

Functional Annotation & Comparative Genomics

Aroon Chande

Lecture 15

Tuesday, February 27, 2018

Outline

- Functional Annotation
 - What is functional annotation?
 - Annotation approaches
 - Tools and techniques

- Comparative Genomics
 - What is comparative genomics?
 - Tools and techniques

Outline

- **Functional Annotation**
 - What is functional annotation?
 - Annotation approaches
 - Tools and techniques

- **Comparative Genomics**
 - What is comparative genomics?
 - Tools and techniques

What is functional annotation

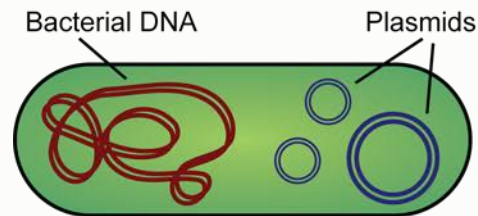
- You have assembled genomes...what now?
- Predict genes!
- Ok...so you have these genomics regions that you think encode genes, which make proteins, or ncRNAs, or other interesting features. What does this do for me?

What is functional annotation

- Let's take a few steps back...

What is functional annotation

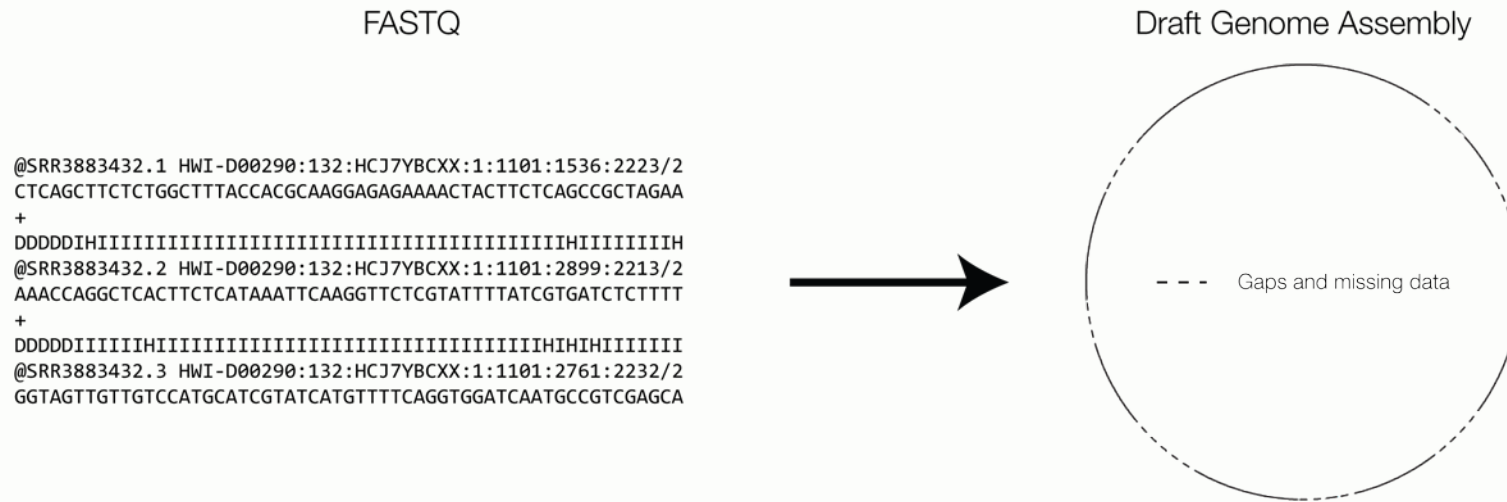
- Let's take a few steps back...



[https://commons.wikimedia.org/wiki/File:Plasmid_\(english\).svg](https://commons.wikimedia.org/wiki/File:Plasmid_(english).svg)

What is functional annotation

- Let's take a few steps back...



What is functional annotation

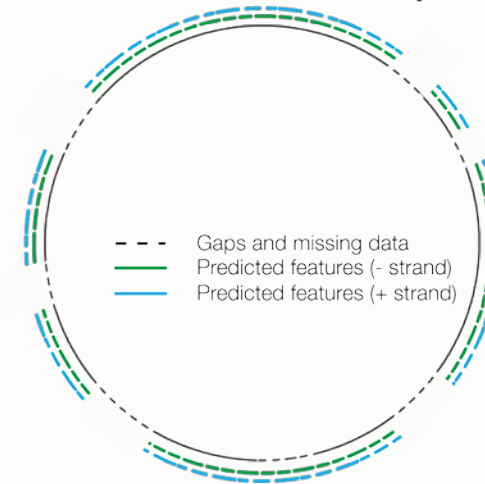
- Let's take a few steps back...

FASTQ

```
@SRR3883432.1 HWI-D00290:132:HCJ7YBCXX:1:1101:1536:2223/2
CTCAGCTTCTCTGGCTTTACCACGCAAGGAGAGAAAACACTTCTCAGCCGCTAGAA
+
DDDDDIHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIIIIIIIH
@SRR3883432.2 HWI-D00290:132:HCJ7YBCXX:1:1101:2899:2213/2
AAACCAGGCTCACTTCTCATAAATCAAGGTTCTCGTATTTTATCGTGATCTCTTTT
+
DDDDDIIIIIIIHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIHIIIIIIII
@SRR3883432.3 HWI-D00290:132:HCJ7YBCXX:1:1101:2761:2232/2
GGTAGTTGTTGTCCATGCATCGTATCATGTTTTTCAGGTGGATCAATGCCGCTCGAGCA
```



Draft Genome Assembly



What is functional annotation

Imagine that your genomes are really books describing the organism

```
>SRA1231 chromosome 1 whole genome shotgun sequence
ACTACTACATCTCCACTCAGCCAGGTGAAGTCTACCACATACACCCTTCTGTGGTCACCACCGACT
CTAACGGACAAACTACTACCAGTCCGATGTCGTATCGTACACTGATTCTGATGGATCGTTGACCAC
TACTACATCTCCACTCAGCCAGGTGAGGGTCCAACAACCTTATACACTTCTGTCTGACCCAGGATTTCT
AACGGACAAACTACTACCAGTCCGATGTCGTATTGTGCAACAGACTCTGATGGATCGTTGACTACGA
CTACCTCTCCTCTTGGTCCAAGTGGTCCAGCCGAGGGTCCAACGACTTACACTACTTCGGTGTCACTAC
CGACTCTAACGGTGAAGTACTCACTTCTGACGTTGTCTTGTGACTACTGATTCTGACGGATCGTTG
ACTACTACCCTTCCCCTCGGTCCACCTTCCAGGTGAGGGTCCAACCACTTACTTACTGATATCG
TGCAACCGATGACAGGGTCACTACCACCTCGTCTGGCGTCTGATTTGTTACCACCTGATTGACGGG
CTCCTTGACCACCAACAGTCTCCTTTGAGCCTTCTGGTCCAACCACTTACACAACCGGATTTGTTACT
ACTGACGATCAGGGTAAACAGTACCAGTCCGATGTCGTATTGTTACTACTGACTCTGACGGTTCAT
TgacaacaacaacctcTCCACTTGGTCCAGGCGGTCCAACCTACACAACCTTCCCTTGTGACCACTGA
TGATCAGGGTCACCAGACAACGAATCTGATGTTGTCATTGTGACTACTGACTCCGATGGTAACTTGATC
AccacaactctcctcttGTCAGGTGGTCCATCTGGCCCAACCACTTACACCACCTCATTGTCACTA
CTGACGACCAAGGCCACAAGACTACTGAGTCCGATGTCGTATCGTTACTACTGACTCTGACGGCACTT
GGTACTACCACCTCTCCTCTTGGTCTGGTATCACGCTGGAGACATCACAGCTTCACTTCAACCTGG
GAGACACCTTCTGACGGCAGCGTGGCTACCGATTCTGGTGTGGTATTGTGACTACTGATACCAAG
GCAACTTGATCACCCTACTTCCCCTTGGCCAGGTGAACAATGGCCCAACCTCCTACACCCTAC
TGTGTCTCAACTGACAAGAATGGCCAGGAGTACCAAGACGGTATTGTGTGAGACTACTGCTCCT
AACGGTCAGTTGACCTTTACACCCTGTGTCTCTGAGACCAACCTTTCGAGACTACCAACAAGGAAG
GTTCCAAAGCACTGTGAGTGTGTGTTGATCGAAACCACTCACGGCGTGGTCACTTACATCTC
GGTATGTCCACCAGCCAAGGAGACTCGTATGTCTCACTTACGAGACCAACAAGGCAAGCGGGAAGT
GAGACTACCCTGTGCGTTGTTGCTGTTGAACTGACGTCGAGGGCAACGTCAGGACTCTACTCTTGTG
CTGAATCGACAGTCTCTGAGGGTCTCAAGGCTCTCAGCAGAGACATCGACCCTGACGCCCCAAG
CGAAGCTCCTAATACCACACTGTTGCGGCCCTTCACTACCGTACGACCTACGAAGGCGCTGGCTCT
CTCCCAAGATCAGctttagctttttcttcCACTTGGACTTTTGTCTTAAAGCAATTTTCTA
GcctttttttcaacCCGGATTCTTATAAATAACCGCTTTGGCTGATAATTTTGGATACCCATT
CGCTTTTCTGTGCTTTATAAATAATGATTTTCACTTAAACCAAGTGTAGATGTAATAACCG
AAAacgttttttttttttttttttttttttttttttaagcCAAGATCATAAGACGGCTTGGTAAGCGAT
ATAGCTCTAAAAAATCCTTTGATTAGCGGAATTTAATCCTTAAAAATAGGTCAAACAATCTATACTATT
GGGAAAGACGTAATAAATCAGTgcttttttcaagagcAGAATGAACAGAGAATAAAAAACATGGCCAT
AGTCAAGCAATTGCAAGTCAATTAAGGAGCCAAAGGCATAATCGAAAAAGGAATTTAGACAAGGACAT
TGGTTTGTCAAGGAGAATGAACACGATAAATGGGGTCAAATCACGAAATAGTCTGTAAAGACAGTCAAC
TTTTGGTAGGGAGAAGTCATACAGTAAGGGCACTAAAAGCTGCAACTACTAAGTAAATGCACATGCAAGAAA
GTACCGCTCTCGTAAATCTTCTGCACAGGCCAACGAGAGTAAAGTTGACTTTGAAAAAGCCACAGGTGGCA
ATAAATCAAGGTGATGATGCTAAATATGTGATTTGTCTCAATGAATCGACTCTGAAGAATTGAG
```



This is the genome of *Klebsiella pneumoniae* strain XYZ, isolated from a patient in the Emory Healthcare. This genome contains many interesting antimicrobial resistance genes, as well other genes for invasion, host evasion, and adherence. Antimicrobial compounds such as beta-lactams, penicillins, and even carbapenems have no effect on me.

What is functional annotation

Some words, or **genomic features**, in the book have special meaning

```
>SRA1231 chromosome 1 whole genome shotgun sequence
ACTACTACATCTCCACTCAGCCAGGTGAAGTCTACCACATACACCCTTCTGTGGTCACCACCGACT
CTAACGGACAAACTACTACCAGTCCGATGTCGTATCGTACACTGATTCTGATGGATCGTTGACCAC
TACTACATCTCCACTCAGCCAGGTGAGGGTCCAACAACCTTATACACTTCTGTCTGACCCAGGATCT
AACGGACAAACTACTACCAGTCCGATGTCGTATTGTGCAACAGACTCTGATGGATCGTTGACTACGA
CTACCTCTCCTTGGTCCAAGTGGTCCAGCCGAGGGTCCAACGACTTACACTACTTCCGTTGCTACTC
CGACTTAACGGTGAGACTATCACTTCTGACGTTGTCACTGTGACTACTGATTCTGACGGATCGTTG
ACTACTACCCTTCCCACTCGGTCCACCTTCTCAGGTGAGGGTCCAACCACTTACTTACTGATATCG
TGACAACCGATGACAGGGTCACTACCACCTCGTCTGGCGTGTCACTTGTACCACGATCTGACGG
CTCCTTGACCACCAACAGTCTCCTTGGAGCTTCTGGTCCAACCACTTACACAACCGGATTTGTTACT
ACTGACGATCAGGGTAAACAGTACCAGTCCGATGTCGTATTGTTACTACTGACTCTGACGGTTCAT
TgacaacaacaacctcTCCACTTGGTCCAGGCGGTCCAACCTACACAACCTTCCCTTGTGACCACTGA
TGATCAGGGTCAACAGACAACCGAATCTGATGTTGTCACTGTGACTACTGACTCCGATGGTAACTGATC
AccacaactctcctcttGTCCAGGTGGTCCATCTGGCCCAACCACTTACACCACCTCATTGTCTACTA
CTGACGACCAAGGCCACAAGACTACTGAGTCCGATGTCGTATCGTTACTACTGACTCTGACGGCAACTT
GGTACTACCACCTCTCCTTGGTCTGGTATCAGCTGGAGACATCACCAGCTTCACTTCAACCTGG
GAGACACCTTCTGACGGCAGCGTGGCTACCGATTCTGGTGTGGTTATTGTGACTACTGATACCAAG
GCAACTTGATCACCCTACTTCCCACTTGGCCAGGTGAACAATGGCCCAACCTCCTACACCCTAC
TGTGTCTCAACTGACAAGAATGGCCAGGAGTCAAGACGGTTATTGTGTGAGACTACTGCTCTCT
AACGGTCAGTTGACTCTTACACCCTGTGTCTGAGACCAACCTTTCGAGACTACCAACAAGGAAG
GTTCCAAAGCAACTGTGAGTGTGTGTTGATCGAAACCACTTCCAGGGCGTGGTCACTTACATCTC
GGTATGTCCACCAGCCAAAGGAGACTCGTATGTGTCCACTTACGAGACCAACAAGGCGAAGCGGAACT
GAGACTACCCTGTGCGTTGTTGCTGTTGAACTGACGTGAGGGCAACGTCAAGACTCTACTTGTCTG
CTGAATCGACAGTCTCTGAGGGTCTCAAGGTCTCAGCAGAGACATCGACCCTGCCAGGCCCAAG
CGAAGCTCTAATAACCAACTGTTGCGGGCCCTTCAAGTACCGTACGACCTCAGGACCTCAAGGCGCTGGCTCT
CTCCCAAGATCAGctttagctttttcttcCACTTGGACTTTTGTCTAAGCAATTTTCTCA
GcctttatTTTTcaacCCGGATTCTTATAAATAACCGCTTTGGCTGATAATTTTGGATACCCATT
CGCTTTTTCGTGTGCTTTATAAATAATGATTTTCACTTCTTAAACCAAGTGTAGATGTAATAACCG
AAAacgttttttttttttttttttttttttttttttaagcCAAGATCATAAGACGGCTGGTAAGCGAT
ATAGCTCTAAAAATTCCTTGTATTAGCGGAATTTAATCCTTAAAAATAGGTCAAAACATCTATACTATT
GGGAAAGACGTAATAAATCAGTgcttttttcaagagcAGAATGAACAGAGAATAAAAAACATGGGCAT
AGTCAAGCAATTGCAAGTCAATTAAGGAGCCAAAGGCATAATCGAAAAAGGAATTAGACAAGGACAT
TGGTTTGTCAAGGAGAATGAACACGATAAATGGGGTCAAAACCCGAATAGTCTGTAAGACACGTCAAC
TTTTGGTAGGGAGAAGTCATACAGTAAGGGCACTAAAAGCTGCAACTACTAAGTAATGCACATGCAAGAAA
GTACCGCTCTCGTAAATCTTGCACAGGCCAACGAGAGTAAAGTTGACTTTGAAAAAGCCACAGGTGGCA
ATAAATCAAGGTGTAGATGATGCTAAATATGTGATTTGTCTCAATGAATCGACTTCTGAAGAAATTGAG
```



This is the genome of *Klebsiella pneumoniae* strain XYZ, isolated from a patient in the Emory Healthcare. This genome contains many interesting **antimicrobial resistance genes**, as well other genes for **invasion**, **host evasion**, and **adherence**. Antimicrobial compounds such as **beta-lactams**, **penicillins**, and even **carbapenems** have no effect on me.

What is functional annotation

This special meaning is because these words have functional consequences

```
>SRA1231 chromosome 1 whole genome shotgun sequence
ACTACTACATCTCCACTCAGCCAGGTGAAGGTCCTACCACATACACCCTTCTGTGGTCACCACCGACT
CTAACGGACAAACTACTACCAGCTCCGATGTCGTATCGTACGACTGATTCTGATGGATCGTTGACCAC
TACTACATCTCCACTCAGCCAGGTGAGGGTCCAACAACCTTATACACTTCTGTCTGACCCAGCGATTCT
AACGGACAAACACTACCAGCTCCGATGTCGTATTGTGCAACAGACTCTGATGGATCGTTGACTACGA
CTACCTCTCCTCTTGGTCCAAGTGGTCCAGCCGAGGGTCCAACGACTTACACTACTTCGGTTGCTACTAC
CGACTTAACGGTGAGACTATCACTTCACTGACGTTGTCACTGTGACTACTGATTCTGACGGATCGTTG
ACTACTACCCTTCCCCTCCGTCACCTTCTCAGGTGAGGGTCCAACCACTTACTTACTGATATCG
TGACAAACCGATGACAGGGTCACTACCACCTCGTCTGGCGTGTCACTTGTACCACCTGATCTGACGG
CTCCTTGACCAACCAACAGCTCTCCTTTGAGCCTTCTGGTCCAACCACTTACACAACCGGATTGTTACT
ACTGACGATCAGGGTAAACAGTACCAGCTCCGATGTCGTATTGTTACTACTGACTCTGACGGTTCAT
TgacaacaacaactcTCCACTTGGTCCAGGCGGTCCAACCTACACAACCTTCCCTTGTCCACTGA
TGATCAGGGTCACCAGACAACCGAATCTGATGTTGTCACTGACTACTGACTCCGATGGTAACTGATC
AccacaactctcctcttGTCCAGGTGGTCCATCTGGCCCAACCACTTACACCACCTCATTGTCTACTA
CTGACGACCAAGGCCACAAGACTACTGAGTCCGATGTCGTATCGTTACTACTGACTCTGACGGCAACTT
GGTACTACCACCTCTCCTTGGTCTGGTATCAGCTGGAGACATCACCAGCTTCACTTCAACCTGG
GAGACACCCCTTCTGACGGCAGCGTGGCTACCATTCTGGTGTGGTTATTGTGACTACTGATACCAAG
GCAACTTGATCACCCTACTTCCCCTTGGCCAGGTGAACAATGGCCCAACCTCCTACACCCTAC
TGTGTGCTCAACTGACAAGAATGGCCAGGAGTACCAAGACGGTATTGTGTGAGACTACTGCTCCT
AACGGTCAGTTGACTCTTACACCCTGTGTGCTGAGACCAACCTTTCGAGACTACCAACAAGGAAG
GTTCCAAAGCAACTGTGAGTGTGTGGTCACTGAAACCACTCACCAGGCGTGGTCACTTACATCTC
GGTATGTCACCAGCCCAAGGAGACTCGTATGTGTCACTTACGAGACCAACAAGGCGAAGCGGAACT
GAGACTACCCTGTGCGTTGTTGCTGTTGAACTGACGTCGAGGGCAACGTCAGACTCTACTCTTGTG
CTGAATCGACAGCTCCTTCTGAGGGTCTCAAGGTCCTCAGCAGAGACATCGACCCTGCCAGCCCAAG
CGAAGCTCTAATAACCAACTGTTGCGGCCCTTCAAGTACCCTCAGCACCTACGAAGGCGCTGGCTCT
CTCCCAAGATCAGctttagctttttcttcCACTTGGACTTTTGTCTAAGCAATTTTCTCA
GcctttttttcaacCCGGATTCTTATAAATAACCGCTTTGGCTGATAATTTTGGATACCCATT
CGCTTTTTCGTGTGCTTTATAAATAATGATTTTCAAGTCTTAAACCAAGTGTAGATTGAATAACCG
AAAacgttttttttttttttttttttttttttttttaagCAAGATCATAAGACGGCTGGTAAGCGAT
ATAGCTTAAAAAATCCTTTGATTAGCGGAATTTAATCCTTAAAAATAGGTCAAACAATCTATACTATT
GGGAAAGACGTAATAAATCAGTgcttttttcaagagcAGAATGAACAGAGAATAAAAAACATGGCAT
AGTCAAGCAATTGCAAGTCAATTAAGGAGCCAAAGGCATAATCGAAAAAGGAATTAGACAAGGACAT
TGGTTTGTCAAGGAGAATGAACACGATAAATGGGGTCAAATCACCGAATAGTCTGTAAGACACGTC AAC
TTTTGGTAGGGAGAAGTCATACAGTAAGGGCACTAAAAGCTGCAACTACTAAGTAATGCACATGCAAGAAA
GTACCGCTCTCGTAAATCTTCTGCACAGGCCAACGAGAGTAAAGTTGACTTTGAAAAAGCCACAGGTGGCA
ATAAATCAAGGTGTAGATGATGCTAAATATGTGATTTGTCTCAATGAATCGACTTCTGAAGAAATTGAG
```



This is the genome of *Klebsiella pneumoniae* strain XYZ, isolated from a patient in the Emory Healthcare. This genome contains many interesting **antimicrobial resistance genes**, as well other genes for **invasion**, **host evasion**, and **adherence**. Antimicrobial compounds such as **beta-lactams**, **penicillins**, and even **carbapenems** have no effect on me.

What is functional annotation

- Taking just one more step back...

What is function?

What is functional annotation

- To a *microbiologist* function might refer to **phenotypes** that they study and observe in the lab, like antimicrobial resistance
- To a *cell biologist*, function might refer to the **network of interactions** in which the protein participates or to its **location** in a certain cellular compartment
- To a *biochemist*, function refers to the **metabolic process** in which a protein is involved or to the **reaction catalyzed** by an enzyme.

What is functional annotation

Functional annotation is defined as the process of collecting information about and describing a [genome feature's] biological identity—its various aliases, molecular function, **biological role(s)**, subcellular location, and its expression domains within the [organism]

What is functional annotation

Functional annotation is for:

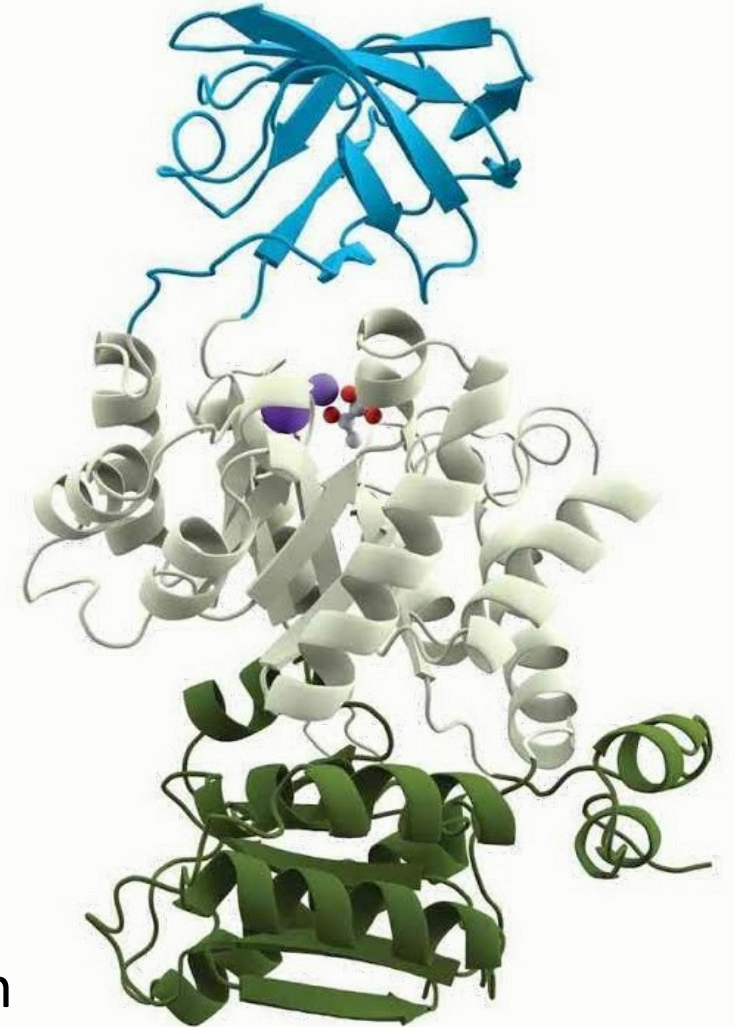
- Proteins
 - Domains/Motifs
 - Signal peptides
 - Transmembrane regions
- Non-coding RNA
 - Ribosomes and tRNA
 - Riboswitches
 - CRISPR
 - sRNA
- Operons
- Other?

What is functional annotation

- Newly synthesized proteins, or pro-peptides, contain amino acids that are not part of the mature, functional protein
- These residues are often molecular signals, or tags, that direct the nascent peptide to its final cellular/extracellular home

What is functional annotation

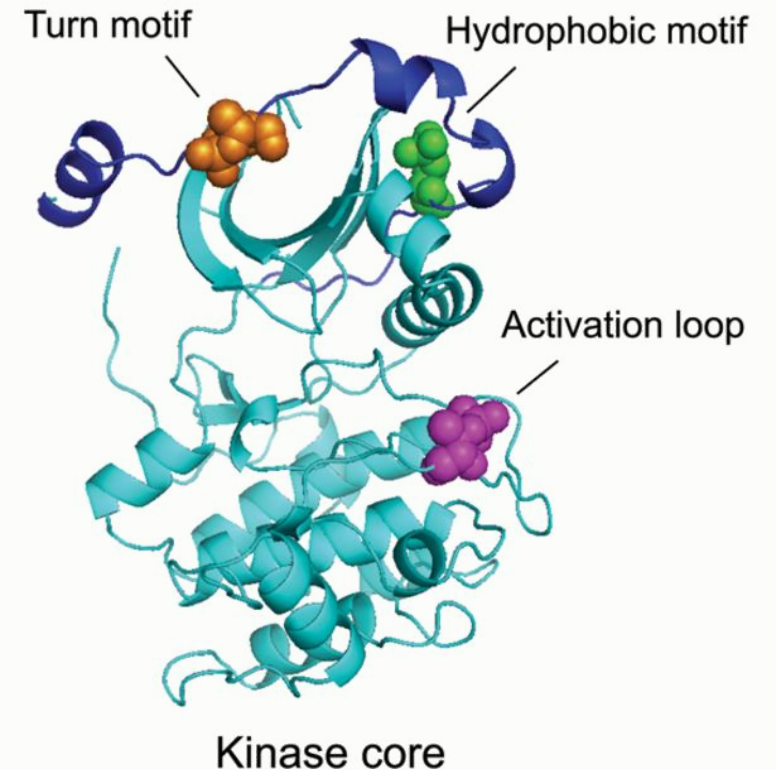
- Mature proteins contain functional regions and nonfunctional regions
- A **domain** is:
 - a discrete structural unit
 - assumed to fold independently of the rest of the protein
 - have its own function
 - be relatively short (20-100 AA)
 - be composed of one or more smaller subdomains with linkers



Pyruvate kinase, a protein with three domains

What is functional annotation

- Mature proteins contain functional regions and nonfunctional regions
- A **motif** is:
 - a short (5-20 AA) conserved region
 - typically the most conserved region in a *domain*
 - critical to the function of a domain



Key functional motifs in the Kinase core domain

What is functional annotation

- **Operons** are a contiguous set of genes of related and/or complimentary function that are regulated together
- Operons are generally transcribed into a single mRNA coding for several genes
- Such mRNA are **polycistronic** (cistron = gene) and only exist in prokaryotes

Outline

- **Functional Annotation**
 - What is functional annotation?
 - **Annotation approaches**
 - Tools and techniques

- **Comparative Genomics**
 - What is comparative genomics?
 - Tools and techniques

Annotation approaches

Ab initio

Takes advantage of **intrinsic characteristics** that a subset of genes/proteins all contain

- Signal peptides (SignalP, LipoP, Phobius...)
- Transmembrane domains (TMHMM, Tmpred...)

Homology based

Transfers information from experimentally derived **prior knowledge** to new data based on similarity

- BLAST
- InterPro

Ab initio approaches

- **Transmembrane (TM)** and **Signaling peptides (SP)** have a distinct pattern of sequence composition
- TM proteins are **membrane bound** receptors and channels that are of particular pharmacological relevance (therapeutic or vaccine target)
- Signal peptides **direct proteins** to their proper subcellular or extracellular location

Homology based approaches

- **Assumption:** *Significant* sequence similarity implies homology or shared ancestry that often leads to shared function
- This assumption is rooted in the principle of **molecular evolution**

Homology based approaches

- **Assumption:** Significant sequence similarity implies homology or shared ancestry that often leads to shared function
- This assumption is rooted in the principle of **molecular evolution**

“Nothing in Biology Makes Sense Except in the Light of Evolution”

~ Theodosius Dobzhansky

Molecular Evolution

- Molecular evolution is a conservative process
- Genes/proteins that evolve to perform some function will retain that function
- Genes/proteins that evolve from the same ancestor are called homologous genes/proteins

Molecular Evolution

- The information of common ancestry is gained from sequence similarity
- Sequences that share *significant* level of similarity are inferred to be homologous
- Knowing this helps in characterizing the function/structure of unknown sequences

Molecular Evolution

Underlying assumptions:

- Evolution is mostly dominated by divergence
- Deleterious mutations are weeded out by purifying selection
- Massive homoplasy is highly unlikely

- Homology will thus entail a high chance of shared origin and shared function

Homology

There are multiple ways to generate homologous sequences through evolution

Two broad types:

1. Orthology:

2. Paralogy:

Homology

There are multiple ways to generate homologous sequences through evolution

Two broad types:

1. Orthology: Shared ancestry due to speciation event
2. Paralogy: Shared ancestry due to gene duplication event

Human and mouse HBA genes are *orthologs*

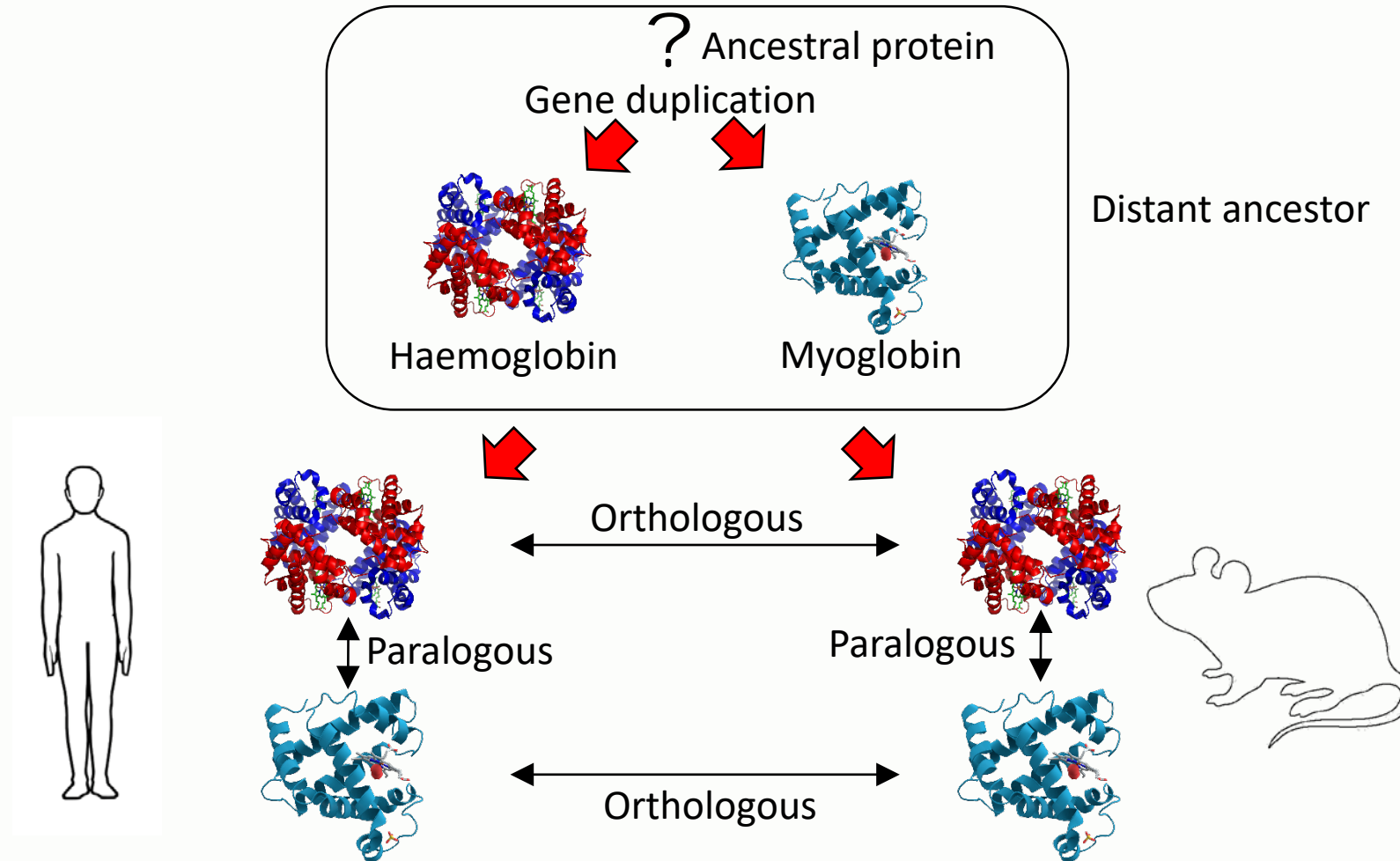
- The *Homo sapiens* and *Mus* HBA genes are orthologous
 - They share the same function in both species
 - They are the result of a *species* divergence
 - They are each others best match across species

- Same for Hs HBB and Mm HBB

Human HBA and Myoglobin genes are *paralogs*

- There are two homologs in the **same** genome
 - They have different (but very similar) functions
 - They may not be each others best match
- They are the result of a gene duplication early on in hemichordate evolution

Orthology and Paralogy



Outline

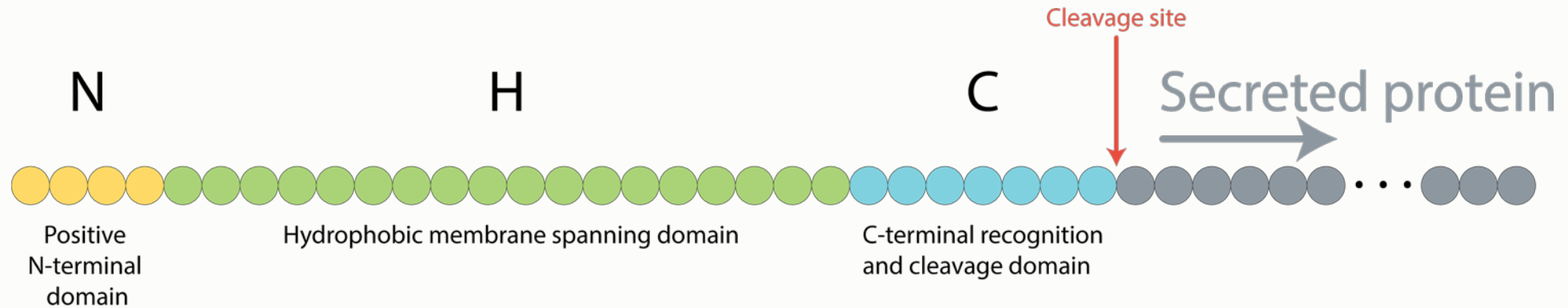
- **Functional Annotation**
 - What is functional annotation?
 - Annotation approaches
 - **Tools and techniques**

- **Comparative Genomics**
 - What is comparative genomics?
 - **Tools and techniques**

Tools and Techniques: *Ab initio*

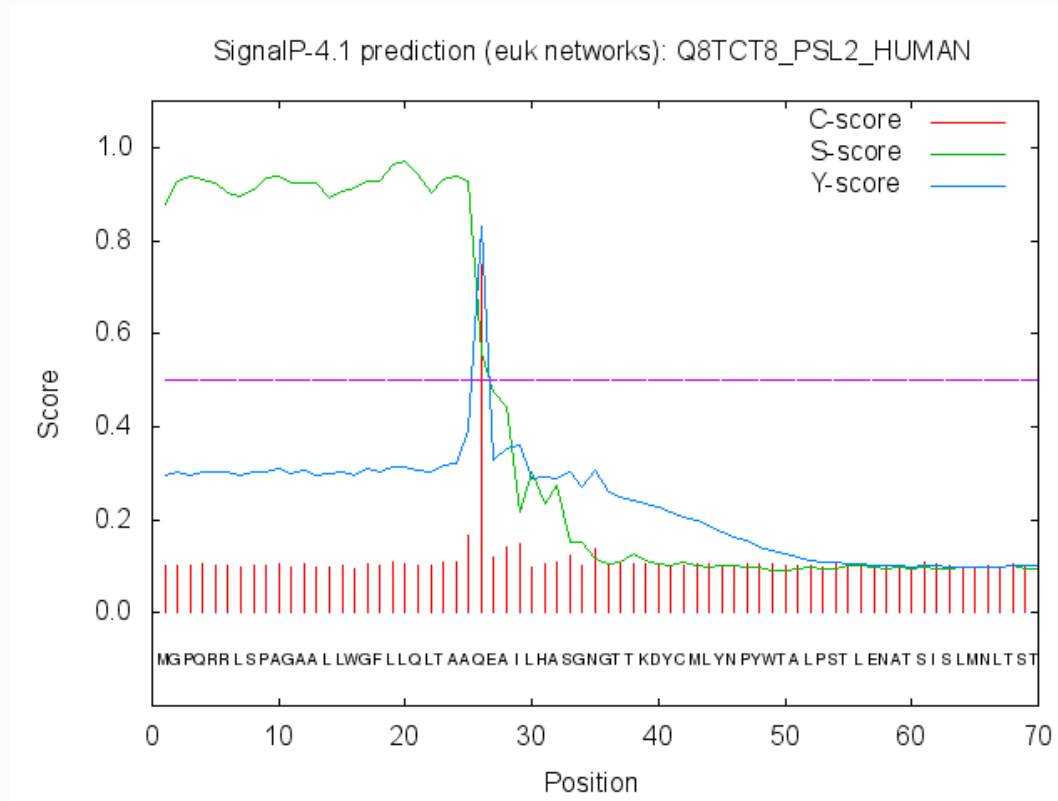
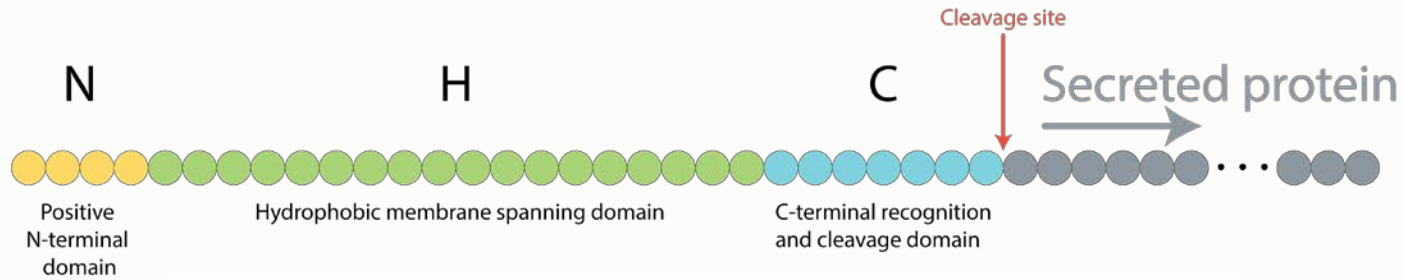
- *Ab initio* tools rely on highly conserved patterns found in functional motifs and domains
- Signal peptides are a classical example, with a conserved N-H-C structure on the N-terminus of pro-peptides

Tools and Techniques: *Ab initio*



- The conserved structural elements can be used to train classifiers
- Tools like SignalP and Phobius use neural networks and markov models, respectively, to find and classify the N-H-C regions in pro-peptides sequences
- Nearly all proteins with a signal peptide are destined to the membrane

Tools and Techniques: *Ab initio*

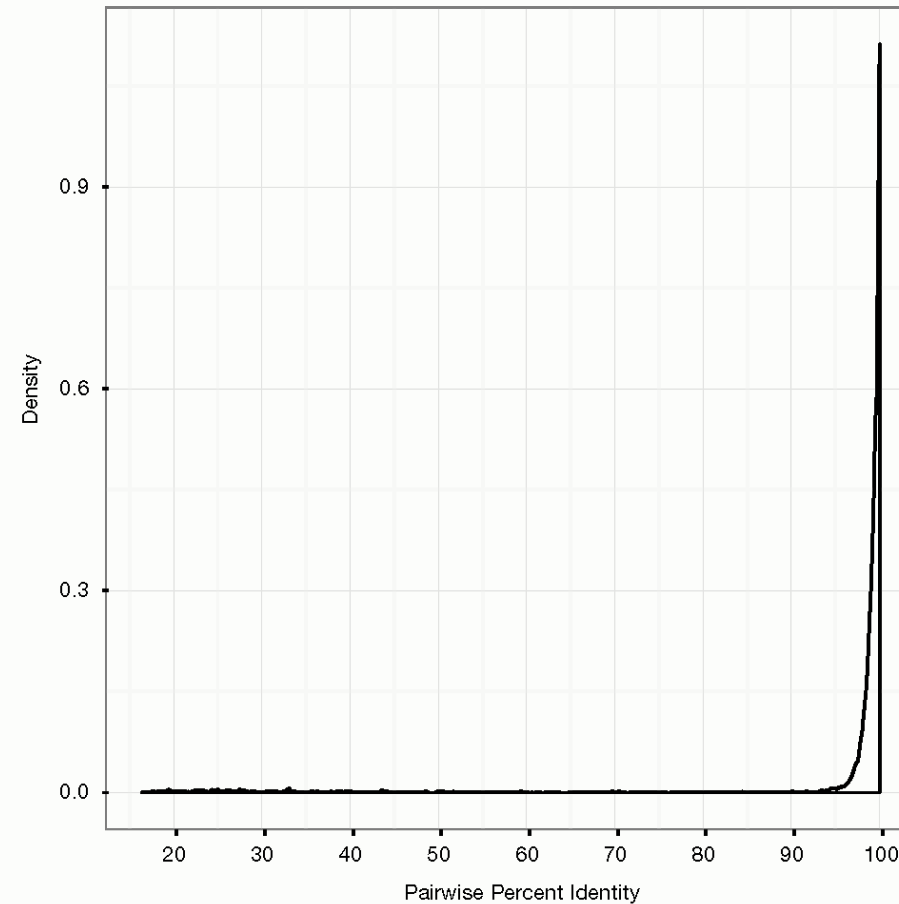


<http://www.cbs.dtu.dk/services/SignalP/>

Tools and Techniques: Homology based

- Homology based tools boil down to sequence comparisons by alignment
- There are two paradigms for sequence alignment
 - Global – end to end alignments of two sequences
 - Local – Motif and domain alignment
- BLAST (or BLAST-like aligners) and the InterPro suite are the two most commonly used

Tools and Techniques: Homology based



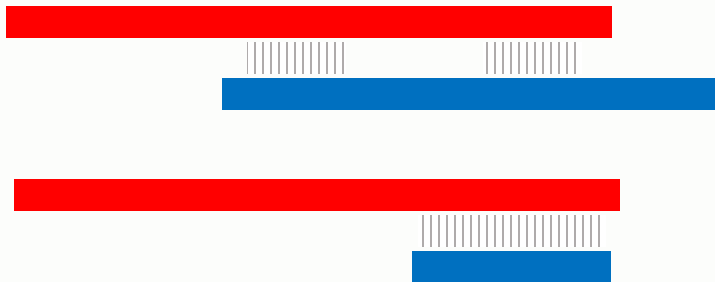
Supplementary Figure 3. **Distribution of sequence identity between same annotations.** Distribution of pairwise sequence identity between two genes with matching annotations between genomes. Approximate 276,000 comparisons are shown; >99.99% of density is represented in the 95-100% identity region.

Tools and Techniques: Homology based



Global alignment

- End to end alignment
- Sequences are of approximately the same length
- More applicable for aligning closely related sequences



Local alignment

- Local level of similarity such as in distantly related sequences
- Sequences don't have to be of the same length
- Ideal for finding motifs and domains

Tools and Techniques: BLAST

- Basic Local Alignment & Search Tool
- Comparing a query nucleotide or protein sequence against a database of nucleotide or protein sequences
- Hosted at NCBI and is a defining tool of bioinformatics

Tools and Techniques: BLAST

BLAST® >> blastp suite Standard Protein BLAST

[blastn](#) [blastp](#) [blastx](#) [tblastn](#) [tblastx](#) BLASTP programs search protein databases using a p

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) [Query subrange](#)

From
To

Or, upload file No file chosen

Job Title
Enter a descriptive title for your BLAST search

[Align two or more sequences](#)

Choose Search Set

Database

Organism [Exclude](#)
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude [Models \(XM/XP\)](#) [Uncultured/environmental sample sequences](#)

Entrez Query [YouTube](#) [Create custom database](#)

Program Selection

Algorithm

- Quick BLASTP (Accelerated protein-protein BLAST)
- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

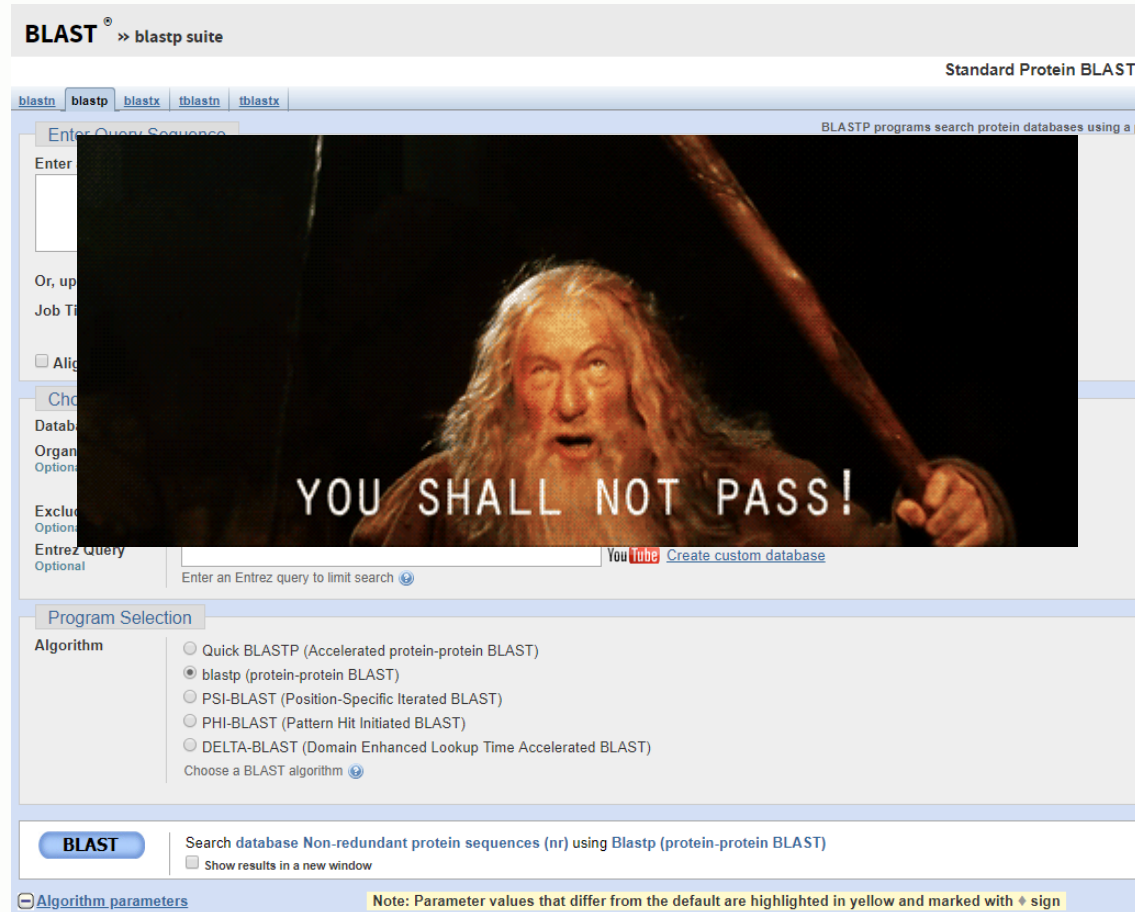
Choose a BLAST algorithm

Search database **Non-redundant protein sequences (nr)** using **Blastp (protein-protein BLAST)**

Show results in a new window

[Algorithm parameters](#) Note: Parameter values that differ from the default are highlighted in yellow and marked with + sign

Tools and Techniques: BLAST



The image shows a screenshot of the BLAST web interface. A large meme featuring Gandalf the White from 'The Lord of the Rings' is overlaid on the page, with the text "YOU SHALL NOT PASS!" in white capital letters across his chest. The background interface includes the BLAST logo, navigation tabs for different BLAST programs (blastn, blastp, blastx, tblastn, tblastx), and a form for entering a query sequence. Below the form, there are options for database selection and program selection. The program selection section shows radio buttons for various BLAST algorithms, with 'blastp (protein-protein BLAST)' selected. At the bottom, there is a 'BLAST' button and a checkbox for 'Show results in a new window'.

BLAST® » blastp suite

Standard Protein BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence

Enter

Or, up

Job T

Ali

Chc

Datab

Organ

Option

Exclud

Option

Entrez Query

Optional

Enter an Entrez query to limit search

YouTube Create custom database

Program Selection

Algorithm

- Quick BLASTP (Accelerated protein-protein BLAST)
- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

Show results in a new window

[Algorithm parameters](#) Note: Parameter values that differ from the default are highlighted in yellow and marked with + sign

Tools and Techniques: BLAST

- Similar sequences can be searched in a database by the use of sequence alignment methods
- A straight-forward approach → perform pairwise sequence alignment against each database sequence
- Not feasible as the size of the sequence and/or database increases

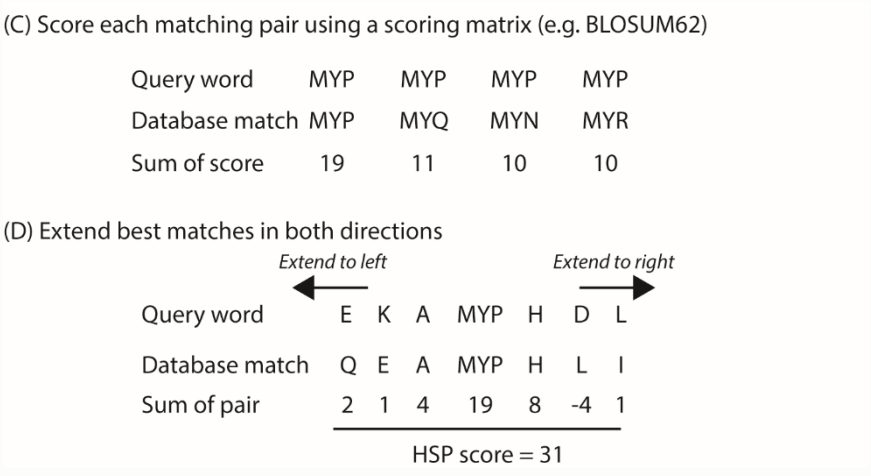
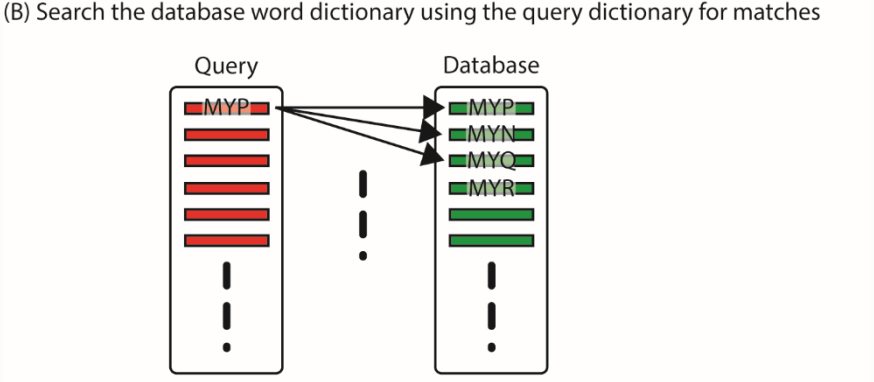
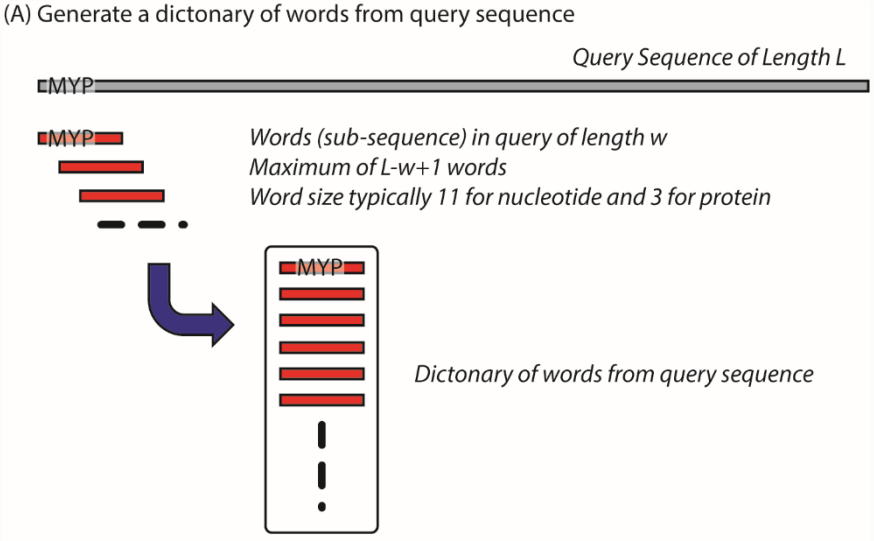
Tools and Techniques: BLAST

- To resolve the exhaustive methods, heuristics have been proposed
- The earliest of these algorithms was FastA proposed by Wilbur and Lipman in 1983
 - This algorithm is obsolete now
- Faster algorithms such as BLAST/BLAST+ are the *de facto* standard for searching sequence databases

BLAST Algorithm

1. Break the input query sequences into smaller words
2. Query each word to search for matches against every database word
3. Score matching words and pick the best pairs
4. Extend both words in both directions to find high scoring pairs (HSPs) with score above a threshold

BLAST Algorithm



BLAST algorithm overview

BLAST Algorithm

- Matches can be exact or inexact
- Extensions can be gapped or ungapped
- The threshold can be strict or flexible

Traditional BLAST programs

BLAST Flavor	Query Type	Database Type	Application
BLASTn	Nucleotide	Nucleotide	Finding similar nucleotide sequence
BLASTp	Protein	Protein	Finding similar protein sequence
BLASTx	Translated nucleotide	Protein	Finding protein functions from a gene sequence
tBLASTn	Protein	Translated nucleotide	Finding the gene sequence of a protein
tBLASTx	Translated nucleotide	Translated nucleotide	Comparing cross species genes

Alternative BLAST programs

BLAST search	Program	Description	Application
blastn	megablast*	Searches for nearly identical sequences	For sequence identification, intra-species comparisons
	discontinuous megablast	Performs cross-species comparison	For cross-species comparison, searching with coding sequences
blastp	PSI-BLAST	Position-Specific Iterative BLAST; Generates a PSSM that it utilizes for searching distantly related sequences	Finding distantly related sequences
	DELTA-BLAST	Domain enhanced lookup time accelerated BLAST; Searches against a database of PSSM, <i>i.e.</i> , profiles in biological sense	Finding domains in a given query sequence
	PHI-BLAST	Pattern Hit Initiated BLAST; Searches for protein sequences using a protein query and a signature (can be represented as a regex)	Narrowing database search matching a given pattern

Statistical Significance of Alignment

- Given an alignment, what is the probability that it is random?
- To answer this, the alignment needs to be quantified and compared to a number of random alignment scores
- This is the idea of the classical testing model for significance

Statistical Significance of Alignment

- The alignment can be scored based on the scoring system used
- This score can then be normalized using a variety of techniques. E.g. a longer “bad” alignment can get a larger score than a short “good” alignment if scores are not normalized
- Random scores are then calculated using random alignments

Statistical Significance of Alignment

- Random alignments can be generated by:
 - Shuffling the input sequence randomly
 - Generating random sequences of comparable length
- A distribution can be plotted based on these random alignment scores
- Empirical analysis show these plots follow Gumble Extreme Value Distribution (EVD)
 - Valid for both gapped and ungapped alignments

Statistical Significance of Alignment

- P -value can thus be calculated for any alignment
- P -value less than 10^{-5} is considered statistically significant (homologs)
- P -value between 10^{-1} and 10^{-5} may indicate distant homologs

Alignment Score

- Alignment score (S) implies the quality of the alignment
- Higher the score, better the alignment
- It is calculated based on the scoring system used, i.e., substitution matrix (BLOSUM, PAM, etc.) and gap penalty
- The score can change if either of the two parameters above is changed

Bit score

- Bit score (S') is the normalized log scaled version of score expressed in bits
- Provides an estimate of “the magnitude of the search space you would have to look through before you would expect to find a score as good as or better than this one by chance”
- Mathematically,

$$S' = \frac{\lambda S - \ln(K)}{\ln(2)}$$

Bit score

- λ & K depend on the substitution matrix and gap penalties
- If the bit-score (S') is 30, you would have to score, on average, about $2^{30} = 1$ billion independent segment pairs to find this score by chance. Each additional bit doubles the size of the search space.

Bit score

- The size of the search space is calculated as:

$$N = n \times m \times K$$

n is the number of bases in the query

m is the number of bases in the database

K is a coefficient

Bit score

- Consider that a sequence of length $n = 130$ aa is searched against a database of size $m = 123,456,789$ aa
- The K for a protein is about 0.13. In this case, the size of search space will be $= 130 \times 123,456,789 \times 0.13 = 2.1$ billion
- In this case, a bit score (S') of 30 (corresponding to a space of 1 billion) may have occurred by chance alone

P -value

- P -value in this context can be defined as the probability of obtaining a score (thus an alignment) greater than S at random

$$P = Ke^{-\lambda S}$$

- Based on Gumbel extreme value distribution (EVD)

E-value

- Expect value or E-value is the correction of P -value for multiple testing
- Defined as the number of distinct alignments, with a score equivalent to or better than S , that are expected to occur in a database search by chance.

$$E = m \times n \times P = mnKe^{-\lambda S}$$

- The lower the E-value, the more significant the score is

E-value

- E-value cutoff of 0.01 is considered to yield significant results and homology
- Higher E-values are not considered statistically significant

E-value and Bit Scores

- E-value is database size dependent
- Increasing database size will proportionally increase E-value of the same match
- Bit scores are independent of the database size and depends only on the alignment
- Bit scores are constant indicators of the database match

Result Interpretation: Table Interpretation

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected: 0

[Alignments](#) [Download](#) [GenBank](#) [Graphics](#) [Distance tree of results](#)

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	Neisseria meningitidis G2136, complete genome	520	1249	100%	2e-147	100%	CP002419.1
<input type="checkbox"/>	Neisseria meningitidis serogroup C FAM18 complete genome	520	885	100%	2e-147	100%	AM421808.1
<input type="checkbox"/>	Nmeningitidis class II pilin (pilS) silent truncated genes, two copies	520	720	100%	2e-147	100%	M19305.1
<input type="checkbox"/>	Neisseria meningitidis alpha710, complete genome	405	2572	99%	1e-112	88%	CP001561.1
<input type="checkbox"/>	Neisseria meningitidis PilE (pilE) and PilS1 cassette (pilS1) genes, complete cds; transposase	385	1767	99%	1e-106	86%	JN681264.1
<input type="checkbox"/>	Neisseria meningitidis PilE (pilE), PilS1 cassette (pilS1), PilS2 cassette (pilS2), and PilS3 cas	385	1195	99%	1e-106	86%	JN681263.1

Max score = highest alignment score (bit score) between the query sequence and the database sequence segment.

Total score = sum of alignment scores of all segments from the same database sequence that match the query sequence (calculated over all segments). This score is different from the max score if several parts of the database sequence match different parts of the query sequence.

Query coverage = percent of the query length that is included in the aligned segments. This coverage is calculated over all segments (cf. total score).

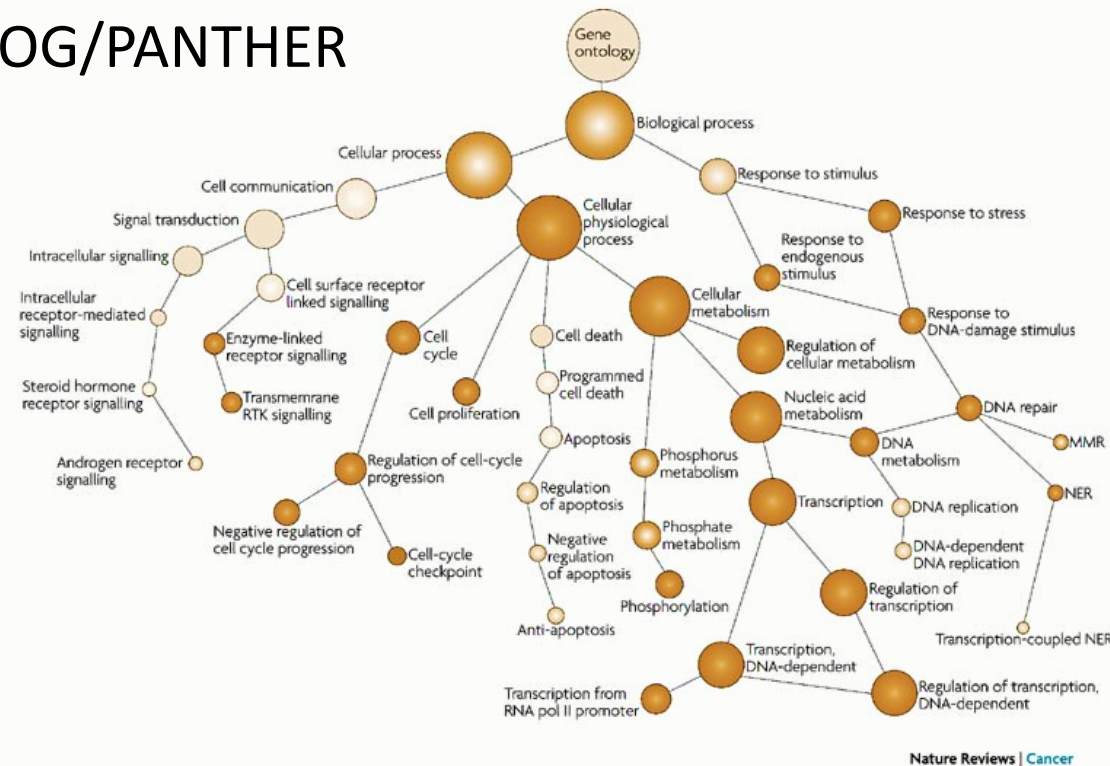
E value = number of alignments expected by chance with a particular score or better. The expect value is the default sorting metric and normally gives the same sorting order as Max score.

Tools and Techniques: HH-Suite

- When you want sensitive BLAST searches and high level structural predictions but don't have the time or resources, HH-suite is your friend
- HHsearch and HHblits allow rapid interrogation of homologous sequences and domains

Tools and Techniques: Ontology

- Annotations can be organized into coherent categories and compared across species, families and even kingdoms using Gene Ontology (GO)
 - Interpro/EggNOG/PANTHER
 - KOBAS



Gene ontology network in cancer cell lines

Final considerations: Selecting a method

1. The method or your implementation **must scale** from 10's to 100's of genomes
2. The software should be **maintained** – avoid old tools without developer support
3. Applicable to **prokaryotic** sequences
4. The method must produce (functionally) **consistent** and **reproducible** results

Final considerations: Gene naming

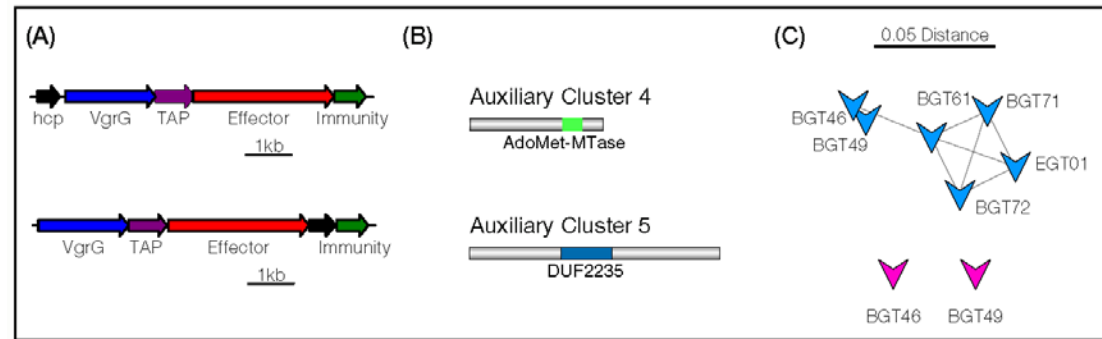
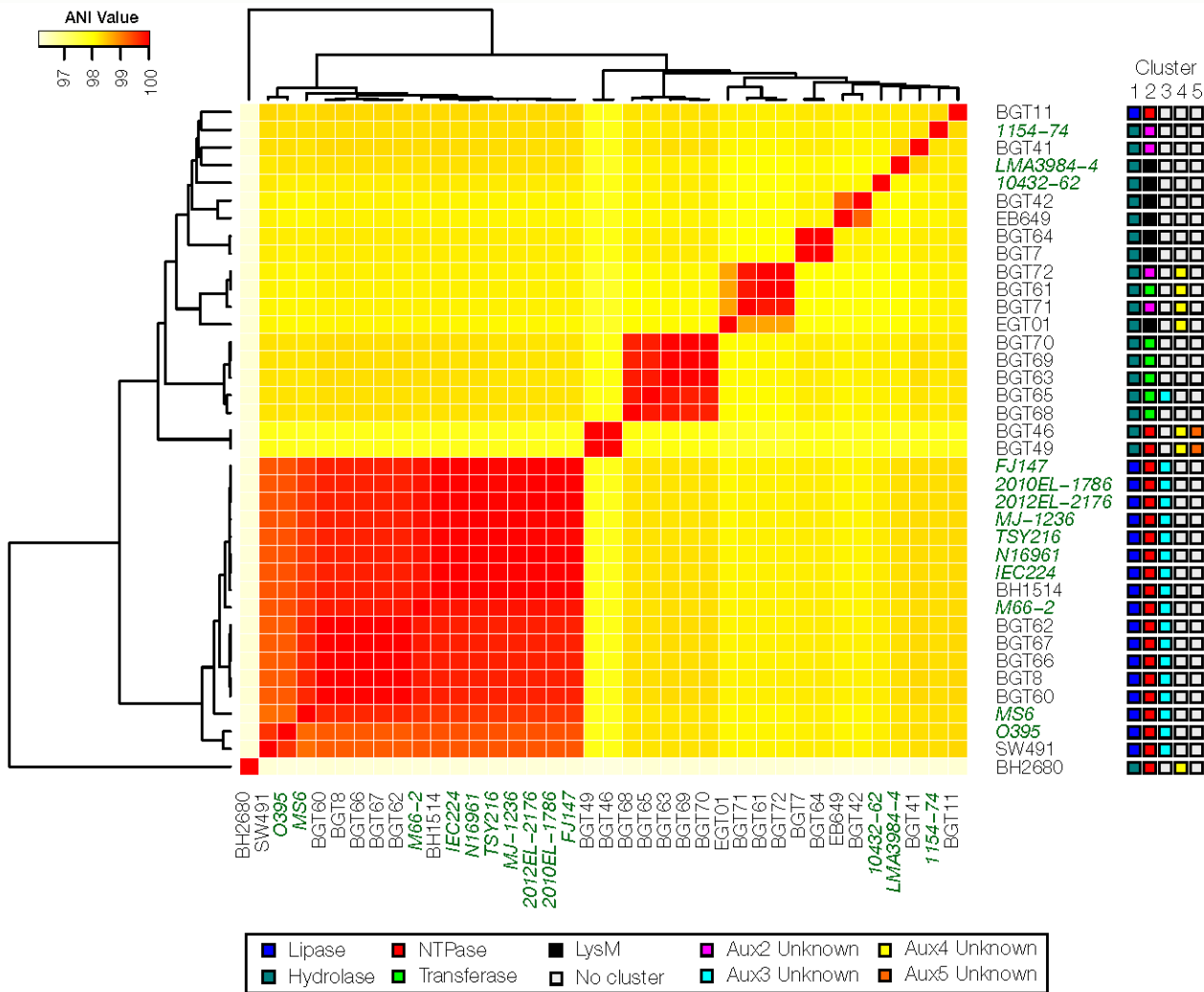
- Genes should be named using **clear logic** with strong evidentiary support
- Your naming scheme should be **consistent**
- A generally accepted scheme is as follows:
 - **High confidence matches** – Directly transfer name and functional annotation
 - **Multiple high confidence matches** – assign to the shared broad protein family
 - **Low confidence matches** – Assign a *putative* function
 - **Match to hypothetical protein** – *conserved hypothetical protein* annotation
 - **No matches** – are you sure it's a gene? If so, *hypothetical protein*

Comparative genomics

Outline

- Functional Annotation
 - What is functional annotation?
 - Annotation approaches
 - Tools and techniques
- Comparative Genomics
 - What is comparative genomics?
 - Tools and techniques

What is comparative genomics



Chande et al. (2018). Computational characterization of Type VI Secretion Systems in *Vibrio cholerae*. In preparation

What is comparative genomics

- Comparative genomics is what it sounds like:

Comparing the **similarities** and **differences** between two or more genomes of the same or different species

What is comparative genomics

Comparisons can happen at many levels:

- Whole genome
 - Structural variation (repeats, inversions, rearrangements)
 - Protein content (presence/absence, core proteins)
- Locus-specific
 - Sequence variation (SNPs and indels in functional regions)
 - Transcriptional / regulatory differences (RNAseq-guided)

What is comparative genomics

We can answer questions...

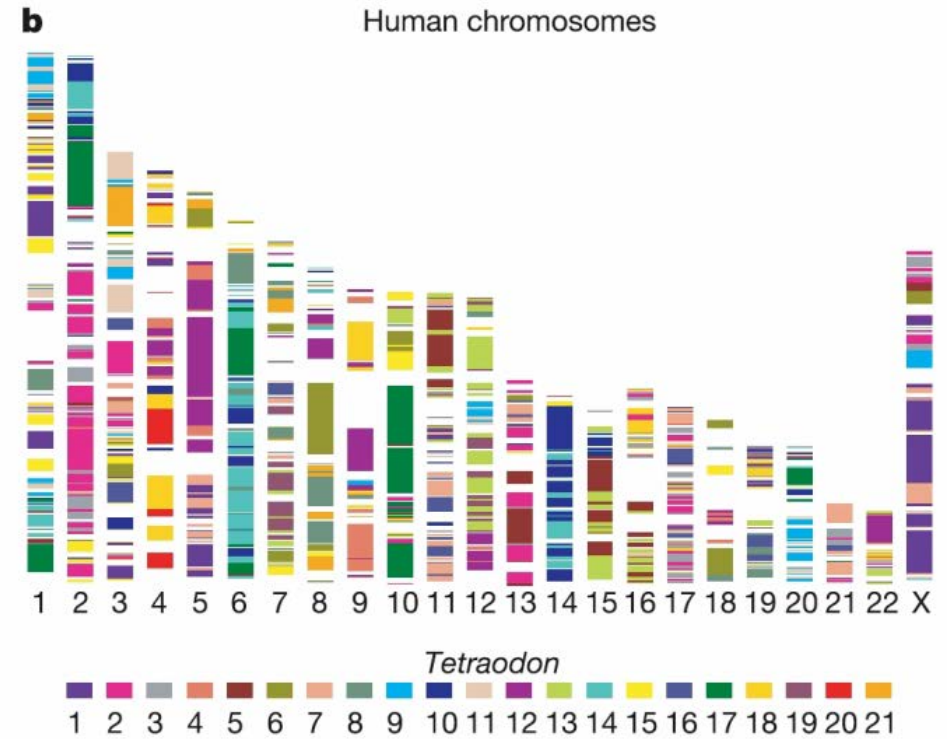
- What's the lifestyle of an organism (virulent or not)?
- Is there horizontal gene transfer?
- How is this organism evolving?

What is comparative genomics

- Biological questions to keep in mind
 - Are there rearrangement?
 - Is the region(s) of interest syntenic across species?
 - Are there genomic features present in some isolates but not others?
 - Are there structural elements that might impact gene expression?
 - Which organisms are more similar? Which are more distant?

What is synteny

- Synteny is shared genetic organization (gene-chromosome localization) between species
- Syntenic loci suggest common evolutionary history
- Functionally related operons are often highly syntenic



Synteny between human and tetraodon chromosomes. Colors blocks represent regions from a particular tetraodon chromosome.

<https://www.nature.com/scitable/topicpage/synteny-inferring-ancestral-genomes-44022>

Outline

- Functional Annotation
 - What is functional annotation?
 - Annotation approaches
 - Tools and techniques

- **Comparative Genomics**
 - What is comparative genomics?
 - **Tools and techniques**

Tools and techniques

- Homology based
 - BLAST/Diamond/MMseqs, Pathway analysis, GO analysis
 - Pan- and core-genome analysis, Protein Clusters
 - Virulence (VFDB, CARD)
- Phylogenetics
 - Horizontal Gene Transfer – Dark Horse, HGTector
 - Gene and genome based trees
- Visualization

Tools and techniques: Pan- and core genome

- Whole genome analyses, like ANI and dot plots, broadly describes the similarity between genomes
- These measures are content agnostic – that is, they treat all segments of the genome the same
- Pan-genome analysis is a finer-grained approach to whole genome comparison that uses gene content

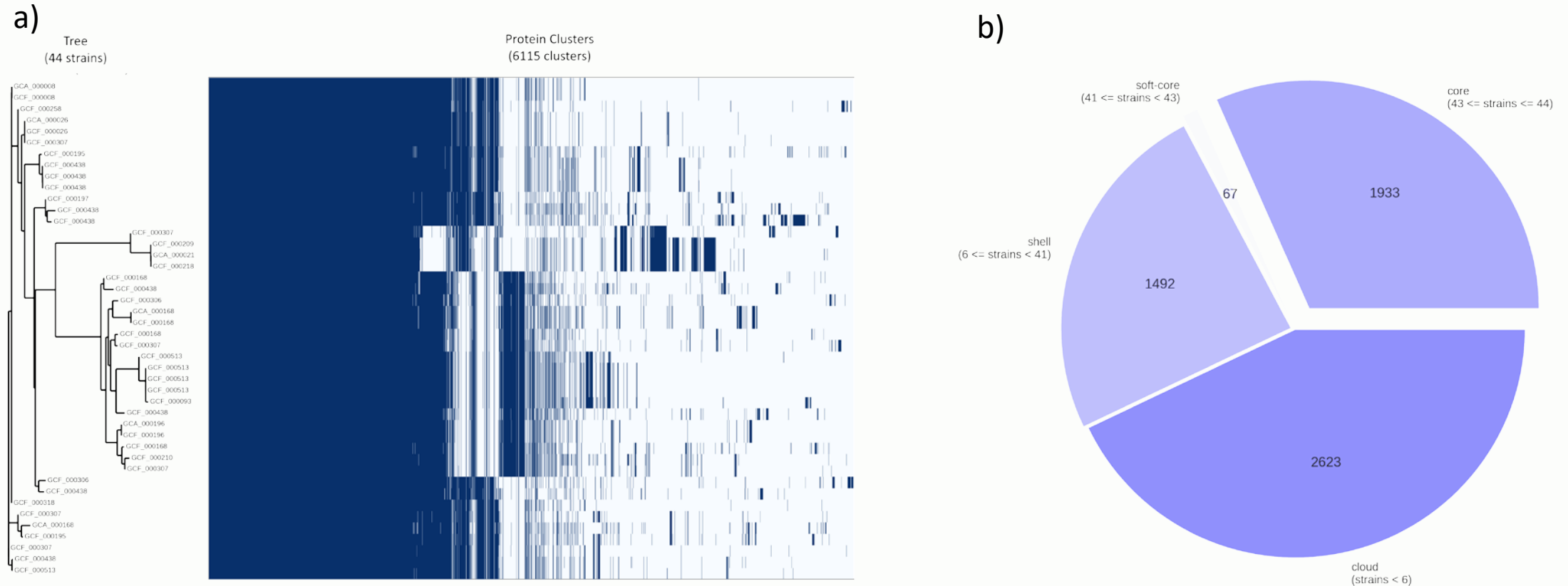
Tools and techniques: Pan- and core genome

- Pan-genome analysis is typically used to compare genomes from the same species but can be applied to genera
- The pan-genome is all the genes found, at least once, in a given sample set
 - For example, all the genes found in the *Listeria monocytogenes* isolates collected by CDC
- The core-genome is all the genes found in every member of the sample set
 - For example, the genes shared between every *L. monocytogenes* isolate collected by CDC
- The accessory genome is the pan-genome minus the core-genome

Tools and techniques: Pan- and core genome

- To make the dataset easier to handle, the pan-genome definition is usually tightened to genes found in at least 5% of the genomes analyzed
 - Excludes genes that may be a result of contamination or are less informative
- Similarly, core-genome is typically relaxed to include all genes found in 98-99% of genomes
 - Includes genes that are present in all but a small number of samples

Tools and techniques: Pan- and core genome



Pan-genome analysis of 44 *Listeria* strains. Presence (blue) and absence (white) of gene clusters from 44 related strains (a). Each column represents one or more functionally related genes present at least one genome. Pie chart depicting the number of gene clusters bellowing to cloud , shell and core genomes of 44 strains (b).

Tools and techniques: Visualization

- Your job as a bioinformatician is to explain, potentially complex, biological phenomenon using computational techniques
- Your results are meaningless unless non-bioinformaticians can understand them
- Effective visualization is fundamental to bioinformatics research
 - Pictures are in fact worth 1000 words

Tools and techniques: Visualization

- Your graphics don't need to be fancy but they must be **clear** and **descriptive**
- Making graphics can take just as long as the analysis
 - Don't go too crazy for this class, you don't have the time
 - Err on the side of getting the message across
 - Bonus points for prettiest picture

Tools and techniques: Visualization

Rules for effective visualization:

1. Create a visual hierarchy
 - Identify and clearly distinguish important information
 - Background and contextual elements should not distract from the result being displayed
2. Use color and typography to guide the viewer
 - **Emphasize** important text
 - Use color to differentiate patterns in your data
3. Create the simplest figure that describes your data
4. Don't make misleading graphics
5. Caption your figures

Final considerations

- Biology is messy, there are lots of exceptions to the rules – follow the data, even when things look a little weird... just don't go down the rabbit hole
- You will need to attempt many analyses, many of them won't be informative
- Don't get discouraged if something doesn't work, iterate and move on

Final considerations

- Send me an email and we can meet/talk if:
 - You want to talk through an analysis idea
 - You're not sure about how to interpret a result
- Don't wait for final results
 - You don't need the final functional annotations or even gene predictions to start working
 - You can refine your knowledge over time with the help of groups before you