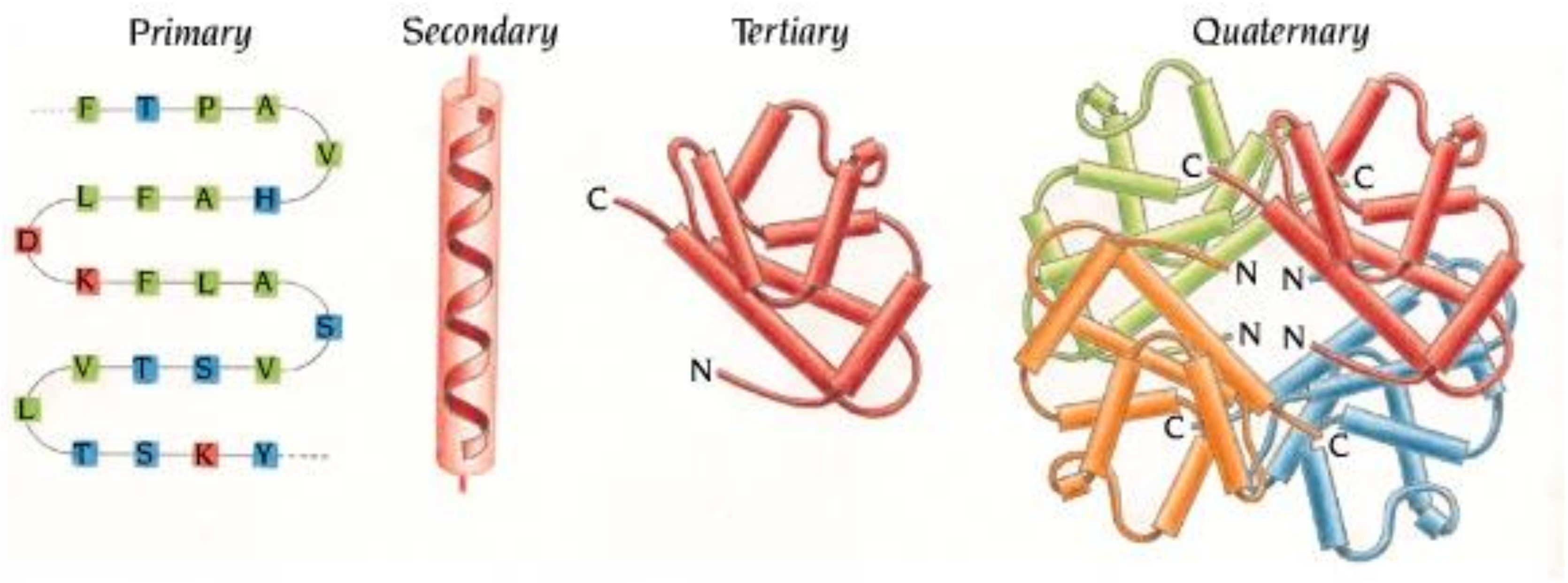
Rifampin



rpoB drug target

Most Proteins Spontaneously Fold

Important to Computational Biologists, because this suggests that all information relating to the correct folding of a protein is contained in it's primary amino acid sequence, but …



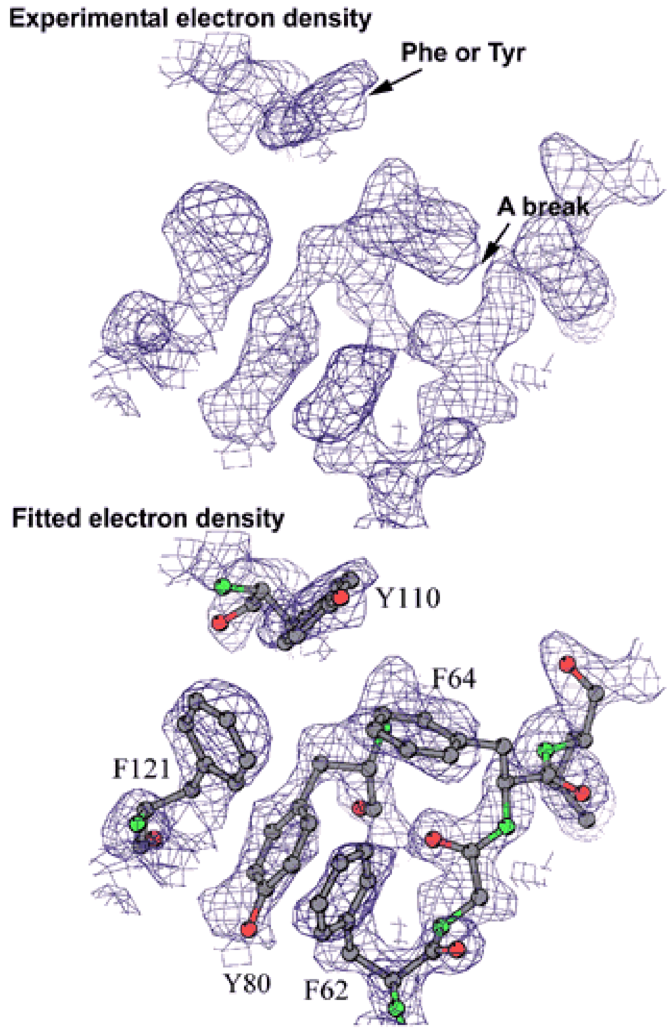
Experimental Methods of Structure Determination

X-ray crystallography

High resolution structure determination

- Intensities and phases of all reflections are combined in a Fourier transform to provide maps of electron density

Phases determined by using heavy metals or selenomethionine (MAD)



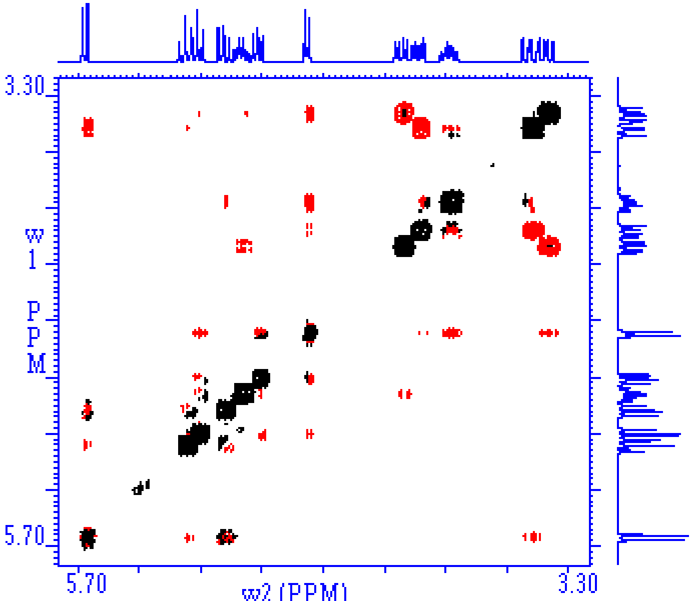
Experimental Methods of Structure Determination

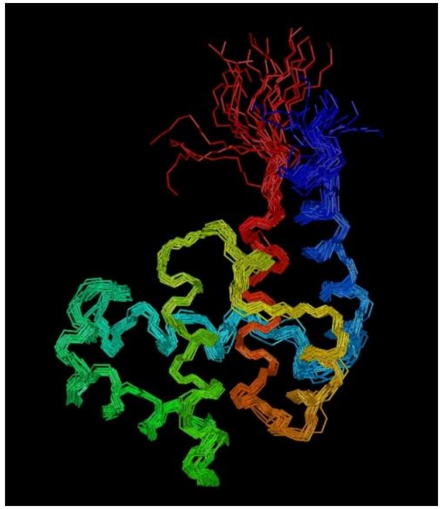
NMR – Nuclear Magnetic Resonance

High resolution structure determination

- Smaller Proteins than X-ray
- Distances between pairs of hydrogen atoms
- Lots of information about dynamics
- Requires soluble, non-aggregating material
- Assignment sometimes difficult

NOESY





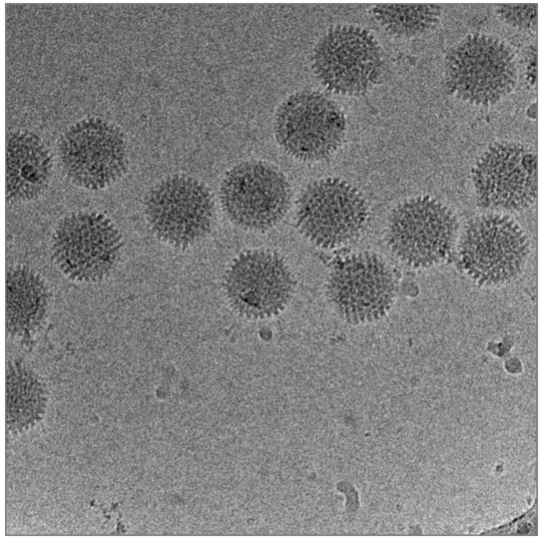
NOE cross-peak if they are within 5.0 Å

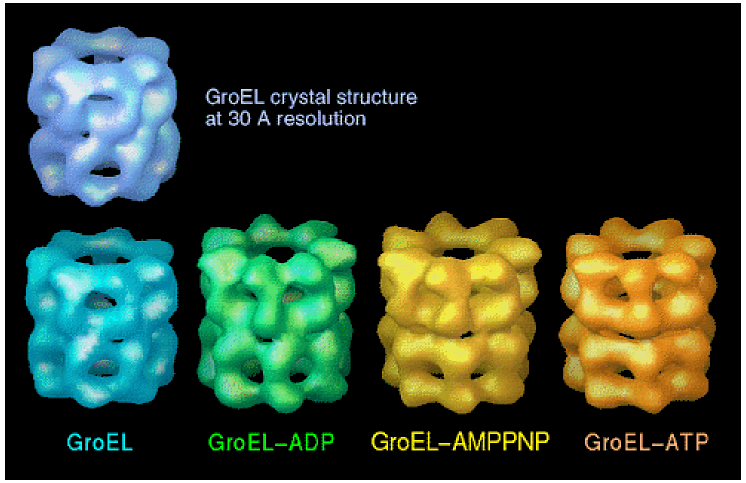
Experimental Methods of Structure Determination

Cryo Electron Microscopy

Low to medium resolution structure determination

- Medium resolution typically ~10-15Å (up to 3.8Å in some special cases)
- Limited information about dynamics
- Can be used for very large molecules and complexes





Rosetta structure prediction

2 phases

1.Low-resolution phase – statistical scoring function and fragment assembly

A. local structure conformations using info from PDB (3 and 9mer stretches)

B. multiple fragment substitution simulated annealing – to find best arrangement of the fragments (Monte Carlo Search)

C. low resolution ensemble of decoy conformations

2. Atomic refinement phase using rotamers and small backbone angle moves (in populated regions of Ramachandran plot)

A. Refinement

B. Then structures clustered based on RMSD

C. Center of the Largest Clusters chosen as representative folds (likely to be correct fold)

The energy model

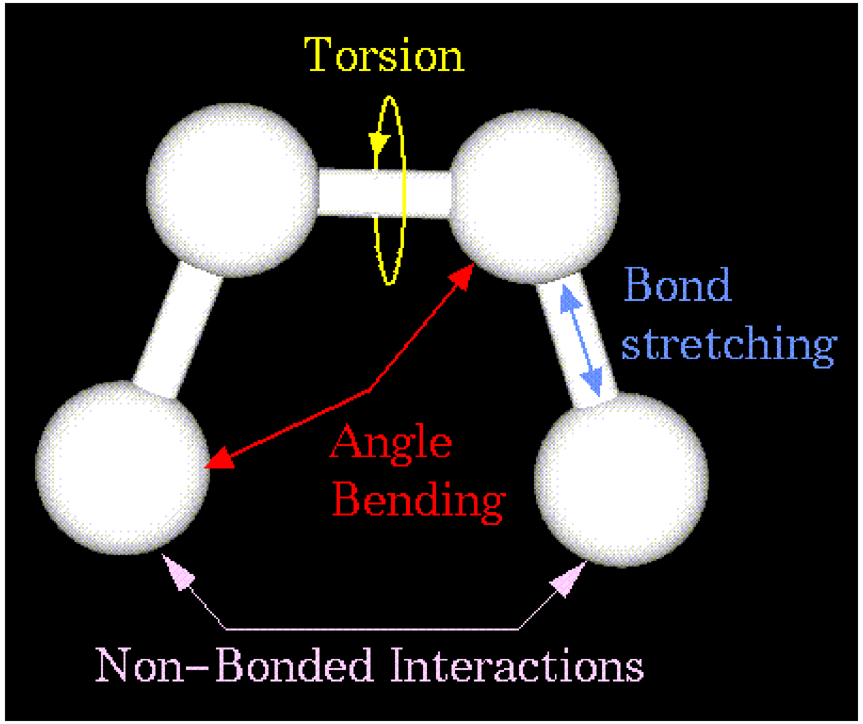
- Proposed by Linus Pauling in the 1930s
- Bond angles and lengths are almost always the same
- Energy model broken up into two parts:

  Covalent terms

  - Bond distances
  - Bond angles
  - Dihedral angles

  Non-covalent terms

  - Forces at a distance between all non-bonded atoms



Microbiome & Metagenomics

- Microbiome: “The ecological community of commensal, symbiotic, and pathogenic microorganisms that share our body space”

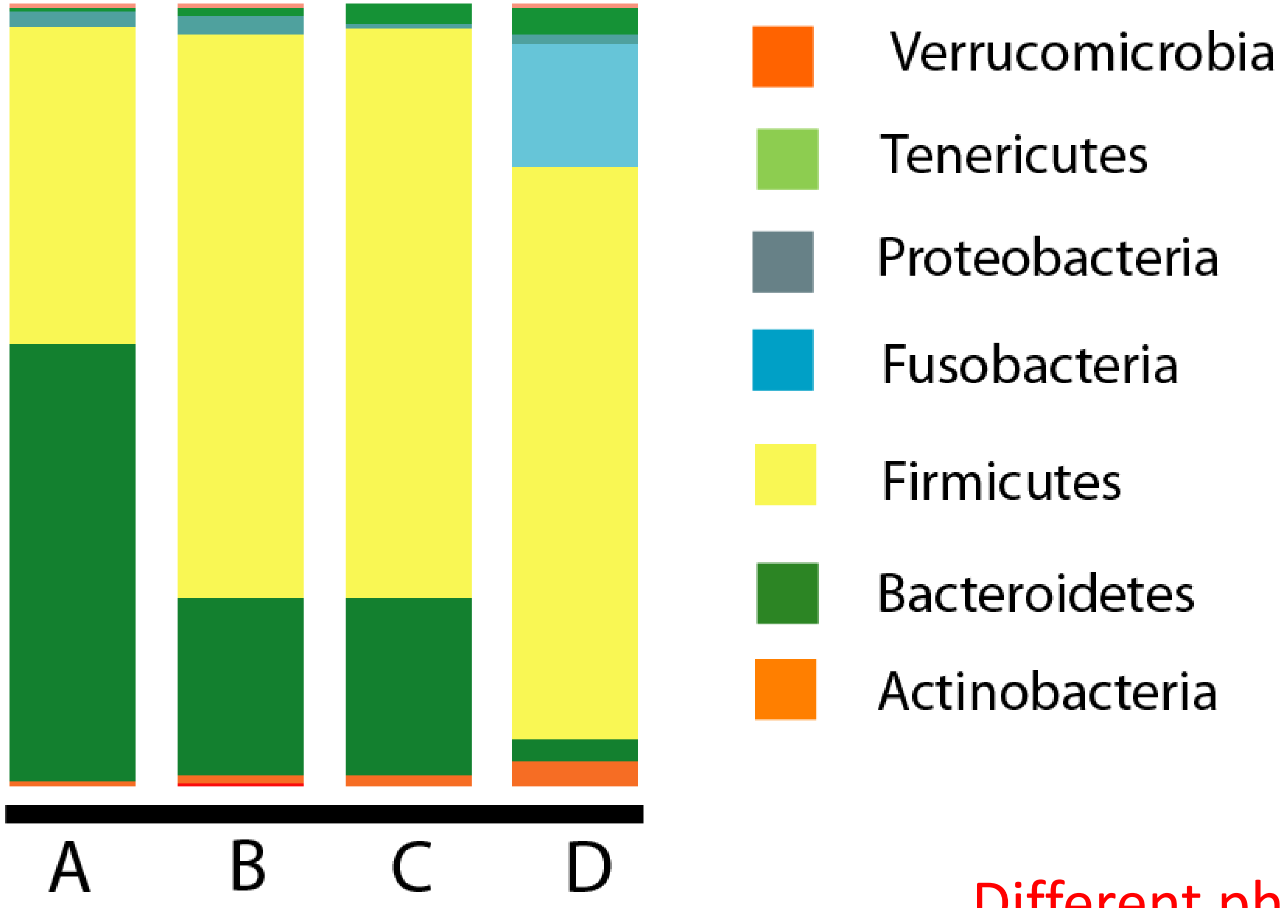
- Metagenomics: In essence, using DNA sequences from a microbial sample to identify the species present

- Slides in this section from: Dr. Catherine Lozupone

What do we want to understand?

- What does a healthy microbiome look like?
  - How diverse is it?
  - What types of bacteria are there?
  - What is their function?
- How variable is the microbiome?
  - Over time within an individual?
  - Across individuals?
  - Functionally?
- What are driving factors of variability?
  - Age, culture, physiological state (pregnancy)
- How do changes affect disease?
  - What properties (taxa, amount of diversity) change with disease?
  - Cause or affect?
  - Functional consequences of dysbiosis
- Host Interactions
  - Evolution/adaptation to the host over time.
  - Immune system

Gut microbiota has simple composition at the phylum level



Different phyla: Animals and plants

Data from: Yatsunenko *et. al.* 2012. Nature.

Small Subunit Ribosomal RNA

- Present in all known life forms
- Highly conserved
- Resistant to horizontal transfer events
- Big database!

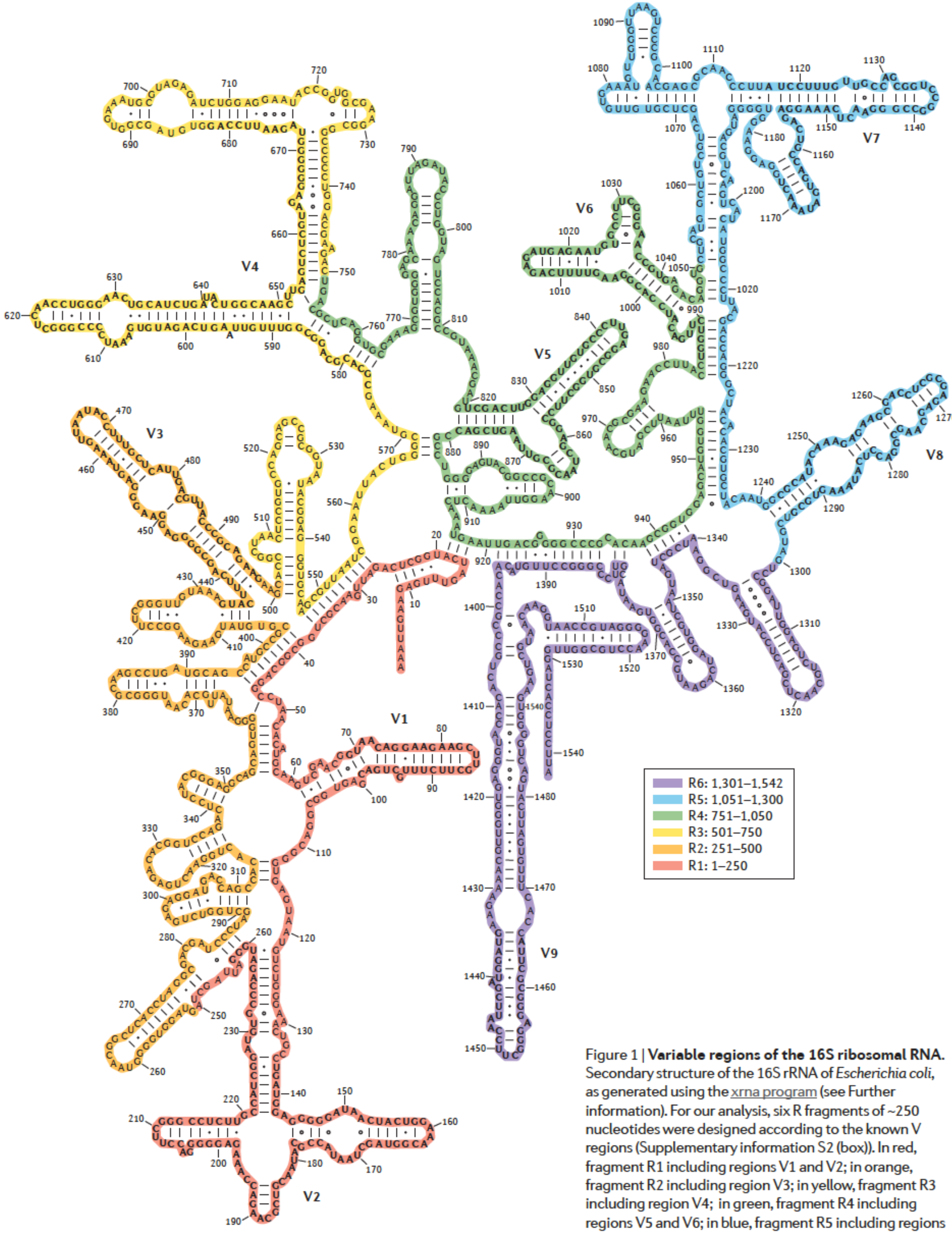
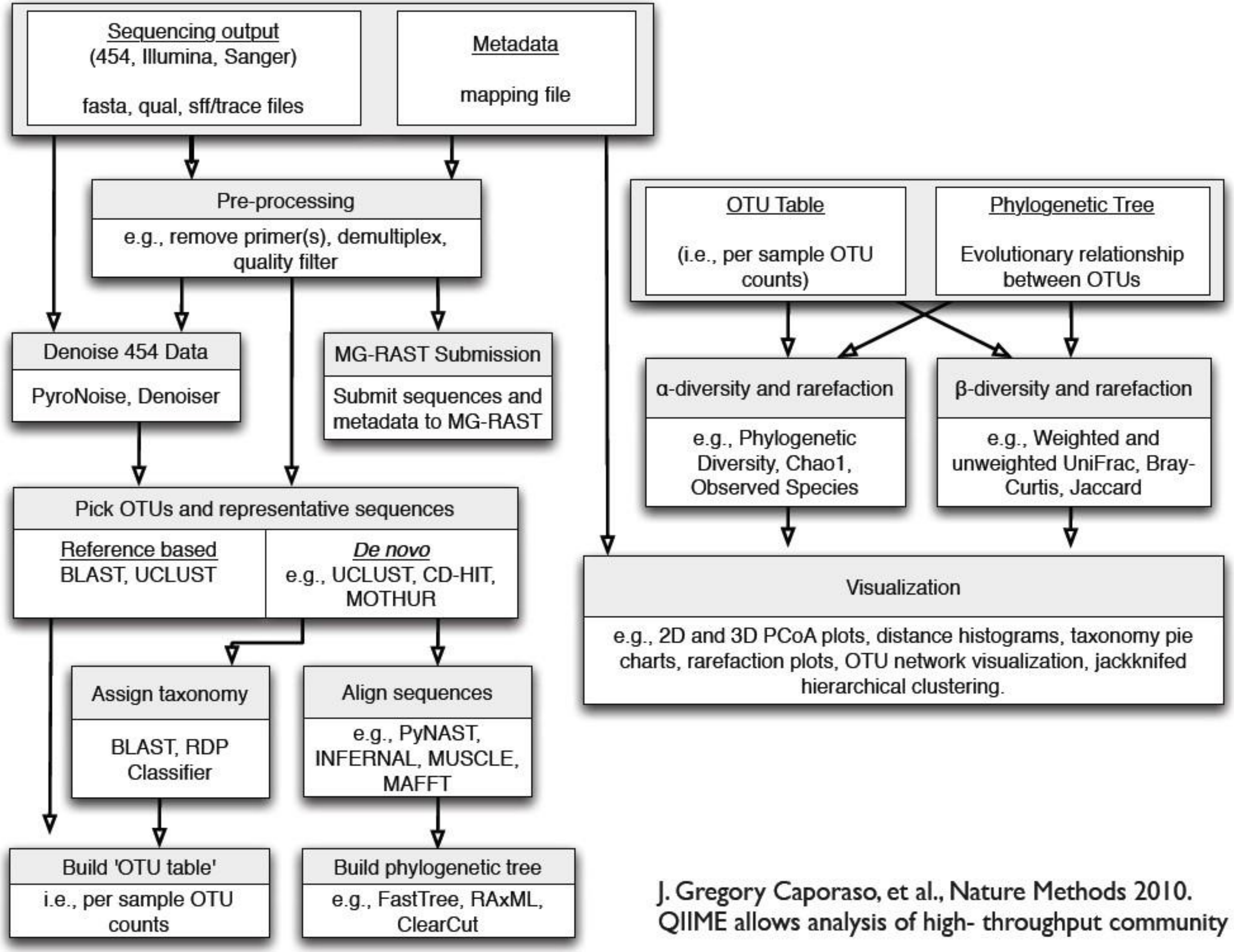


Figure 1 | **Variable regions of the 16S ribosomal RNA.** Secondary structure of the 16S rRNA of *Escherichia coli*, as generated using the xrna program (see Further information). For our analysis, six R fragments of ~250 nucleotides were designed according to the known V regions (Supplementary information S2 (box)). In red, fragment R1 including regions V1 and V2; in orange, fragment R2 including region V3; in yellow, fragment R3 including region V4; in green, fragment R4 including regions V5 and V6; in blue, fragment R5 including regions V7 and V8; and in purple, fragment R6 including region V9.

qiime

Quantitative Insights Into Microbial Ecology



Sequencing output
(454, Illumina, Sanger)
fasta, qual, sff/trace files

Metadata
mapping file

Pre-processing
e.g., remove primer(s), demultiplex, quality filter

Denoise 454 Data
PyroNoise, Denoiser

MG-RAST Submission
Submit sequences and metadata to MG-RAST

Pick OTUs and representative sequences
Reference based
BLAST, UCLUST
De novo
e.g., UCLUST, CD-HIT, MOTHUR

Assign taxonomy
BLAST, RDP Classifier

Align sequences
e.g., PyNAST, INFERNAL, MUSCLE, MAFFT

Build 'OTU table'
i.e., per sample OTU counts

Build phylogenetic tree
e.g., FastTree, RAxML, ClearCut

OTU Table
(i.e., per sample OTU counts)

Phylogenetic Tree
Evolutionary relationship between OTUs

α-diversity and rarefaction
e.g., Phylogenetic Diversity, Chao1, Observed Species

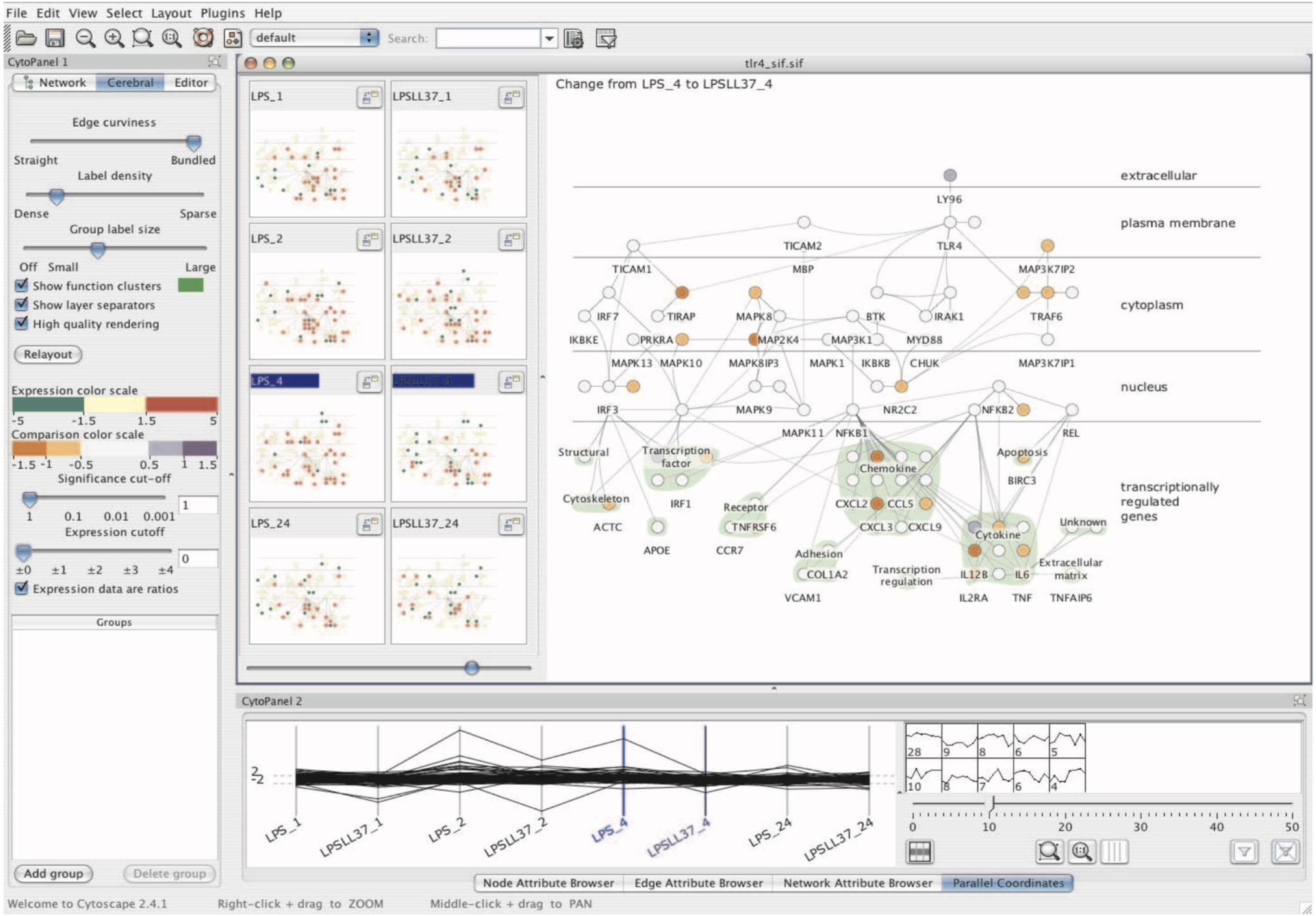
β-diversity and rarefaction
e.g., Weighted and unweighted UniFrac, Bray-Curtis, Jaccard

Visualization
e.g., 2D and 3D PCoA plots, distance histograms, taxonomy pie charts, rarefaction plots, OTU network visualization, jackknifed hierarchical clustering.

J. Gregory Caporaso, et al., Nature Methods 2010.
QIIME allows analysis of high- throughput community sequencing data

BioViz

- How should one visualize all of this complex bio data?
  - Sequences/genes
  - Pathways
  - Multiple –omics
- Why visualize?
  - Help the patient understand their treatment
  - Help clinicians make critical decisions
  - Understand the data and transformations as a bioinformatisist
- Information in this section from: Dr. Carsten Göerg



☑ Show function clusters

☑ Show layer separators

☑ High quality rendering

Significance cut-off 1

Expression cutoff 0

☑ Expression data are ratios

From Barsky, Aaron et. al. (2008)

Knowledge-based Analysis

- Doing a bio study is costly
    - Patient consent and the IRB (prepare to wait)
    - Chemicals (reagents) used in some experiments are expensive
    - How to avoid running redundant experiments and use existing data?
- Many databases exist for bio data
    - Many are small and specialized
- Many papers are submitted to PubMed each year
    - Difficult to parse all of them, even in a narrow field

- Slides in this section from Dr. Larry Hunter

Knowledge-based Analysis

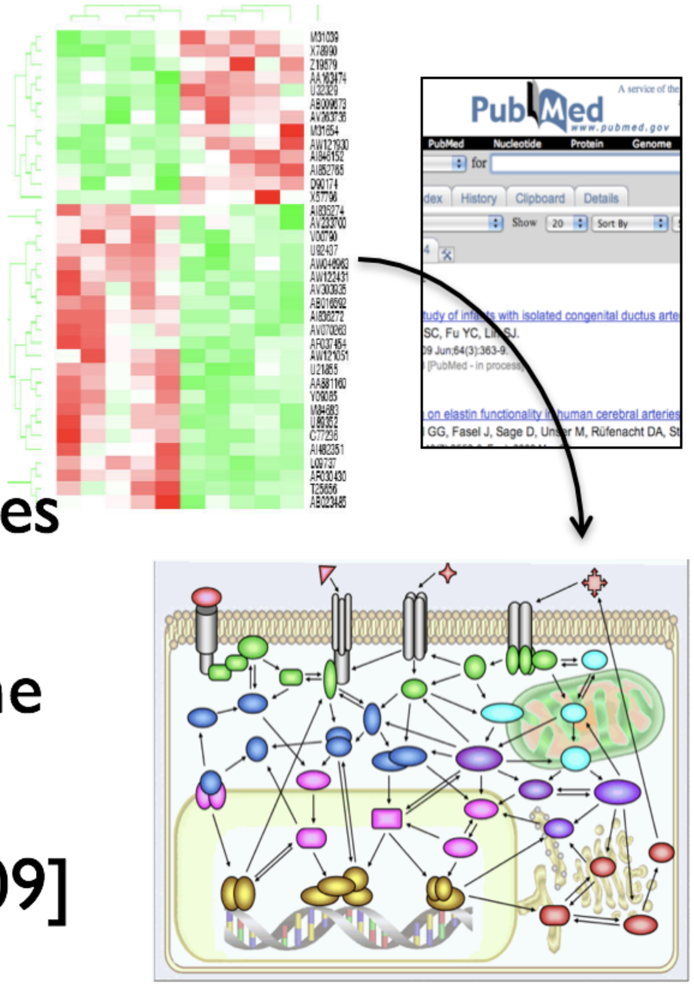
- Build a database that crawls the existing works and builds a graph of **knowledge** about a certain keyword/phrase
- Aggregate the many data sources into one
- Link data through knowledge parsed from PubMed articles
- (Similar to W3’s semantic web)

Analysis is the hard part

- “We are close to having a $1,000 genome sequence, but this may be accompanied by a $1,000,000 interpretation.”
  - Bruce Korf, president American College of Medical Genetics
- Not only is the cost of sequencing essentially free, but big computers and big storage are cheap, too. What will keep us busy for the next 50 years is understanding the data”
  - Russ Altman, chair of Biomedical Engineering at Stanford

Hanalyzer PoC

- Goal: Mixed human/computer approach to knowledge-based approach to analyzing genome-scale datasets.
- Uses graphs to align experimental results with knowledge about genes extracted from many databases (and, increasingly, directly from the biomedical literature).
- [Leach, et al., PLoS Comp Bio 2009] http://hanalyzer.sourceforge.org Search YouTube, for "Hanalyzer"



tudy of infants with isolated congenital ductus arte
SC, Fu YC, Lin SJ.
09 Jun;64(3):363-9.
[PubMed - in process]

on elastin functionality in human cerebral arteries
GG, Fasel J, Sage D, Unser M, Rüfenacht DA, St