VEuPathDB

Eukaryotic Pathogen, Vector & Host Informatics Resources

Crash Course in Omics
Terminology, Concepts & Data Types

Jessie C Kissinger
2023

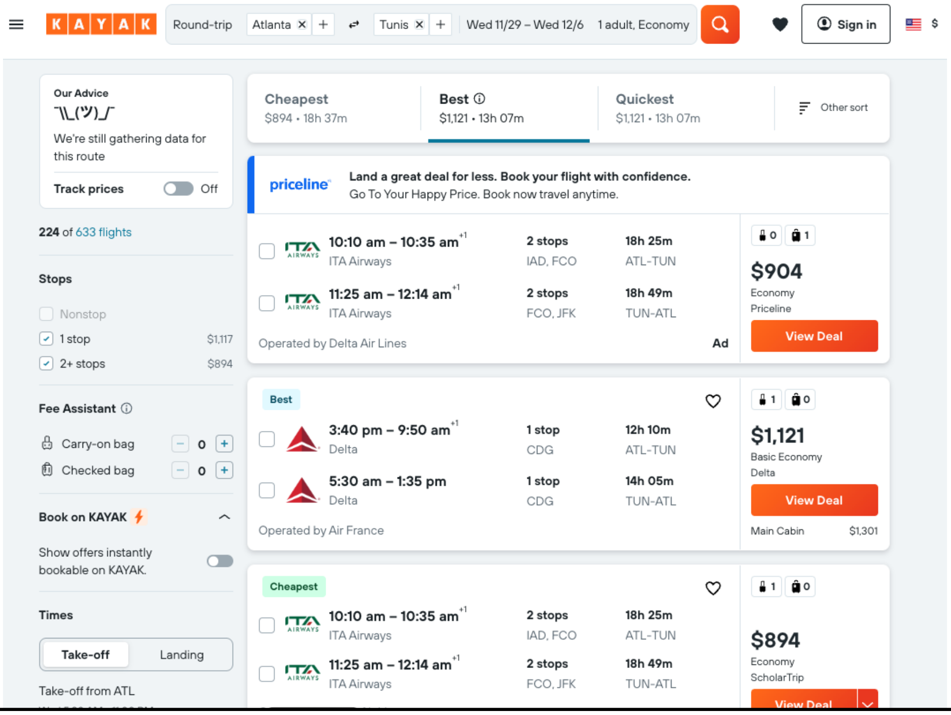
@veupathdb
@jcklab
jkissing@uga.edu

UNIVERSITY OF GEORGIA
Center for Tropical & Emerging Global Diseases

Institute of Bioinformatics
UNIVERSITY OF GEORGIA

1

1



KAYAK | Round-trip | Atlanta × + | ⇄ | Tunis × + | Wed 11/29 – Wed 12/6 | 1 adult, Economy | Sign in | $

Our Advice
¯_(ツ)_/¯
We're still gathering data for this route

Track prices ○ Off

224 of 633 flights

Stops

□ Nonstop
☑ 1 stop $1,117
☑ 2+ stops $894

Fee Assistant ⓘ

Carry-on bag	− 0 +	
Checked bag	− 0 +	

Book on KAYAK

Show offers instantly bookable on KAYAK.

Times

Take-off | Landing

Take-off from ATL

Cheapest	Best ⓘ	Quickest	Other sort
$894 • 18h 37m	$1,121 • 13h 07m	$1,121 • 13h 07m	

priceline — Land a great deal for less. Book your flight with confidence.
Go To Your Happy Price. Book now travel anytime.

10:10 am – 10:35 am +1 ITA Airways	2 stops IAD, FCO	18h 25m ATL-TUN	0 1 $904 Economy Priceline View Deal
11:25 am – 12:14 am +1 ITA Airways	2 stops FCO, JFK	18h 49m TUN-ATL	

Operated by Delta Air Lines — Ad

Best

3:40 pm – 9:50 am +1 Delta	1 stop CDG	12h 10m ATL-TUN	1 0 $1,121 Basic Economy Delta View Deal
5:30 am – 1:35 pm Delta	1 stop CDG	14h 05m TUN-ATL	Main Cabin $1,301

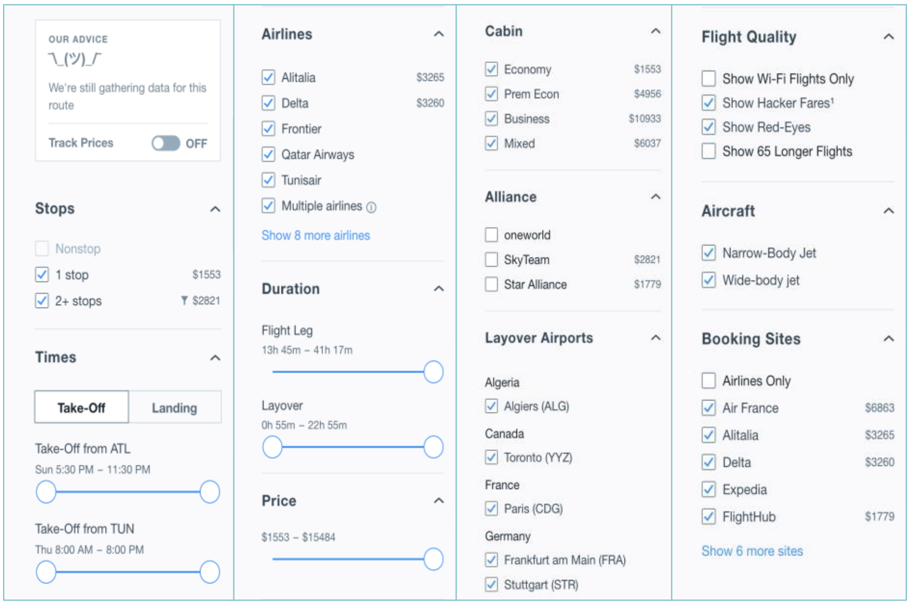
Operated by Air France

Cheapest

10:10 am – 10:35 am +1 ITA Airways	2 stops IAD, FCO	18h 25m ATL-TUN	1 0 $894 Economy ScholarTrip View Deal
11:25 am – 12:14 am +1 ITA Airways	2 stops FCO, JFK	18h 49m TUN-ATL	

2

The Travel Site has Very Useful Data Filters!



OUR ADVICE
¯_(ツ)_/¯
We're still gathering data for this route

Track Prices OFF

Stops

- □ Nonstop
- ☑ 1 stop $1553
- ☑ 2+ stops $2821

Times

Take-Off | Landing

Take-Off from ATL
Sun 5:30 PM – 11:30 PM

Take-Off from TUN
Thu 8:00 AM – 8:00 PM

Airlines

- ☑ Alitalia $3265
- ☑ Delta $3260
- ☑ Frontier
- ☑ Qatar Airways
- ☑ Tunisair
- ☑ Multiple airlines

Show 8 more airlines

Duration

Flight Leg
13h 45m – 41h 17m

Layover
0h 55m – 22h 55m

Price

$1553 – $15484

Cabin

- ☑ Economy $1553
- ☑ Prem Econ $4956
- ☑ Business $10933
- ☑ Mixed $6037

Alliance

- □ oneworld
- □ SkyTeam $2821
- □ Star Alliance $1779

Layover Airports

Algeria
- ☑ Algiers (ALG)

Canada
- ☑ Toronto (YYZ)

France
- ☑ Paris (CDG)

Germany
- ☑ Frankfurt am Main (FRA)
- ☑ Stuttgart (STR)

Flight Quality

- □ Show Wi-Fi Flights Only
- ☑ Show Hacker Fares[1]
- ☑ Show Red-Eyes
- □ Show 65 Longer Flights

Aircraft

- ☑ Narrow-Body Jet
- ☑ Wide-body jet

Booking Sites

- □ Airlines Only
- ☑ Air France $6863
- ☑ Alitalia $3265
- ☑ Delta $3260
- ☑ Expedia
- ☑ FlightHub $1779

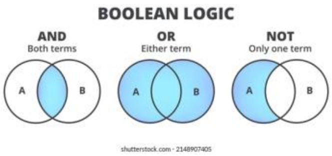
Show 6 more sites

3

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

4

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

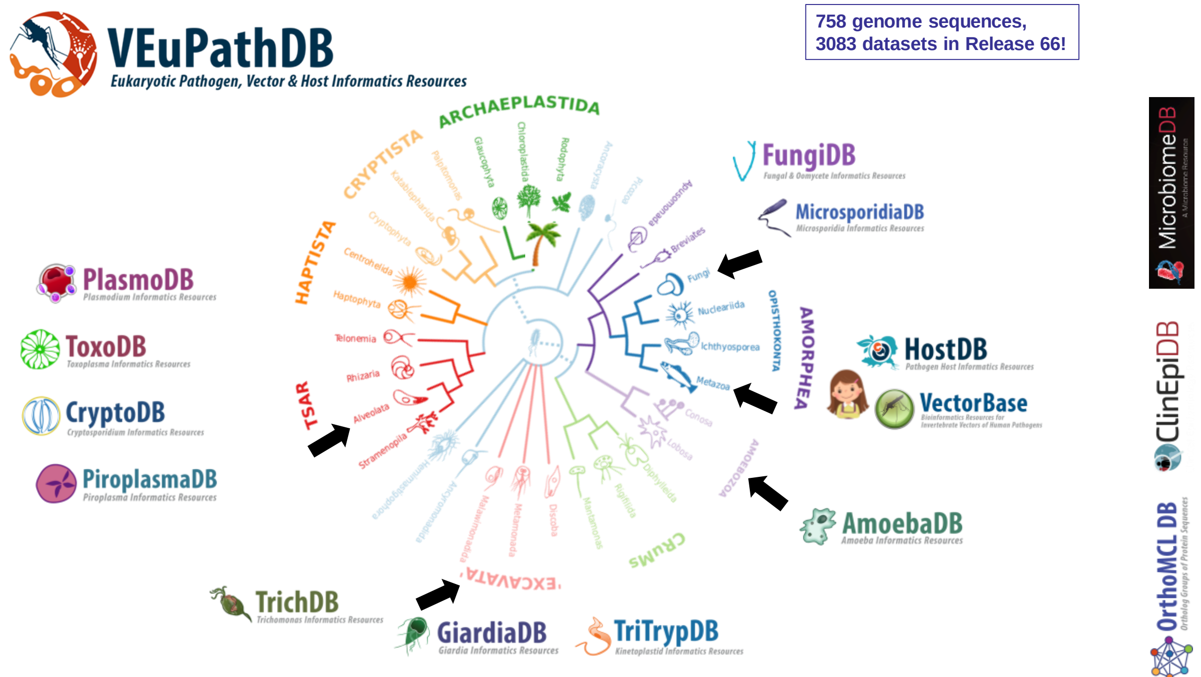
- Find all genes that….
  - That are near centromeres
  - That encode a predicted signal protein
  - That encode the amino acid motif CC..CC
- Which have evidence of expression …
  - In developmental stage X
  - After treatment with drug Y
- That are phosphorylated in proteomic studies
- That show evidence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

5

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

6



758 genome sequences, 3083 datasets in Release 66!

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

7

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

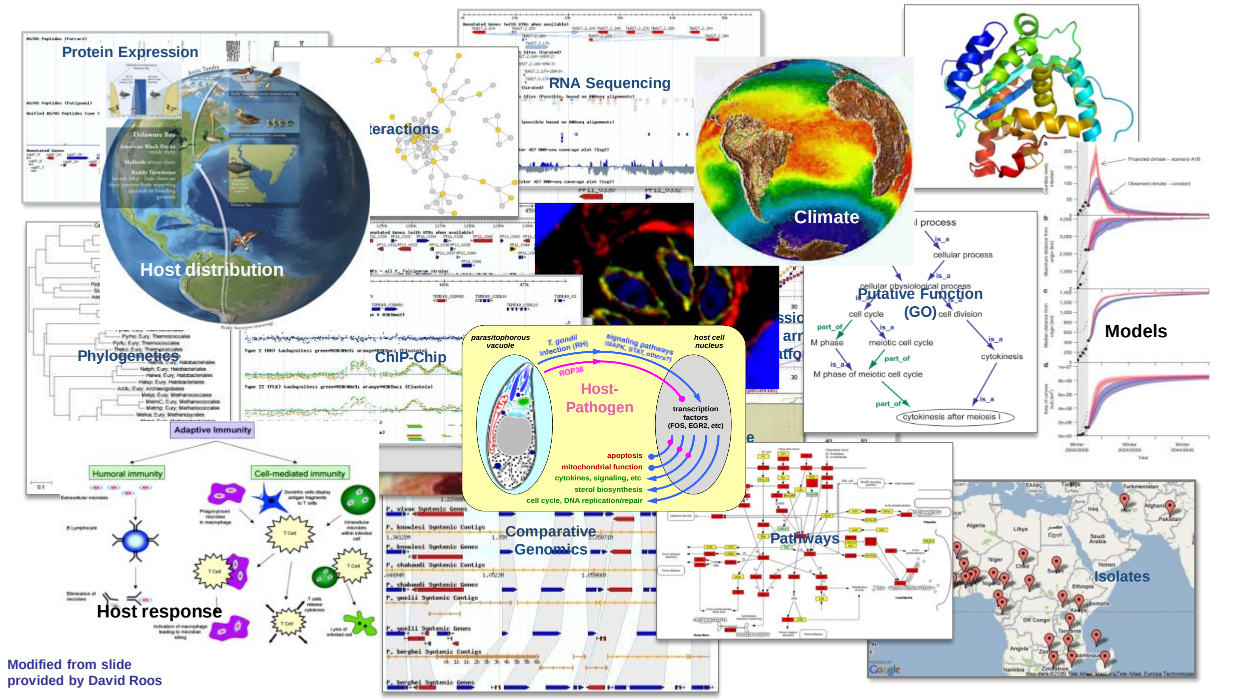
Metabolites: chemical compounds, enzymes, pathways, flux

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

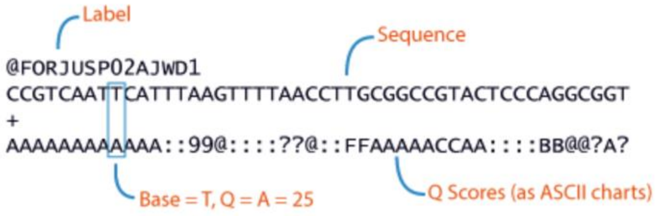
8



Modified from slide provided by David Roos

Genome Assembly 30,000 ft View

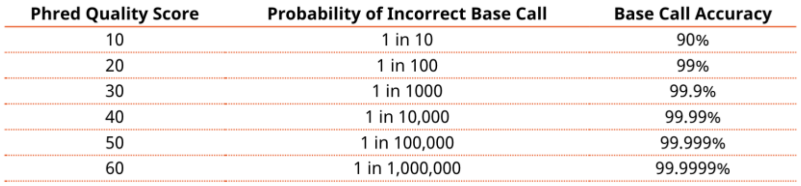
FASTQ format for reads



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@::FFAAAAACCAA::::BB@@?A?
```

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%

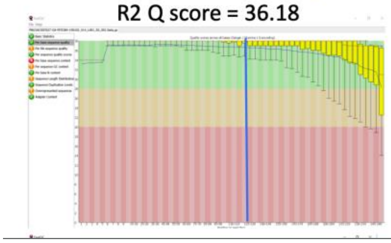
Figure 3 – Phred quality score chart

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

FastQC Analysis – Passing Q Scores

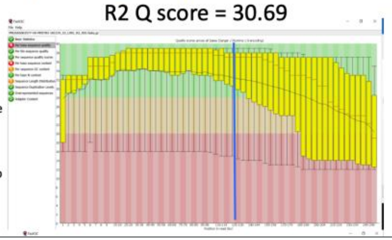
Per Base Sequence Quality & Per Sequence Quality Scores

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

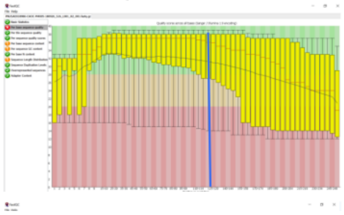
R2 Q score = 30.69



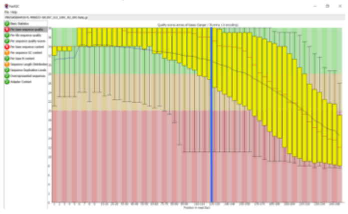
FastQC Analysis – Suboptimal Q Scores (pass with extra coverage)

Per Base Sequence Quality & Per Sequence Quality Scores

R2 Q score = 29.56



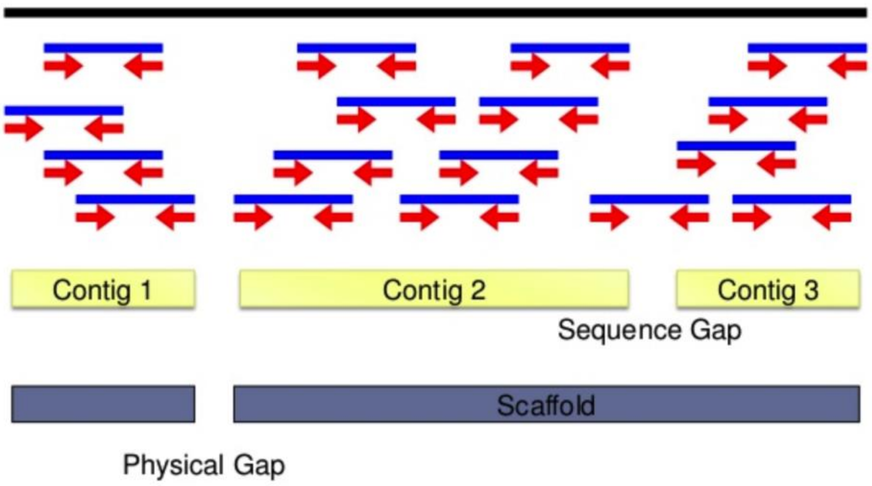
R2 Q score = 28.56



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

11

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



DNA Chromosome e.g. “truth”

Paired-End Reads, Note the direction of the arrows

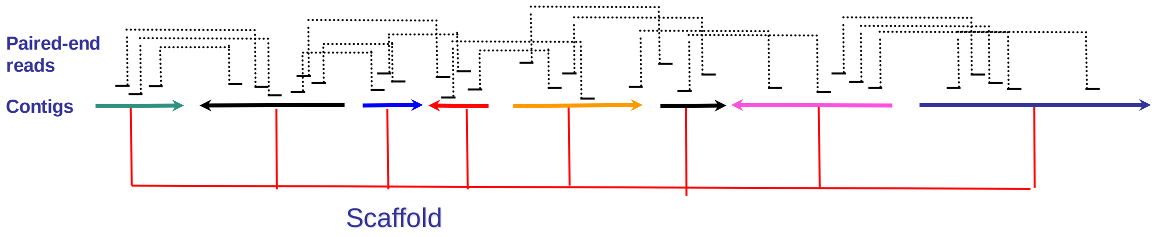
Assembled Contigs

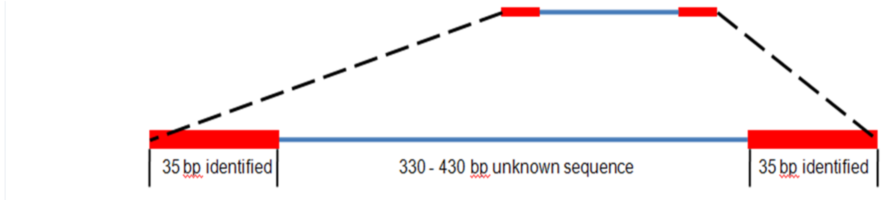
Assembled Scaffolds

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

12

Paired-End Reads can Yield Order & Orientation of Contigs

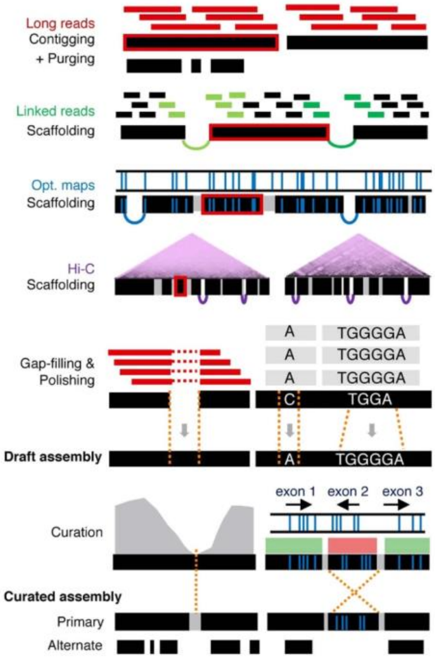




https://www.biostars.org/p/104218/

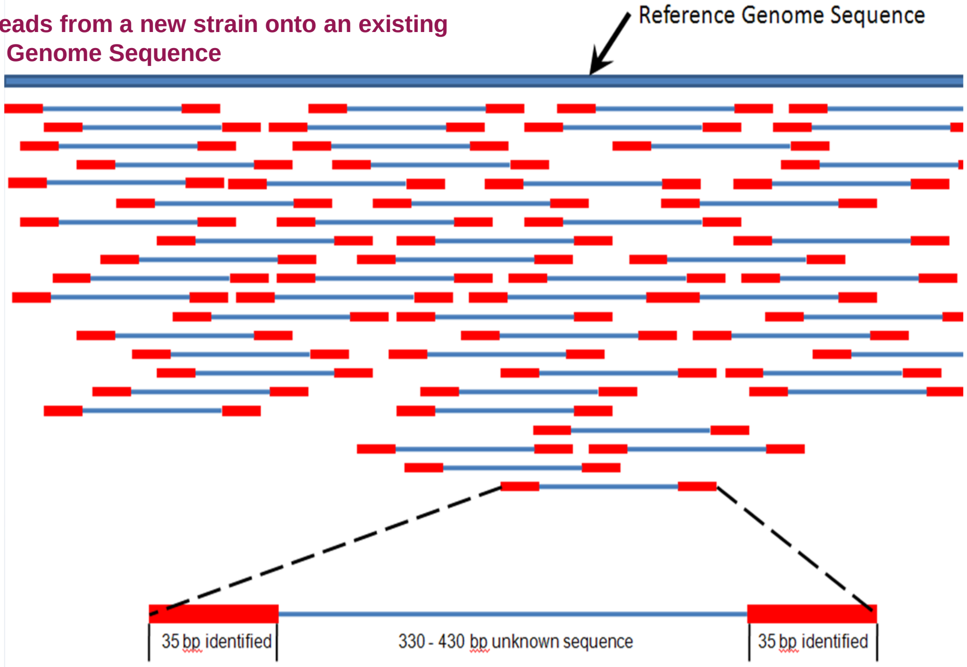
De novo assembly of a complex genome sequence from scratch requires many technologies:

- **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 gaps 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

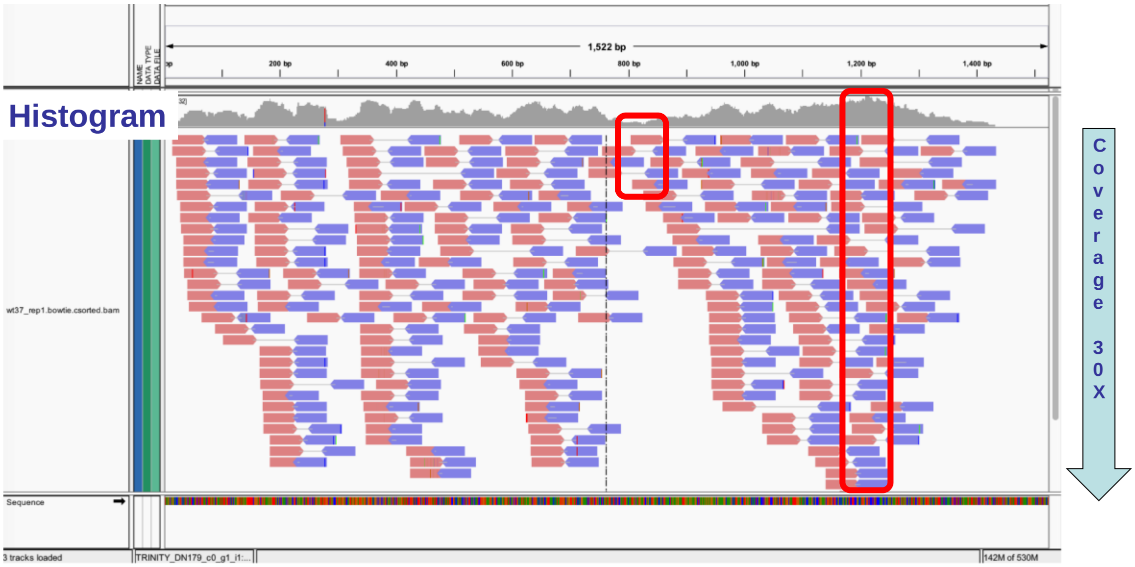
Mapping reads from a new strain onto an existing Reference Genome Sequence



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

15

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

16

Hybrid Assembly = short + long reads from differing technologies

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

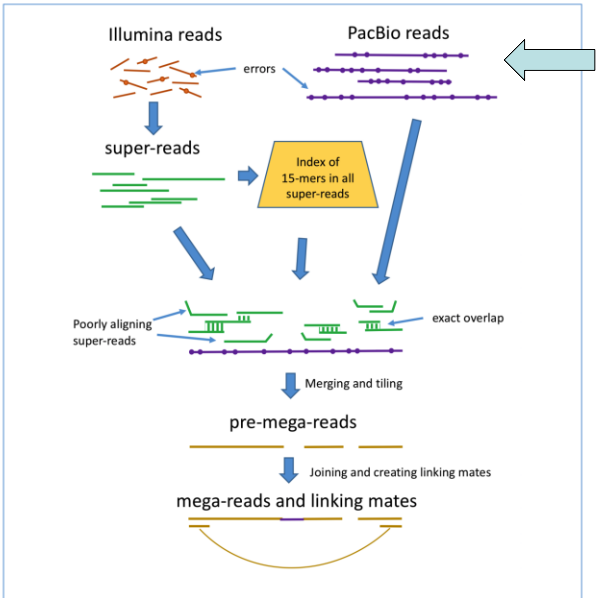

Illumina reads
PacBio reads
errors
super-reads
Index of 15-mers in all super-reads
Poorly aligning super-reads
exact overlap
Merging and tiling
pre-mega-reads
Joining and creating linking mates
mega-reads and linking mates

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

17

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

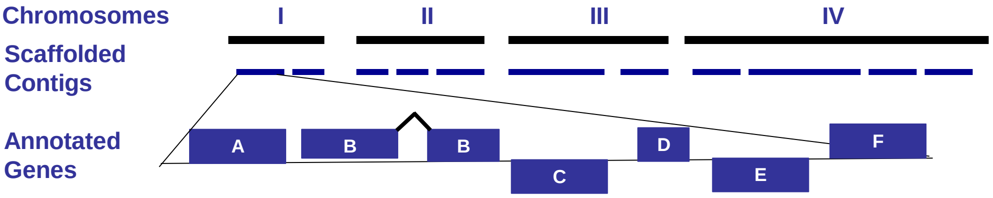
- Read length
  - Short
  - Long
- Coverage
  - 5X
  - 100X
- Read Quality
  - 20
  - 30
  - 40
  - Bias?

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)

18

30,000 ft View – Genome Annotation



The Genome Sequence

```
AAGCTTCGCCAGGCTGTAAATCCCGTGAGTCGTCCTCACAAATCATCAAGCAGGTGTCCTCAGGGAGACTGCCTGACTGAGTTATGCTAATTCCTTTCTACTTTGGCGTGGTCACGTGTA
ACCATATCCGAATCATTTCTCTAGCCCTACGAACAGGTAAGAGCGCTAGGGATGTCCGTGGAGTAGTGTGCTTACTCGATAATATTCAGTTGGGACTACCAGCGAGGCGCTCGCTTTGCT
CACGCAATGCCTGAGA CAGTTGCAGAATGAATGGTAACCGACAAACGCGTTCATATGCGTTTTCAAACTTAGTAGACGCGTACTGTCTGAAACTGGCGGTCACAGGCACCAGATAACGCC
CTTGGCATCGGCATGTCTCGTACAGAGGTCCGTATGTAGTGCCACGACTTCTAAATCCGGCGACAGGCTGGTCTTTTGTCTTACCACGTATTAGCCCGCGTGCGATTTCTCGGAGCGCAC
CTGTTCAACACTAGAAAACGGAGTTTCCTGAT CGAGAAGCCACCACCTTTCCAGAAGTTGAACGCTAGCATGTCATTCGATTTTCACCCCCCGCGTAGTTCCTGTGTGTCATTCGTTGTC
GAGACAACTCTGTCCCGCCCCGGTGCTGTTCCATATGCGTGACTTTCCCGCAATTTTTTCAGACTTTCAGGAAAGACAGGCTCCGGAACGATCTCGTCCATGACTGGTAAATCCACGACA
CCGCAATGGCCCCCAGCACCTCTATCTCTCGTGCCAGGGGACTAACGT TGTATGCGTCTGCGTCTTGTCTTTTTGCATTCGCTTTCCAAAAAAGAGAGCCATCCGTTCCCCCGCACATTC
AACGCCGCGAGTGCGGTTTTTGTCTTTTTTGAGTGGTAGGACGCTTTTCATGCGCGAACTACGTGGACATTAAGTTCCATTCTCTTTTTCGACAGCACGAAACCTTGCATTCAAACCCGC
CCGCGGAAGATCCGATCTTGCTGCTGTTCGCAGTCCCAGTAGCGTCCTGTCGGCCGCGCCGTCT  CTGTTGGTGGGCAGCCGCTACACCTGTTATCTGACTGCCGTGCGCGAAAATGACGC
CATTTTTGGGAAAATCGGGGAACTTCATTCTTTAAAAGTATGCGGAGGTTTCCTTTTTCTTCTGTTCGTTTCTTTTTCTCGGGTTTGATAACCGTGTTCGATGTAAGCACTTTCCGTCTC
TCCTCCGTGCTTTGTTCGACATCGAGACCAGGTGTGCAGATCCTTCGCTTGTCGATCCGGAGACGCGTGTCTCGTAGAAC  CTTTTCATTTTACCACACGGCAGTGCGGAGCACTGCTCTG
AGTGCAGCAGGGACGGGTGAAGTTTCGCTTTAGTAGTGCGTTTCTGCTCTACGGGGCGTTGTCGTGTCTGGGAAGATGCAGAAACCGGTGTGTCTGGTCGTCGCGATGACCCCCAAGAGG
GGCATCGGCATCAACAACGGCCTCCCGTGGCCCCACTTGACCACAGATTTCAAACACTTTTCTCGTGTGACAAAAACGACGCCCGAAGAAGCCAGT  CGCCTGAACGGGTGGCTTCCCAGG
AAATTTGCAAAGACGGGCGACTCTGGACTTCCCTCTCCATCAGTCGGCAAGAGATTCAACGCCGTTGTCATGGGACGGAAAACCTGGGAAAGCATGCCTCGAAAGTTTAGACCCCTCGTG
GACAGATTGAACATCGTCGTTTCCTCTTCCCTGTGAGCACACACAGTAGTCGCCACACGCTGTTTGAGACGTGTCAATCTCCAAGAGTGTGGACGCTGTTCCACGTCTTCAA  ATGTTTCC
CAACATCCGTCGTCTAGTAGACACACCAACAAAAAGCACACGGCGAATCTGCTCATCGGAGGGAGGAGCCGGGGGGCACACAACTATCCTCAACTCTCGAACGAACATATCCGGGGCCGC
GAAGACGTCCAGTCTCTCAAATCCAACCCGGAACGCAAACATTTCTGCATCAAGTCACGATTGCGCCGGTACCTCCATGTGTAAGCAGTTCCATGAAACCTCCGATATTACACACGACTG
TGGATATGAATTATATGCAGATGCATATATACTGAGACGCCGATGCAACTATAGGTTTCCTGGCCCTCCATGGATATTTCAGACCTTCCTCTCACATTTGGTTTGCCCGTACACCTCCGT
TACGCTTTTTTTCTGGCTTTCTTCTTCGTCTCTGTTTATCAGCAAAGAAGAAGACATTGCGGCGGAGAAGCCTCAAGCTGAAGGCCAGCAGCGCGTCCGAGTCTGTGCTTCACTCCCAGC
AGCTCTCAGCCTTCTGGAGGAAGA GTACAAGGATTCTGTCGACCAGATTTTTGTCGTGGGTATGTTGTCCTAAACTCCTTGGAACTCCATTCTTGGTCAGAAACGTACTGAAACTGTATA
CATGTATATACAGATGTATGGATAATATCTAGAGAAGATACAGGGAAGACTGGCAAGGATGAAAAGACATGCAGCTTTAACGAAGCAGAGGGCATTGGCGAGAGGGACGCCCGTTATGCT
GTGTGATGTGGCTGTGAATCTTACCTCGCCGTTTGACTTG CTGCAGCGCTTTGTCCACTTGAACGTGACTTCTTGTTTCTACCTTCCCCAACGCCTTCTATTCCCTTCACTGCGAAAGCG
CGCTCAGTGGGCCGTCACCGAACACCCTTGGTTCTTTCGTTCAGCTGTTGTCCTCTTTCTCGCGTTGCTTCCTGTGGCGTCGTGGCTCGGCTTCTCTCTCTTTCCTGTTGGTGCGTCCAG
ACTATGTCGCCTGTTTCCCCACCCTTCTCGGCTTGTGCTTTCAGGAGGAGCGGGAC  TGTACGAGGCAGCGCTGTCTCTGGGCGTTGCCTCTCACCTGTACATCACGCGTGTAGCCCGCGA
GTTTCCGTGCGACGTTTTCTTCCCTGCGTTCCCCGGAGATGACATTCTTTCAAACAAATCAACTGCTGCGCAGGCTGCAGCTCCTGCCGAGTCTGTGTTCGTTCCCTTTTGTCCGGAGCT
CGGAAGAGAGAAGGACAATGAAGCGACGTATCGACCCATCTTCATTTCCAAGACCTTCTCAGACAACGGGGT  ACCCTACGACTTTGTGGTTCTCGAGAAGAGAAGGAAGACTGACGACGC
AGCCACTGCGGAACCGGTAAGAGGCAACCGAAGCGCGTAGATAAGAAAAACAACAAAGAGAAGGTGAAACACGAAGAGAAGGGAAAATGCGGAGAAACCGTGGATTTACAAAGATATCAA
GAGCAATGCTTTGTGGAGATTTTTTTAATTCAGTAGAGACACCCGCCGTGCGAGGTGTGTAGAAATAACTGCGACCCTGGAGACAGA  GATGCCGCGAGTACACCACTTGTCGTTTTTCC
TCCTATGTTCATGACGGGTGCTGAACGTCTATCGTACTTAATTGGAGGAGTCGTCTCCGAAGCAGCTTTGGCTGGCCATCCGTGTGTTTGCCTTGTTCCTGAAAAGCCAGAAGGCGCTCC
ACAGTGAGGCGATATACAGGGACGCCTACCGGAGCCCCGTTTTCTGCCTTTGTCGACTCTTGCAGAGCAACGCAATGAGCTCCTTGACGTCCACGAGGGAGACA  ACTCCCGTGCACGGGT
TGCAGGCTCCTTCTTCGGCCGCAGCCATTGCCCCGGTGTTGGCGTGGATGGACGAAGAAGACCGGAAAAAACGCGAGCAAAAGGAACTGATTCGGGCCGTTCCGCATGTTCACTTTAGAG
GCCATGAAGAATTCCAGTACCTTGATCTCATTGCCGACATTATTAACAATGGAAGGACAATGGATGACCGAACGGGTAACGGCGACTGCGAGAAAAAGCCACACCGTTTTCTCCTGTGAT
TCTGTCCGCAAGCCCTCTTTTGCTTCATCCACCCTTTGCTATTCTCCGCCGCCTTCCTTTTCTGCTCCATGTTCAATTCGTTCGCTTCTTCAGTCTTTCCATCTTCCCCTGTTACCTCTG
TCATTCGTTTTCTTGCCTCTATTTAACTGTGTTCTACTCACAGTCTGCATTCCGCGATAGACGAGCTTCCACGTCTTGCGTCTCGACAAGCAACTGTCATTTGTACGCGCCTCCCTCCAC
CGTGAATCGGATTGTC GGTTCGCCGGTTCCTGGGTCAGAAAAGGCCTGCGCCAGTATTCTGAATAATACCCTTCGCCATTGTAAAGAGGCGAAGGAACAAAGAGATATTTCGGCGCATCT
TTTGTGCGGCGCGTTTCCTCGTGCTTCACACCGATGCCCTTCTGTGCATGTCTTCTGCTCCTCGTCCTTCTCTCTTTTTCCCTGTTTAGGCGTTGGTGTCATCTCCAAATTCGGCTGCAC
TATGCGCTACTCGCTGGATCAGGCCTTTCCAC TTCTCACCACAAAGCGTGTGTTCTGGAAAGGGTAAGGGCGTCTTCAGTGAATGCATATATTTGACTTCAGACATTCTTAACTGTTTGA
CAACCAACGTACAAATTTGTTTGTCCGTGTGCGTGTTCGACATGTCAAGTATGTGAAGAGTCGCTACTGTAGACTAACGCACGAACCAGATTTGTTTATCTGCATGCGCTGTGCACCCGT
TTCTGAGTGTCTGGAGTTTCCGCAACCTTCCTTTGAATTTCTGGGTTC GTTTTTTTATGCGCGCACTGGTTTGCATGTGGCCTGAGAGAGCACAGATCGAAGGTGGGGTGATGTGGCGTC
GCTGCAGAGAAACTCCGGCGAAGGCGACAGATAAAGGAGAGTGGAAATCATTGAACAGTGTCGGTCGTCTGTTGTTTCGCAGGGTCCTCGAAGAGTTGCTGTGGTTCATTCGCGGCGACA
CGAACGCAAACCATCTTTCTGAGAAGGGCGTGAAGGCAAGTCTACGTTGTACCTCTTGTCTCTG  CCGAAGCTCAGATGTCTCCACGGCGTTGGTTTCTTTTCGTTTTTGCTTTCGTGGCA
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)

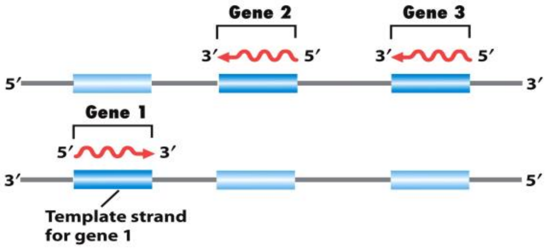
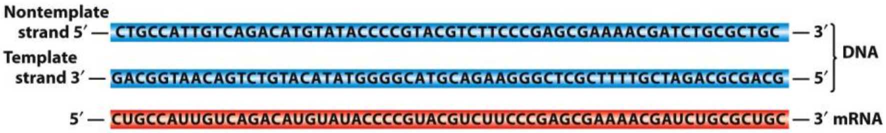


Figure 8-3
Introduction to Genetic Analysis, Ninth Edition
© 2008 W.H. Freeman and Company



Nontemplate strand 5′ — CTGCCATTGTCAGACATGTATACCCCGTACGTCTTCCCGAGCGAAAACGATCTGCGCTGC — 3′

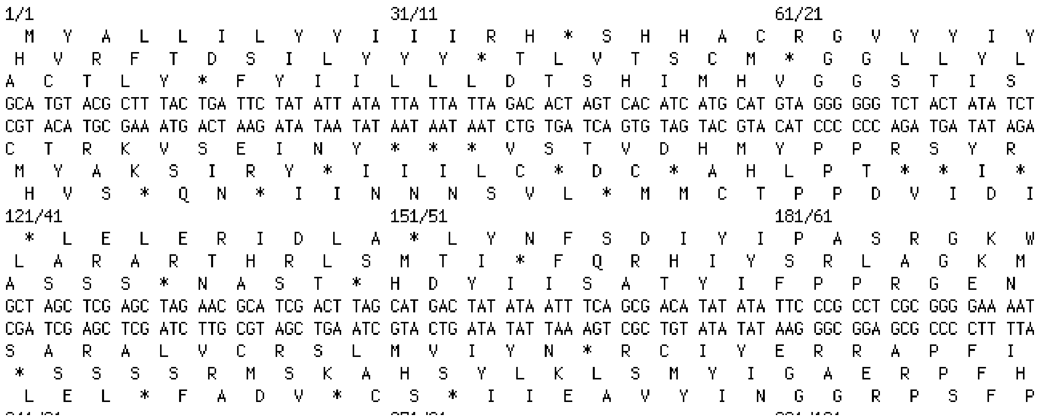
Template strand 3′ — GACGGTAACAGTCTGTACATATGGGGCATGCAGAAGGGCTCGCTTTTGCTAGACGCGACG — 5′

DNA

5′ — CUGCCAUUGUCAGACAUGUAUACCCCGUACGUCUUCCCGAGCGAAAACGAUCUGCGCUGC — 3′ mRNA

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

Six Frame Translation
Looking for Open Reading Frames, ORFs



```
1/1                                     31/11                                   61/21
  M   Y   A   L   L   I   L   Y   Y   I   I   I   R   H   *   S   H   H   A   C   R   G   V   Y   Y   I   Y
 H   V   R   F   T   D   S   I   L   Y   Y   Y   *   T   L   V   T   S   C   M   *   G   G   L   L   Y   L
A   C   T   L   Y   *   F   Y   I   I   L   L   L   D   T   S   H   I   M   H   V   G   G   S   T   I   S
GCA TGT ACG CTT TAC TGA TTC TAT ATT ATA TTA TTA TTA GAC ACT AGT CAC ATC ATG CAT GTA GGG GGG TCT ACT ATA TCT
CGT ACA TGC GAA ATG ACT AAG ATA TAA TAT AAT AAT AAT CTG TGA TCA GTG TAG TAC GTA CAT CCC CCC AGA TGA TAT AGA
C   T   R   K   V   S   E   I   N   Y   *   *   *   V   S   T   V   D   H   M   Y   P   P   R   S   Y   R
 M   Y   A   K   S   I   R   Y   *   I   I   I   L   C   *   D   C   *   A   H   L   P   T   *   *   I   *
  H   V   S   *   Q   N   *   I   I   N   N   N   S   V   L   *   M   M   C   T   P   P   D   V   I   D   I
121/41                                  151/51                                  181/61
  *   L   E   L   E   R   I   D   L   A   *   L   Y   N   F   S   D   I   Y   I   P   A   S   R   G   K   W
 L   A   R   A   R   T   H   R   L   S   M   T   I   *   F   Q   R   H   I   Y   S   R   L   A   G   K   M
A   S   S   S   *   N   A   S   T   *   H   D   Y   I   I   S   A   T   Y   I   F   P   P   R   G   E   N
GCT AGC TCG AGC TAG AAC GCA TCG ACT TAG CAT GAC TAT ATA ATT TCA GCG ACA TAT ATA TTC CCG CCT CGC GGG GAA AAT
CGA TCG AGC TCG ATC TTG CGT AGC TGA ATC GTA CTG ATA TAT TAA AGT CGC TGT ATA TAT AAG GGC GGA GCG CCC CTT TTA
S   A   R   A   L   V   C   R   S   L   M   V   I   Y   N   *   R   C   I   Y   E   R   R   A   P   F   I
 *   S   S   S   S   R   M   S   K   A   H   S   Y   L   K   L   S   M   Y   I   G   A   E   R   P   F   H
  L   E   L   *   F   A   D   V   *   C   S   *   I   I   E   A   V   Y   I   N   G   G   R   P   S   F   P
```

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

```
AAGCTTCGCCAGGCTGTAAATCCCGTGAGTCGTCCTCACAAATCATCAAGCAGGTGTCCTCAGGGAGACTGCCTGACTGAGTTATGCTAATTCCTTTCTACTTTGGCGTG
GTCACGTGTAACCATATCCGAATCATTTCTCTAGCCCTACGAACAGGTAAGAGCGCTAGGGATGTCCGTGGAGTAGTGTGCTTACTCGATAATATTCAGTTGGGACTACC
AGCGAGGCGCTCGCTTTGCTCACGCAATGCCTGAGACAGTTGCAGAATGAATGGTAACCGACAAACGCGTTCATATGCGTTTTCAAACTTAGTAGACGCGTACTGTCTGA
AACTGGCGGTCACAGGCACCAGATAACGCCCTTGGCATCGGCATGTCTCGTACAGAGGTCCGTATGTAGTGCCACGACTTCTAAATCCGGCGACAGGCTGGTCTTTTGTC
TTACCACGTATTAGCCCGCGTGCGATTTCTCGGAGCGCACCTGTTCAACACTAGAAAACGGAGTTTCCTGATCGAGAAGCCACCACCTTTCCAGAAGTTGAACGCTAGCA
TGTCATTCGATTTTCACCCCCCGCGTAGTTCCTGTGTGTCATTCGTTGTCGAGACAACTCTGTCCCGCCCCGGTGCTGTTCCATATGCGTGACTTTCCCGCAATTTTTC
AGACTTTCAGGAAAGACAGGCTCCGGAACGATCTCGTCCATGACTGGTAAATCCACGACACCGCAATGGCCCCCAGCACCTCTATCTCTCGTGCCAGGGGACTAACGTTG
TATGCGTCTGCGTCTTGTCTTTTTGCATTCGCTTTCCAAAAAAGAGAGCCATCCGTTCCCCCGCACATTCAACGCCGCGAGTGCGGTTTTTGTCTTTTTTGAGTGGTAGG
ACGCTTTTCATGCGCGAACTACGTGGACATTAAGTTCCATTCTCTTTTTCGACAGCACGAAACCTTGCATTCAAACCCGCCCGCGGAAGATCCGATCTTGCTGCTGTTCG
CAGTCCCAGTAGCGTCCTGTCGGCCGCGCCGTCTCTGTTGGTGGGCAGCCGCTACACCTGTTATCTGACTGCCGTGCGCGAAAATGACGCCATTTTTGGGAAAATCGGGG
AACTTCATTCTTTAAAAGTATGCGGAGGTTTCCTTTTTCTTCTGTTCGTTTCTTTTTCTCGGGTTTGATAACCGTGTTCGATGTAAGCACTTTCCGTCTCTCCTCCGTGC
TTTGTTCGACATCGAGACCAGGTGTGCAGATCCTTCGCTTGTCGATCCGGAGACGCGTGTCTCGTAGAACCTTTTCATTTTACCACACGGCAGTGCGGAGCACTGCTCTG
AGTGCAGCAGGGACGGGTGAAGTTTCGCTTTAGTAGTGCGTTTCTGCTCTACGGGGCGTTGTCGTGTCTGGGAAGATGCAGAAACCGGTGTGTCTGGTCGTCGCGATGAC
CCCCAAGAGGGGCATCGGCATCAACAACGGCCTCCCGTGGCCCCACTTGACCACAGATTTCAAACACTTTTCTCGTGTGACAAAAACGACGCCCGAAGAAGCCAGTCGCC
TGAACGGGTGGCTTCCCAGGAAATTTGCAAAGACGGGCGACTCTGGACTTCCCTCTCCATCAGTCGGCAAGAGATTCAACGCCGTTGTCATGGGACGGAAAACCTGGGAA
AGCATGCCTCGAAAGTTTAGACCCCTCGTGGACAGATTGAACATCGTCGTTTCCTCTTCCCTGTGAGCACACACAGTAGTCGCCACACGCTGTTTGAGACGTGTCAATCT
CCAAGAGTGTGGACGCTGTTCCACGTCTTCAAATGTTTCCCAACATCCGTCGTCTAGTAGACACACCAACAAAAAGCACACGGCGAATCTGCTCATCGGAGGGAGGAGCC
GGGGGGCACACAACTATCCTCAACTCTCGAACGAACATATCCGGGGCCGCGAAGACGTCCAGTCTCTCAAATCCAACCCGGAACGCAAACATTTCTGCATCAAGTCACGA
TTGCGCCGGTACCTCCATGTGTAAGCAGTTCCATGAAACCTCCGATATTACACACGACTGTGGATATGAATTATATGCAGATGCATATATACTGAGACGCCGATGCAACT
ATAGGTTTCCTGGCCCTCCATGGATATTTCAGACCTTCCTCTCACATTTGGTTTGCCCGTACACCTCCGTTACGCTTTTTTTCTGGCTTTCTTCTTCGTCTCTGTTTATC
AGCAAAGAAGAAGACATTGCGGCGGAGAAGCCTCAAGCTGAAGGCCAGCAGCGCGTCCGAGTCTGTGCTTCACTCCCAGCAGCTCTCAGCCTTCTGGAGGAAGAGTACAA
GGATTCTGTCGACCAGATTTTTGTCGTGGGTATGTTGTCCTAAACTCCTTGGAACTCCATTCTTGGTCAGAAACGTACTGAAACTGTATACATGTATATACAGATGTATG
GATAATATCTAGAGAAGATACAGGGAAGACTGGCAAGGATGAAAAGACATGCAGCTTTAACGAAGCAGAGGGCATTGGCGAGAGGGACGCCCGTTATGCTGTGTGATGTG
GCTGTGAATCTTACCTCGCCGTTTGACTTGCTGCAGCGCTTTGTCCACTTGAACGTGACTTCTTGTTTCTACCTTCCCCAACGCCTTCTATTCCCTTCACTGCGAAAGCG
CGCTCAGTGGGCCGTCACCGAACACCCTTGGTTCTTTCGTTCAGCTGTTGTCCTCTTTCTCGCGTTGCTTCCTGTGGCGTCGTGGCTCGGCTTCTCTCTCTTTCCTGTTG
GTGCGTCCAGACTATGTCGCCTGTTTCCCCACCCTTCTCGGCTTGTGCTTTCAGGAGGAGCGGGACTGTACGAGGCAGCGCTGTCTCTGGGCGTTGCCTCTCACCTGTAC
ATCACGCGTGTAGCCCGCGAGTTTCCGTGCGACGTTTTCTTCCCTGCGTTCCCCGGAGATGACATTCTTTCAAACAAATCAACTGCTGCGCAGGCTGCAGCTCCTGCCGA
GTCTGTGTTCGTTCCCTTTTGTCCGGAGCTCGGAAGAGAGAAGGACAATGAAGCGACGTATCGACCCATCTTCATTTCCAAGACCTTCTCAGACAACGGGGTACCCTACG
ACTTTGTGGTTCTCGAGAAGAGAAGGAAGACTGACGACGCAGCCACTGCGGAACCGGTAAGAGGCAACCGAAGCGCGTAGATAAGAAAAACAACAAAGAGAAGGTGAAAC
ACGAAGAGAAGGGAAAATGCGGAGAAACCGTGGATTTACAAAGATATCAAGAGCAATGCTTTGTGGAGATTTTTTTTAATTCAGTAGAGACACCCGCCGTGCGAGGTGTG
TAGAAATAACTGCGACCCTGGAGACAGAGATGCCGCGAGTACACCACTTGTCGTTTTTCCTCCTATGTTCATGACGGGTGCTGAACGTCTATCGTACTTAATTGGAGGAG
TCGTCTCCGAAGCAGCTTTGGCTGGCCATCCGTGTGTTTGCCTTGTTCCTGAAAAGCCAGAAGGCGCTCCACAGTGAGGCGATATACAGGGACGCCTACCGGAGCCCCGT
TTTCTGCCTTTGTCGACTCTTGCAGAGCAACGCAATGAGCTCCTTGACGTCCACGAGGGAGACAACTCCCGTGCACGGGTTGCAGGCTCCTTCTTCGGCCGCAGCCATTG
CCCCGGTGTTGGCGTGGATGGACGAAGAAGACCGGAAAAAACGCGAGCAAAAGGAACTGATTCGGGCCGTTCCGCATGTTCACTTTAGAGGCCATGAAGAATTCCAGTAC
CTTGATCTCATTGCCGACATTATTAACAATGGAAGGACAATGGATGACCGAACGGGTAACGGCGACTGCGAGAAAAAGCCACACCGTTTTCTCCTGTGATTCTGTCCGCA
AGCCCTCTTTTGCTTCATCCACCCTTTGCTATTCTCCGCCGCCTTCCTTTTCTGCTCCATGTTCAATTCGTTCGCTTCTTCAGTCTTTCCATCTTCCCCTGTTACCTCTG
TCATTCGTTTTCTTGCCTCTATTTAACTGTGTTCTACTCACAGTCTGCATTCCGCGATAGACGAGCTTCCACGTCTTGCGTCTCGACAAGCAACTGTCATTTGTACGCGC
CTCCCTCCACCGTGAATCGGATTGTCGGTTCGCCGGTTCCTGGGTCAGAAAAGGCCTGCGCCAGTATTCTGAATAATACCCTTCGCCATTGTAAAGAGGCGAAGGAACAA
AGAGATATTTCGGCGCATCTTTTGTGCGGCGCGTTTCCTCGTGCTTCACACCGATGCCCTTCTGTGCATGTCTTCTGCTCCTCGTCCTTCTCTCTTTTTCCCTGTTTAGG
CGTTGGTGTCATCTCCAAATTCGGCTGCACTATGCGCTACTCGCTGGATCAGGCCTTTCCACTTCTCACCACAAAGCGTGTGTTCTGGAAAGGGTAAGGGCGTCTTCAGT
GAATGCATATATTTGACTTCAGACATTCTTAACTGTTTGACAACCAACGTACAAATTTGTTTGTCCGTGTGCGTGTTCGACATGTCAAGTATGTGAAGAGTCGCTACTGT
AGACTAACGCACGAACCAGATTTGTTTATCTGCATGCGCTGTGCACCCGTTTCTGAGTGTCTGGAGTTTCCGCAACCTTCCTTTGAATTTCTGGGTTCGTTTTTTTATGC
```

The “Coding Sequence” - CDS

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

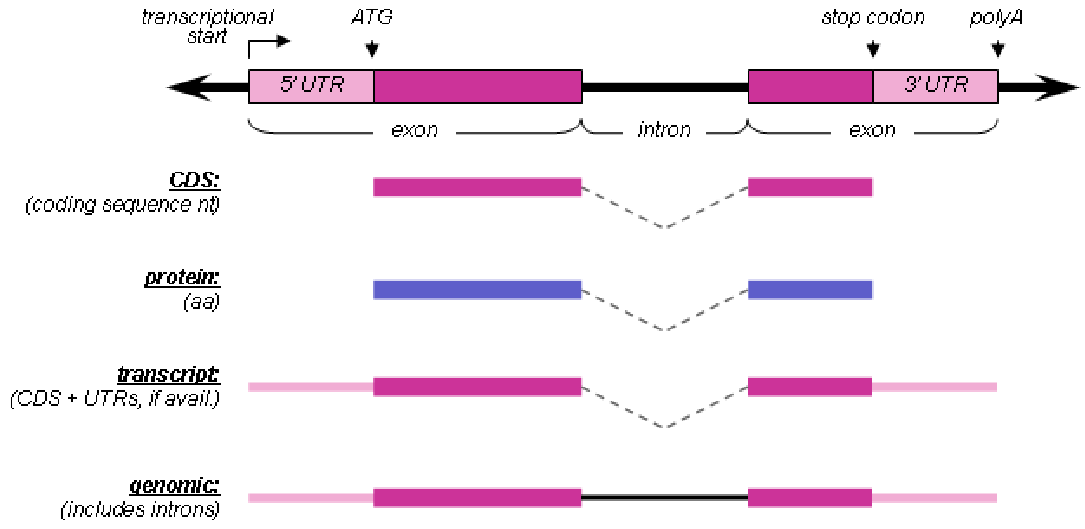
The Protein

```
>Translation Frame 1

MQKPVCLVVAMTPKRGIGINNGLPWPHLTTDFKHFSRVTKTTPEEASRLN

GWLPRKFAKTGDSGLPSPSVGKRFNAVVMGRKTWESMPRKFRPLVDRLNI

VVSSSLKEEDIAAEKPQAEGQQRVRVCASLPAALSLLEEEYKDSVDQIFV

VGGAGLYEAALSLGVASHLYITRVAREFPCDVFFPAFPGDDILSNKSTAA

QAAAPAESVFVPFCPELGREKDNEATYRPIFISKTFSDNGVPYDFVVLEK

RRKTDDAATAEPSNAMSSLTSTRETTPVHGLQAPSSAAAIAPVLAWMDEE

DRKKREQKELIRAVPHVHFRGHEEFQYLDLIADIINNGRTMDDRT
```

Green = UTRs **Red = CDS** **Pink = Intron**

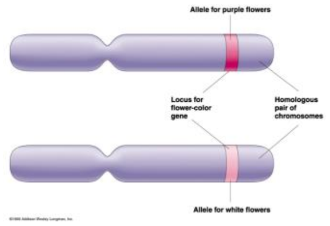
23

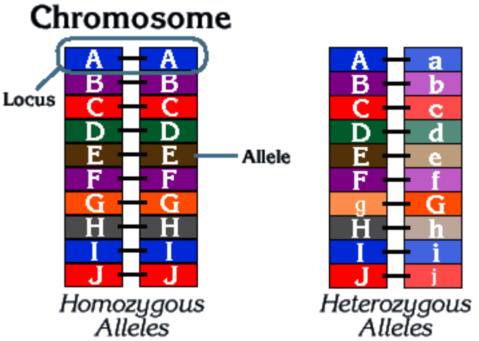
Terminology



24

Evolution Homologous chromsomes (in a diploid)


Allele for purple flowers
Locus for flower-color gene
Homologous pair of chromosomes
Allele for white flowers
©1999 Addison Wesley Longman, Inc.



```
A    AAGCCTCATC

a    ACGCCTCATC
```

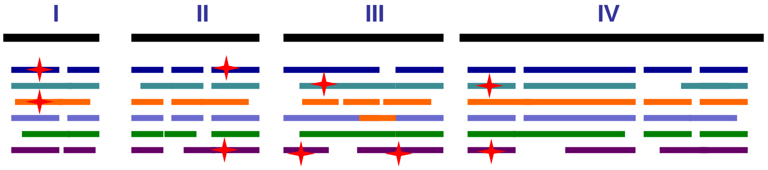
SNP =Single Nucleotide Polymorphism (a variant)

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



NGS sequence reads from different clinical/environmental isolates

✦ = SNP
= SNV

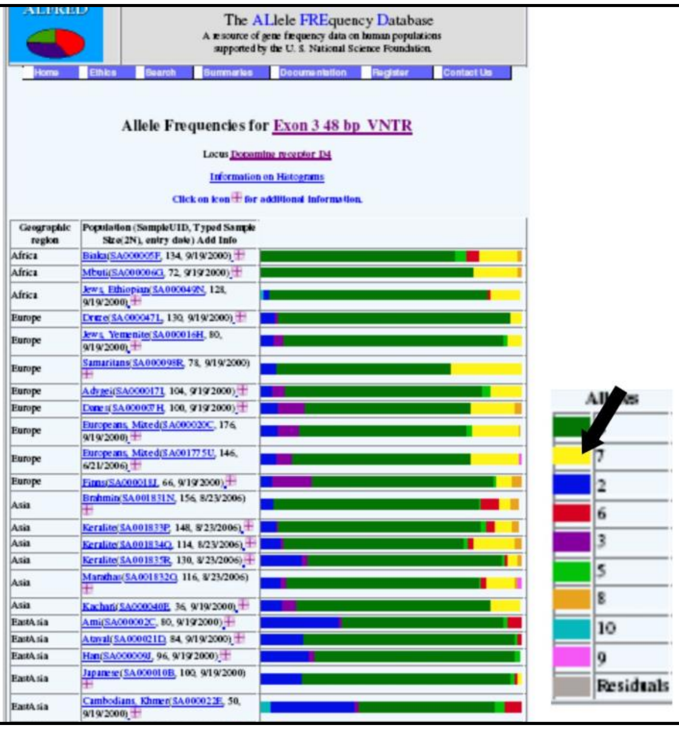
```
   reference   TGGTGATACT AAGCTGGGAA CTCCACTTCT TTTTCTACTG CGGTGCTTCA
       303.1   TGGTGATACT AAGCTGGGAA CTCCACTTCT TTTTCTACTG CGGTGCTTTA
       309.1   TGATAATNCT AAACTGGGAA CTCCACTTCC TTTTCTACTG CAGTGCTTCA
     RV_3600   TGGTGATACT AAACTGGGAA CTCCACTTCT TTTTCTACTG CGGTGCTTCA
     RV_3606   TGATAATNCT 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)

27



The ALlele FREquency Database
A resource of gene frequency data on human populations supported by the U. S. National Science Foundation.

Allele Frequencies for Exon 3 48 bp VNTR

Locus Dopamine receptor D4

Information on Histograms

Click on icon ➕ for additional information.

Geographic region	Population (SampleUID, Typed Sample Size(2N), entry date) Add Info
Africa	Biaka(SA000005F, 134, 9/19/2000) ➕
Africa	Mbuti(SA000006G, 72, 9/19/2000) ➕
Africa	Jews, Ethiopian(SA000049N, 128, 9/19/2000) ➕
Europe	Druze(SA000047L, 130, 9/19/2000) ➕
Europe	Jews, Yemenite(SA000016H, 80, 9/19/2000) ➕
Europe	Samaritans(SA000098R, 78, 9/19/2000) ➕
Europe	Adygei(SA000017I, 104, 9/19/2000) ➕
Europe	Danes(SA000007H, 100, 9/19/2000) ➕
Europe	Europeans, Mixed(SA000020C, 176, 9/19/2000) ➕
Europe	Europeans, Mixed(SA001775U, 146, 6/21/2006) ➕
Europe	Finns(SA000018J, 66, 9/19/2000) ➕
Asia	Brahmin(SA001831N, 156, 8/23/2006) ➕
Asia	Keralite(SA001833P, 148, 8/23/2006) ➕
Asia	Keralite(SA001834Q, 114, 8/23/2006) ➕
Asia	Keralite(SA001835R, 130, 8/23/2006) ➕
Asia	Marathas(SA001832O, 116, 8/23/2006) ➕
Asia	Kachari(SA000040E, 36, 9/19/2000) ➕
EastAsia	Ami(SA000002C, 80, 9/19/2000) ➕
EastAsia	Atayal(SA000021D, 84, 9/19/2000) ➕
EastAsia	Han(SA000009J, 96, 9/19/2000) ➕
EastAsia	Japanese(SA000010B, 100, 9/19/2000) ➕
EastAsia	Cambodians, Khmer(SA000022E, 50, 9/19/2000) ➕

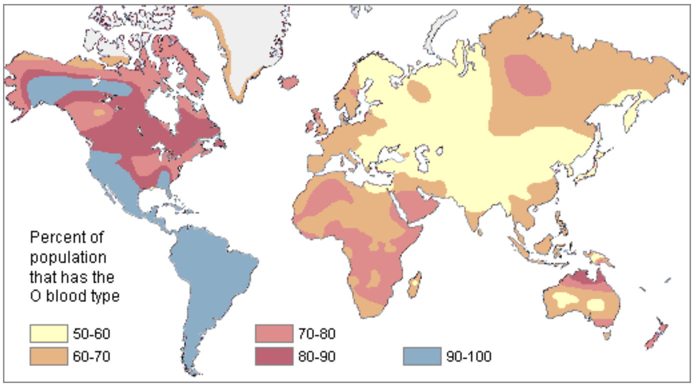
Alleles	
🟩	
🟨	7
🟦	2
🟥	6
🟪	3
🟩	5
🟧	8
▨	10
▨	9
▨	Residuals

Alleles have frequencies in different populations

28

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



29

Population variation data

Data

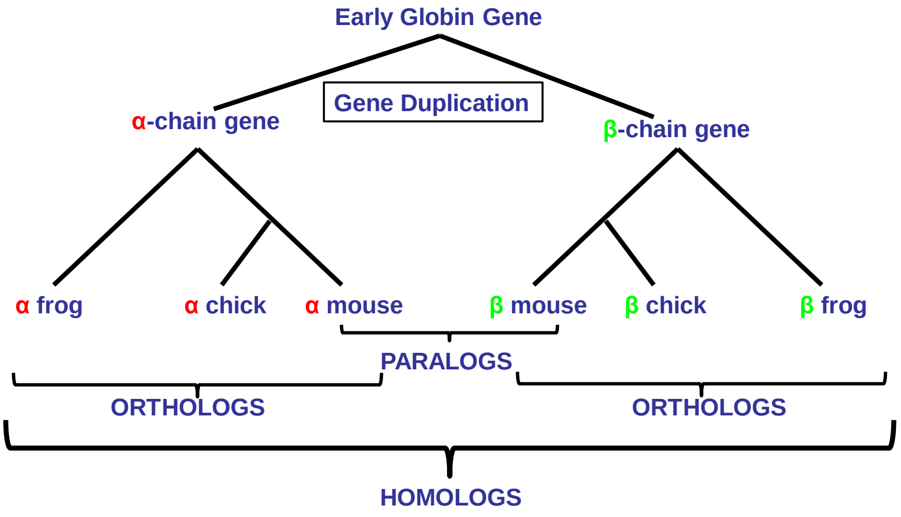
- Single Nucleotide Polymorphisms, SNPs. SNVs
- Rearrangements
- Alleles
- Allele frequency
- Haplotypes (an organism's collection of variants)

Technology

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

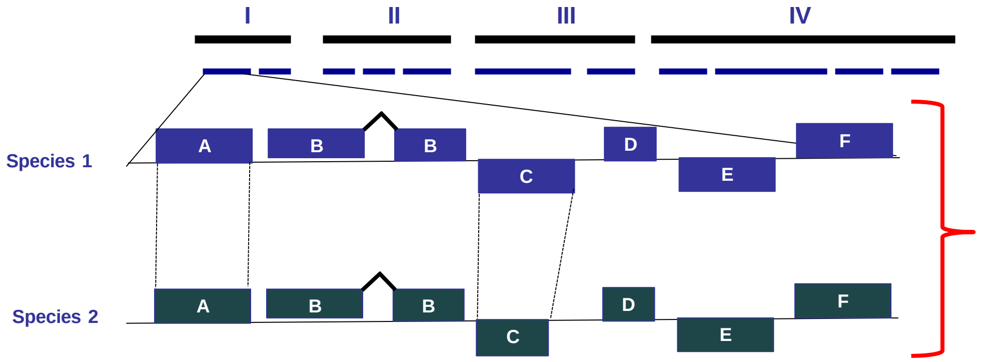
30

Homology – a vocabulary for different types of evolutionary relationships



31

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

32

Synteny among *Plasmodium* species

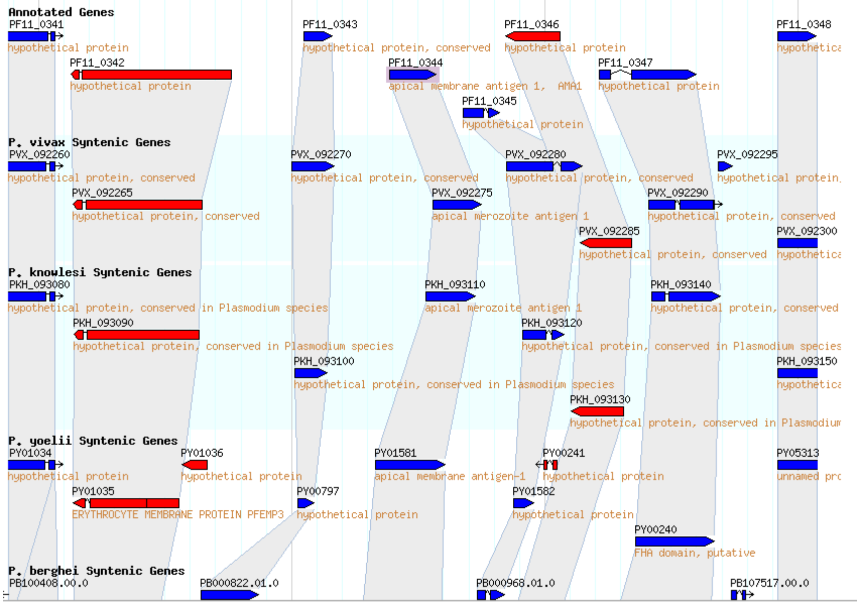
P. falciparum

P. vivax

P. knowlesi

P. yoelii

P. berghei



33

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.

34

For Example:

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

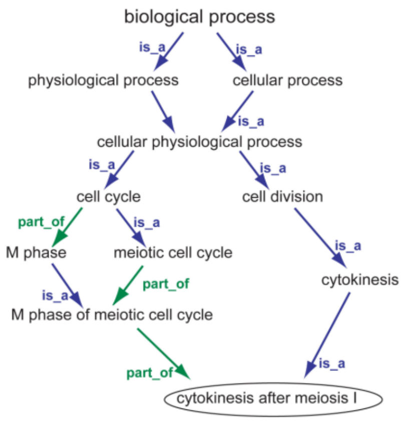
Both can be annotated with GO term:

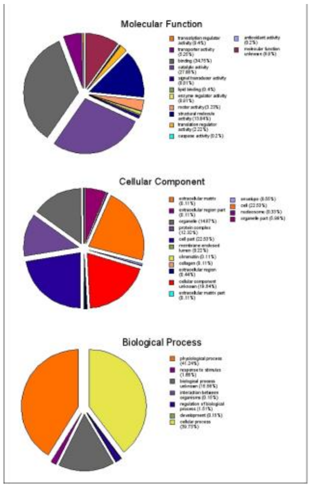
GO:0003705 (RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity)

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

35

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





36

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

37

RNA expression

Bulk sequencing from many cells

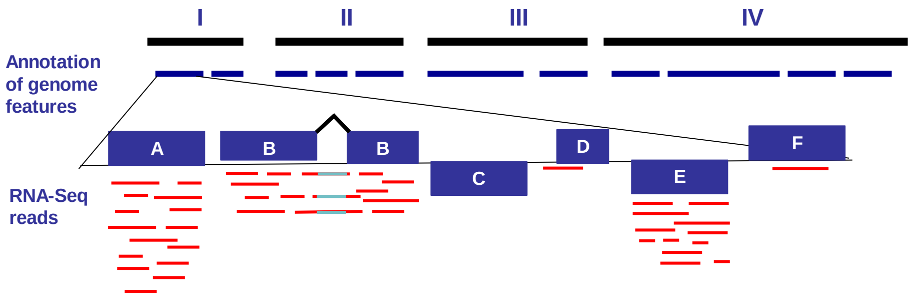
- RNA-Seq (NGS)
  - Little sequence bias
  - Quantitative
  - Usually are strand-specific
- PacBio ISO-seq
  - Full-length transcripts from single molecules
- ONT Direct seq
  - Single-molecule, direct sequencing of RNA (or can sequence cDNA)
- All of these methods can be used to identify UTR's and exon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

38

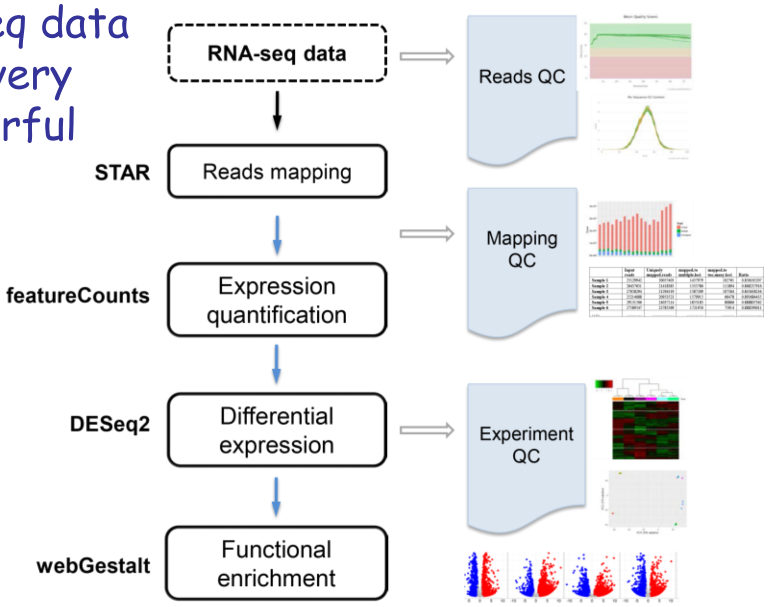
30,000 ft View – RNA-Seq



FPKM = Fragments per kilobase of exon per million fragments mapped (old calculation)
TPM = Transcripts per kilobase million (counts per length of transcript (kb) per million reads mapped

39

RNA-seq data are very powerful



RNA-seq data

STAR — Reads mapping

featureCounts — Expression quantification

DESeq2 — Differential expression

webGestalt — Functional enrichment

Reads QC

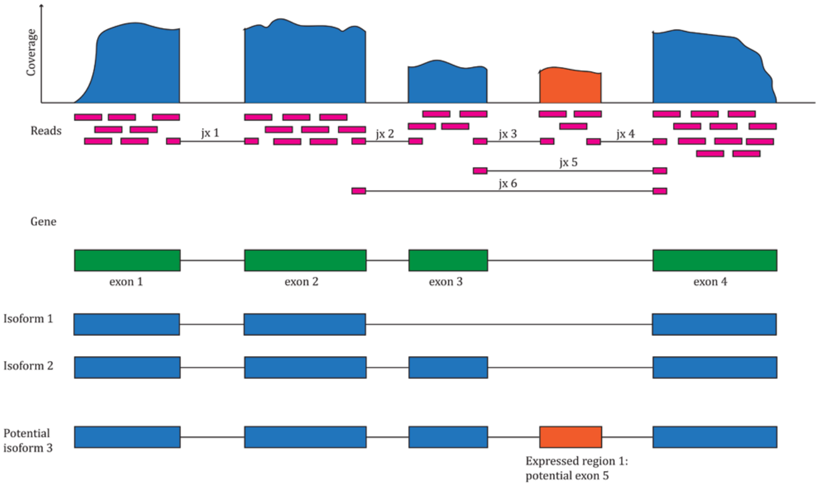
Mapping QC

Experiment QC

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

40

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?

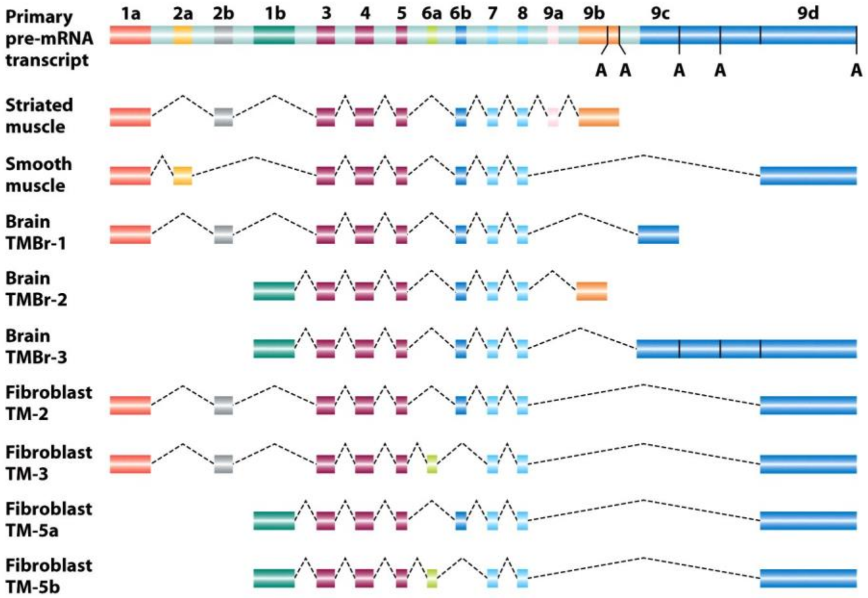
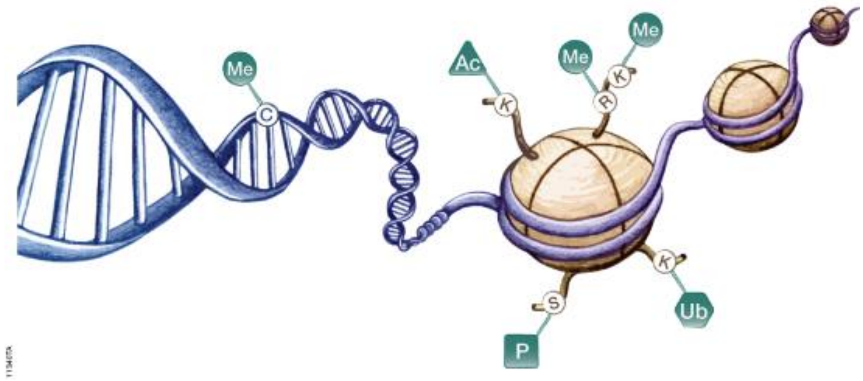


Figure 8-14
Introduction to Genetic Analysis, Ninth Edition
© 2008 W. H. Freeman and Company

Chromatin Status and Epigenetic Gene Regulation



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

- 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

43

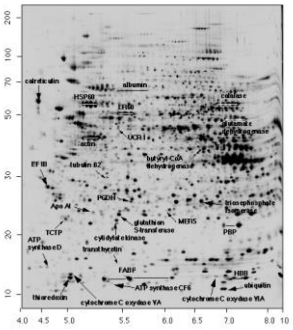
Protein Expression/Sequence

Data

- MW-Isoelectric point
- MW
- Sequence/spans

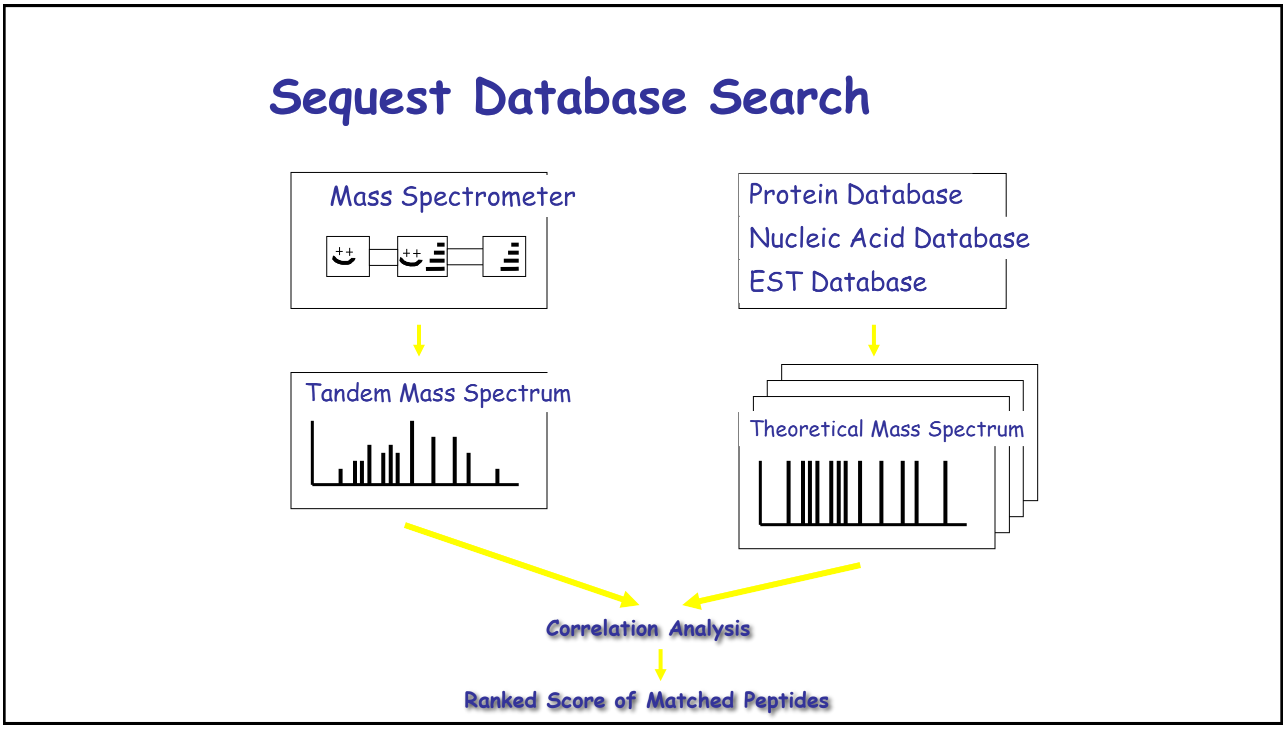
Technology

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



Typical 2 D gel

44



Sequest Database Search

Mass Spectrometer

Protein Database

Nucleic Acid Database

EST Database

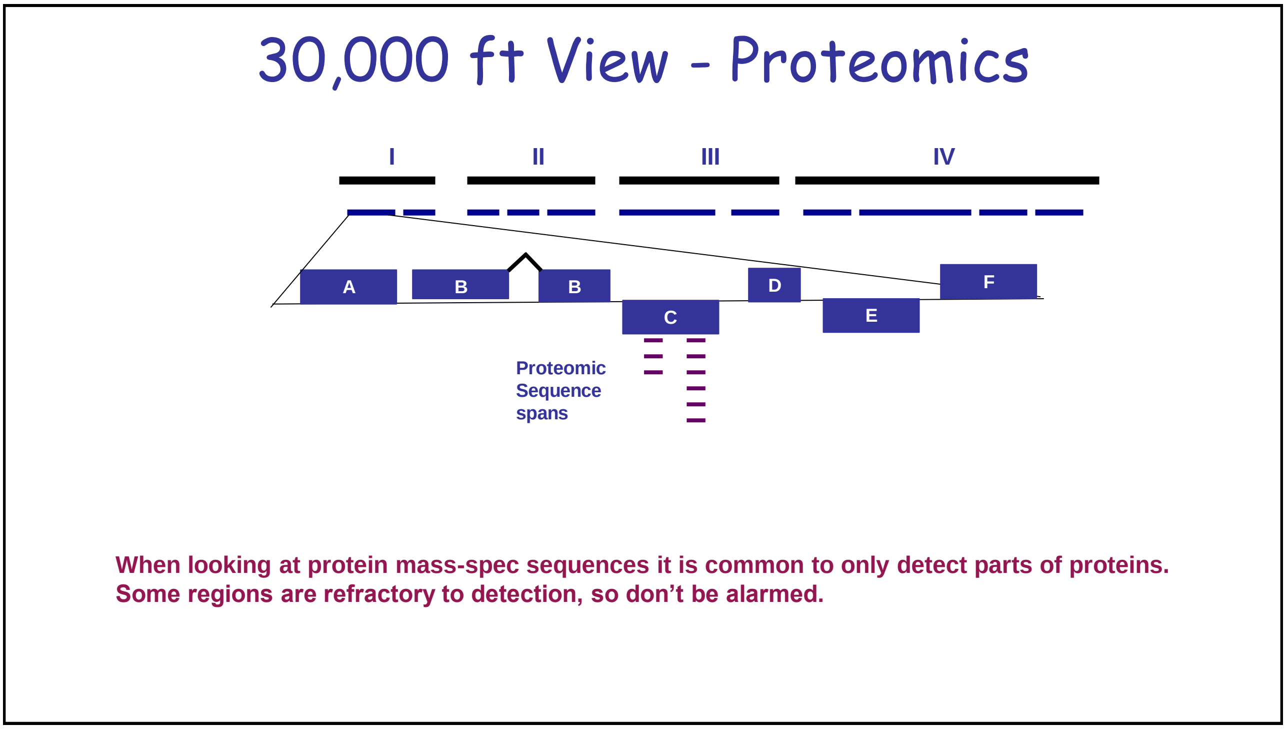
Tandem Mass Spectrum

Theoretical Mass Spectrum

Correlation Analysis

Ranked Score of Matched Peptides

45

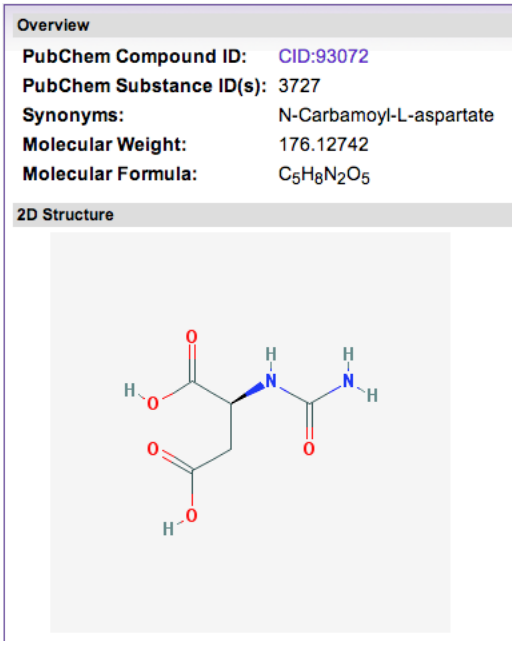


30,000 ft View - Proteomics

Proteomic Sequence spans

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.

46



Overview

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

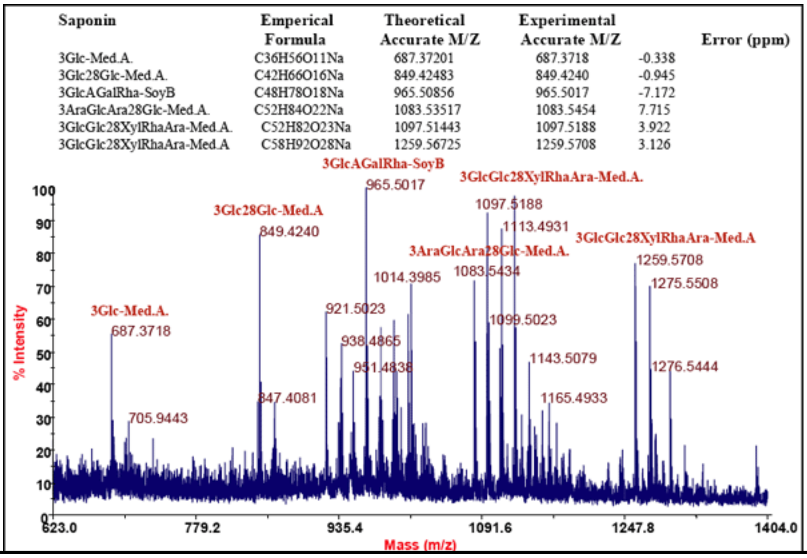
2D Structure

Metabolites

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

47

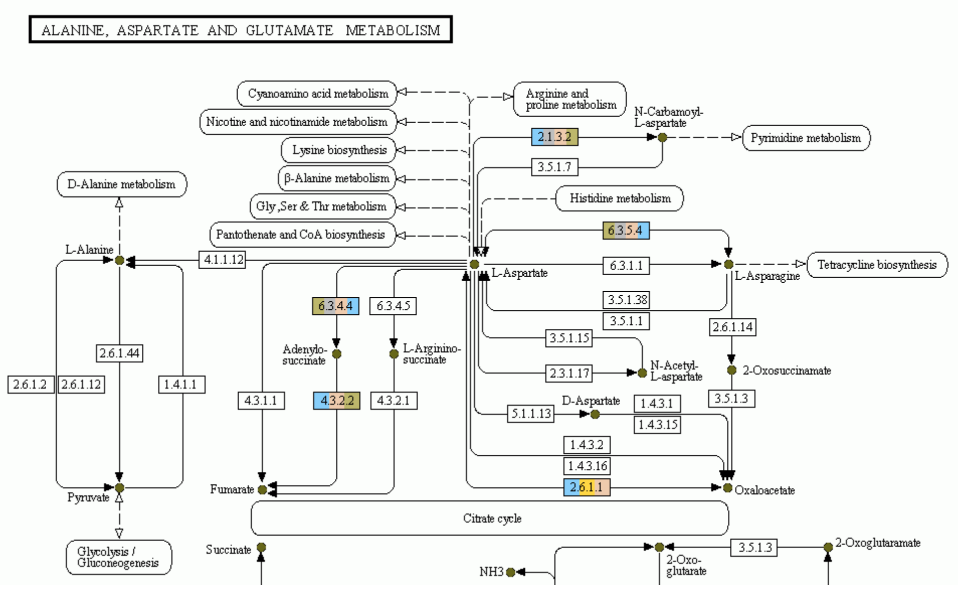
Complex mixtures can be analyzed and interpreted



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

48

Metabolites can be linked to metabolic pathways and enzymes

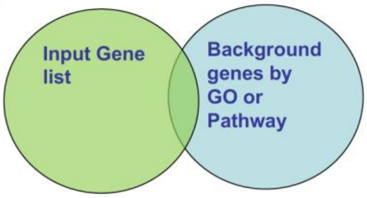


Gene & Pathway Enrichment

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



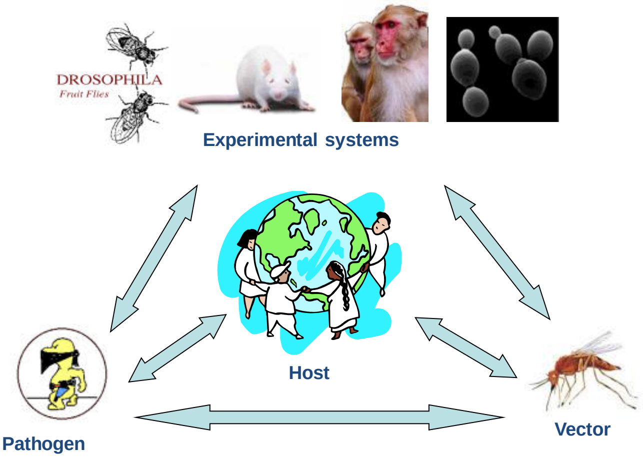
Result: GO:ID or Pathway ID that is enriched

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

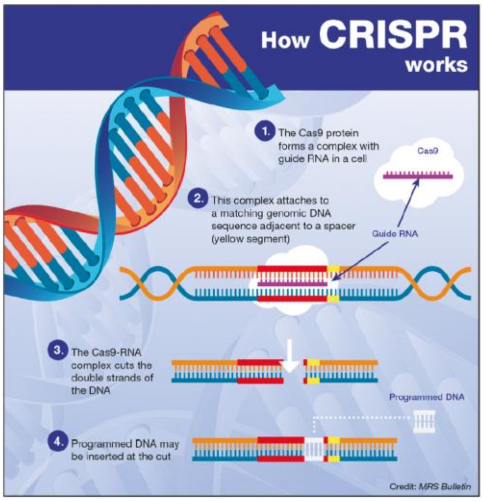
Host

Pathogen

Vector

Mutant analysis

CRISPR-CAS



1. The Cas9 protein forms a complex with guide RNA in a cell
2. This complex attaches to a matching genomic DNA sequence adjacent to a spacer (yellow segment)
3. The Cas9-RNA complex cuts the double strands of the DNA
4. Programmed DNA may be inserted at the cut

Credit: *MRS Bulletin*

- 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

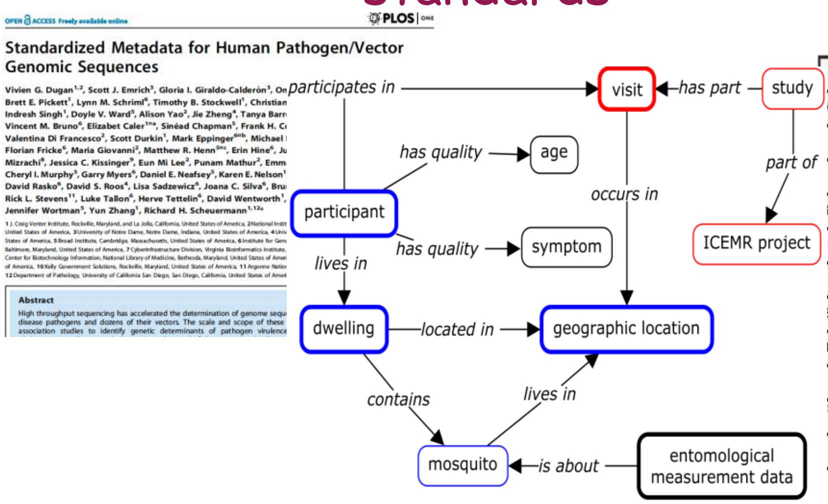
Ball et al., MRS Bulletin November 2016

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

53

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[1na], Sinéad Chapman[5], Frank H. C Valentina Di Francesco[2], Scott Durkin[1], Mark Eppinger[6nb], Michael Florian Fricke[6], Maria Giovanni[2], Matthew R. Henn[5nc], 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 Gen 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 Amer of America, 10 Kelly Government Solutions, Rockville, Maryland, United States of America, 11 Argonne Natio 12 Department of Pathology, University of California San Diego, San Diego, California, United States of Ame

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

	Core Project	Core Sample	Project Specific	Pathogen Specific	Sequencing Assay
Investigation	■		■		
Host Characterization		■	■	■	
Specimen Isolation		■	■	■	
Pathogen Characterization		■	■	■	
Specimen Processing		■	■	■	
Pathogen Detection		■			
Pathogen Isolation			■	■	
Sample Management			■	■	
Data Transformation				■	■
Sample Shipment					■
Sequencing Sample Preparation					■
Sequencing Assay					■

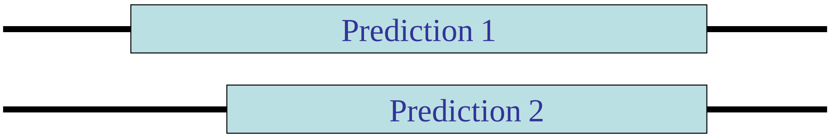
54

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.

55

Different algorithms often generate different results



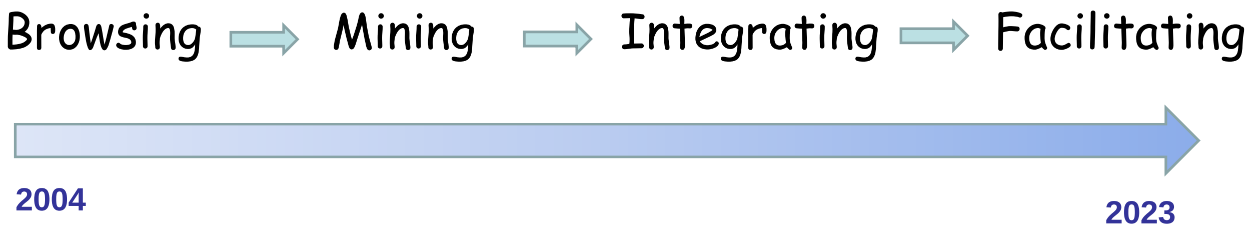
We provide lots evidence so that you can decide or design an experiment to confirm!

56

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



Browsing → Mining → Integrating → Facilitating

2004

2023

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.

59

VEuPathDB

Eukaryotic Pathogen, Vector & Host Informatics Resources

NIH National Institute of Allergy and Infectious Diseases

wellcome

Thank you to the data providers, participants and community for their feedback

Project Leadership:
David Roos – UPENN (joint-PI)
Mary Ann McDowell – Notre Dame (Joint-PI)
Andrew Jones – Liverpool
Jessie Kissinger – UGA
Sarah Dyer – EBI
Kathryn Crouch – 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 Notre Dame

University of Liverpool

University of Georgia

EMBL-EBI

University of Glasgow

Imperial College London

60

60