VEuPathDB

Eukaryotic Pathogen, Vector & Host Informatics Resources

Crash Course in Omics
Terminology, Concepts & Data Types

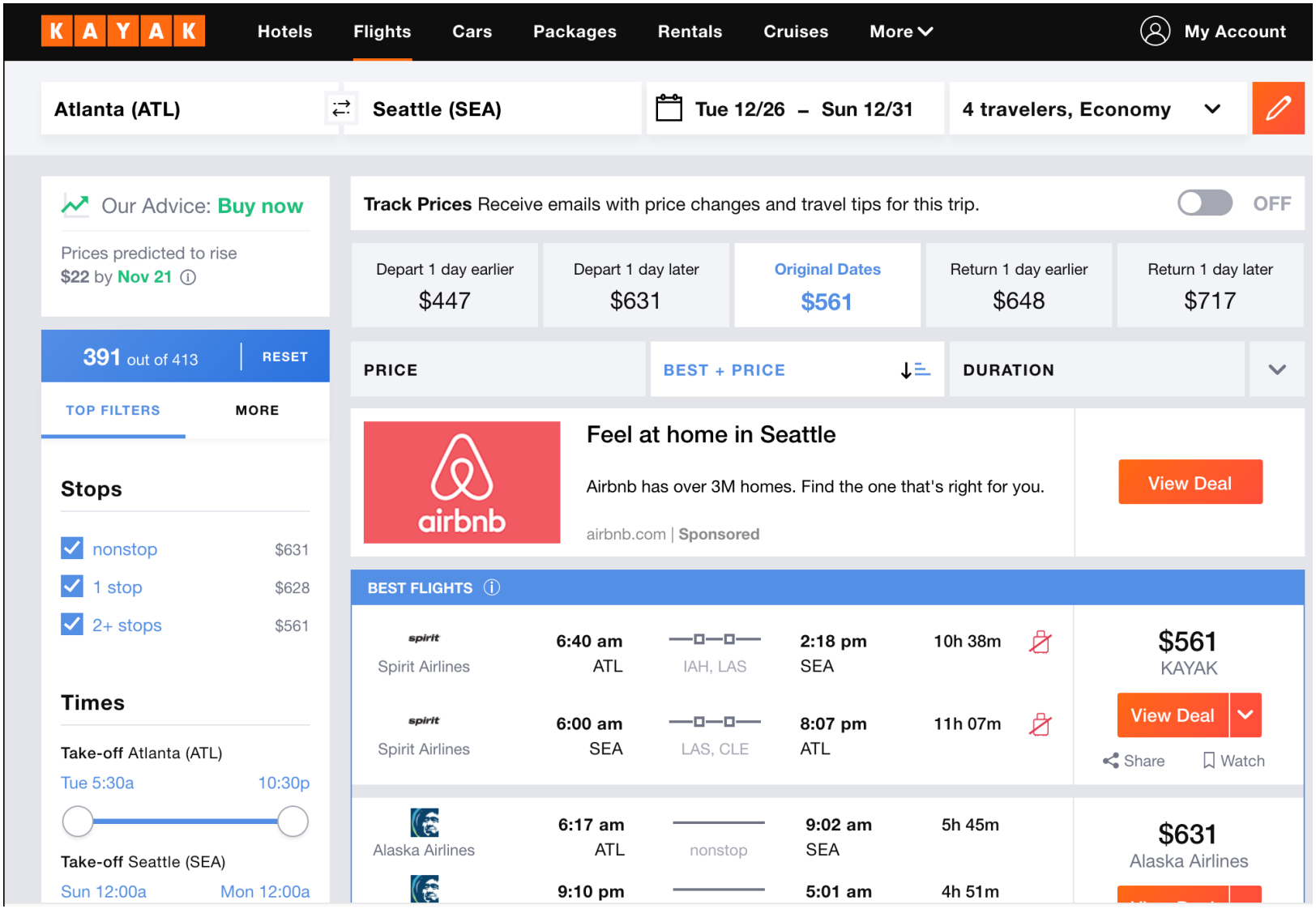
Jessie C Kissinger

2023

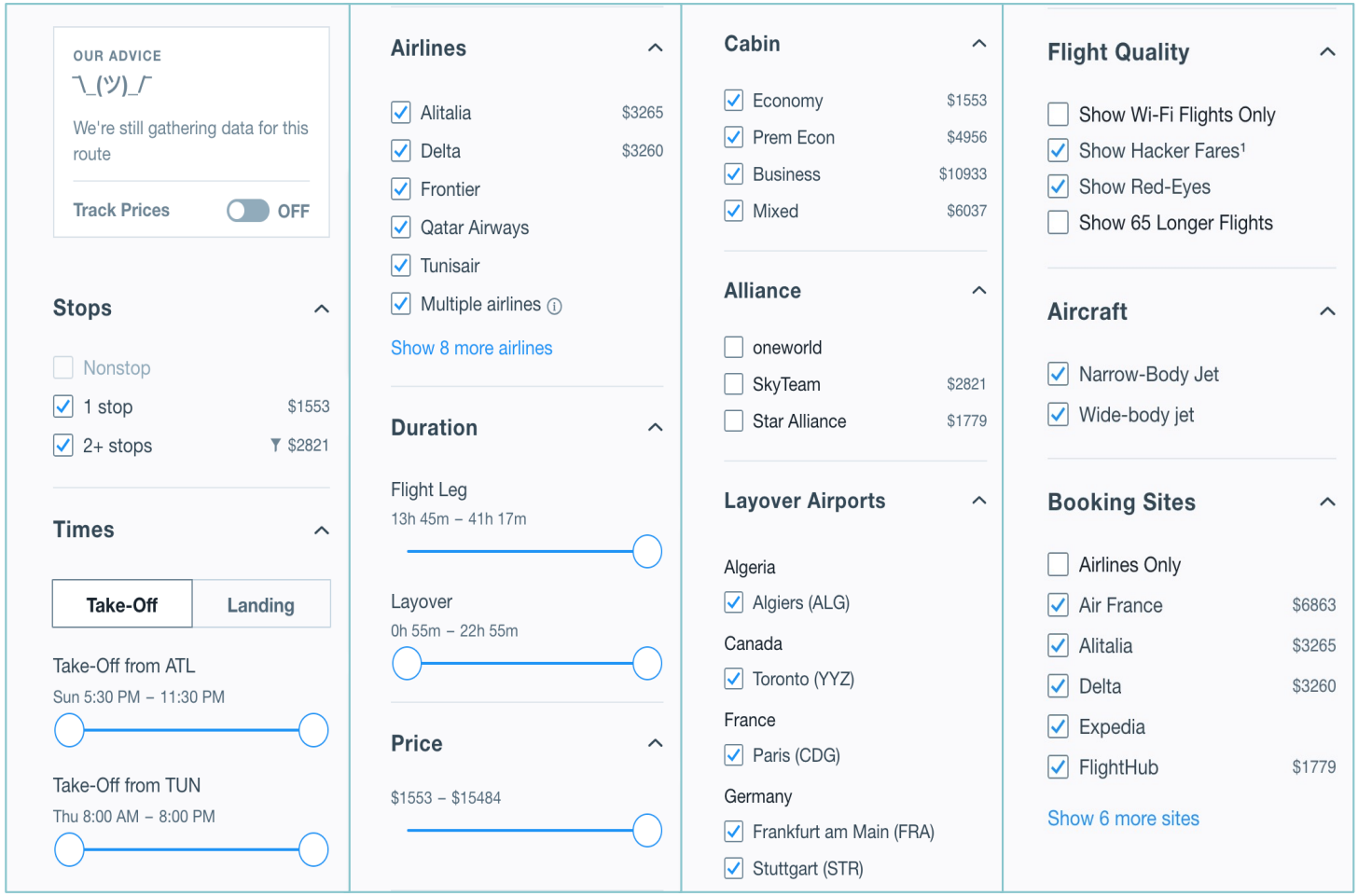
@veupathdb
@jcklab
jkissing@uga.edu

UNIVERSITY OF GEORGIA
Center for Tropical & Emerging Global Diseases

Institute of Bioinformatics
UNIVERSITY OF GEORGIA


KAYAK
Hotels
Flights
Cars
Packages
Rentals
Cruises
More
My Account
Atlanta (ATL)
Seattle (SEA)
Tue 12/26 – Sun 12/31
4 travelers, Economy
Our Advice: Buy now
Prices predicted to rise
$22 by Nov 21
Track Prices Receive emails with price changes and travel tips for this trip.
OFF
Depart 1 day earlier
$447
Depart 1 day later
$631
Original Dates
$561
Return 1 day earlier
$648
Return 1 day later
$717
391 out of 413
RESET
TOP FILTERS
MORE
Stops
nonstop $631
1 stop $628
2+ stops $561
Times
Take-off Atlanta (ATL)
Tue 5:30a
10:30p
Take-off Seattle (SEA)
Sun 12:00a
Mon 12:00a
PRICE
BEST + PRICE
DURATION
airbnb
Feel at home in Seattle
Airbnb has over 3M homes. Find the one that's right for you.
airbnb.com | Sponsored
View Deal
BEST FLIGHTS
Spirit Airlines
6:40 am
ATL
IAH, LAS
2:18 pm
SEA
10h 38m
Spirit Airlines
6:00 am
SEA
LAS, CLE
8:07 pm
ATL
11h 07m
$561
KAYAK
View Deal
Share
Watch
Alaska Airlines
6:17 am
ATL
nonstop
9:02 am
SEA
5h 45m
$631
Alaska Airlines
9:10 pm
5:01 am
4h 51m

The Travel Site has Very Useful Data Filters!



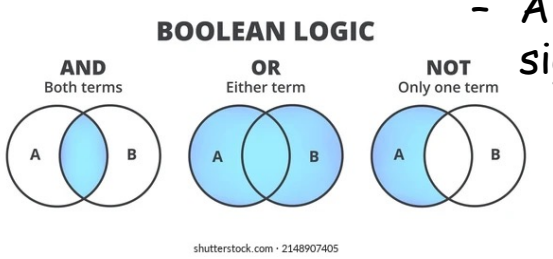
Filters vs Boolean operators

Filters – are very useful, but…

- Can only narrow down the original search
- They only return a subset of the original data
- Examples:
  - All genes on chromosome 4
  - All genes with "kinase in their name
  - All genes from *Trypanosoma cruzi*

Boolean operators (and, or & not)

- Intersect, union, subtract
- They can operate on two different searches!
- They can narrow down or expand the original search
- Examples:
  - All genes on Chr 4 that have kinase in their name
  - All genes on chr 4 or chr 8
  - All genes in *T. cruzi* that also have a signal peptide



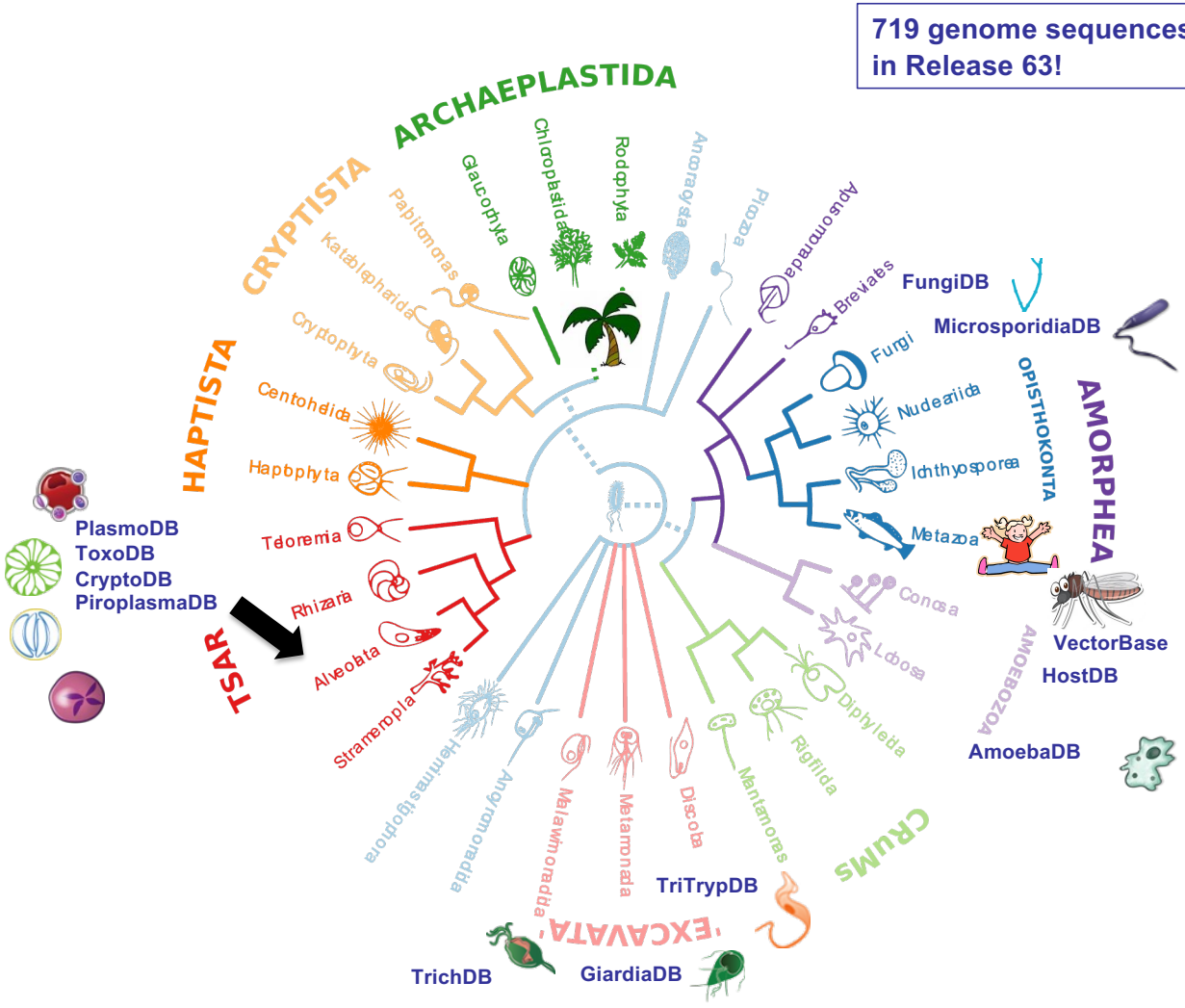
The Biological Equivalent of Travel Search Engine with Filters and Boolean Logic

- Find all genes that….
  - That are near centromeres
  - That encode a predicted signal protein
  - That encode the amino acid motif *CC..CC*
- Which have evidence of expression …
  - In developmental stage X
  - In tissue Y
- That is phosphorylated in proteomic studies
- That show evidence of diversifying selection in population studies

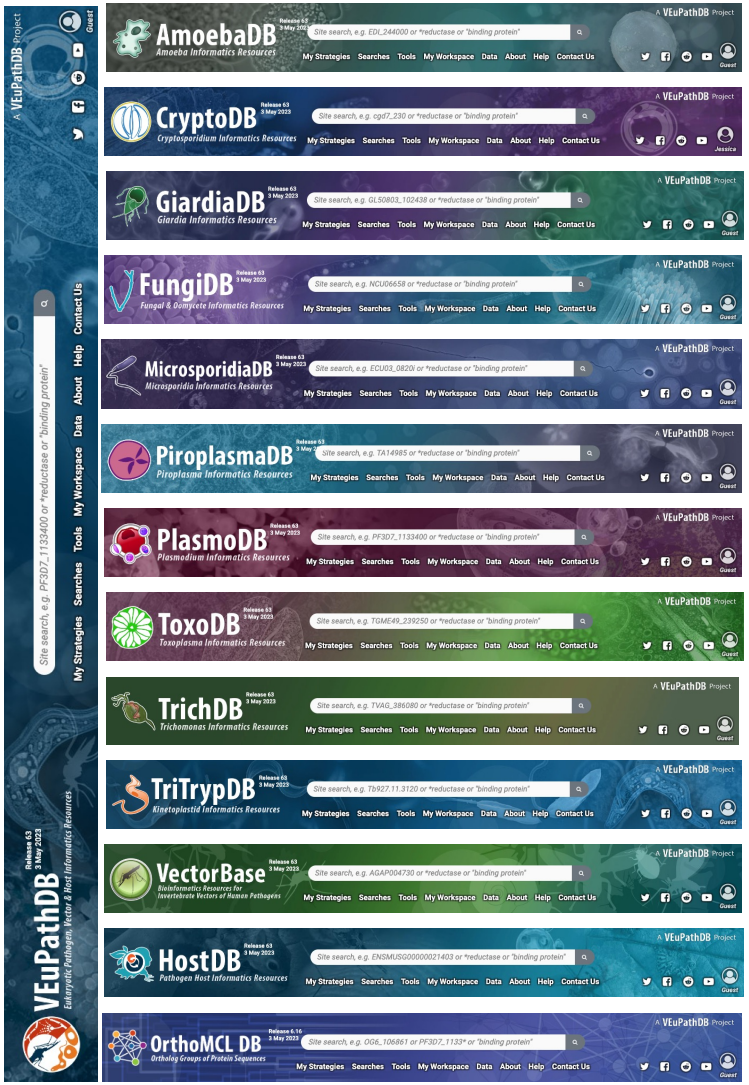
Searching biological data is difficult because there are so many different technologies!

- Each technology e.g. genomics, transcriptomics, proteomics, metabolomics, etc.. has its own vocabulary that is more complicated than selecting a window or aisle seat.
- So,…to use the databases efficiently, you do not need to be a bioinformatician, rather you need to be an expert on the technologies related to the data you will mine so you can use the filters and Boolean operators well and interpret your results.
- Since nobody can keep up with all of the technologies and terminologies, and because we come from so many different backgrounds, we have created this crash course in omics

719 genome sequences in Release 63!



https://commons.wikimedia.org/wiki/File:Eukaryote_Tree_of_Life_2020.svg



Most Genomic terminology in VEuPathDB refers to the following biological concepts:

Genome assembly: Reads, contigs, scaffolds, chromosomes, genome sequences, gaps, indels rearrangements, sequence

Genome annotation: Genes, sequence, coding and non-coding, intergenic regions, untranslated regions, introns, Promoters

Evolution: Sequence differences, SNPs, SNV, InDels, synonymous, non-synonymous, orthologs, paralogs, homology

Chromatin status: Epigenetics, Methylation, open chromatin, closed chromatin

Gene expression: Transcripts, splicing, alternative splicing, differential expression, expression levels (relative or absolute), transcript modifications. Analyses can bulk on a tissue or population of cells/organisms or can be single-cell

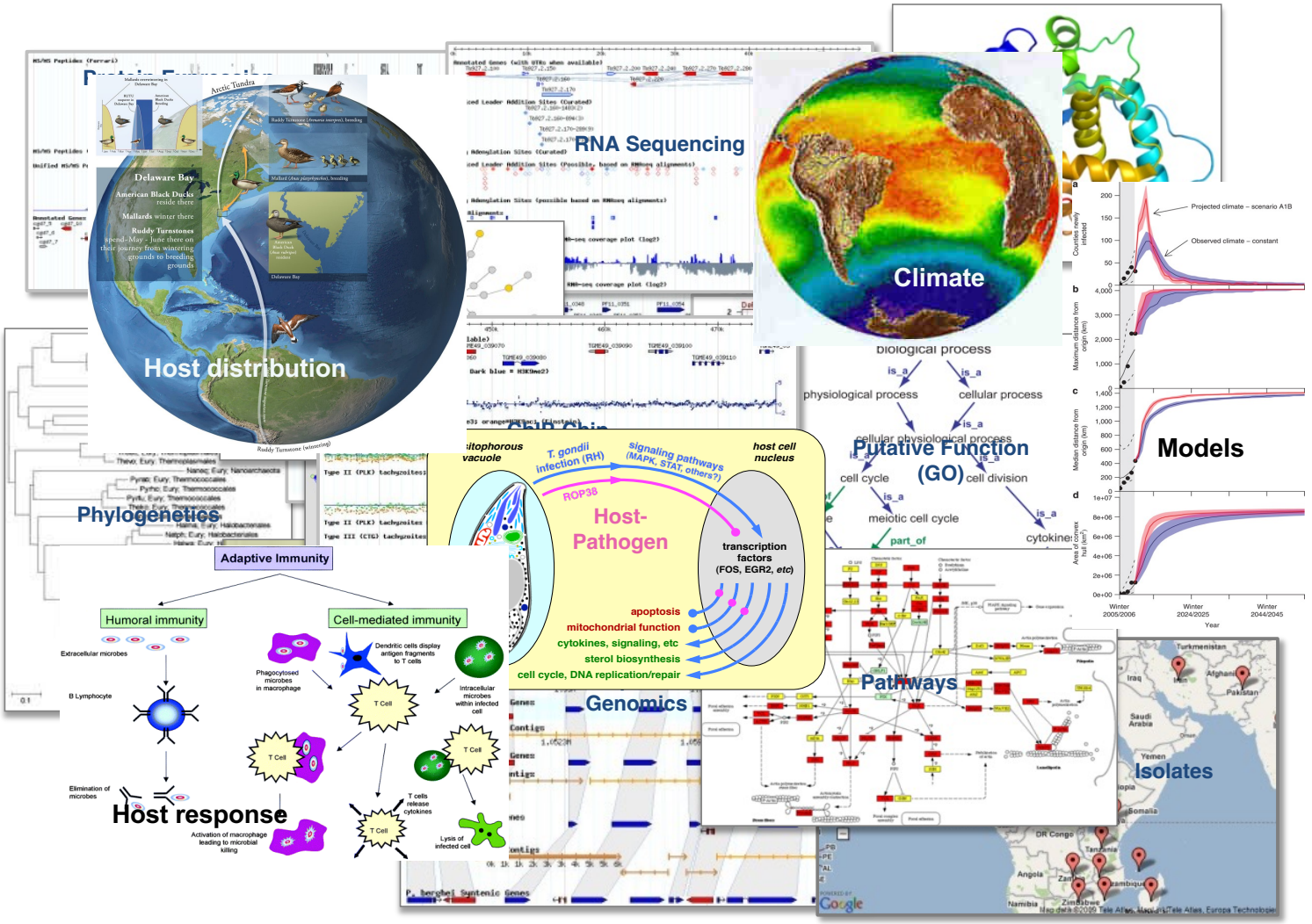
Proteins: sequence, protein features (motifs, signal peptides, TM domains: chemical properties, chemical modifications (phosphorylation, glycosylation), expression, processing, localization

Metabolites: chemical compounds, enzymes, pathways, flux

Host(s): Host response, immune responses, gene regulation responses, metabolic responses

Mutant analysis: phenotypic response to gene knock-down or knock out, e.g. via CRISPR or other approach, or specific mutations

Metadata: data about the data, e.g. the patient, source, environment or experimental condition



Modified from slide provided by David Roos

Genome Assembly 30,000 ft View

FASTQ format for reads

Figure 2 – Flowchart of an NGS workflow

$$Q = -10 \log_{10} P$$

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

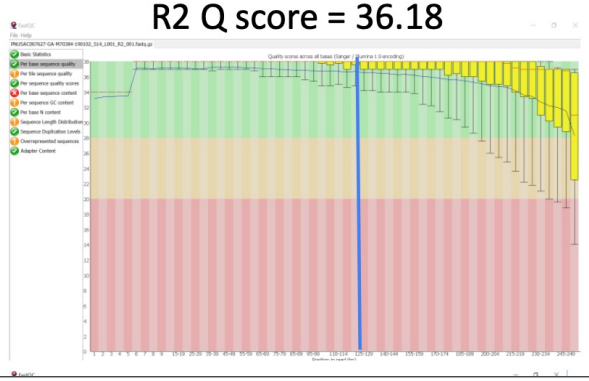
Figure 3 – Phred quality score chart

https://www.abmgood.com/marketing/knowledge_base/next_generation_sequencing_data_analysis.php

FastQC Analysis – Passing Q Scores

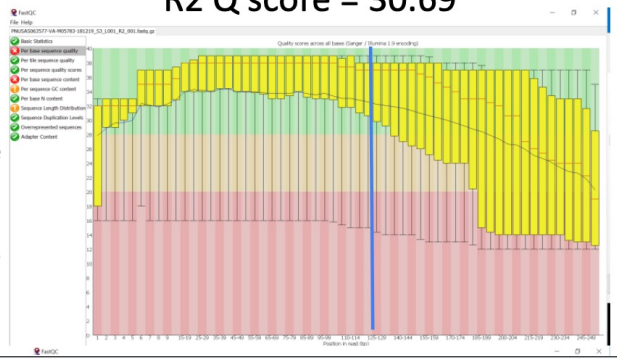
Per Base Sequence Quality & Per Sequence Quality Scores

R2 Q score = 36.18



- Horizontal red line: median Q score
- Horizontal blue line: mean Q score
- Yellow boxes: 50% of the reads
- Whiskers: 80% of the reads
- Vertical blue line: 125 bp of the read

R2 Q score = 30.69

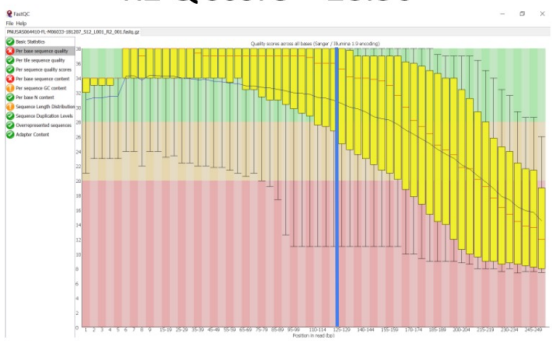


FastQC Analysis – Suboptimal Q Scores (pass with extra coverage)

Per Base Sequence Quality & Per Sequence Quality Scores

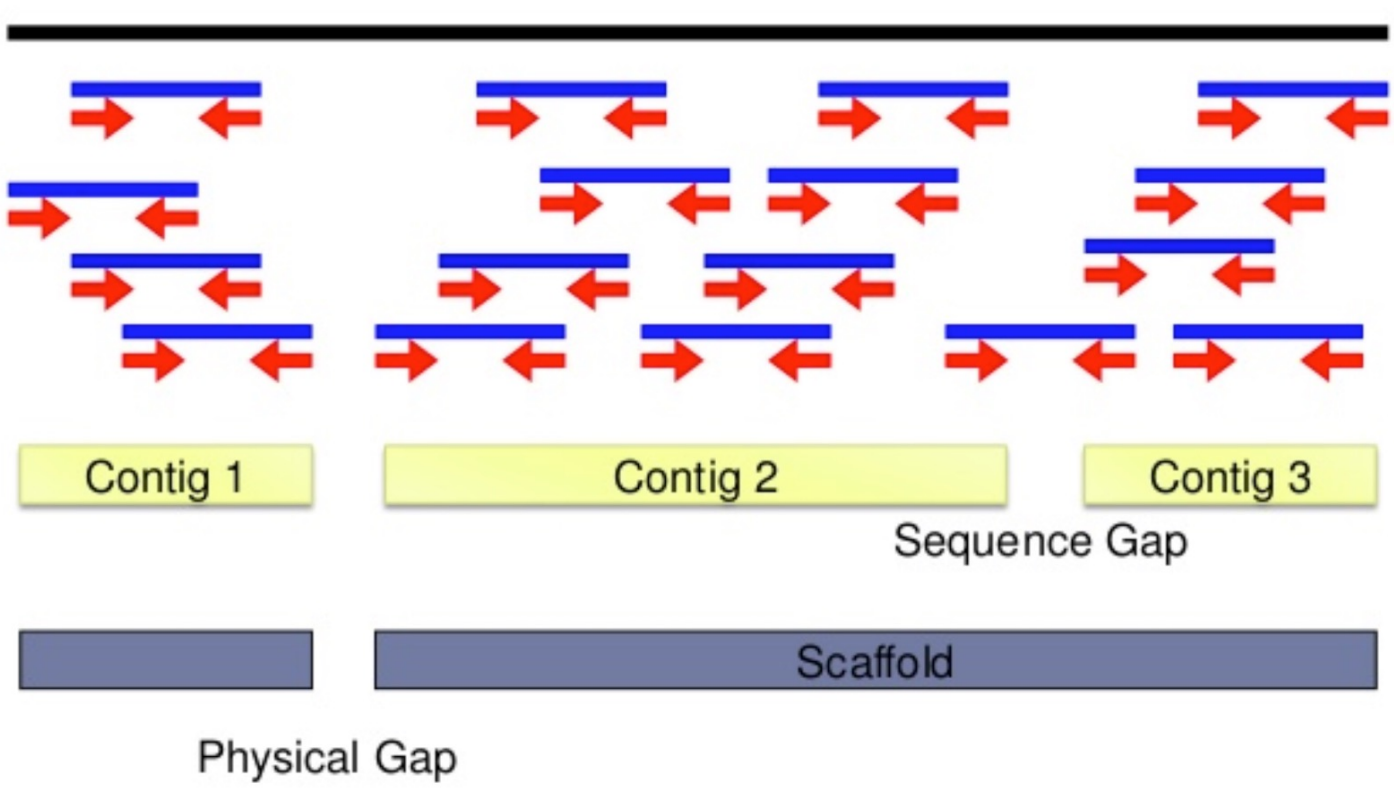
R2 Q score = 29.56

R2 Q score = 28.56



https://www.aphl.org/conferences/proceedings/Documents/2018/4_Eija%20Trees.pdf

A *de novo* Short-Read Paired-End Genome Assembly



DNA Chromosome e.g. “truth”

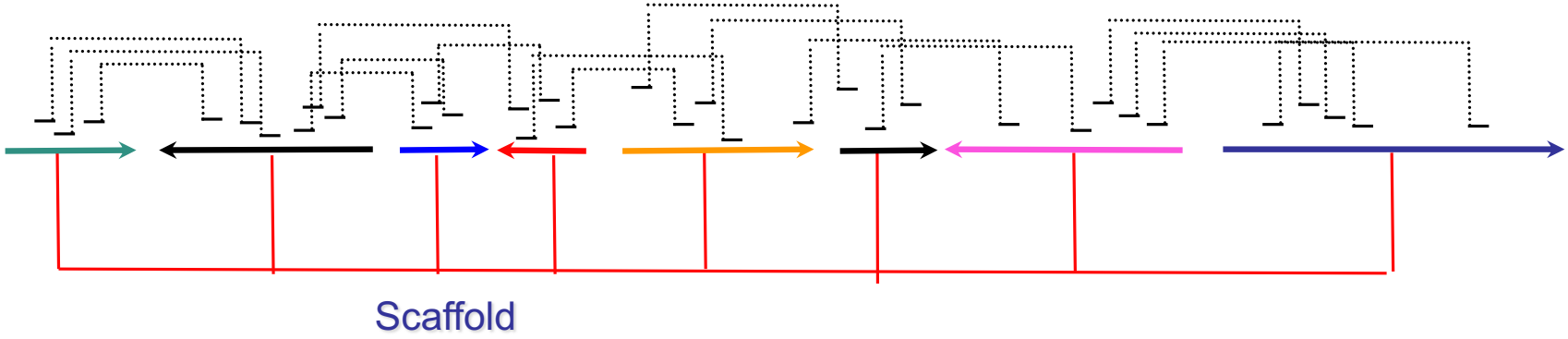
Paired-End Reads, Note the direction of the arrows

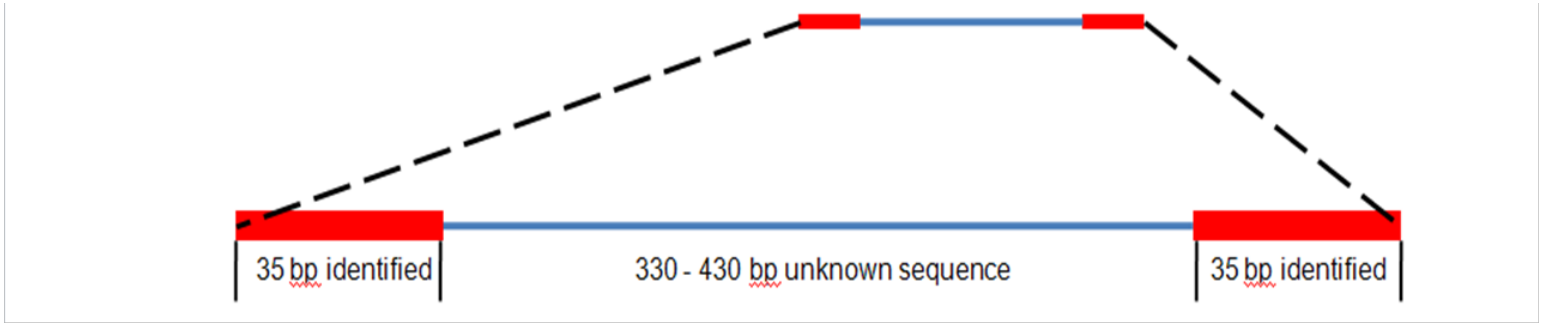
Assembled Contigs

Assembled Scaffolds

https://github.com/Ecological-and-Evolutionary-Genomics/eeg2016/wiki/Mar-21-Exercise-7----SPAdes-assembler

Paired-End Reads can Yield Order & Orientation of Contigs





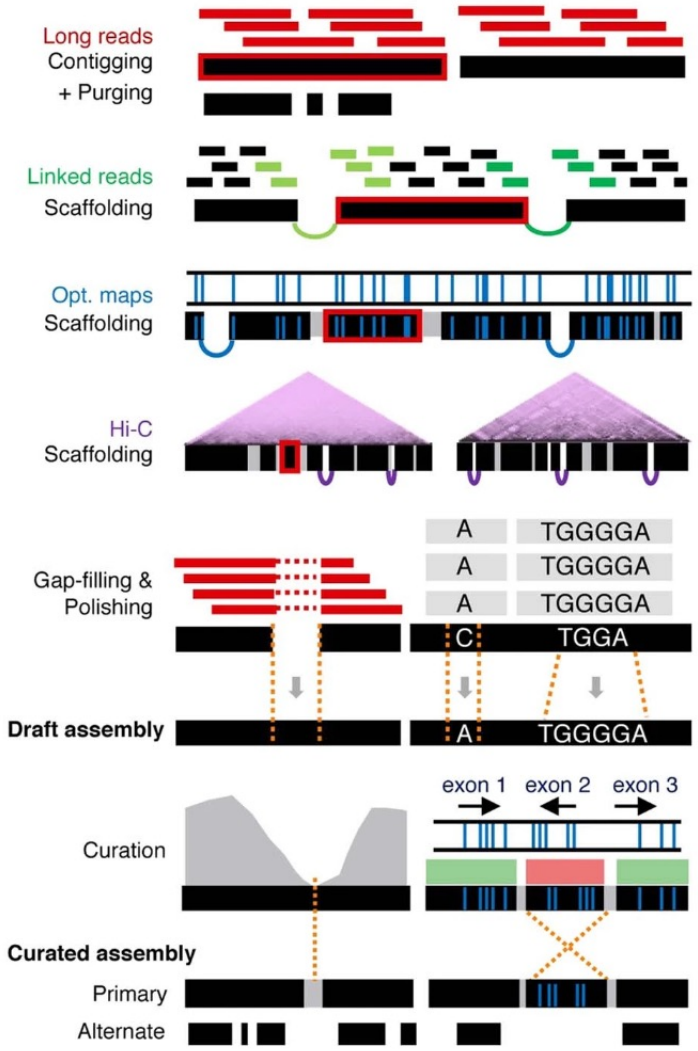
https://www.biostars.org/p/104218/

De novo assembly of a complex genome sequence from scratch requires many technologies:

- **Deep long reads or Long reads and Illumina short-reads**
- **Some form of physical mapping, can be genetic or optical mapping for chromosome interactions captured with Hi-C**

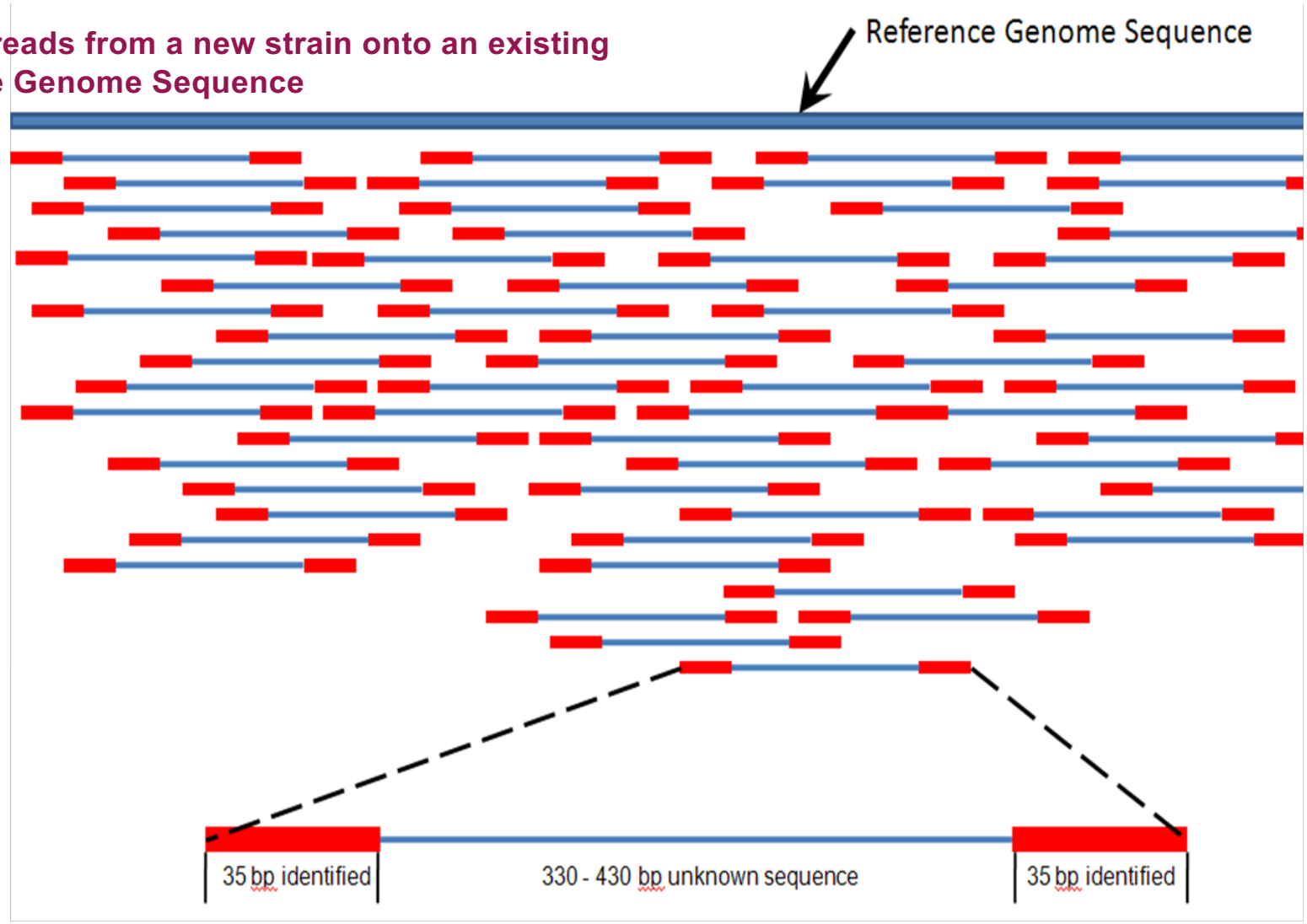
- **All assemblies have gap and these need to be filled and/or corrected this phase is called polishing.**

- **Assemblies should be curated by a human to catch errors of mis-assembly (often apparent when read coverage is low as in the example**



Rhie et al 2021 Nature

Mapping reads from a new strain onto an existing Reference Genome Sequence



https://www.biostars.org/p/104218/

RNA or DNA reads mapped to a reference genome sequence provide insight into “coverage”, the number of reads mapping to a specific region

https://github.com/trinityrnaseq/NaplesWorkshop2016/wiki/Day_2

Hybrid Assembly = short + long reads from differing technologies

It is a VERY useful approach for "correcting" and completing telomere to telomere (T2T) genome assemblies

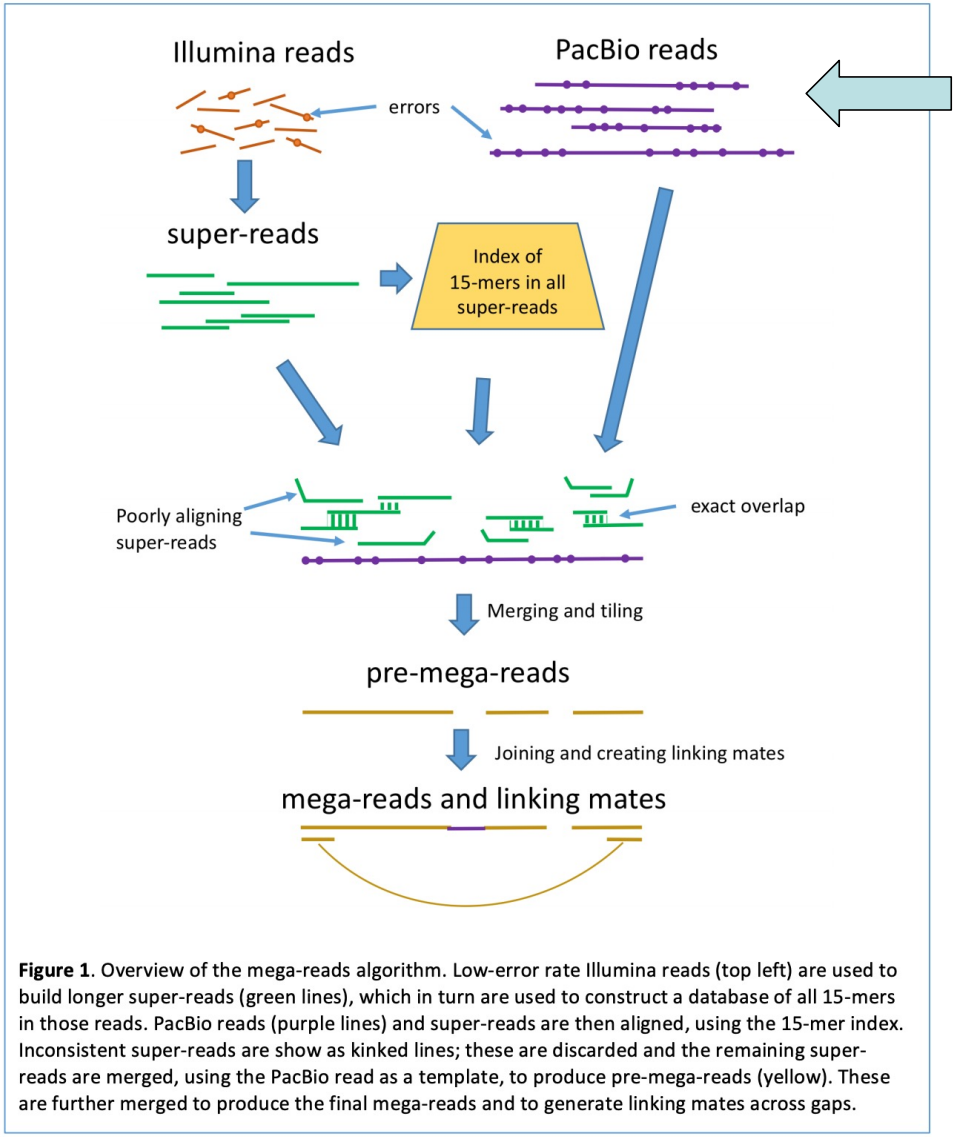


Figure 1. Overview of the mega-reads algorithm. Low-error rate Illumina reads (top left) are used to build longer super-reads (green lines), which in turn are used to construct a database of all 15-mers in those reads. PacBio reads (purple lines) and super-reads are then aligned, using the 15-mer index. Inconsistent super-reads are show as kinked lines; these are discarded and the remaining super-reads are merged, using the PacBio read as a template, to produce pre-mega-reads (yellow). These are further merged to produce the final mega-reads and to generate linking mates across gaps.

Long reads can be PacBio or Oxford Nanopore, ONT

https://genome.cshlp.org/content/early/2017/01/27/gr.213405.116.full.pdf

Genomes:Important Considerations for Assembly and Interpretation

Biological

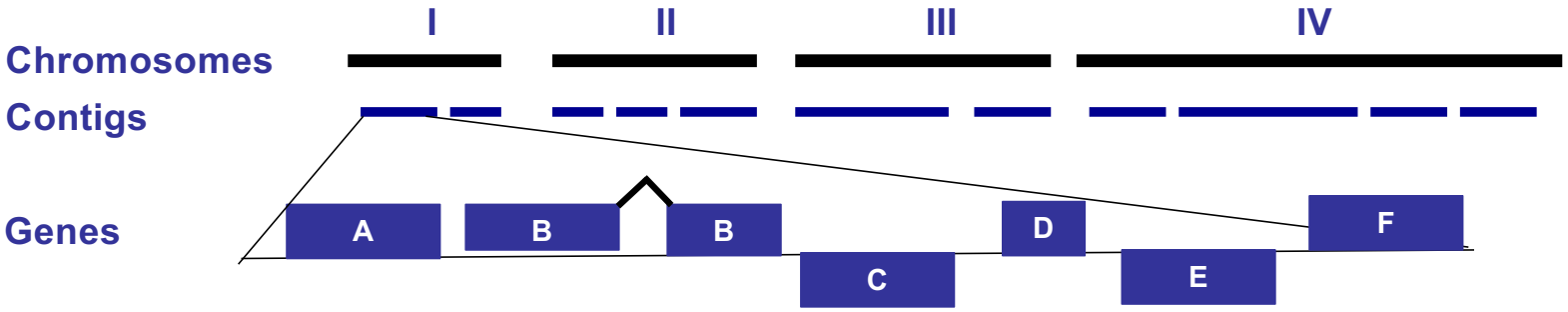
- Size
  - Mb
  - Gb
- Ploidy
  - Haploid
  - Diploid
  - Tetraploid
- Repeat content
  - Retrotransposons
  - Big gene families
  - AT content
- Clone vs population

Technical

- Read length
  - Short
  - Long
- Coverage
  - 5X
  - 100X
- Read Quality
  - 20
  - 30
  - 40
  - Bias?

DNA sequencing technologies: 2006–2016 Elaine R Mardis NATURE PROTOCOLS | VOL.12 NO.2 | 2017 | 213 Seq types chart download (11 MB)
https://www.dropbox.com/s/kfkkft5glmxd68z/ForAllYouSeqMethods.pdf?dl=0 (PDF figures for next two slides)

30,000 ft View – Genome Annotation



The Genome Sequence

AAGCTTCGCCAGGCTGTAAATCCCGTGAGTCGTCCTCACAAATCATCAAGCAGGTGTCCTCAGGGAGACTGCCTGACTGAGTTATGCTAATTCCTTTCTACTTTGGCGTGGTCACGTGTA
ACCATATCCGAATCATTTCTCTAGCCCTACGAACAGGTAAGAGCGCTAGGGATGTCCGTGGAGTAGTGTGCTTACTCGATAATATTCAGTTGGGACTACCAGCGAGGCGCTCGCTTTGCT
CACGCAATGCCTGAGACAGTTGCAGAATGAATGGTAACCGACAAACGCGTTCATATGCGTTTTCAAACTTAGTAGACGCGTACTGTCTGAAACTGGCGGTCACAGGCACCAGATAACGCC
CTTGGCATCGGCATGTCTCGTACAGAGGTCCGTATGTAGTGCCACGACTTCTAAATCCGGCGACAGGCTGGTCTTTTGTCTTACCACGTATTAGCCCGCGTGCGATTTCTCGGAGCGCAC
CTGTTCAACACTAGAAAACGGAGTTTCCTGATCGAGAAGCCACCACCTTTCCAGAAGTTGAACGCTAGCATGTCATTCGATTTTCACCCCCCGCGTAGTTCCTGTGTGTCATTCGTTGTC
GAGACAACTCTGTCCCGCCCCGGTGCTGTTCCATATGCGTGACTTTCCCGCAATTTTTTCAGACTTTCAGGAAAGACAGGCTCCGGAACGATCTCGTCCATGACTGGTAAATCCACGACA
CCGCAATGGCCCCCAGCACCTCTATCTCTCGTGCCAGGGGACTAACGTTGTATGCGTCTGCGTCTTGTCTTTTTGCATTCGCTTTCCAAAAAAGAGAGCCATCCGTTCCCCCGCACATTC
AACGCCGCGAGTGCGGTTTTTGTCTTTTTTGAGTGGTAGGACGCTTTTCATGCGCGAACTACGTGGACATTAAGTTCCATTCTCTTTTTCGACAGCACGAAACCTTGCATTCAAACCCGC
CCGCGGAAGATCCGATCTTGCTGCTGTTCGCAGTCCCAGTAGCGTCCTGTCGGCCGCGCCGTCTCTGTTGGTGGGCAGCCGCTACACCTGTTATCTGACTGCCGTGCGCGAAAATGACGC
CATTTTTGGGAAAATCGGGGAACTTCATTCTTTAAAAGTATGCGGAGGTTTCCTTTTTCTTCTGTTCGTTTCTTTTTCTCGGGTTTGATAACCGTGTTCGATGTAAGCACTTTCCGTCTC
TCCTCCGTGCTTTGTTCGACATCGAGACCAGGTGTGCAGATCCTTCGCTTGTCGATCCGGAGACGCGTGTCTCGTAGAACCTTTTCATTTTACCACACGGCAGTGCGGAGCACTGCTCTG
AGTGCAGCAGGGACGGGTGAAGTTTCGCTTTAGTAGTGCGTTTCTGCTCTACGGGGCGTTGTCGTGTCTGGGAAGATGCAGAAACCGGTGTGTCTGGTCGTCGCGATGACCCCCAAGAGG
GGCATCGGCATCAACAACGGCCTCCCGTGGCCCCACTTGACCACAGATTTCAAACACTTTTCTCGTGTGACAAAAACGACGCCCGAAGAAGCCAGTCGCCTGAACGGGTGGCTTCCCAGG
AAATTTGCAAAGACGGGCGACTCTGGACTTCCCTCTCCATCAGTCGGCAAGAGATTCAACGCCGTTGTCATGGGACGGAAAACCTGGGAAAGCATGCCTCGAAAGTTTAGACCCCTCGTG
GACAGATTGAACATCGTCGTTTCCTCTTCCCTGTGAGCACACACAGTAGTCGCCACACGCTGTTTGAGACGTGTCAATCTCCAAGAGTGTGGACGCTGTTCCACGTCTTCAAATGTTTCC
CAACATCCGTCGTCTAGTAGACACACCAACAAAAAGCACACGGCGAATCTGCTCATCGGAGGGAGGAGCCGGGGGGCACACAACTATCCTCAACTCTCGAACGAACATATCCGGGGCCGC
GAAGACGTCCAGTCTCTCAAATCCAACCCGGAACGCAAACATTTCTGCATCAAGTCACGATTGCGCCGGTACCTCCATGTGTAAGCAGTTCCATGAAACCTCCGATATTACACACGACTG
TGGATATGAATTATATGCAGATGCATATATACTGAGACGCCGATGCAACTATAGGTTTCCTGGCCCTCCATGGATATTTCAGACCTTCCTCTCACATTTGGTTTGCCCGTACACCTCCGT
TACGCTTTTTTTCTGGCTTTCTTCTTCGTCTCTGTTTATCAGCAAAGAAGAAGACATTGCGGCGGAGAAGCCTCAAGCTGAAGGCCAGCAGCGCGTCCGAGTCTGTGCTTCACTCCCAGC
AGCTCTCAGCCTTCTGGAGGAAGAGTACAAGGATTCTGTCGACCAGATTTTTGTCGTGGGTATGTTGTCCTAAACTCCTTGGAACTCCATTCTTGGTCAGAAACGTACTGAAACTGTATA
CATGTATATACAGATGTATGGATAATATCTAGAGAAGATACAGGGAAGACTGGCAAGGATGAAAAGACATGCAGCTTTAACGAAGCAGAGGGCATTGGCGAGAGGGACGCCCGTTATGCT
GTGTGATGTGGCTGTGAATCTTACCTCGCCGTTTGACTTGCTGCAGCGCTTTGTCCACTTGAACGTGACTTCTTGTTTCTACCTTCCCCAACGCCTTCTATTCCCTTCACTGCGAAAGCG
CGCTCAGTGGGCCGTCACCGAACACCCTTGGTTCTTTCGTTCAGCTGTTGTCCTCTTTCTCGCGTTGCTTCCTGTGGCGTCGTGGCTCGGCTTCTCTCTCTTTCCTGTTGGTGCGTCCAG
ACTATGTCGCCTGTTTCCCCACCCTTCTCGGCTTGTGCTTTCAGGAGGAGCGGGACTGTACGAGGCAGCGCTGTCTCTGGGCGTTGCCTCTCACCTGTACATCACGCGTGTAGCCCGCGA
GTTTCCGTGCGACGTTTTCTTCCCTGCGTTCCCCGGAGATGACATTCTTTCAAACAAATCAACTGCTGCGCAGGCTGCAGCTCCTGCCGAGTCTGTGTTCCGTTCCCTTTTGTCCGGAGCT
CGGAAGAGAGAAGGACAATGAAGCGACGTATCGACCCATCTTCATTTCCAAGACCTTCTCAGACAACGGGGTACCCTACGACTTTGTGGTTCTCGAGAAGAGAAGGAAGACTGACGACGC
AGCCACTGCGGAACCGGTAAGAGGCAACCGAAGCGCGTAGATAAGAAAAACAACAAAGAGAAGGTGAAACACGAAGAGAAGGGAAAATGCGGAGAAACCGTGGATTTACAAAGATATCAA
GAGCAATGCTTTGTGGAGATTTTTTTTAATTCAGTAGAGACACCCGCCGTGCGAGGTGTGTAGAAATAACTGCGACCCTGGAGACAGAGATGCCGCGAGTACACCACTTGTCGTTTTTCC
TCCTATGTTCATGACGGGTGCTGAACGTCTATCGTACTTAATTGGAGGAGTCGTCTCCGAAGCAGCTTTGGCTGGCCATCCGTGTGTTTGCCTTGTTCCTGAAAAGCCAGAAGGCGCTCC
ACAGTGAGGCGATATACAGGGACGCCTACCGGAGCCCCGTTTTCTGCCTTTGTCGACTCTTGCAGAGCAACGCAATGAGCTCCTTGACGTCCACGAGGGAGACAACTCCCGTGCACGGGT
TGCAGGCTCCTTCTTCGGCCGCAGCCATTGCCCCGGTGTTGGCGTGGATGGACGAAGAAGACCGGAAAAAACGCGAGCAAAAGGAACTGATTCGGGCCGTTCCGCATGTTCACTTTAGAG
GCCATGAAGAATTCCAGTACCTTGATCTCATTGCCGACATTATTAACAATGGAAGGACAATGGATGACCGAACGGGTAACGGCGACTGCGAGAAAAAGCCACACCGTTTTCTCCTGTGAT
TCTGTCCGCAAGCCCTCTTTTGCTTCATCCACCCTTTGCTATTCTCCGCCGCCTTCCTTTTCTGCTCCATGTTCAATTCGTTCGCTTCTTCAGTCTTTCCATCTTCCCCTGTTACCTCTG
TCATTCGTTTTCTTGCCTCTATTTAACTGTGTTCTACTCACAGTCTGCATTCCGCGATAGACGAGCTTCCACGTCTTGCGTCTCGACAAGCAACTGTCATTTGTACGCGCCTCCCTCCAC
CGTGAATCGGATTGTCGGTTCGCCGGTTCCTGGGTCAGAAAAGGCCTGCGCCAGTATTCTGAATAATACCCTTCGCCATTGTAAAGAGGCGAAGGAACAAAGAGATATTTCGGCGCATCT
TTTGTGCGGCGCGTTTCCTCGTGCTTCACACCGATGCCCTTCTGTGCATGTCTTCTGCTCCTCGTCCTTCTCTCTTTTTCCCTGTTTAGGCGTTGGTGTCATCTCCAAATTCGGCTGCAC
TATGCGCTACTCGCTGGATCAGGCCTTTCCACTTCTCACCACAAAGCGTGTGTTCTGGAAAGGGTAAGGGCGTCTTCAGTGAATGCATATATTTGACTTCAGACATTCTTAACTGTTTGA
CAACCAACGTACAAATTTGTTTGTCCGTGTGCGTGTTCGACATGTCAAGTATGTGAAGAGTCGCTACTGTAGACTAACGCACGAACCAGATTTGTTTATCTGCATGCGCTGTGCACCCGT
TTCTGAGTGTCTGGAGTTTCCGCAACCTTCCTTTGAATTTCTGGGTTCGTTTTTTTATGCGCGCACTGGTTTGCATGTGGCCTGAGAGAGCACAGATCGAAGGTGGGGTGATGTGGCGTC
GCTGCAGAGAAACTCCGGCGAAGGCGACAGATAAAGGAGAGTGGAAATCATTGAACAGTGTCGGTCGTCTGTTGTTTCGCAGGGTCCTCGAAGAGTTGCTGTGGTTCATTCGCGGCGACA
CGAACGCAAACCATCTTTCTGAGAAGGGCGTGAAGGCAAGTCTACGTTGTACCTCTTGTCTCTGCCGAAGCTCAGATGTCTCCACGGCGTTGGTTTCTTTTCGTTTTTGCTTTCGTGGCA
TTACCATCGAGTCACCACTCATAGTTGCGTGTGTCTACATGTTTTCTAGAACGTCCGTTGTGTTGCCTCGTGGCGACCGGCGCGAGTGTATGTACCCTGCGCTGTCAGAAGTTGATCCTT

Genes can be located on either DNA strand Convention - Gene location = non-template strand, i.e. the sequence of the gene is the same as the mRNA (except U = T in DNA)

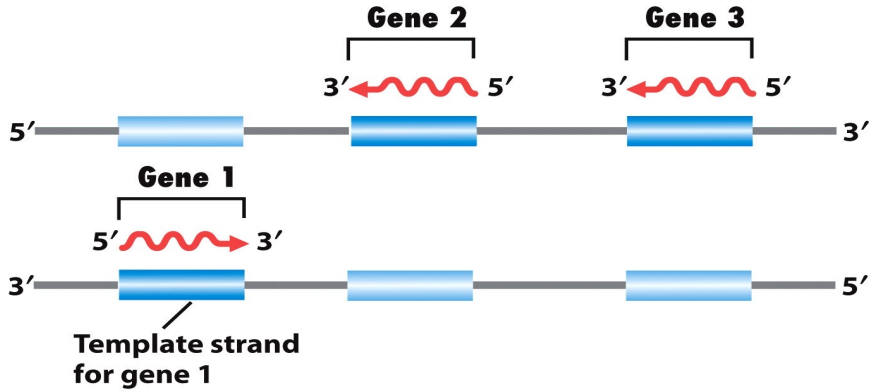


Figure 8-3
Introduction to Genetic Analysis, Ninth Edition
© 2008 W. H. Freeman and Company

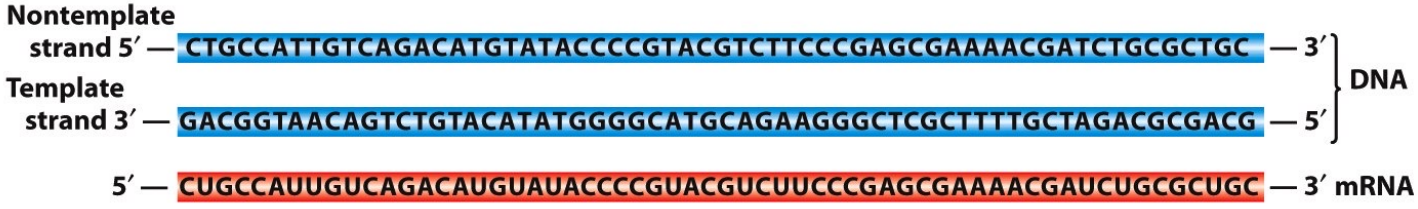
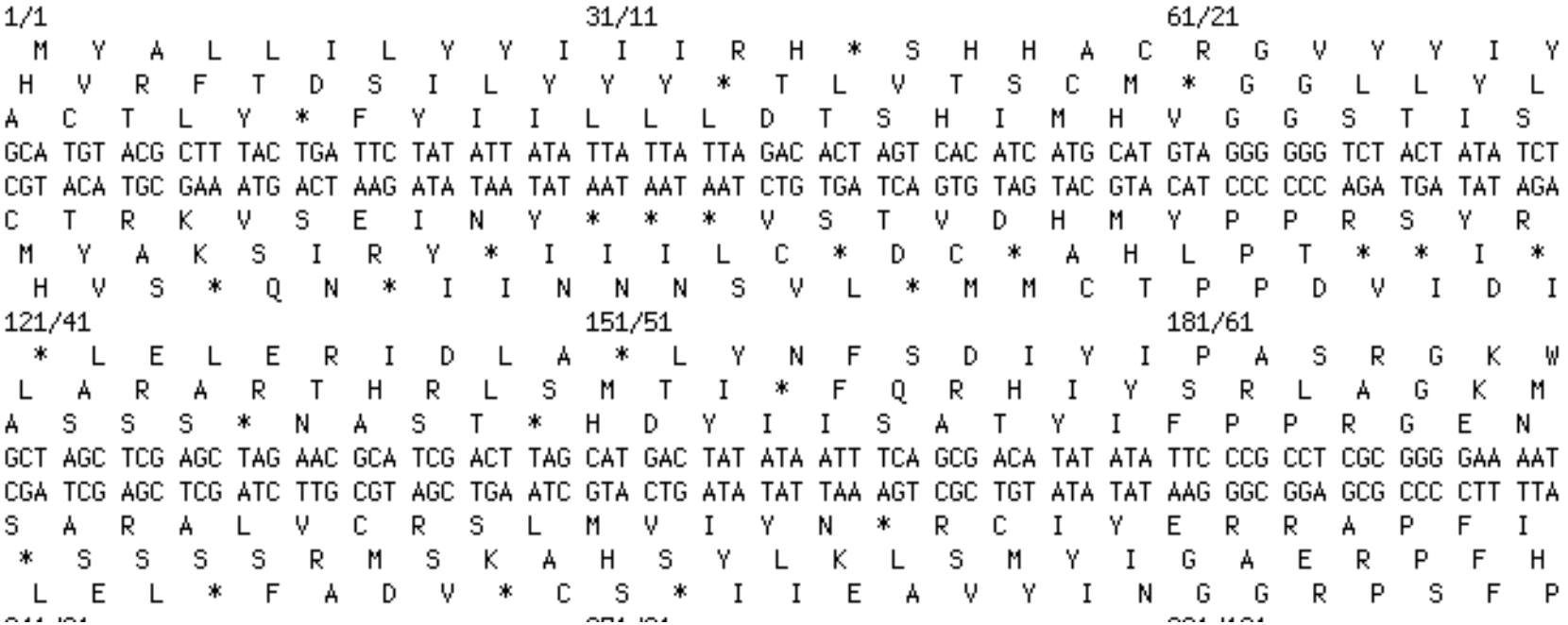


Figure 8-6
Introduction to Genetic Analysis, Ninth Edition
© 2008 W. H. Freeman and Company

Six Frame Translation
Looking for Open Reading Frames, ORFs



ORFs ≠ Genes – but they can be part of a gene

Green = UTRs Red = CDS Pink = Intron

The "Coding Sequence" - CDS

ATGCAGAAACCGGTGTGTCTGGTCGTCGCGATGACCCCCAAGAGGGGCATCGGCATCAACAACGGCCTCCCGTGGCCCC
ACTTGACCACAGATTTCAAACACTTTTCTCGTGTGACAAAAACGACGCCCGAAGAAGCCAGTCGCCTGAACGGGTGGCT
TCCCAGGAAATTTGCAAAGACGGGCGACTCTGGACTTCCCTCTCCATCAGTCGGCAAGAGATTCAACGCCGTTGTCATG
GGACGGAAAACCTGGGAAAGCATGCCTCGAAAGTTTAGACCCCTCGTGGACAGATTGAACATCGTCGTTTCCTCTTCCC
TCAAAGAAGAAGACATTGCGGCGGAGAAGCCTCAAGCTGAAGGCCAGCAGCGCGTCCGAGTCTGTGCTTCACTCCCAGC
AGCTCTCAGCCTTCTGGAGGAAGAGTACAAGGATTCTGTCGACCAGATTTTTGTCGTGGGAGGAGCGGGACTGTACGAG
GCAGCGCTGTCTCTGGGCGTTGCCTCTCACCTGTACATCACGCGTGTAGCCCGCGAGTTTCCGTGCGACGTTTTCTTCC
CTGCGTTCCCCGGAGATGACATTCTTTCAAACAAATCAACTGCTGCGCAGGCTGCAGCTCCTGCCGAGTCTGTGTTCGT
TCCCTTTTGTCCGGAGCTCGGAAGAGAGAAGGACAATGAAGCGACGTATCGACCCATCTTCATTTCCAAGACCTTCTCA
GACAACGGGGTACCCTACGACTTTGTGGTTCTCGAGAAGAGAAGGAAGACTGACGACGCAGCCACTGCGGAACCGAGCA
ACGCAATGAGCTCCTTGACGTCCACGAGGGAGACAACTCCCGTGCACGGGTTGCAGGCTCCTTCTTCGGCCGCAGCCAT
TGCCCCGGTGTTGGCGTGGATGGACGAAGAAGACCGGAAAAAACGCGAGCAAAAGGAACTGATTCGGGCCGTTCCGCAT
GTTCACTTTAGAGGCCATGAAGAATTCCAGTACCTTGATCTCATTGCCGACATTATTAACAATGGAAGGACAATGGATG
ACCGAACGG

The Protein

>Translation Frame 1

MQKPVCLVVAMTPKRGIGINNGLPWPHLTTDFKHFSRVTKTTPEEASRLN

GWLPRKFAKTGDSGLPSPSVGKRFNAVVMGRKTWESMPRKFRPLVDRLNI

VVSSSLKEEDIAAEKPQAEGQQRVRVCASLPAALSLLEEEYKDSVDQIFV

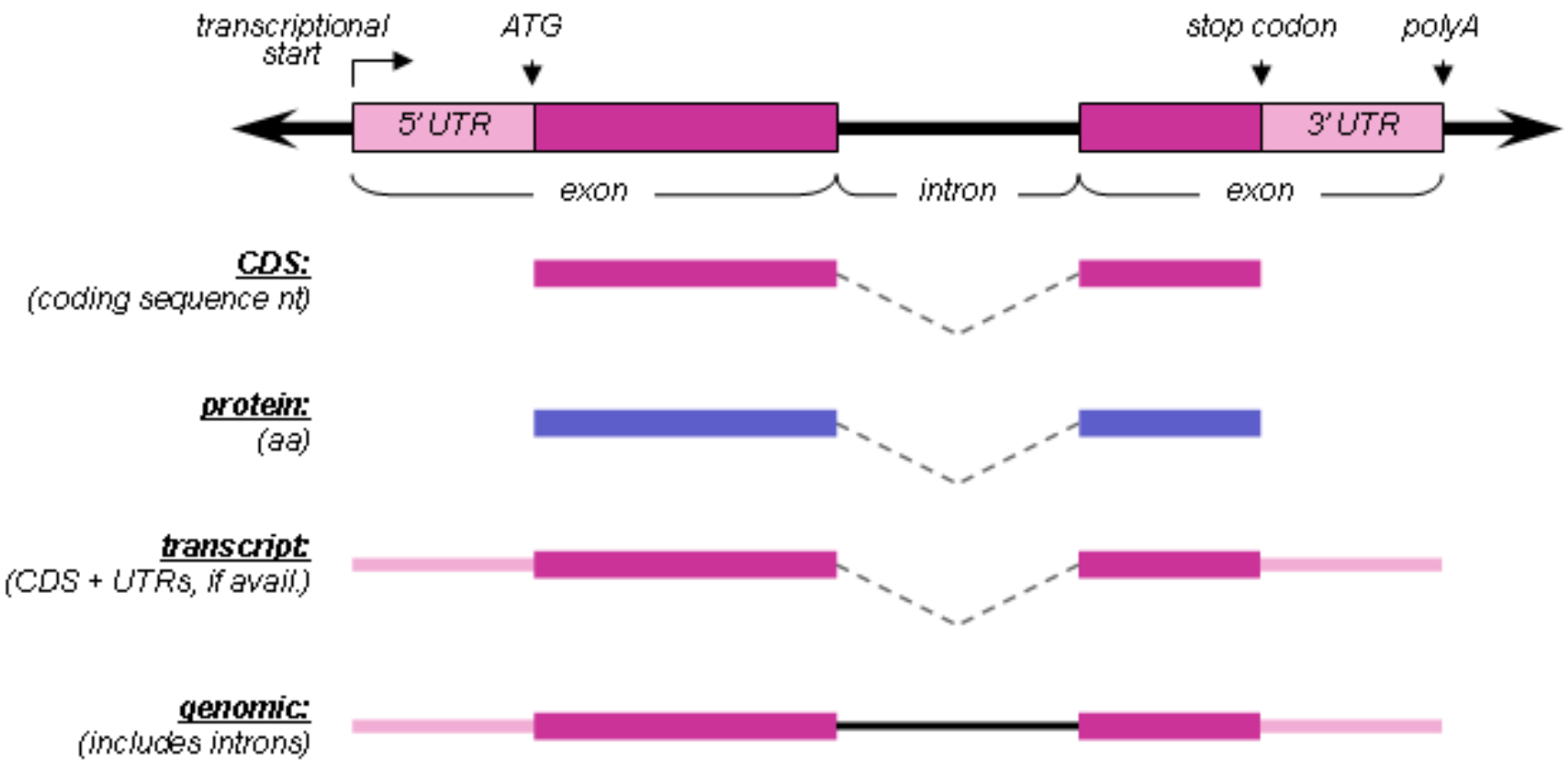
VGGAGLYEAALSLGVASHLYITRVAREFPCDVFFPAFPGDDILSNKSTAA

QAAAPAESVFVPFCPELGREKDNEATYRPIFISKTFSDNGVPYDFVVLEK

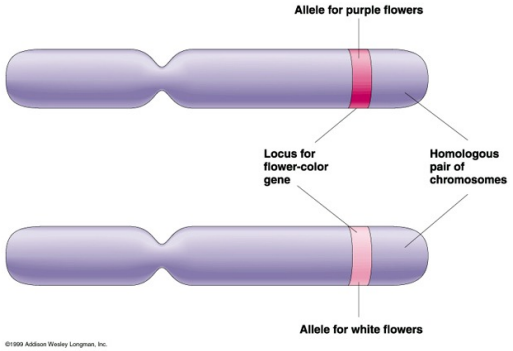
RRKTDDAATAEPSNAMSSLTSTRETTPVHGLQAPSSAAAIAPVLAWMDEE

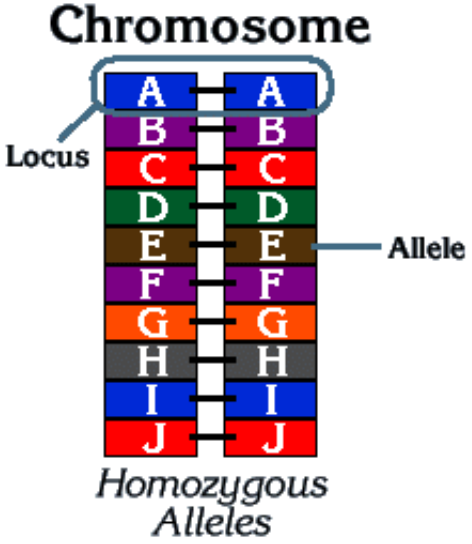
DRKKREQKELIRAVPHVHFRGHEEFQYLDLIADIINNGRTMDDRT

Terminology



Evolution Homologous chromsomes (in a diploid)





A AAGCCTCATC

a ACGCCTCATC

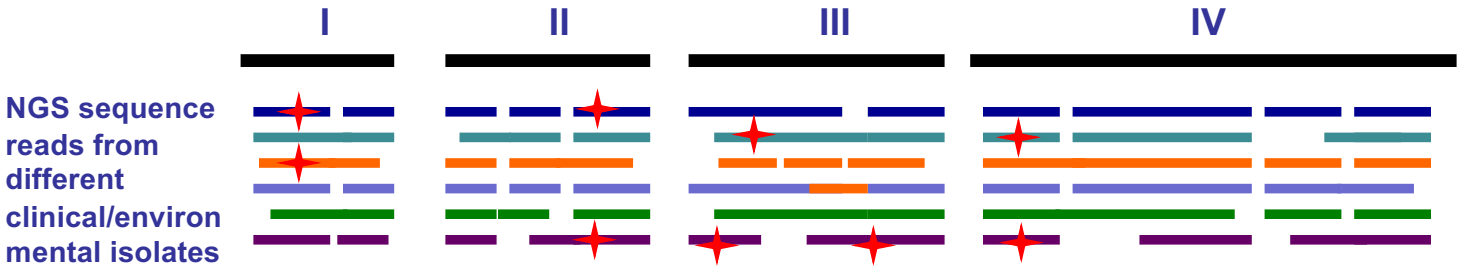
SNP =Single Nucleotide Polymorphism (a variant)

Alleles and Phenotype

- Some phenotypes are caused by a single locus in the genome and a single allele at that locus (e.g. some flower colors, or *Drosophila* eye color)

- Other phenotypes (Type-I diabetes, heart disease are multi-locus or "complex" (i.e. many genes are involved, each potentially with many alleles)

30,000 ft View- NGS SNPs



= SNP
= SNV

```
reference     TGGTGATACT AAGCTGGGAA CTCCACTTCT TTTTCTACTG CGGTGCTTCA
  303.1       TGGTGATACT AAGCTGGGAA CTCCACTTCT TTTTCTACTG CGGTGCTTTA
  309.1       TGATAATNCT AAACTGGGAA CTCCACTTCC TTTTCTACTG CAGTGCTTCA
 RV_3600      TGGTGATACT AAACTGGGAA CTCCACTTCT TTTTCTACTG CGGTGCTTCA
 RV_3606      TGATAATNCT AAACTGGGAA CTCCACTTCC TTTTCTACTG CAGTGCTTCA
 RV_3610      TGATGATTCT AAACTGGGAA CTCCACTTCC TTTTCTACTG CAGTGCTTCA
SenT119.09    TGGTGATACT AAACTGGGAA CTCCACTTCT TTTTCTACTG CGGTGCTTCA
SenT123.09    TGATRATTCT AAACTGGGAA CTCCACTTCC TTTTCTACTG CAGTGCTTCA
SenT140.08    TGGTGATACT AAACTGGGAA CTCCACTTCC TTTTCTACTG CGGTGCTTCA
SenT142.09    TGGTGATACT AAACTGGGAA CTCCACTTCC TTTTCTACTG CAGTGCTTCA
SenT175.08    TGGTGATACT AAACTGGGAA CTCCACTTCT TTTTCTACTG CGGTGCTTTA
```

Reference = A
6 isolate seq = A
2 isolate seq = T
2 isolate seq = N (no call)
% with base call = 80
Major allele = A
Major allele freq = 75% (6/8)

Reference = G
9 isolate seq = A
1 isolate seq = G
% with base call = 100
Major allele = A
Major allele freq = 90% (9/10)

Reference = C
8 isolate seq = C
2 isolate seq = T
% with base call = 100
Major allele = C
Major allele freq = 80% (8/10)

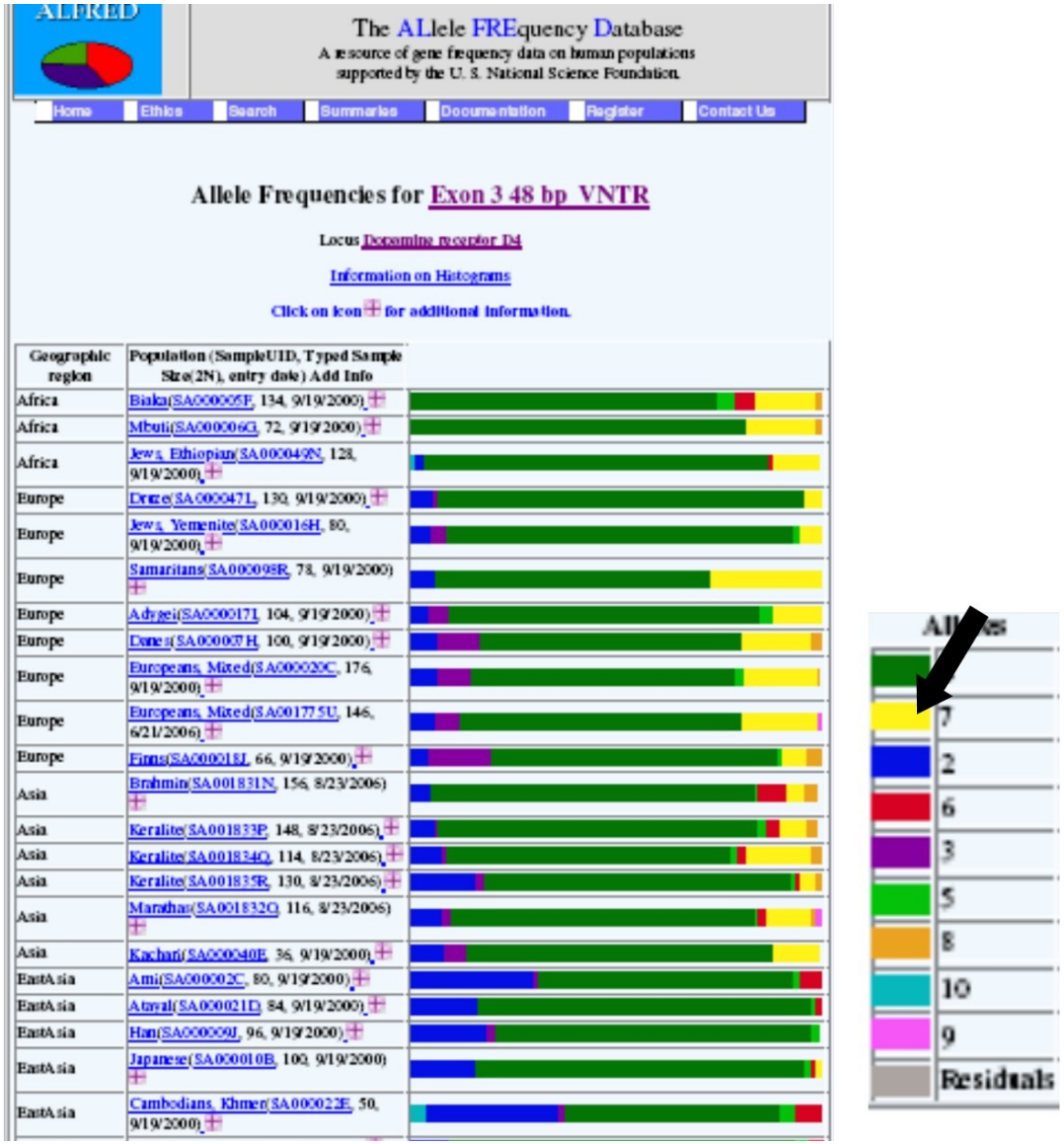
Population variation data

Data

- Single Nucleotide Polymorphisms, SNPs. SNVs
- Rearrangements
- Alleles
- Allele frequency
- Haplotypes (an organism's collection of variants)

Technology

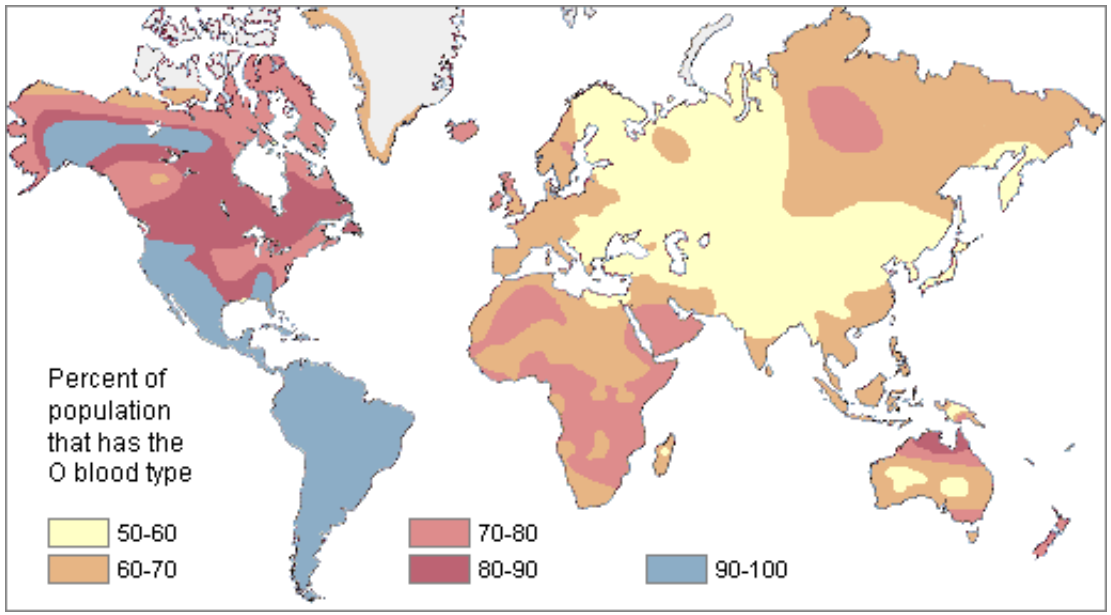
- Next Generation Sequencing, NGS
- Synteny (conserved positions on chromosomes)



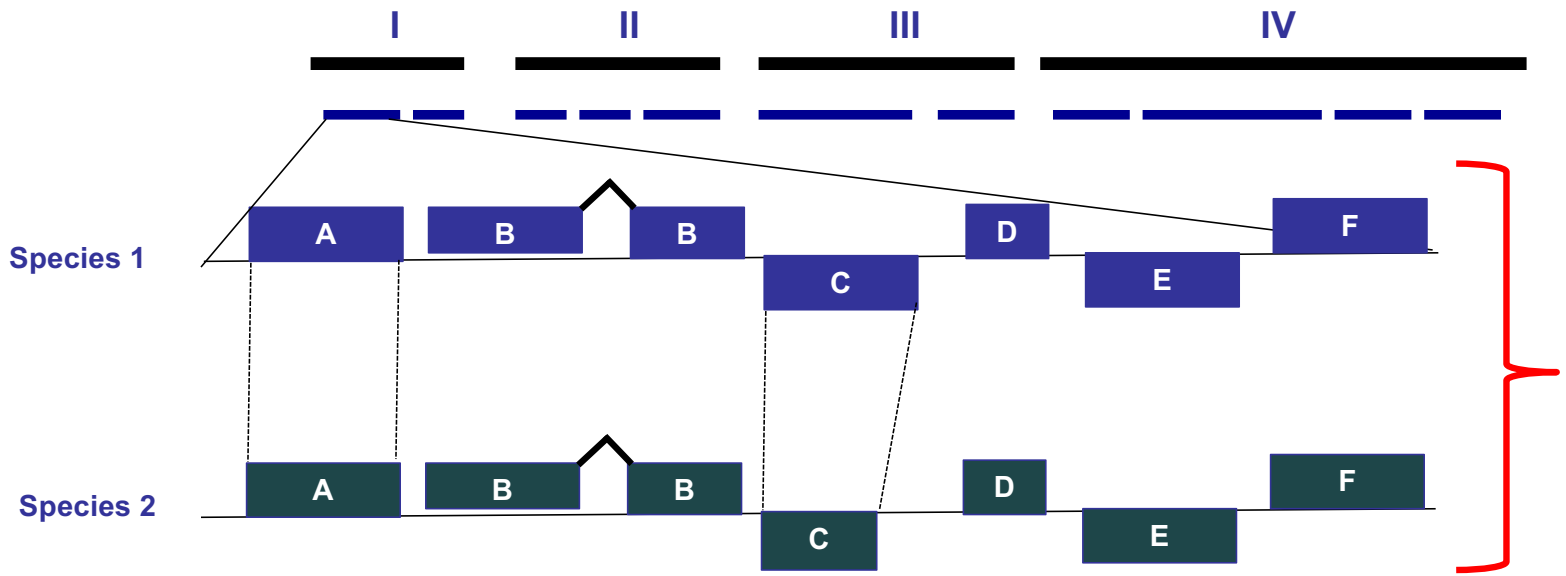
Alleles have frequencies in different populations

Populations and alleles can have geographic boundaries

A parasite isolate comes from a particular population, a particular location and will have a specific haplotype (e.g. representation of alleles) often characterized via SNPs



30,000 ft View - Synteny



Synteny = the majority of the same genes are present in the same order and orientation in another species. The chromosomal regions are evolutionarily related

Synteny among *Plasmodium* species

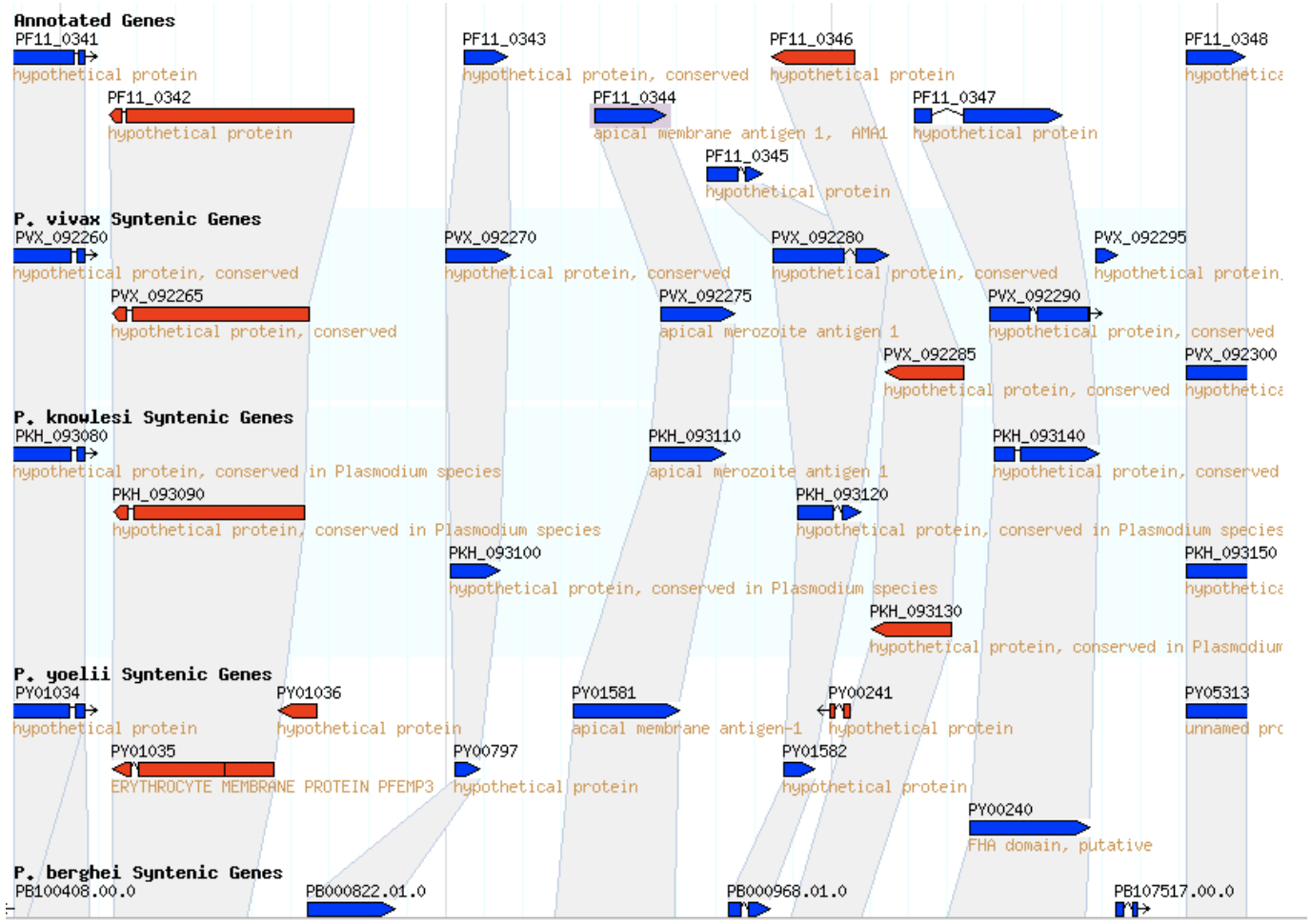
P. falciparum

P. vivax

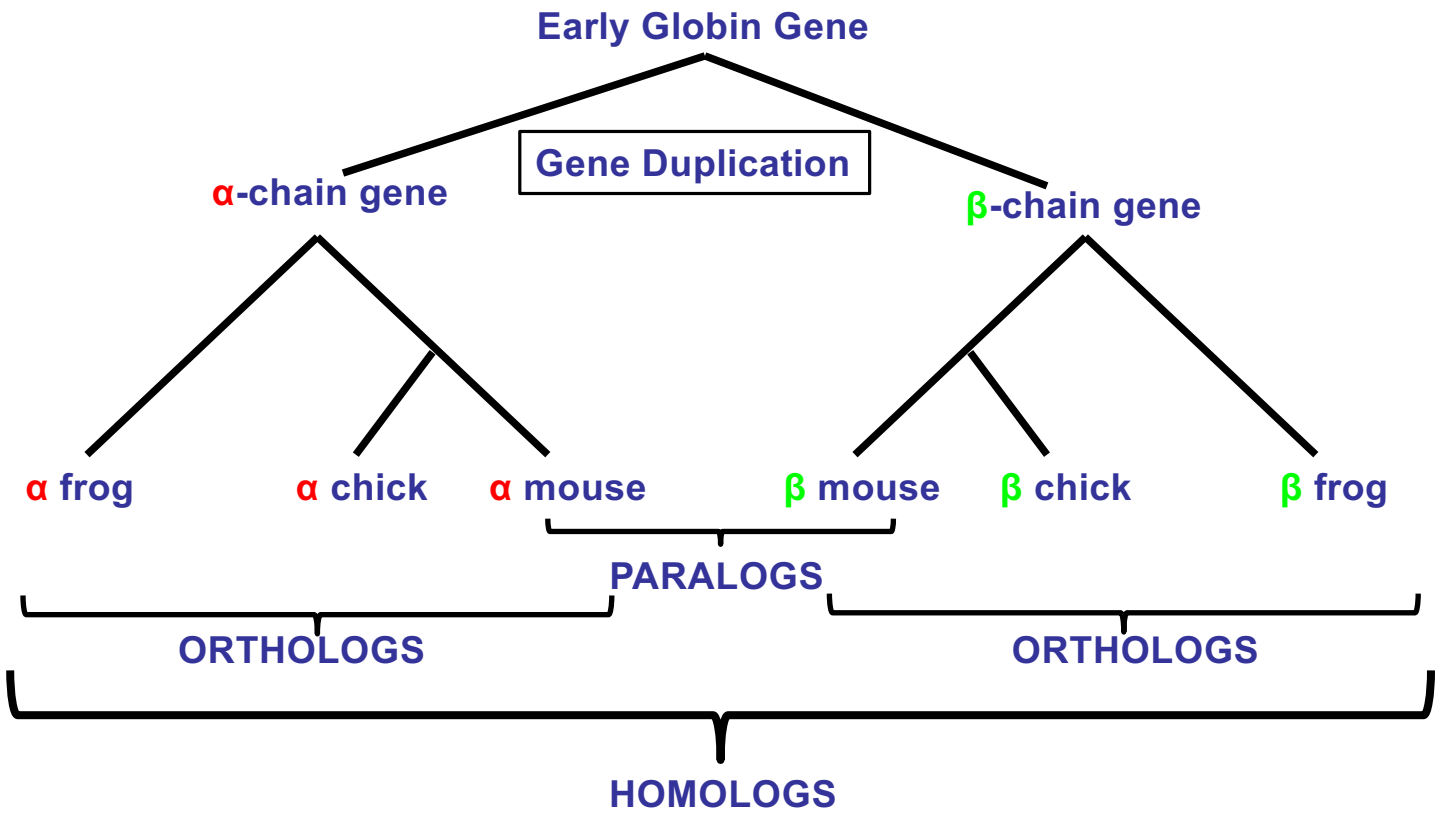
P. knowlesi

P. yoelii

P. berghei



Homology – a vocabulary for different types of evolutionary relationships



Synteny shows relationships in positioning: Ontologies show relationships in meaning

- The Gene Ontology – GO provides terms to link genes with similar functions and/or locations in the cell.

- An ontology was needed because the cultural traditions in different organisms led to different gene naming schemes that made it difficult to identify orthologous genes with the same function.

For Example:

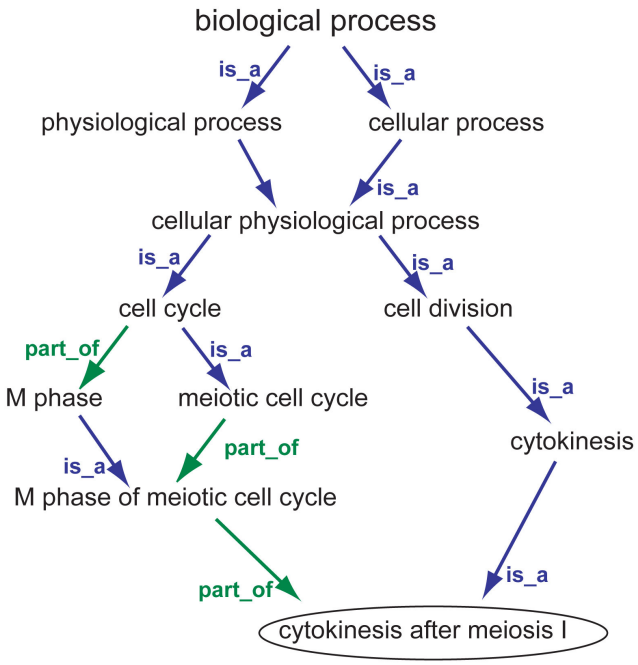
D. melanogaster gene CG3340 annotated as: "*Kruppel*" and *P. falciparum* gene PF3D7_1209300 annotated a "putative KROX1"

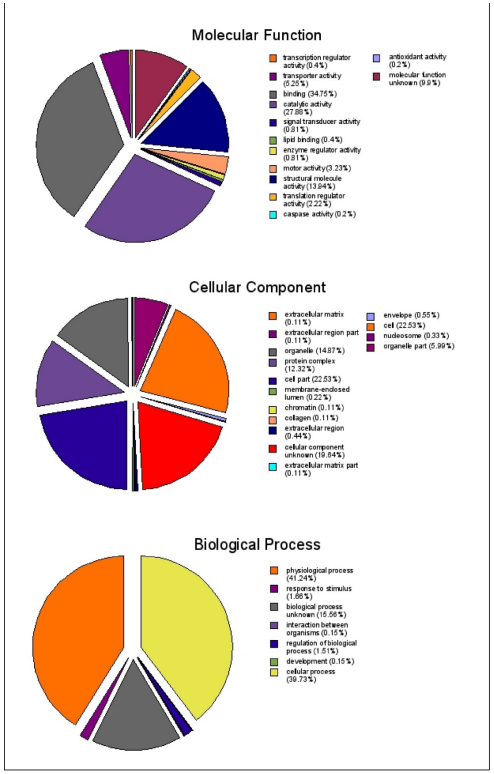
Both can be annotated with GO term:

GO:0003705 (RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity)

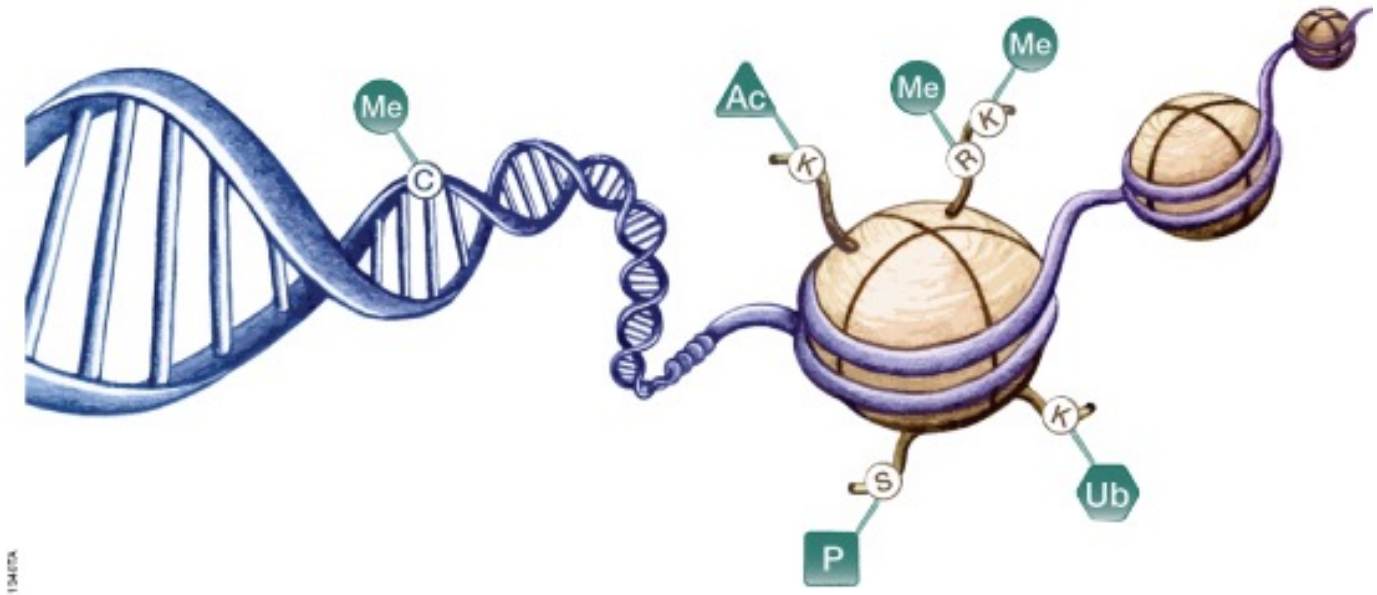
Both proteins, functionally, are Zinc Fingers despite their different names

Note that the Gene Ontologies themselves contain only information about terms in the ontology and their relationships to other terms





Chromatin Status and Epigenetic Gene Regulation



- DNA methylation at CpG islands
- Bisulfite sequencing is a common assay
- H3K4me3 = transcriptionally active chromatin
- H3K27me3 = compact chromatin
- There are MANY other histone modifications
- ChIP-Seq (Chromatin ImmunoPrecipitation) is a common assay for histone markers

https://www.promega.com/resources/guides/nucleic-acid-analysis/introduction-to-epigenetics/

Gene expression

Expression Profiles (RNA and Protein)

- The pattern of expression of one or more genes over time or a set of experimental conditions, e.g. during development or a drug treatment or in a genetic mutant such as a gene knock-out.
- Always… has a time and location component, much like a photograph

RNA expression

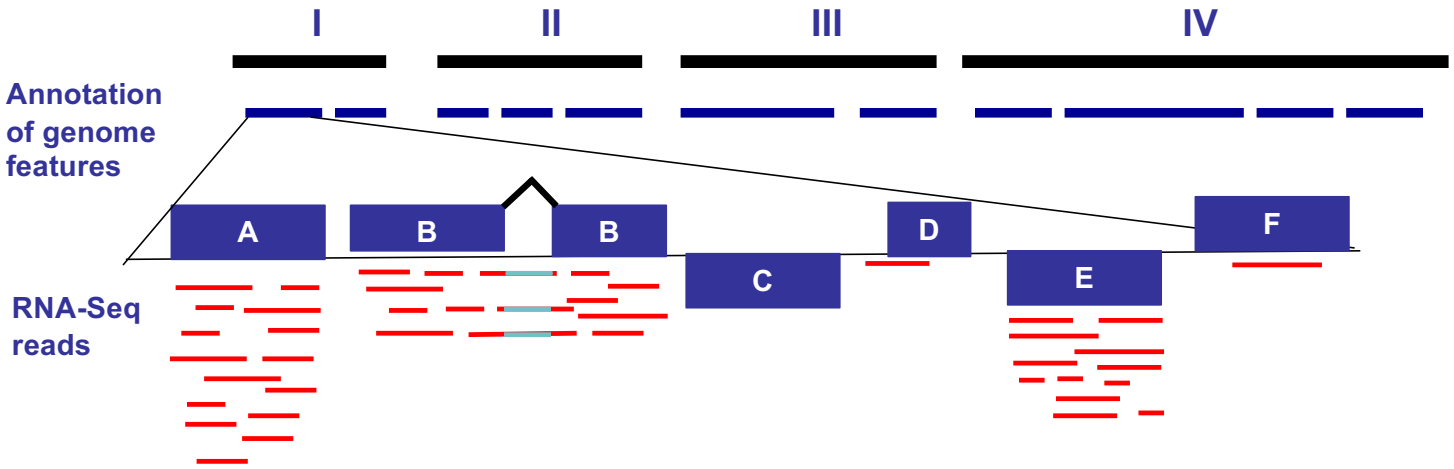
Bulk sequencing from many cells

- RNA-Seq (NGS)
  - Little sequence bias
  - Quantitative
  - Usually are strand-specific
- PacBio ISO-seq
  - Full-length transcripts from single molecules
- ONT Direct seq
  - Single-molecule, direct sequencing of RNA (or can sequence cDNA)
- All of these methods can be used to identify UTR's and exon splice junctions

Single-Cell Sequencing

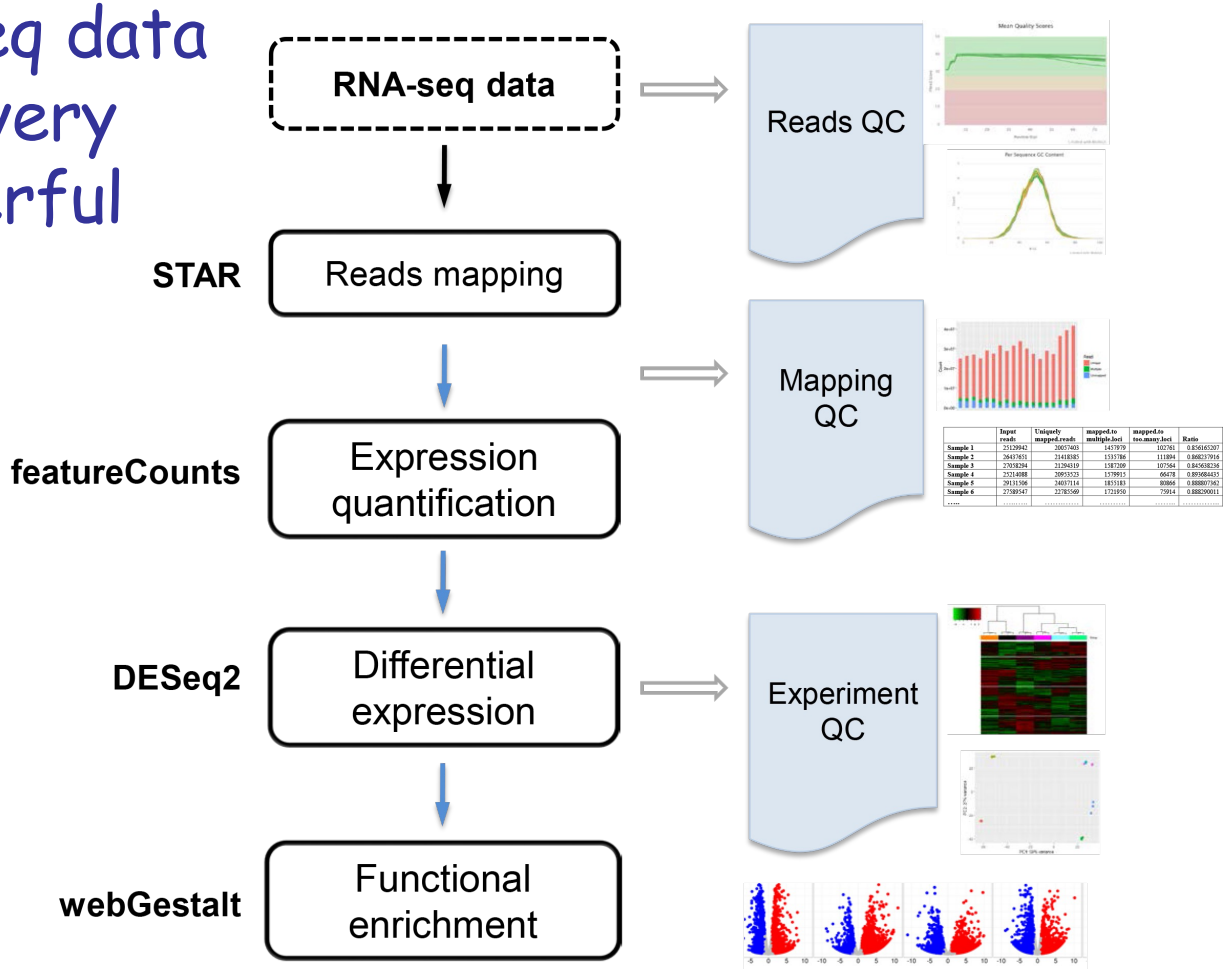
- Examines the transcriptome inside each cell analyzed
- Excellent for detecting cellular heterogeneity or differentiation
- Often only detects a fraction of the transcripts within a cell
- Often analyzed with tSNE plots to categorize cells that have similar transcriptional profiles.

30,000 ft View – RNA-Seq



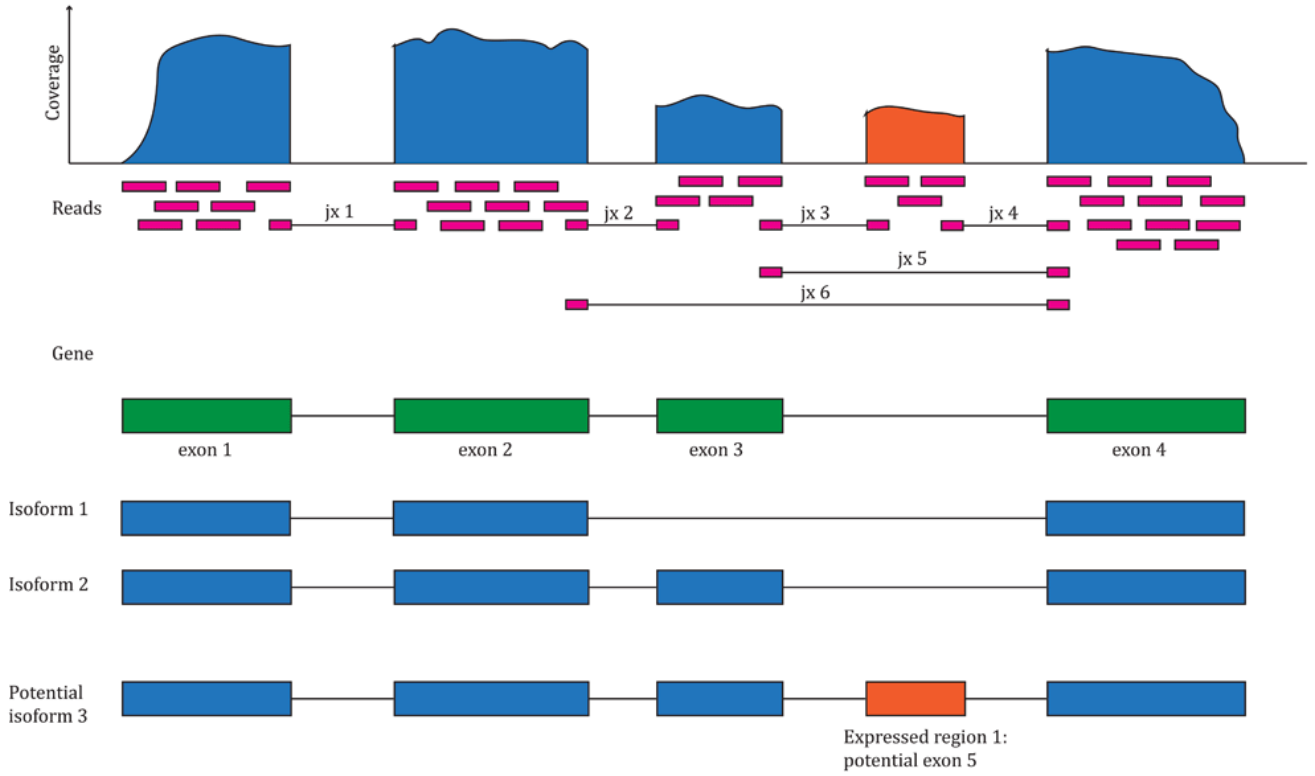
FPKM = Fragments per kilobase of exon per million fragments mapped

RNA-seq data are very powerful



http://bioinfo.vanderbilt.edu/vangard/services-rnaseq.html

RNA-seq identifies splice junctions if present (remember context dependent)



https://bioconductor.org/packages/devel/workflows/vignettes/recountWorkflow/inst/doc/recount-workflow.html

Complex patterns of eukaryotic mRNA splicing: What is a Gene?

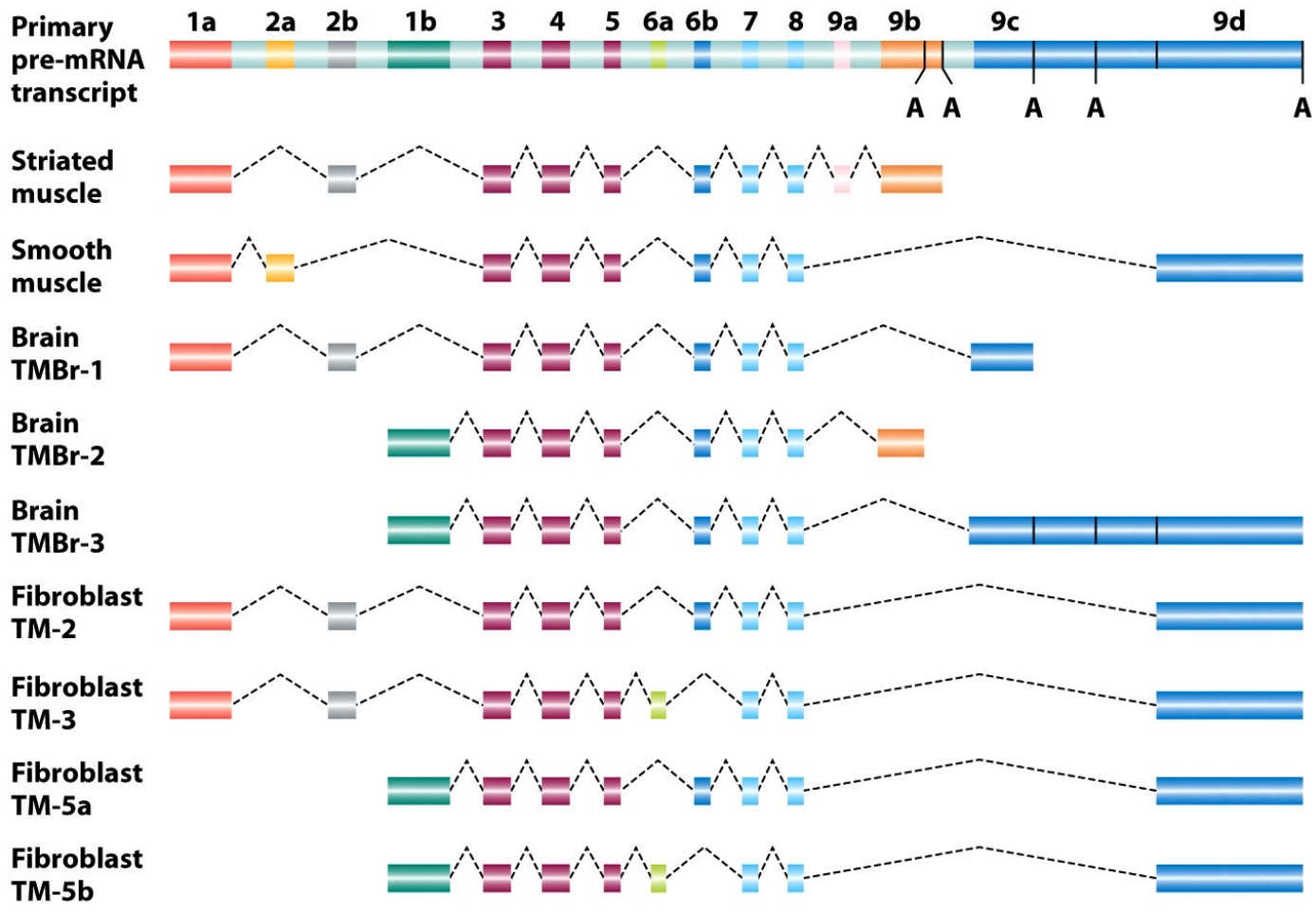
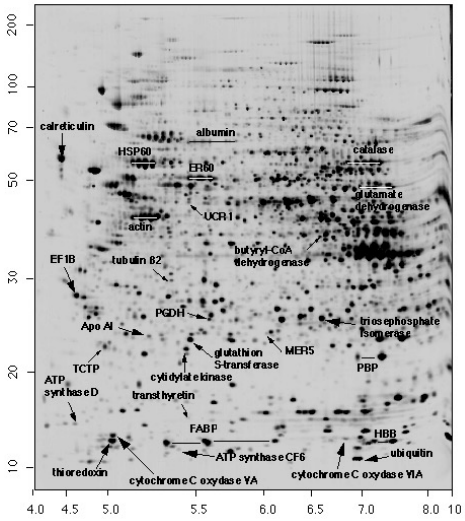


Figure 8-14
Introduction to Genetic Analysis, Ninth Edition

Protein Expression/Sequence

Data

- MW-Isoelectric point
- MW
- Sequence/spans

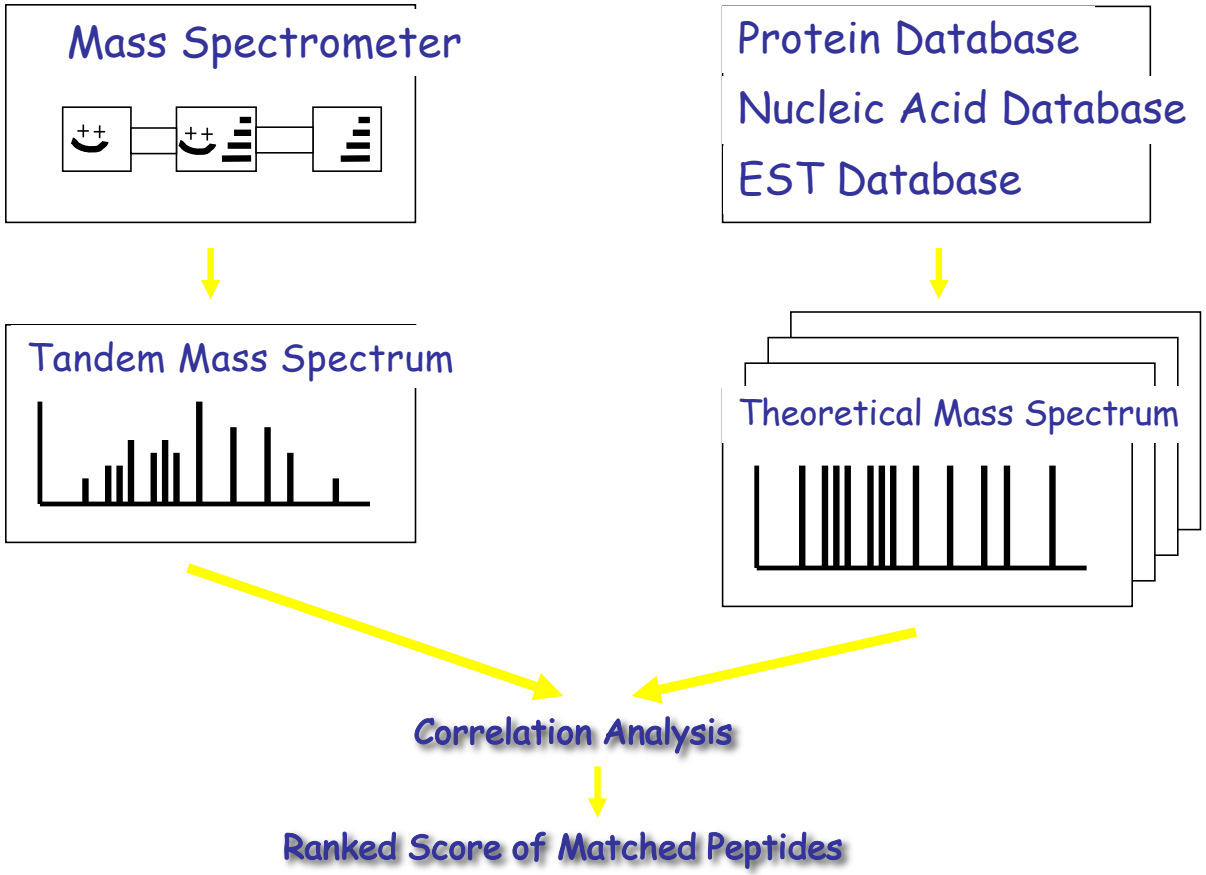


Technology

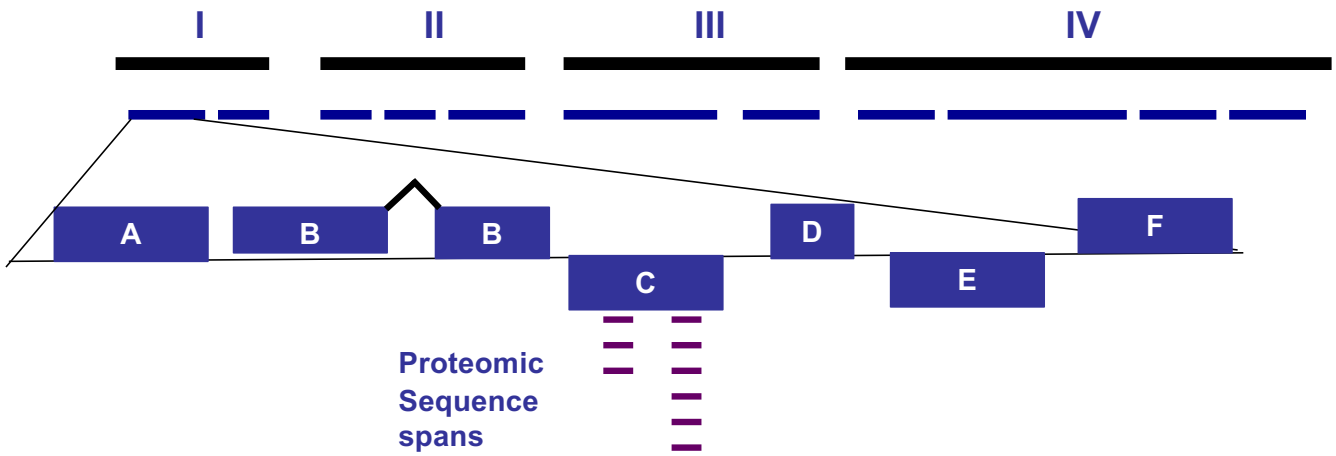
- 2D gel electrophoresis
- Mass spectrometry
- Tandem MS (MS-MS, LC MS-MS etc)

Typical 2 D gel

Sequest Database Search



30,000 ft View - Proteomics



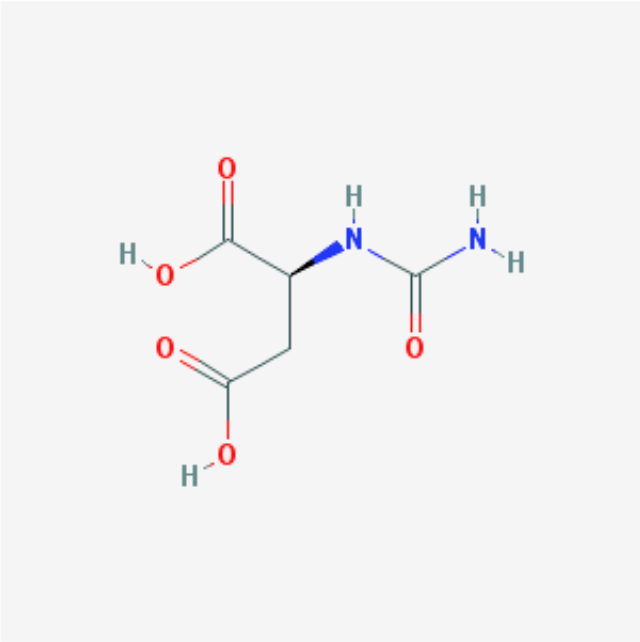
When looking at protein mass-spec sequences it is common to only detect parts of proteins. Some regions are refractory to detection, so don't be alarmed.

Metabolites

Overview

PubChem Compound ID:	CID:93072
PubChem Substance ID(s):	3727
Synonyms:	N-Carbamoyl-L-aspartate
Molecular Weight:	176.12742
Molecular Formula:	$C_5H_8N_2O_5$

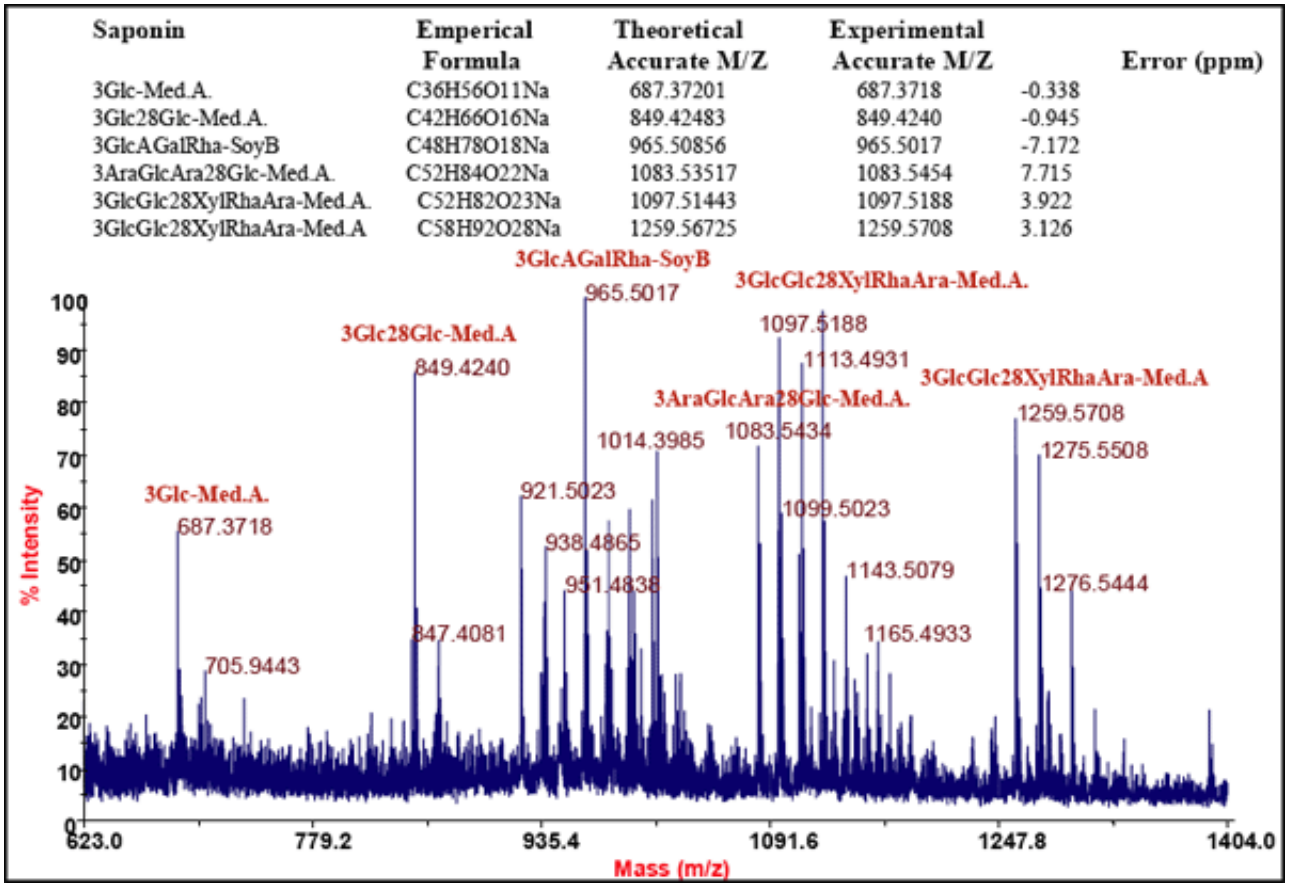
2D Structure



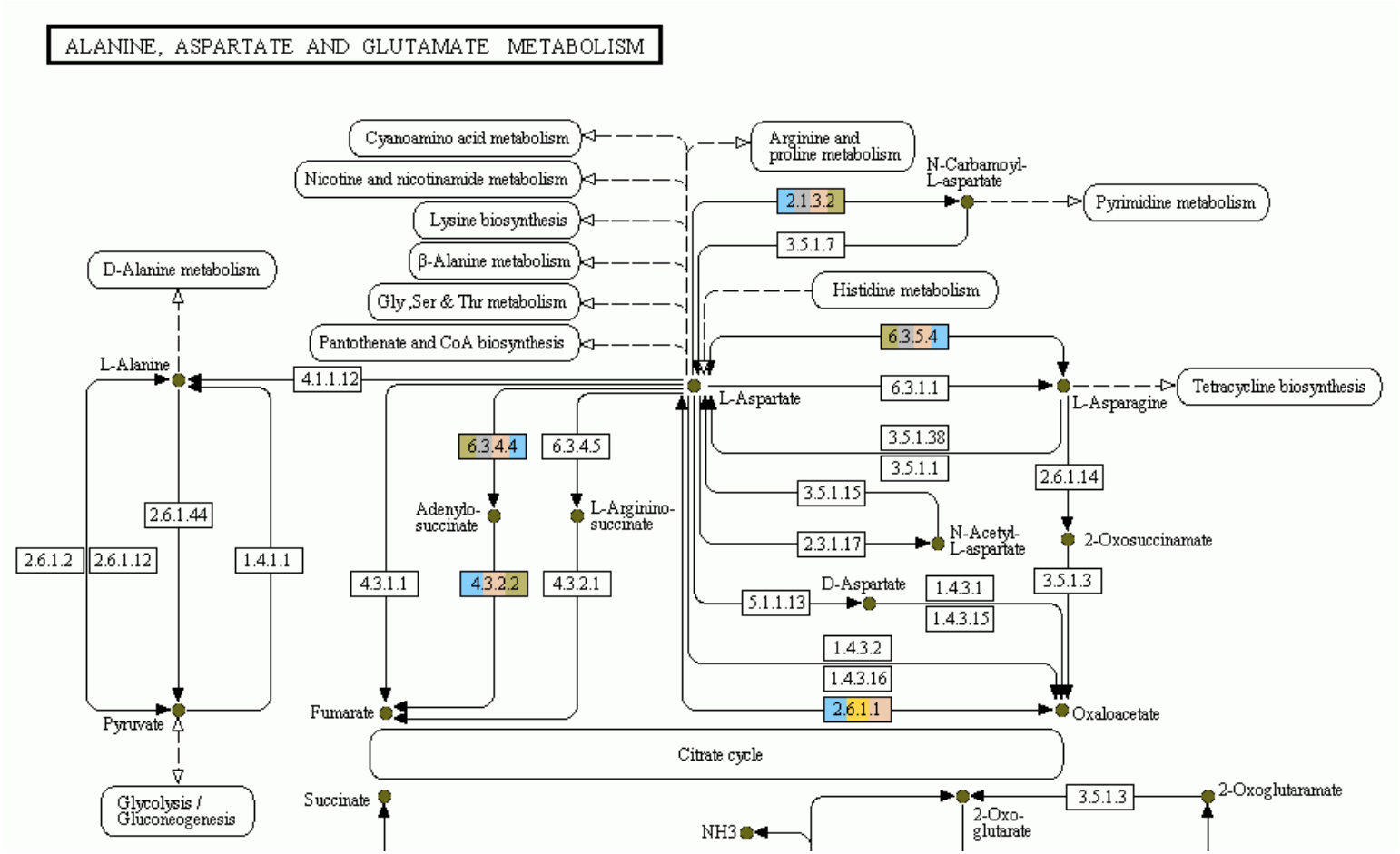
Mass Spectrometry can be used to measure metabolic and other chemical compounds

Complex mixtures can be analyzed and interpreted

Saponin	Emperical Formula	Theoretical Accurate M/Z	Experimental Accurate M/Z	Error (ppm)
3Glc-Med.A.	C36H56O11Na	687.37201	687.3718	-0.338
3Glc28Glc-Med.A.	C42H66O16Na	849.42483	849.4240	-0.945
3GlcAGalRha-SoyB	C48H78O18Na	965.50856	965.5017	-7.172
3AraGlcAra28Glc-Med.A.	C52H84O22Na	1083.53517	1083.5454	7.715
3GlcGlc28XylRhaAra-Med.A.	C52H82O23Na	1097.51443	1097.5188	3.922
3GlcGlc28XylRhaAra-Med.A	C58H92O28Na	1259.56725	1259.5708	3.126



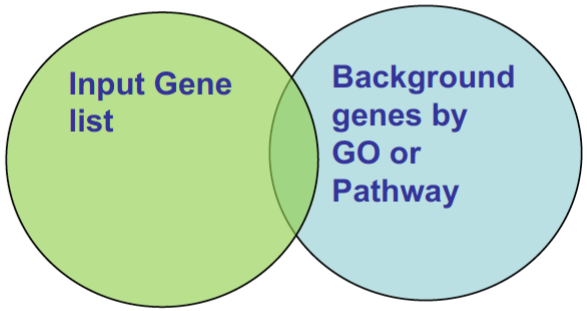
Metabolites can be linked to metabolic pathways and enzymes



Gene & Pathway Enrichment

Gene list:

Up/Down-regulated based on some experiment, e.g. RNA-Seq



Background-Pathway information: All genes known to be involved in some process, e.g. glycolysis or cell signaling. ALL pathways are examined

Result: GO:ID or Pathway ID that is enriched

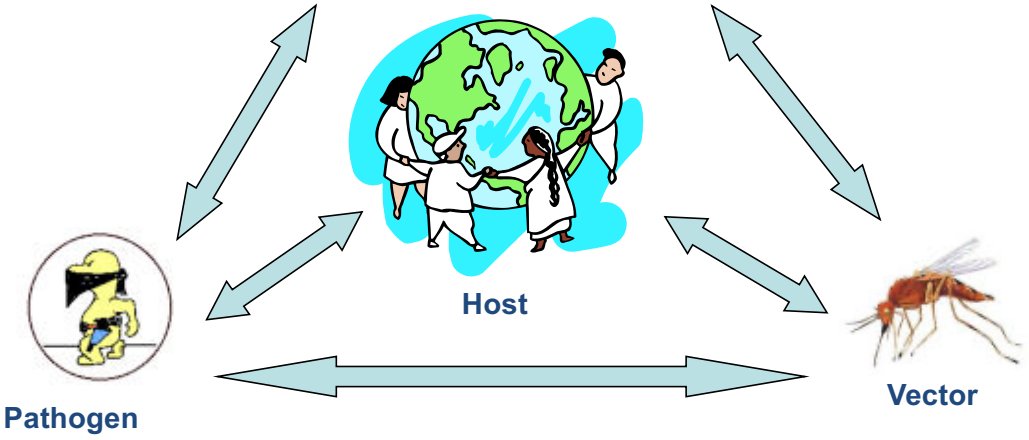
Statistics: Are more genes observed than expected (P-value)
Multiple hypothesis testing (Bonferroni, Benjamini-Hochberg)

Atul Butte Review: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002375

Host(s)

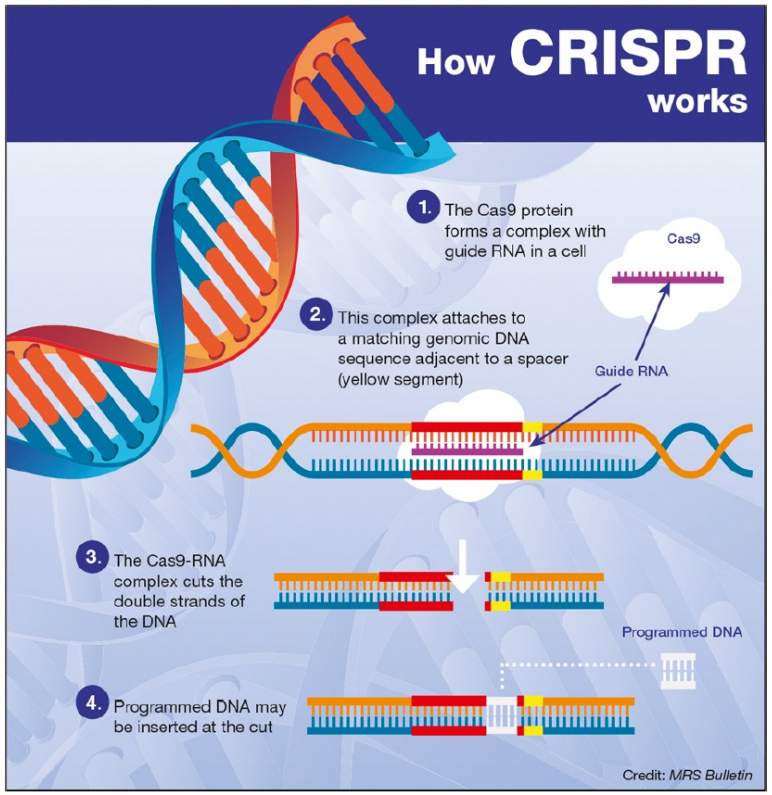
Infectious Disease Paradigm of Host-Pathogen Interactions

Experimental systems



Mutant analysis

CRISPR-CAS



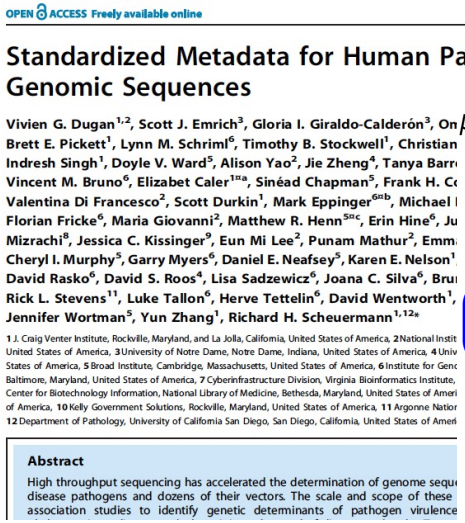
- Need to provide both the enzyme and the guide RNA to the cell
- Need to design the guide RNA to the gene of interest, ideally at multiple target locations per gene

Ball et al., MRS Bulletin November 2016

Metadata – The next Frontier

- Data about the data are critical
- What makes a data set valuable? (The reason it was generated…but often this is missing)
- Introducing the "data set"
- How can you find the data set you need? Pull down Menu? A search of data set properties?
  - Person and technology that generated the data
  - Clinical outcome
  - Geographic location
  - Phenotype

Data sharing standards



OPEN ACCESS Freely available online

PLOS | ONE

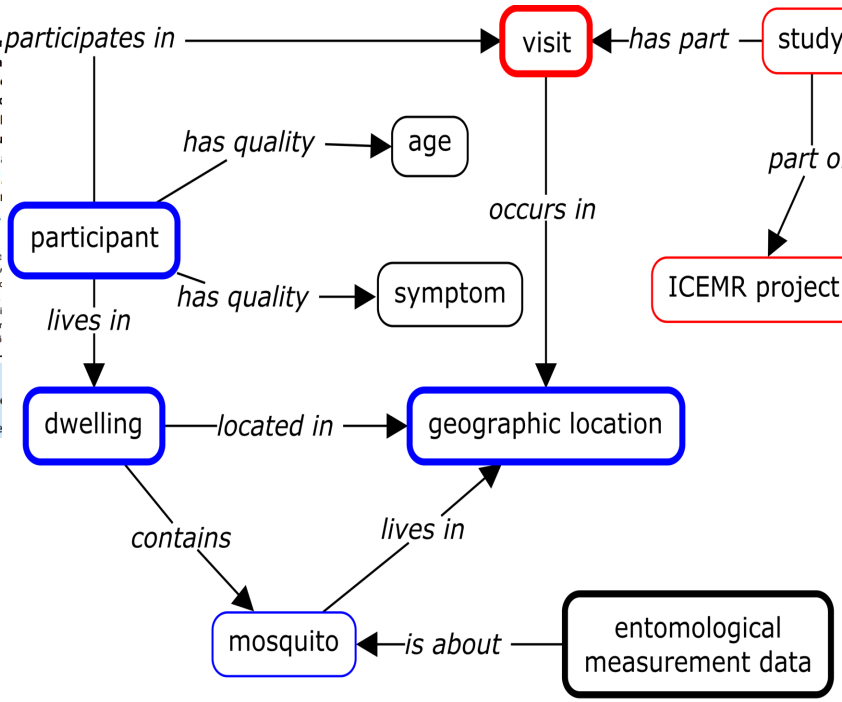
Standardized Metadata for Human Pathogen/Vector Genomic Sequences

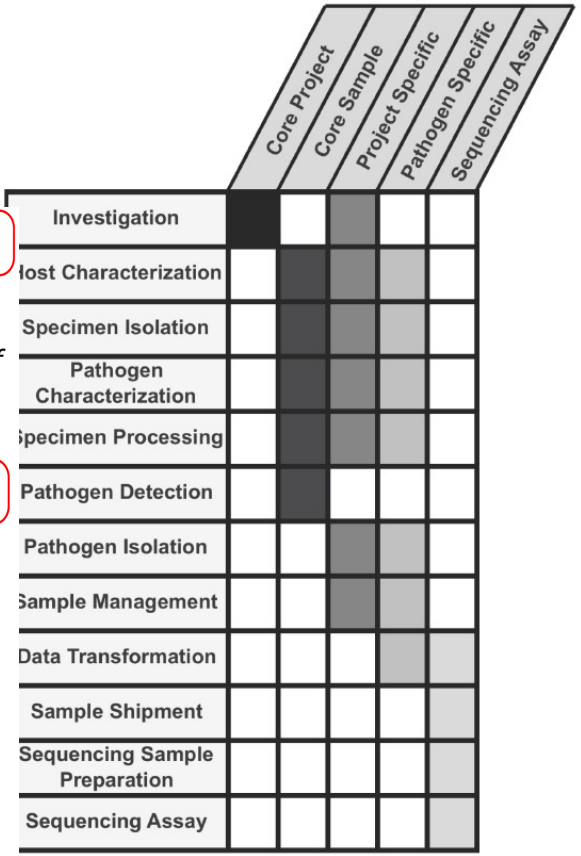
Vivien G. Dugan[1,2], Scott J. Emrich[3], Gloria I. Giraldo-Calderón[3], On
Brett E. Pickett[1], Lynn M. Schriml[6], Timothy B. Stockwell[1], Christian
Indresh Singh[1], Doyle V. Ward[5], Alison Yao[2], Jie Zheng[4], Tanya Barr
Vincent M. Bruno[6], Elizabet Caler[1¤a], Sinéad Chapman[5], Frank H. Cc
Valentina Di Francesco[2], Scott Durkin[1], Mark Eppinger[6¤b], Michael
Florian Fricke[6], Maria Giovanni[2], Matthew R. Henn[5¤c], Erin Hine[6], Ju
Mizrachi[8], Jessica C. Kissinger[9], Eun Mi Lee[2], Punam Mathur[2], Emm
Cheryl I. Murphy[5], Garry Myers[6], Daniel E. Neafsey[5], Karen E. Nelson[1]
David Rasko[6], David S. Roos[4], Lisa Sadzewicz[6], Joana C. Silva[6], Bru
Rick L. Stevens[11], Luke Tallon[6], Herve Tettelin[6], David Wentworth[1],
Jennifer Wortman[5], Yun Zhang[1], Richard H. Scheuermann[1,12*]

1 J. Craig Venter Institute, Rockville, Maryland, and La Jolla, California, United States of America, 2 National Instit
United States of America, 3 University of Notre Dame, Notre Dame, Indiana, United States of America, 4 Univ
States of America, 5 Broad Institute, Cambridge, Massachusetts, United States of America, 6 Institute for Geno
Baltimore, Maryland, United States of America, 7 Cyberinfrastructure Division, Virginia Bioinformatics Institute,
Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland, United States of Ameri
of America, 10 Kelly Government Solutions, Rockville, Maryland, United States of America, 11 Argonne Natior
12 Department of Pathology, University of California San Diego, San Diego, California, United States of Ameri

Abstract

High throughput sequencing has accelerated the determination of genome sequ
disease pathogens and dozens of their vectors. The scale and scope of these
association studies to identify genetic determinants of pathogen virulence

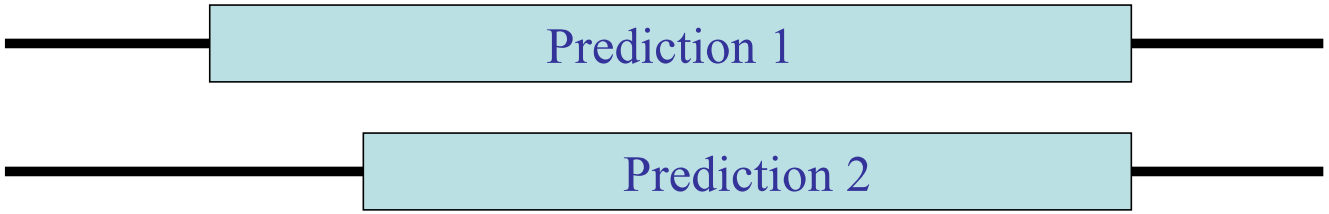




Bioinformatics uses algorithms

- Algorithms are sets of rules for solving problems or identifying patterns
- Algorithms can be general or case specific and often need to be trained
- Computational analysis, like wet-bench analyses are only as good as the tools, techniques and material allow, and all interpretations come with caveats (like the experimental conditions, often call parameters in bioinformatics.

Different algorithms often generate different results

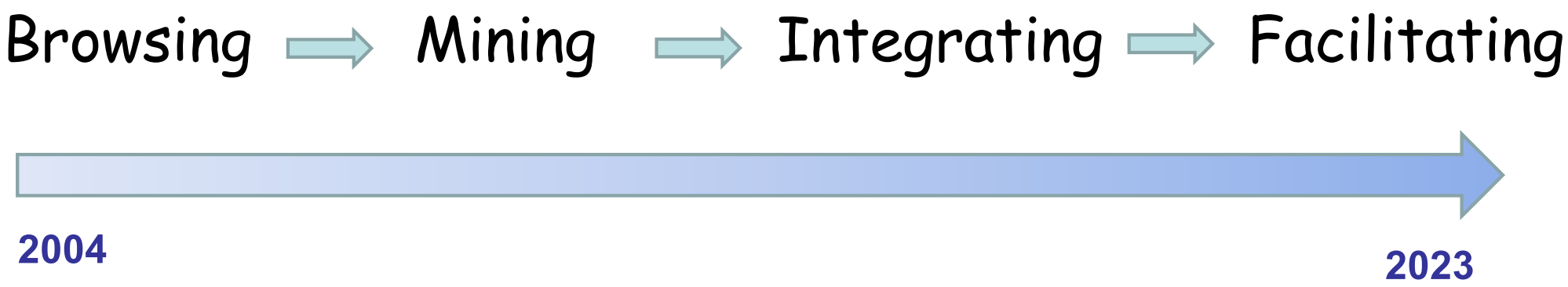


We provide lots evidence so that you can decide or design an experiment to confirm!

Garbage in Garbage out!

- The algorithms will almost always return a result, it is up to you, the scientist to evaluate if it has made a mistake. Much of the data in the archival databases have errors. Not intentional errors but errors
- If you can't find the gene or answer it does NOT mean that it does not exist. It may be in a gap, or never have been annotate, or discovered after the annotation e.g. lncRNAs. Interpret carefully

Bioinformatics Resource Center Community Evolution



The End

- If you have questions, I and the other instructors will be around and we are happy to talk to you.
- These slides are available to you as a PDF on the workshop web site.

VEuPathDB
Eukaryotic Pathogen, Vector & Host Informatics Resources

NIH National Institute of Allergy and Infectious Diseases

wellcome

Project Leadership:
David Roos – UPENN
Mary Ann McDowell – Notre Dame
Andrew Jones – Liverpool
Jessie Kissinger – UGA
Sarah Dyer – EBI
Kathryn Crough – Glasgow
George Christophides - Imperial

Our goal: enabling end users in the lab, field & clinic to make effective and appropriate use of large-scale datasets, expediting discovery research and translational application by making data FAIR

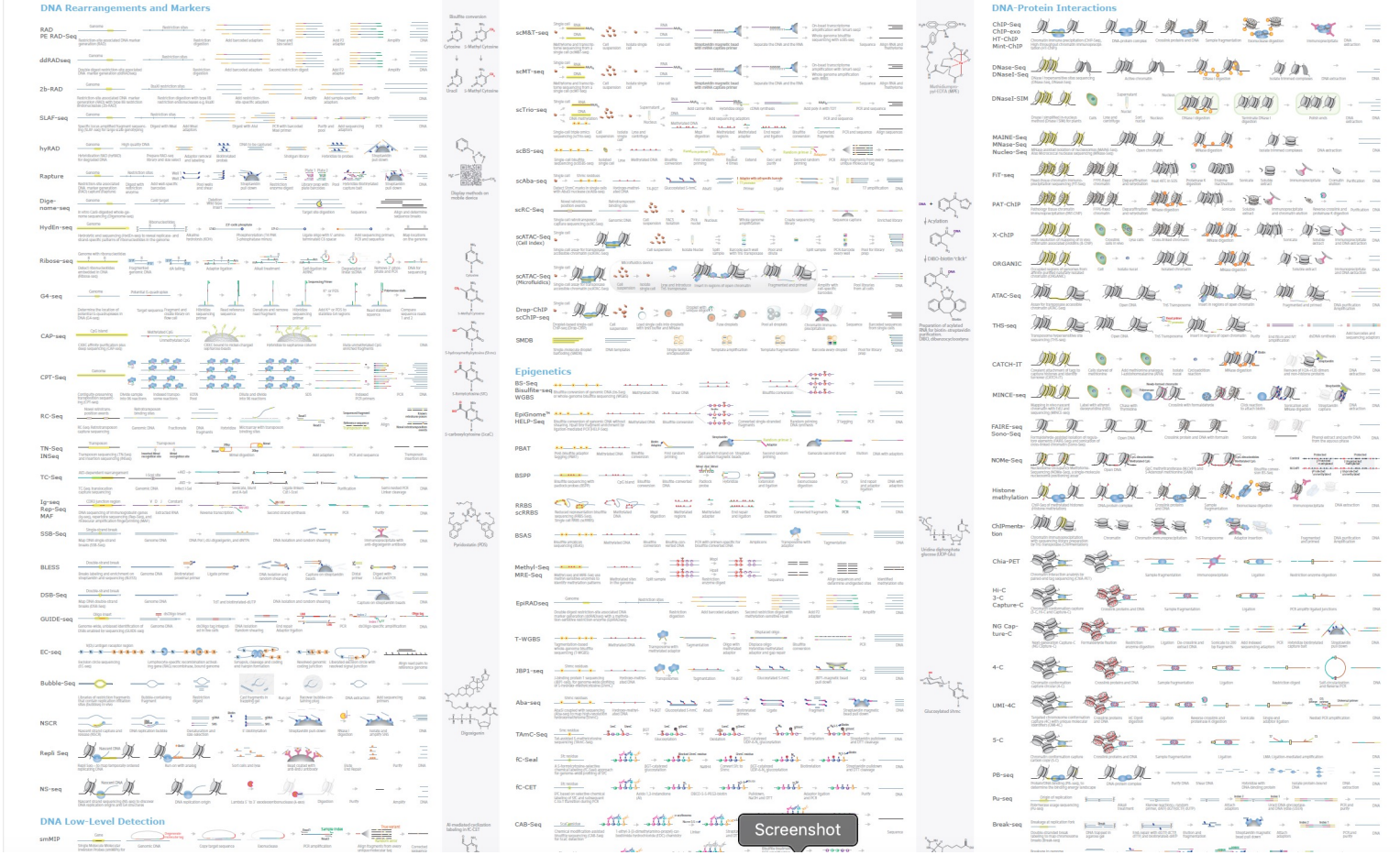
Penn University of Pennsylvania

University of Liverpool

University of Georgia Center for Tropical & Emerging Global Diseases

For all you seq...

DNA



For all you seq...

RNA