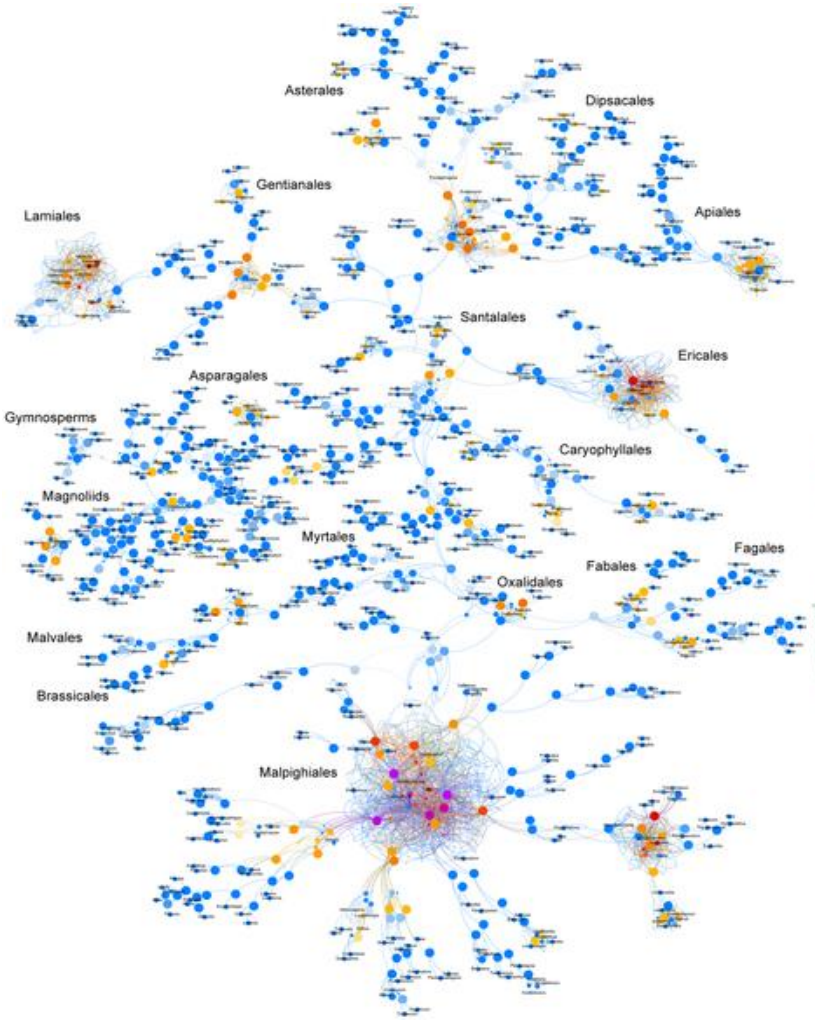
LOMA LINDA UNIVERSITY

Sequence data exploration/ Interrogation

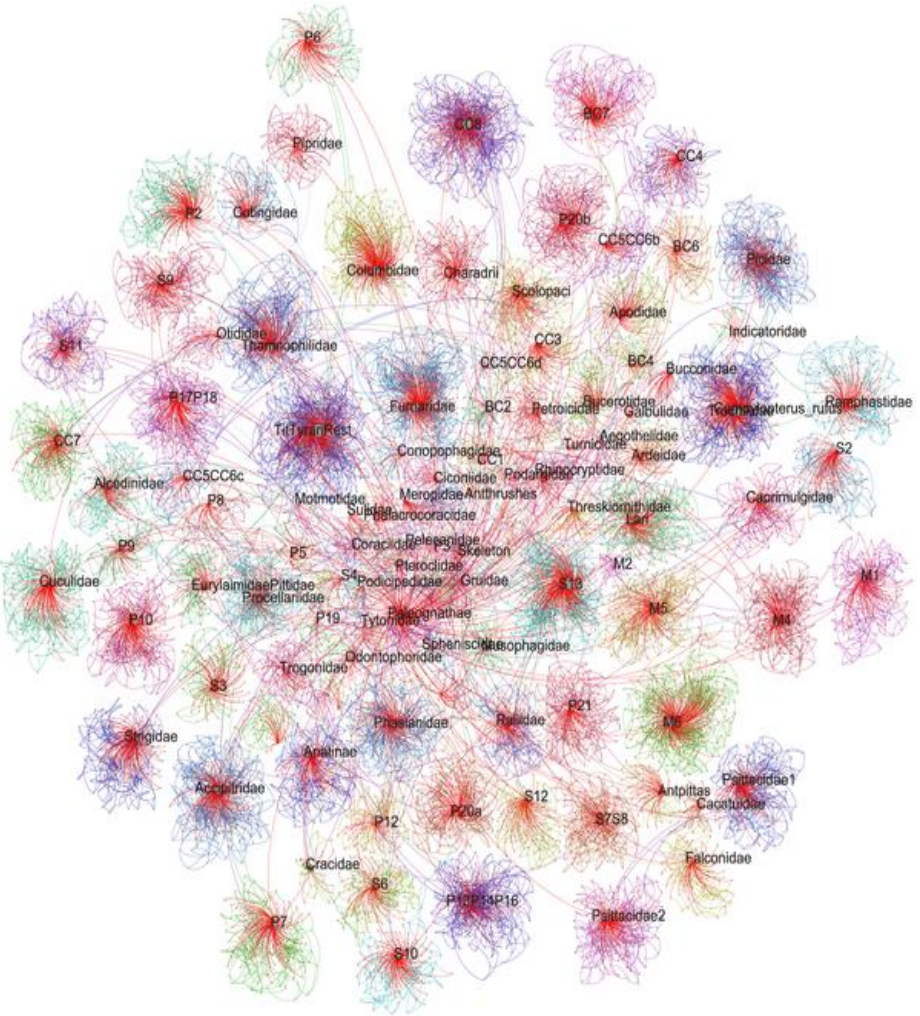
Aruni Wilson

Assistant Research Professor
Division of Microbiology and Molecular Genetics

Angiosperm Tree of Life with effective parents.



Avian tree alignment graph.

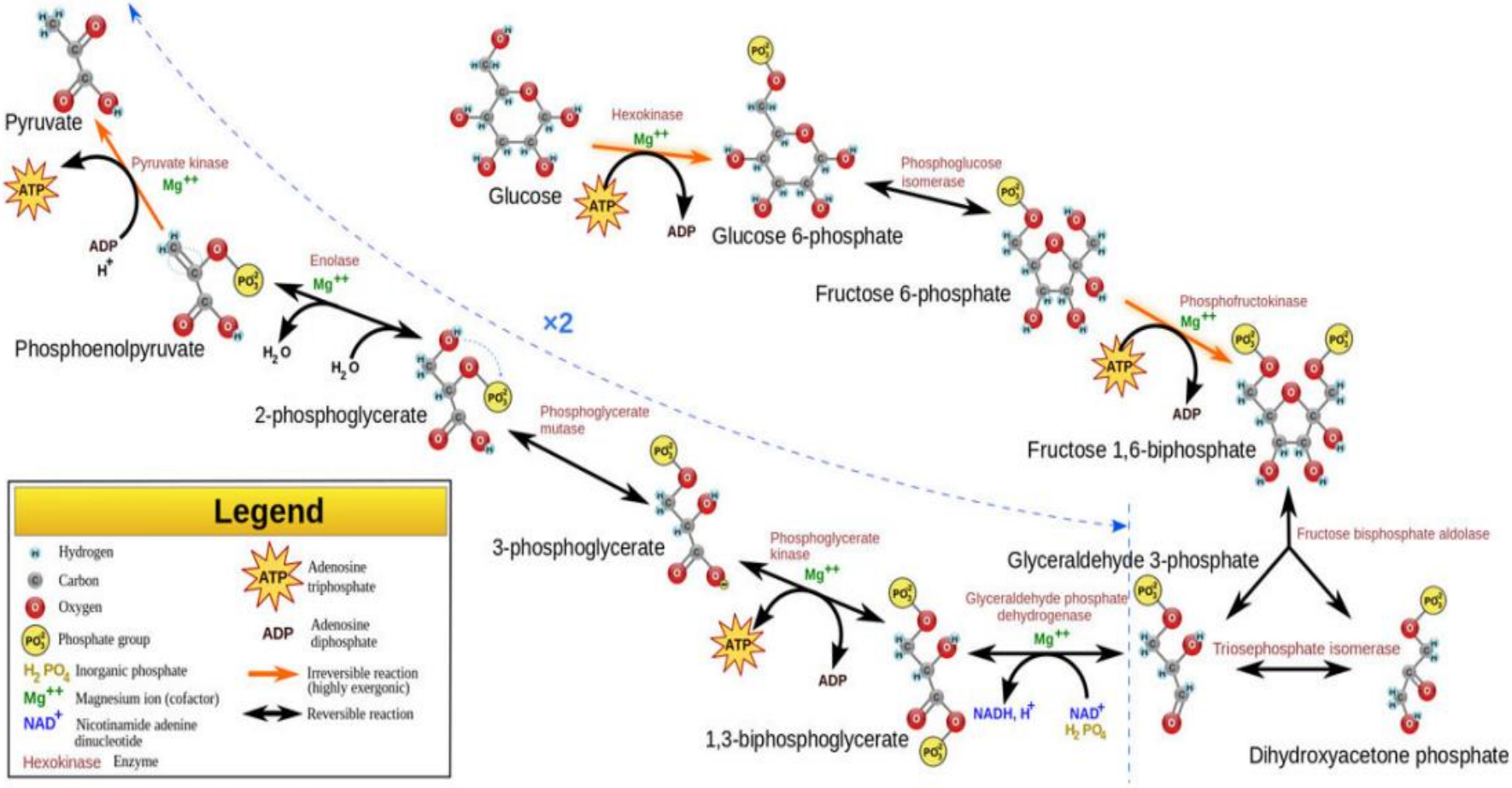


Smith SA, Brown JW, Hinchliff CE (2013). PLoS Comput Biol 9(9): e1003223. doi:10.1371/journal.pcbi.1003223

From GENOME to OMES

DNA	RNA	PROTEIN	PATHWAY
Genome	*Transcriptome*	*Proteome*	*Metabolome*
NSG	*Microarray/ RNA seq*	*MS NMR*	*Systems biology*
MSA Phylogenetic Analysis	*Heat map*	*CDD Motif search Modeling Docking*	*simulations*

Need for data exploration and integration



NCBI Mutation database , SNP db – Gene variants

Gene expression Omnibus (GEO), Array express – mRNA expression levels

MIAME (minimal information about a microarray experiment) &

LIMS (Laboratory information management systems)

BRENDA - enzyme kinetics

EMBL , Uniprot – proteome

Data integration

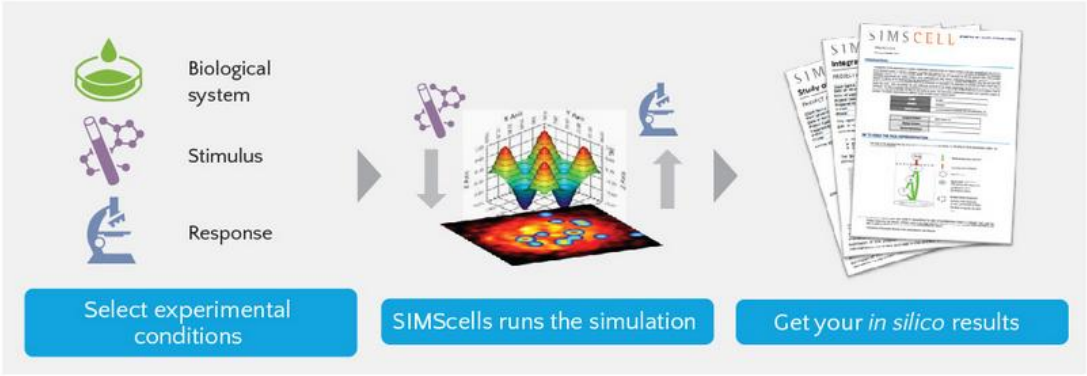
- Cataloging the elements of life
  - Data acquisition
- Deposit and share data
- Data source discovery –finding appropriate data
- Data mining – Knowledge based discovery
- Data exploitation (meta analysis- GWAS)
  - Exploitation of previous knowledge
  - KEGG, REACTOME
  - Gene set enrichment analysis (GSEA)
  - GREAT (analysis of genome regions)
- Data integration
  - Mathematical and relational models

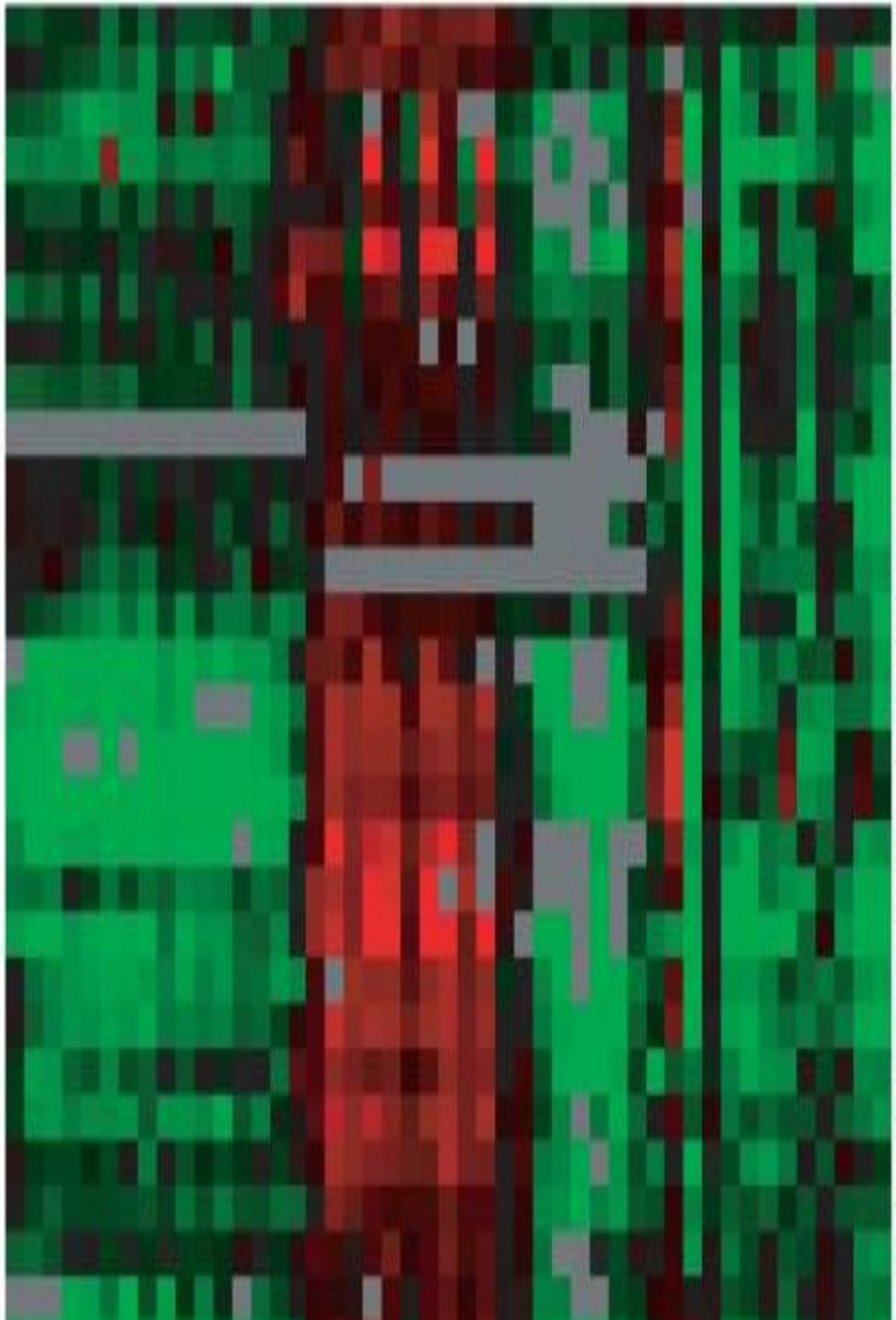
Data sources and integration

- HuGo – Human genome project
- 1000 genome project – genetic variants
- Encyclopedia of DNA elements project (ENCODE)
- The cancer genome atlas project (TCGA)
- Immunological genome project (ImmGen)
- Ingenuity pathway analysis
- Biomax
- Anaxomics - SIMScells
- Life Map

Metabolomics

✓XCMS
✓METLIN
✓Massbank
✓KEGG
✓HMDB
✓LIPIDMAPS





YOR124C	UBP2	protein degradation	ubiquitin-specific protease	cytoplasm
YBR082C	UBC4	protein degradation	ubiquitin conjugating enzyme	unknown
YPR158W	—	unknown	unknown	unknown
YGR142W	BTN2	unknown	unknown	unknown
YOL032W	—	unknown	unknown	unknown
YDR171W	HSP42	stress response	chaperone	cytoskeleton
YKR011C	—	unknown	unknown	unknown
YGL087C	MMS2	protein degradation	ubiquitin conjugating enzyme	unknown
YFR003C	—	unknown	unknown	unknown
YFR010W	UBP6	protein degradation	ubiquitin-specific protease	unknown
YKL213C	DOA1	protein degradation	unknown	unknown
YGR136W	—	unknown	unknown	unknown
YIL076W	SEC28	vesicle coat assembly	vesicle transporter	coatomer
YOR007C	SGT2	unknown	unknown	unknown
YNL007C	SIS1	protein folding	unknown	unknown
YOR027W	STI1	protein folding	unknown	unknown
YPL240C	HSP82	protein folding	chaperonin	unknown
YMR186W	HSC82	protein folding	chaperone	unknown
YER103W	SSA4	protein folding	chaperone	cytoplasm
YDR258C	HSP78	protein folding	chaperone	mitochondrial matrix
YLL026W	HSP104	protein folding	cochaperone	cytoplasm
YLR217W	—	unknown	unknown	unknown
YLR216C	CPR6	protein folding	peptidylprolyl isomerase	cytoplasm
YLR259C	HSP60	protein folding	chaperonin	mitochondrion
YDR214W	—	protein folding	unknown	unknown
YOR020C	HSP10	protein folding	chaperone	mitochondrial matrix
YNL281W	HCH1	protein folding	unknown	unknown
YJL021C	—	unknown	unknown	unknown
YAL028W	—	unknown	unknown	unknown
YJL034W	KAR2	protein folding	chaperone	endoplasmic reticulum lumen

Integrative platforms

Company / Institution	Type of Solution	Website
Appistry	Appistry's high-performance big data platform combines self-organizing computational storage with optimized and distributed high-performance computing to provide secure, HIPAA-complaint accurate on-demand analysis of omics data in association with clinical information	www.appistry.com
BGI	Beijing Genomics Institute (BGI)'s solution serves as a solid foundation for large-scale bioinformatics processing. BGI computing platform is an integrated service composed of versatile software and powerful hardware applied to life sciences	www.genomics.cn/en
CLC Bio	CLC Bio bioinformatics has a platform where both desktop and server software are integrated and optimized for best performance. CLC Bio utilize proprietary algorithms, based on published methods, in order to successfully accelerate data calculations to achieve remarkable improvements in big data analytics	www.clcbio.com
DNAnexus	DNAnexus provides solutions for NGS by using cloud computing infrastructure with scalable systems and advanced bioinformatics in a web-based platform to solve data management and the challenges in analysis that are common in unified systems	www.dnanexus.com
Genome International Corporation	Genome International Corporation (GIC) is a research-driven company that provides innovative bioinformatics products and custom research solutions for corporate, government, and academic laboratories in life sciences	www.genome.com
GNS Healthcare	GNS Healthcare is a big data analytics company that has developed a scalable approach to deal with big data solutions that could be applied across the healthcare industry	www.gnshealthcare.com
Foundation Medicine	Foundation Medicine is a molecular information company on the forefront of bringing comprehensive cancer genomic analysis to routine clinical care. Foundation Medicine is pioneering the development of a comprehensive cancer diagnostic test combining omics data, clinical information and big data analytics applied to cancer research	www.foundationmedicine.com
Knome	Knome analyzes whole genome data using software-based tests simultaneously to examine and compare many genes, gene networks, and genomes as well as integrate other forms of molecular and non-molecular data. Knome provides a platform and tools to help researchers and doctors develop next generation, software-based tests and make clinical decisions.	www.knome.com
NextBio	NextBio's big data technology enables users to systematically integrate and interpret public and proprietary molecular data and clinical information from individual patients, population studies and model organisms applying genomic data in useful ways both in scientific and medical research.	www.nextbio.com

From

Sequence to Function

DNA Sequence Comparison: First Success Story

- Finding sequence similarities with genes of known function is a common approach to infer a newly sequenced gene's function.

- In 1984 Russell Doolittle and colleagues found similarities between cancer-causing gene (v-sys simian viral oncogene) and normal growth factor (PDGF) gene.

- Identifying the similarity between PDGF and the viral oncogene helped lead to a modern hypothesis of cancer. A normal growth gene switched on at the wrong time causes cancer !

- Identifying the similarity between ATP binding ion channels and the Cystic Fibrosis gene led to a modern hypothesis of CF.

- **Comparing sequences can yield biological insight.**
- **Similar genes may have similar functions.**
- **Similar genes may have similar origins.**
- **Similarity between sequences can be quantified**

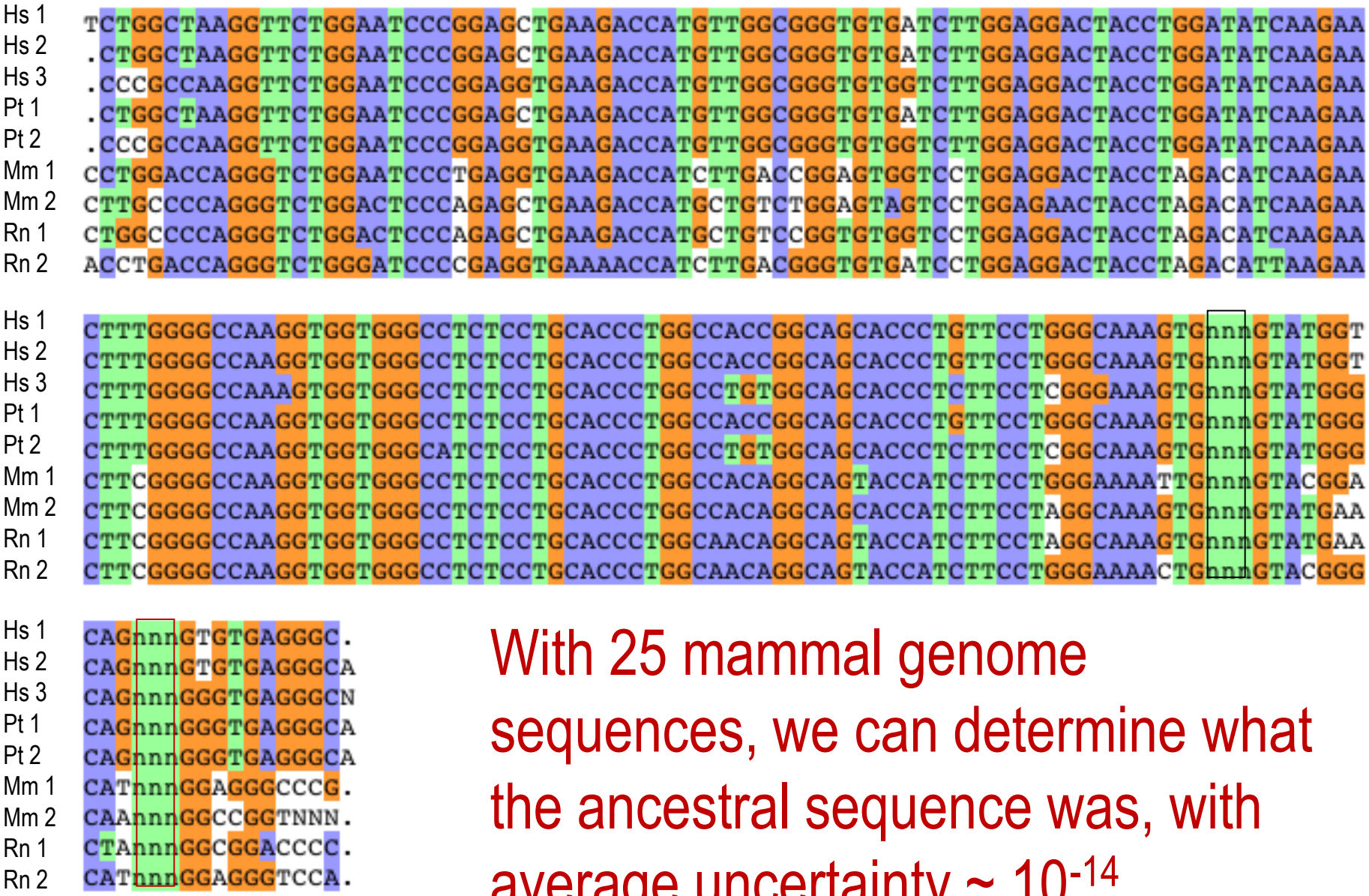
Species diversity

- In a new organism biochemistry is not completely reconfigured and new functionality isn't created by sudden appearance of whole new genes.
- incremental modifications give rise to genetic diversity and novel function in the system.

What do we infer from sequence analysis

- IDENTITY
- CONSERVATION
- DIVERGENCE
- CONVERGENCE
- DELETIONS

Example: Human, Chimp, Mouse, Rat



With 25 mammal genome sequences, we can determine what the ancestral sequence was, with average uncertainty ~ 10^{-14}.

Multiple Sequence Alignment

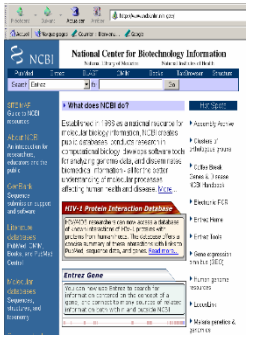
```
ACTAACCGGGAGATTTCAGA   human
AAGTTCCGGGAGATTTCCA    chimp
TAGTTATCCGGGAGATTAGA   mouse
AAAACCGGTAGATTTCAGG    rat
```

```
AC--TAACCGGGAGATTTCAGA   human
AAGTT--CCGGGAGATTTCC-A   chimp
TAGTTATCCGGGAGATT--AGA   mouse
AA---AACCGGTAGATTTCAGG   rat
```

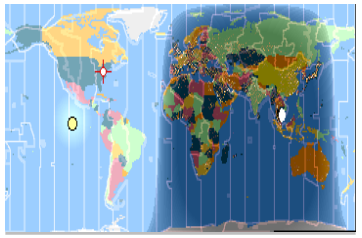
Relevance of sequence alignment

- Discovering functional, structural and microevolutionary (change from the creation) information in biological sequences
- Eases further tasks like:
- -conserved sequence motifs
- – annotation of new sequences
- – modeling of protein structures
- – design and analysis of gene expression experiments

Central server



http://www.ncbi.nlm.nih.gov/

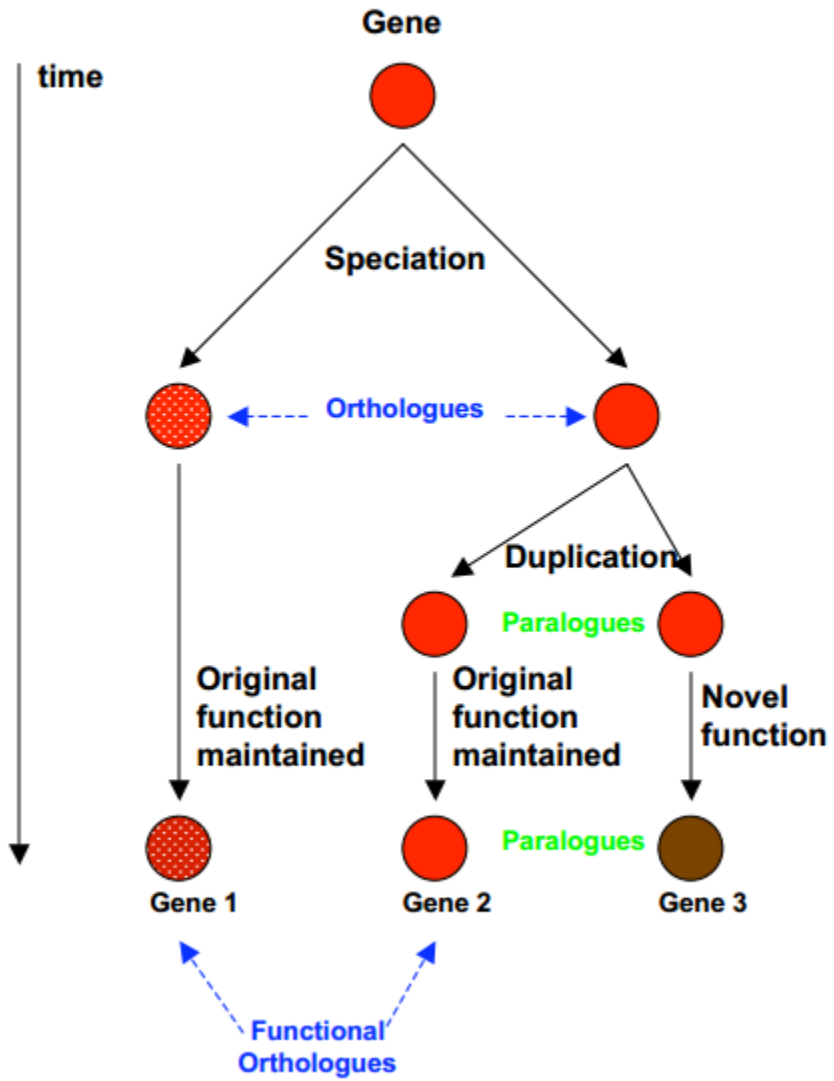


Web-based

http://www.ebi.ac.uk/

Local computer

Identifying Homologous Genes



Homologues - Genes derived from common ancestral gene

Orthologues – Genes in different species that are derived from the same gene in last common ancestor

Paralogues – Gene families that have diverged within a single species, often by duplication

The concept

- Alignment - mutual arrangement of two sequences.
- It exhibits where the two sequences are similar and where they differ.
- An 'optimal' alignment - exhibits the most correspondences, and the least differences.
- sequences that are similar probably have the same function

Sequence alignment

- Arranging the sequences of DNA, RNA or PROTEIN to identify regions of similarity.
- Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a **matrix**.
- **Gaps** are inserted between the residues
- So that identical or similar characters are aligned in **successive columns.**
- **F A T A L**
- **F A T E -**
- **If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutation and gaps as indels (that is, insertion or deletion mutations) introduced in one or both lineages.**

(Tagle *et al*. 1988)

Functional regions of DNA evolve slower than nonfunctional ones.

1. Consider a set of corresponding DNA sequences from related species.
2. Identify unusually well conserved subsequences (i.e., ones that have not mutated much over the course of time): "motifs"

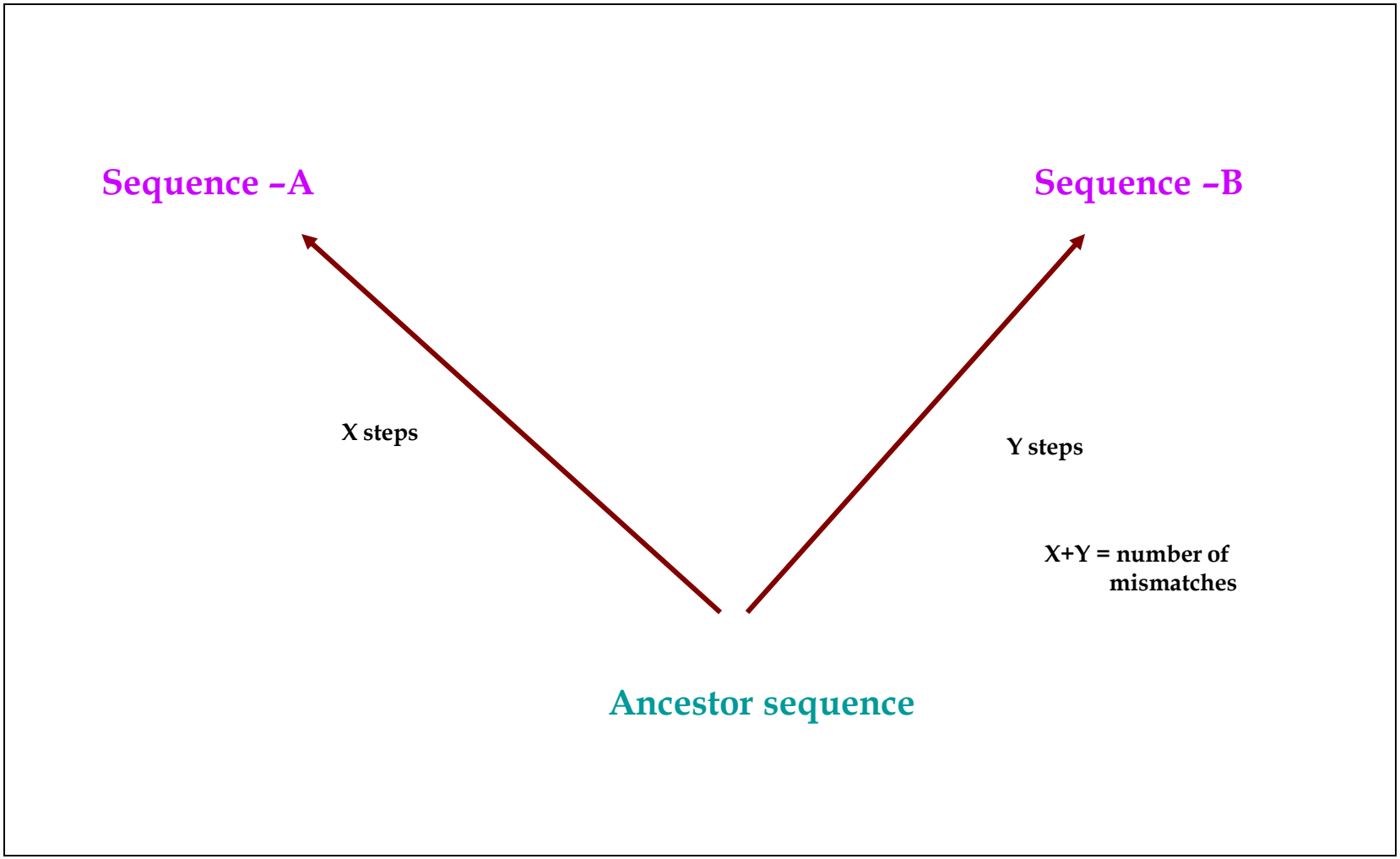
ACTAACCGGGAGATTTCAGA
human

AAGTTCCGGGAGATTTCCA
chimp

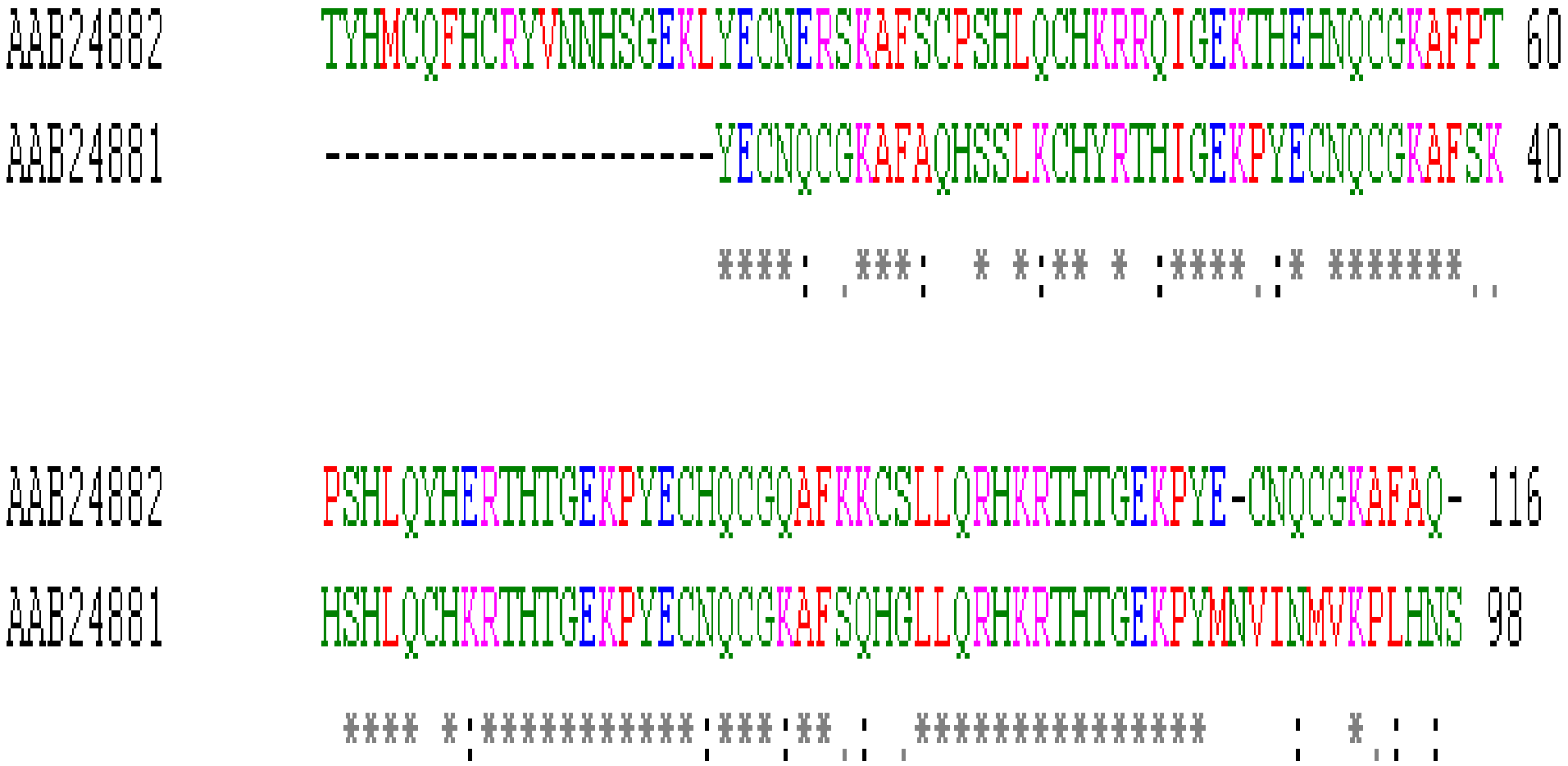
TAGTTATCCGGGAGATTAGA
mouse

AAAACCGGTAGATTTCAGG
rat

Genetic Diversity



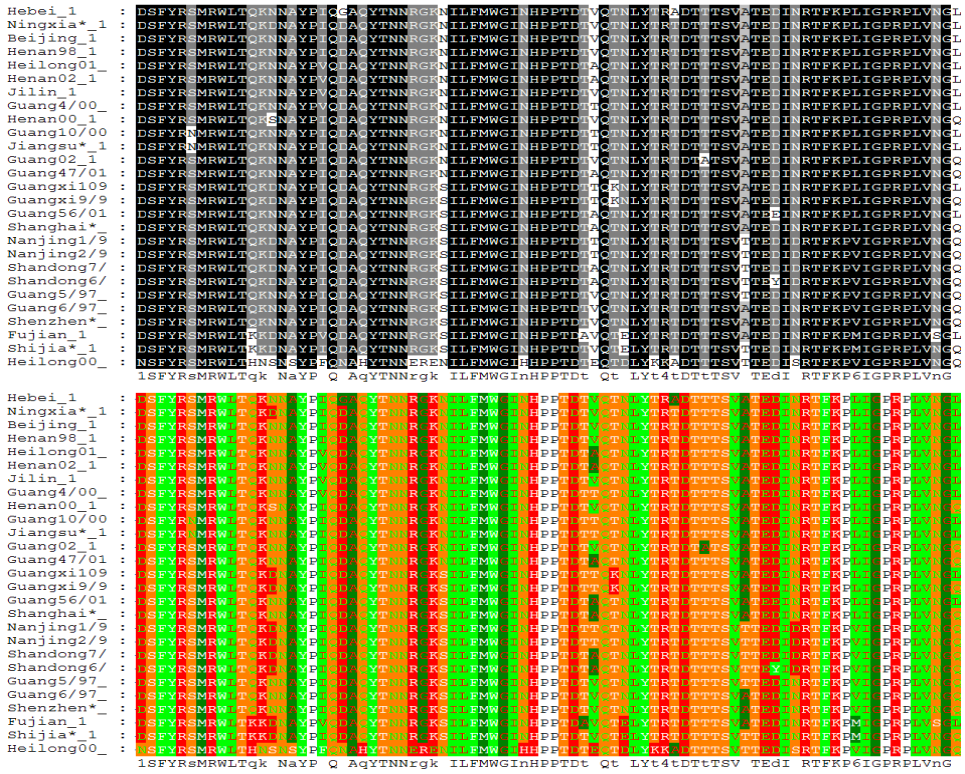
Sequence alignment



Global and local alignment

- **Global alignment**
- Aligns every residue in every sequence.
- Query and base sequence are in equal size.
- Needle-Wunsch algorithm
- **Local alignment**
- Useful for dissimilar sequences
- Smith –Waterman algorithm
- **Hybrid methods**
- Glocal methods
- **Pairwise alignments**
- **Dynamic programming**
- **Progressive methods**
- **Iterative methods**
- **Motif finding**

```
Global  FTFTALILLAVAV
        F--TAL-LLA-AV

Local   FTFTALILL-AVAV
        --FTAL-LLAAV--
```



ALIGNMENT

GLOBAL ALIGNMENT

G - A T E S
G R A T E D

LOCAL ALIGNMENT

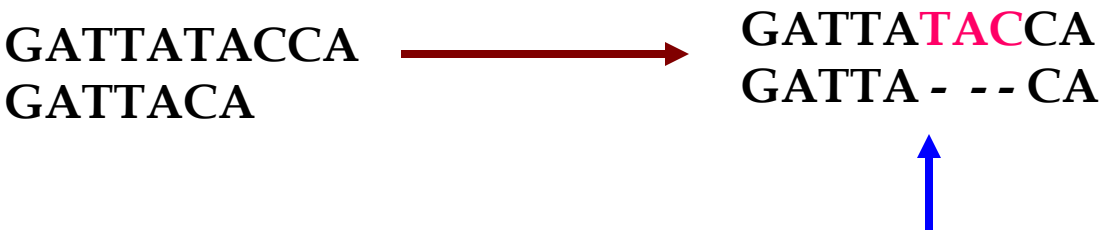
DO NOT NEED TO ALIGN ALL THE BASES IN ALL SEQUENCES

ALIGN **BILL GATES LIKES CHEESE AND GRATED CHEESE**

G- ATES LIKESCHEESE OR G-ATES & CHEESE

GRATED - - - - -CHEESE GRATED & CHEESE

Gap & Gap penalty



Insertions and deletions (INDEL) are represented by gaps in alignments

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

Gap penalties contribute to the overall score of alignments

GAP PENALTY

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

A . C . C . . C . . T

- Introduction of gaps into sequence alignments allows the alignment to be extended into regions
- where one sequence may have lost or gained sequence characters not found in the other.
- If the gap penalty is too low, then a high sequence alignment score is achievable even between unrelated or random sequences

NEEDLEMAN / WUNSCH ALGORITHM

SEQUENCE #1: GAATTCAGTTA; M = 11 (LETTERS)
SEQUENCE #2: GGATCGA; N = 7 (LETTERS)

SCORING SCHEME:

Si,j = 1 IF POS I OF #1 IS THE SAME AS POS J OF #2
Si,j = 0 IF MISMATCH SCORE
w = 0 (GAP PENALTY)

STEPS:

- INITIALIZATION
- MATRIX FILL
- TRACEBACK

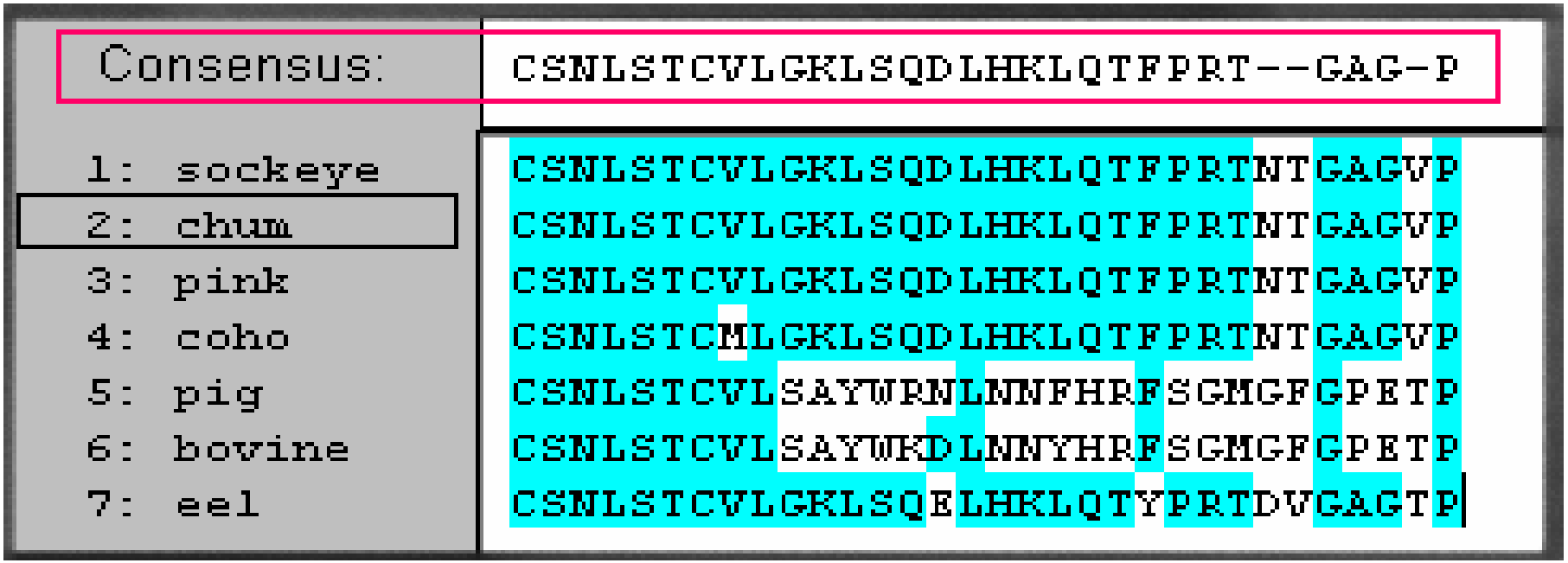
Sequence alignment

- Sequence alignment is the identification of residue – residue correspondences.
- Basic tool of bioinformatics

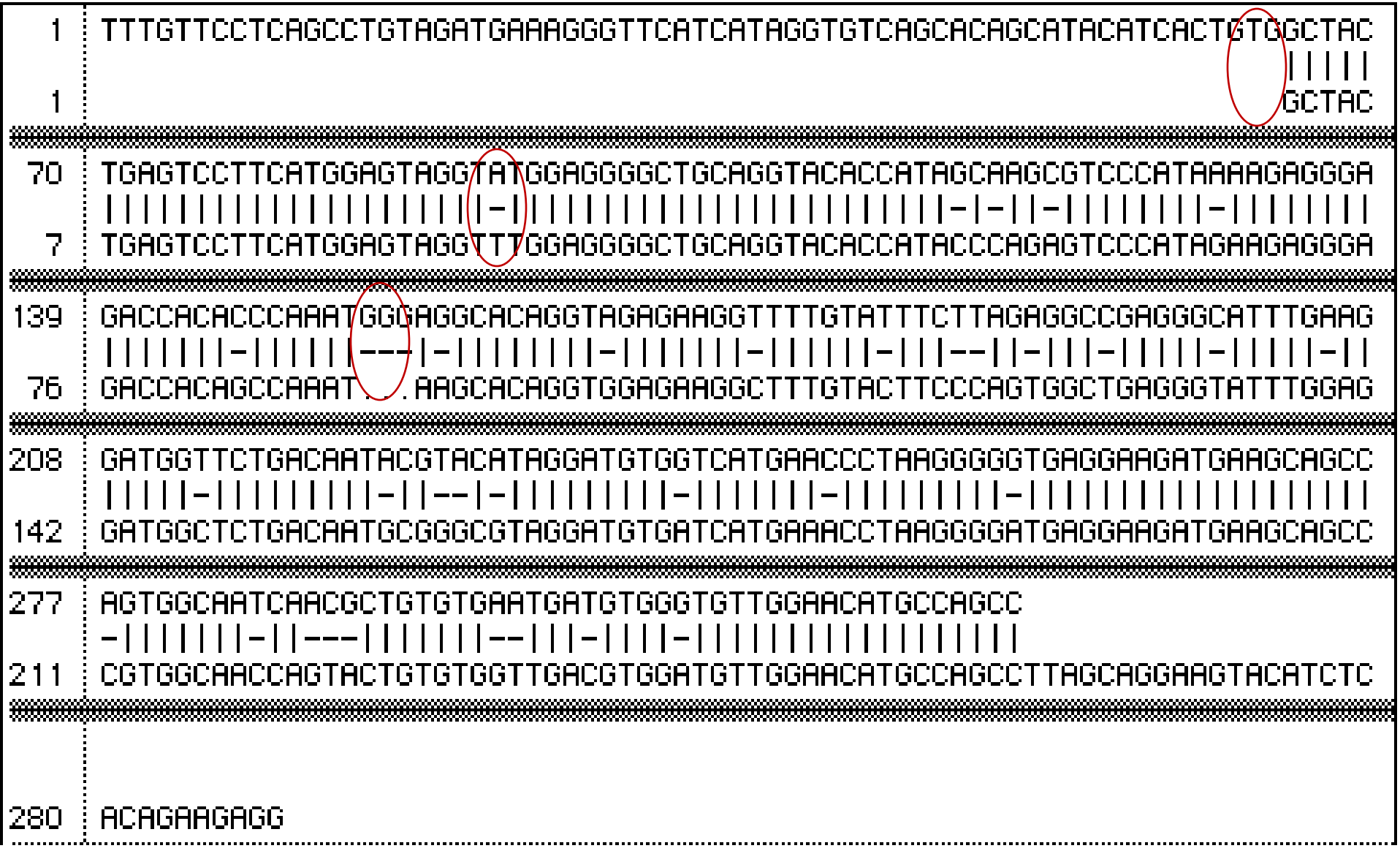
- ACCD(B to C) ABD(C deleted)

- ACCD (or) ACCD
  AB -D A -BD

Consensus sequence



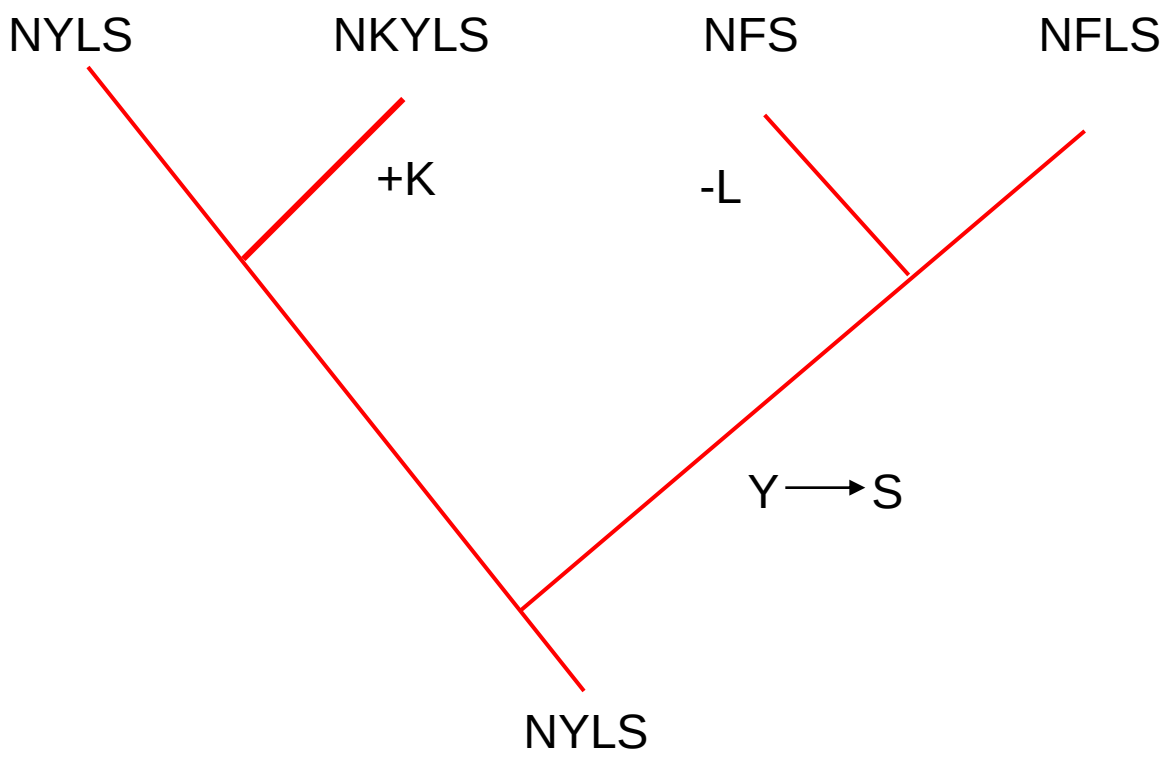
A good pair wise aligment with end gaps, indels and substitutions



Multiple sequence alignment

Alignment

N – Y L S
N K Y L S
N - F - S
N - F L S

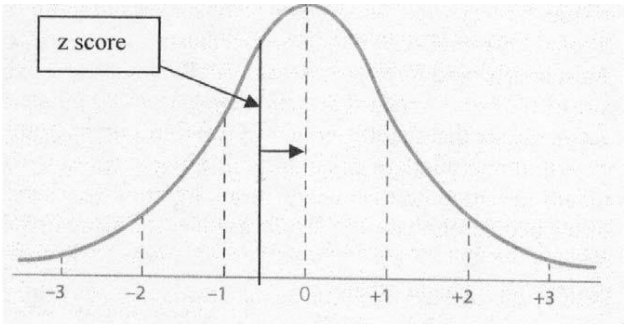


Alignment and search statistics

ALIGNMENT SCORE

SUM OF THE WEIGHTS OF EVERY PAIR OF THE ALIGNMENT

Z-SCORE (*STANDARD DEVIATION FROM THE MEAN*)



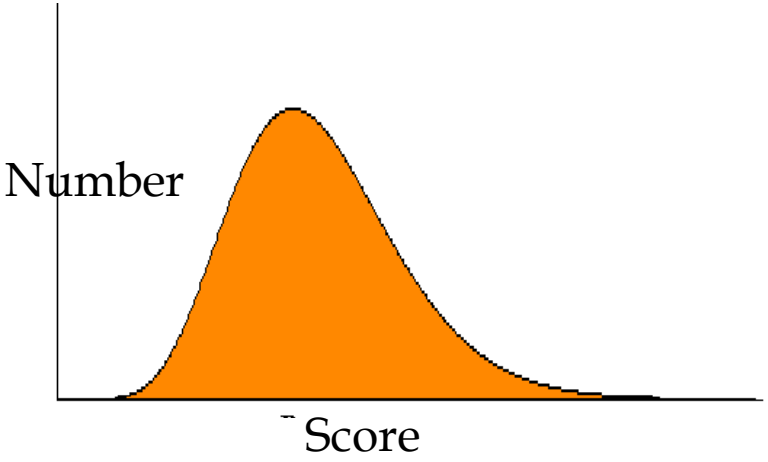
Z-SCORE = (s-m)/e

s = INITIAL SCORE
m = MEAN OF THE RANDOM SCORES
e = DEVIATION OF THE RANDOM SCORES

Alignment and search statistics

EXPECT VALUE

E = NUMBER OF DATABASE HITS YOU EXPECT TO FIND BY CHANCE



$$E = Kmne^{-\lambda S}$$

K = scale for search space
λ = scale for scoring system
S′ = bitscore = (λS - lnK)/ln2
m = effective length of query
n = effective length of database

Scoring matrix

- It is assumed that the sequences have an ancestral sequence in common with the query sequence.
- The best guess at the actual path of change is the path that requires the fewest change events.
- All substitutions are not equally likely and should be weighted to account for this.
- Insertions and deletions are less likely than substitutions and should be weighted to account for this.
- A substitution is more likely to occur between amino acids with similar biochemical properties
- hydrophobic amino acids Isoleucine(I) and valine(V) get a positive score on matrices adding weight to the likeliness that one will substitute for another.
-
- hydrophobic amino acid isoleucine has a negative score with the hydrophilic amino acid cystine(C)
- Thus matrices are used to estimate how well two residues of given types would match if they were aligned in a sequence alignment.
- 300 – BLOSUM50 85-300-BLOSUM62
- 50-85-BLOSUM80 >300 –PAM250 85-300-PAM120

BLAST

- The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences and statistical significance of matches.
- Nblast-Search a **nucleotide** database using a **nucleotide** query

- Pblast-Search **protein** database using a **protein** query
- BastX-Search **protein** database using a **translated nucleotide** query
- Tblastn-Search **translated nucleotide** database using a **protein** query (prevents manual translation)
- Tblastx-Search **translated nucleotide** database using a **translated nucleotide** query
- COBALT-Constraint based Multiple alignment tool-computes multiple protein sequence alignment using conserved domain and local sequence similarity information.
- Reverse Position Specific BLAST (RPS-BLAST) is a more sensitive way of identifying conserved domains in proteins than standard BLAST searching
- PSI Blast-This program is used to find distant relatives of a protein. First, a list of all closely related proteins is created. Then the cycle is continued.

  Used to arrive at distant evolutionary relationships
- BaltZ- A version designed for comparing multiple large genomes or chromosomes .

- **Percent identity** gives the percent of exact matches between the query sequence and the database sequence. PI also shows the number of nucleotide bases or amino acid residues which are matched in the database sequence.

- **E value (Expectation value):** This value shows that number of times your match would be expected to occur **by chance** during the searching.

- if the match is wanted to be reliable, e-value score must be low Lower e-value shows that similarities between the sequence and the database are high.

- **Score (Bit score):** A nucleotide is like a letter in an alphabet for triplet codon which later turn into an amino acid. Likewise an amino acid is like a letter for a polypeptide. While a sequence is matching with a proper database, any letter in an alignment is given a **score** depending on whether it matches or not. A negative score is given to no match and likewise a match takes positive score. When all scores are summed up for each line, total scores show up.

- Sequence -A T T C T T A G T
- Database - A A C T T C G A A
- Score 1 -1 1 1 1-1 -1-1 1
- Total 1

- **Gap value** shows the percent of the alignment sequence which has been gapped in the particular alignment.
- Gap value also includes negative scores. If gap value is high, it means that success of comparing the sequence with database is low which can not be used as a scientific data.

- To calculate this value, a formula is used:
- **Gap value = -(opening cost + (length of gap x continuing cost))**

- Opening cost: is standard number which is given =5
- Continuing cost: is also stardard number which is given = 2

- For Example: to calculate gap value for a gap for 10 nucleotides;
- **GV = -(5+(10x(2)) = -25**

Name	URL
MSA	http://www.ibc.wustl.edu/ibc/msa.html
DCA (requires MSA)	http://bibiserv.techfak.uni-biefield.de/dca
OMA	http://bibiserv.techfak.uni-biefield.de/oma
ClustalW, ClustalX	**ftp://ftp-igbmc.u-strasbg.fr/pub/clustalW or clustalX**
MultAlign	http://www.toulouse.inra.fr/multalin.html
Dialaign	http://www.gsf.de/biodv/dialign.html
ComAlign	http://www.daimi.au.df/~ocaprani
T-Coffee	**http://igs-server.cnrs-mrs.fr/~cnotred**
Praline Iterative/progressive	jhering@nimr.mrc.ac.uk
IterAlign Iterative	http://giotto.Stanford.edu/~luciano/iteralign.html
Prrp	ftp://ftp.genome.ad.jp/pub/genome/saitama-cc/
SAM	rph@cse.ucsc.edu
HMMER	http://hmmer.wustl.edu/
SAGA	http://igs-server.cnrs-mrs.fr/~cnotred
GA	czhang@watnow.uwaterloo.ca

MSA Programs

Allall	Blast	Blocks
DiAlign	Dalign	DCA
Dali	Clustalw	ClustalX
ComAlign	GA	HMMER
IterAlign	MAVID	MAFFT
MSA	MultAlign	MultAlin
Musca	Museqal	Oma
T-Coffee	ToPLign	TreeAlign
Pileup(GCG)	POA	Praline
PRRP	SAM	SAGA

Phylogenetics

• Phylogenetic trees illustrate the relationships among groups of organisms, or among a family of related nucleic acid or protein sequences

• E.g., how might have this family been derived.

Phylogeny Applications

• Analyzing changes that have occurred during the course of time on different organisms.

• Phylogenetic relationships among genes can help predict which ones might have similar functions (e.g., ortholog detection)

• Follow changes occurring in rapidly changing species (e.g., HIV virus)

- Goal: infer *past history* that produced a set of modern *characters* (sequences, typically).
- Ingredients:
  - Characters: e.g. sequence differences
  - Micro evolutionary model
    - Distance metric
    - Probabilistic model of microevolution

Interpretation

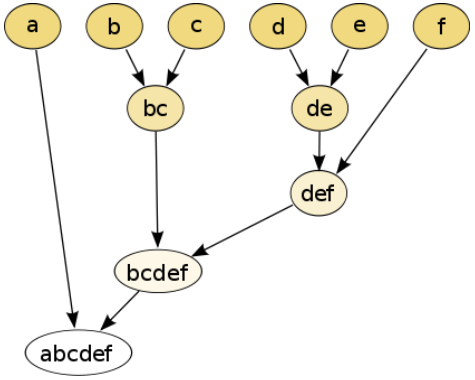
- *Any* set of objects (species; genes; proteins) can be analyzed for their relationships. - "operational taxonomic units" (OTU).

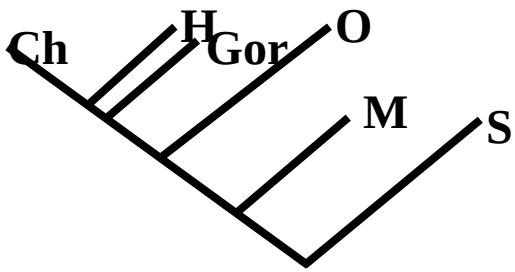
- Phylogeny is based on differences between the OTUs: *characters*. These could be sequence differences; anatomical differences, etc.

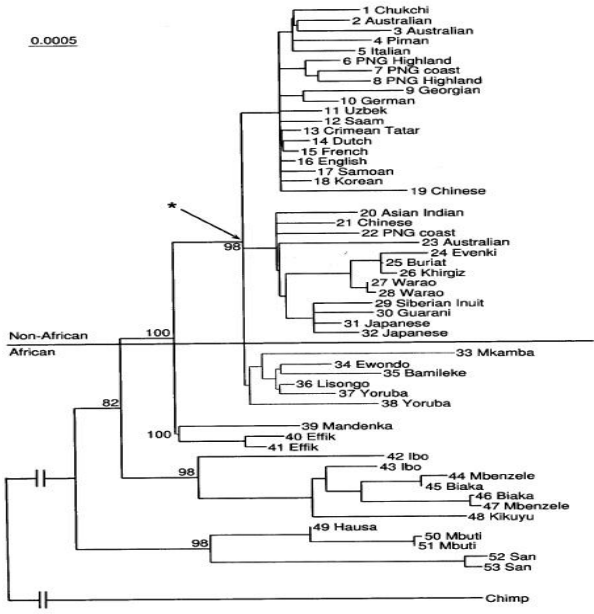
Phylogeny: Standard Assumptions

- Sequences diverge by bifurcation events (no trifurcations etc.). (Model forms a tree).
- Sequences are essentially independent once they diverge from their common ancestor.
- The probability of observing nucleotide *k* at site *j* in the future depends only on the current nucleotide at site *j*. (Markov Chain assumption).
- Different sites (characters) within a sequence evolve independently.

Phylogenetic tree



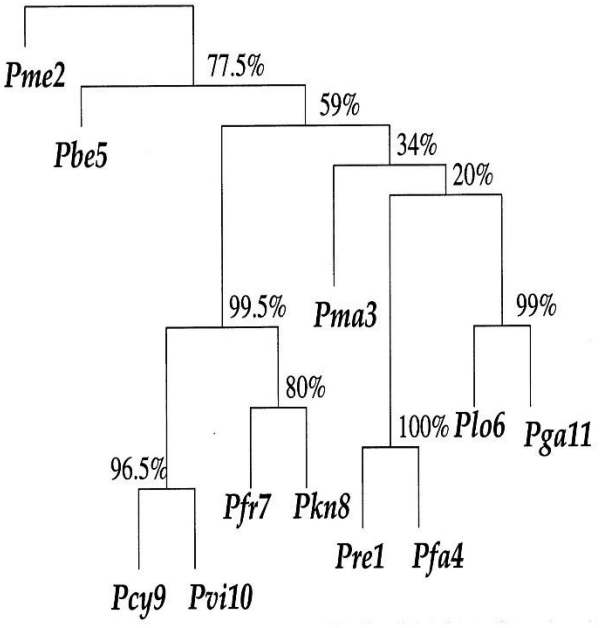




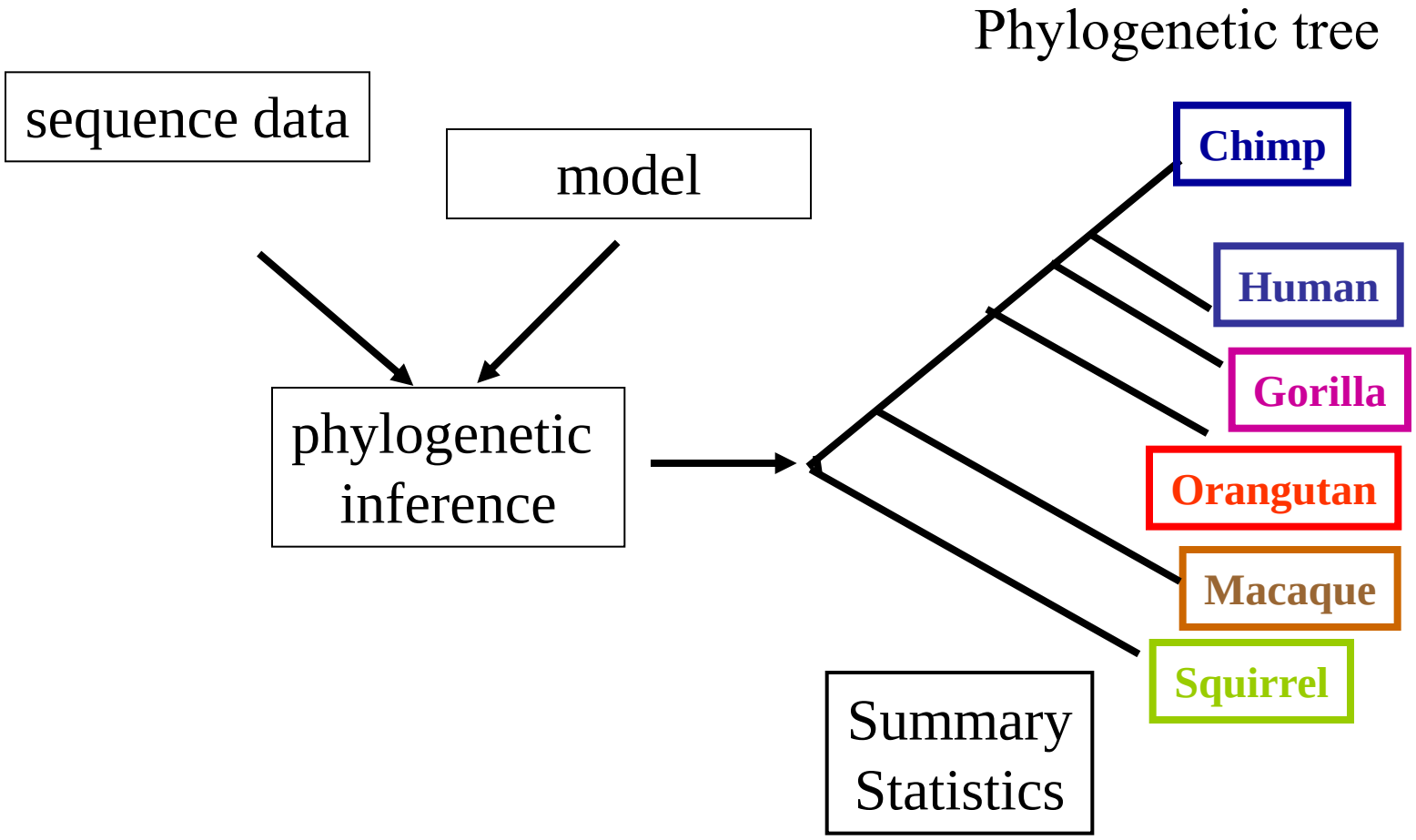
A **dendrogram** is a broad term for the diagrammatic representation of a phylogenetic tree.

A **cladogram** is a tree formed using cladistic methods. This type of tree only represents a branching pattern, i.e., its branch lengths do not represent time.

A **phylogram** is a phylogenetic tree that explicitly represents
number of character changes through its branch lengths.

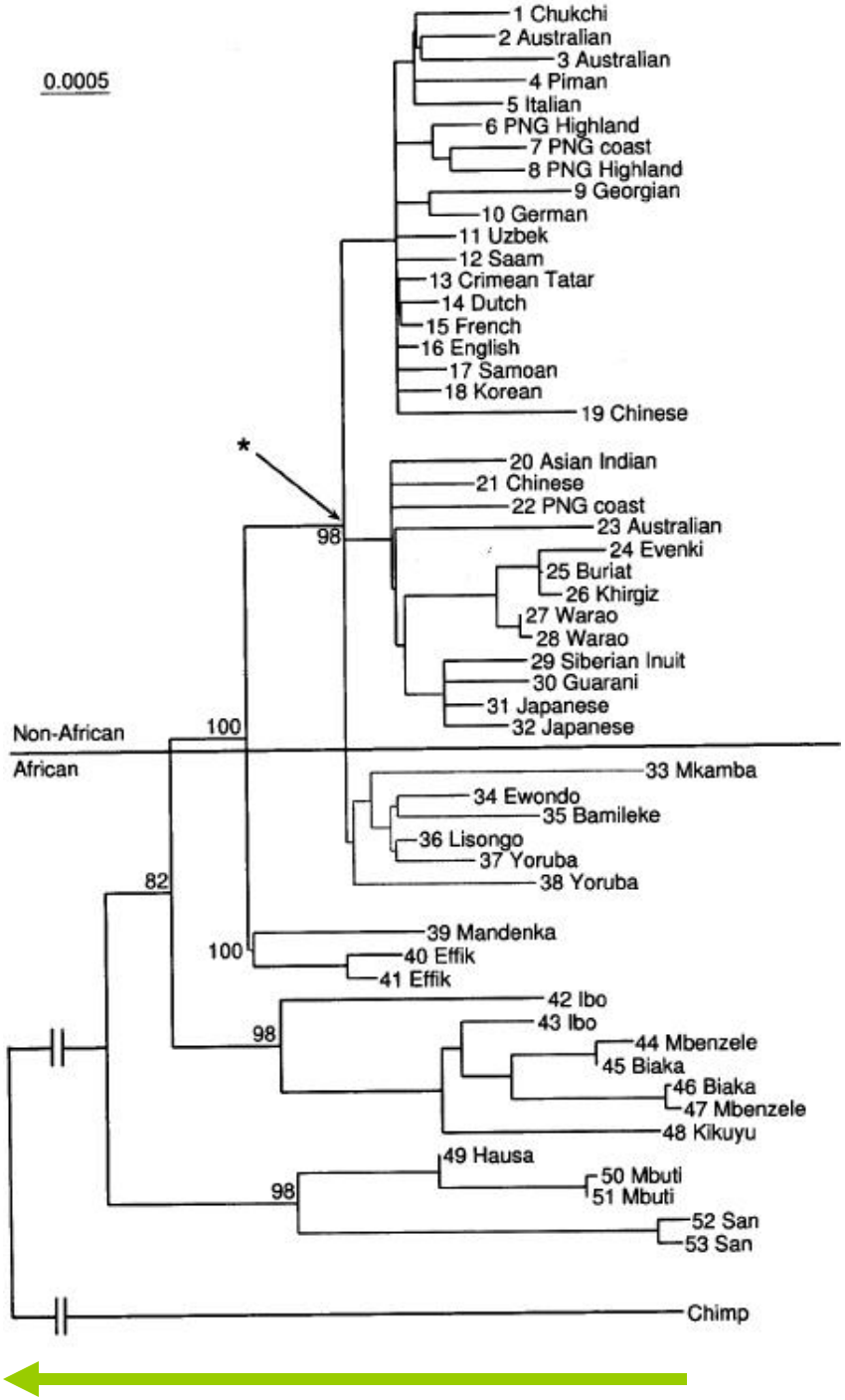


A **chronogram** is a phylogenetic tree that explicitly represents time scale through its branch lengths.


Phylogenetic tree
sequence data
model
phylogenetic inference
Chimp
Human
Gorilla
Orangutan
Macaque
Squirrel
Summary Statistics

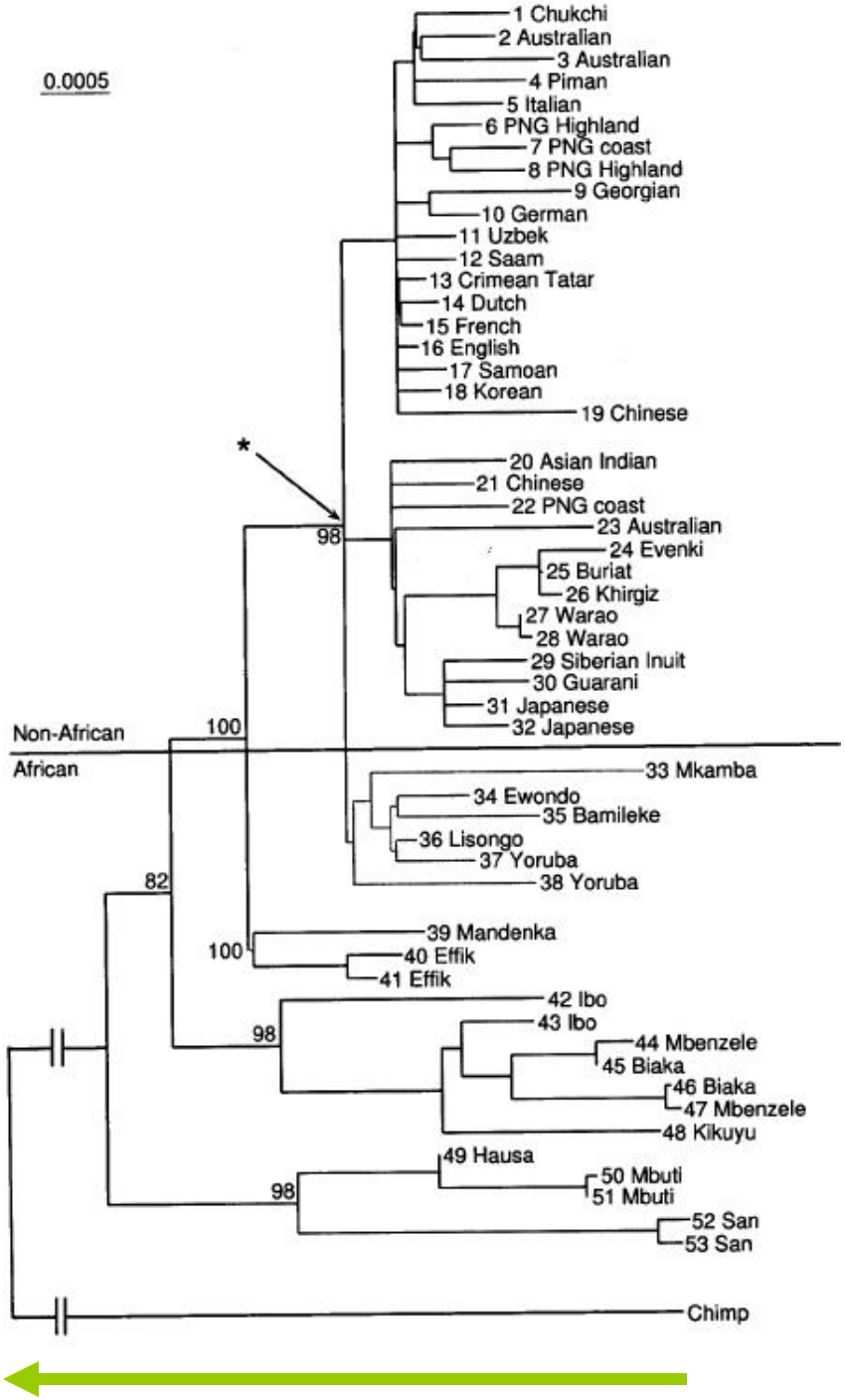
Human mtDNA Phylogeny

- *Vertical layout* is relatively meaningless. E.g. Swapping any 2-way branch has no effect on tree meaning.
- **distance** –time scale (horizontal scale, in this diagram) is the most important information in the phylogeny, and is reflected in the tree structure (grouping).
- Phylogeny is *not* classification, but ***distance***, i.e. there is no answer to "how many clusters?"



Human mtDNA Phylogeny

- *Classification*: "I am English", "You are French". Asserts that the data can be grouped into *k* non-overlapping clusters (e.g. K-means).
- "I am English", could equally well say "I am African"
- What is "African"? Everybody!



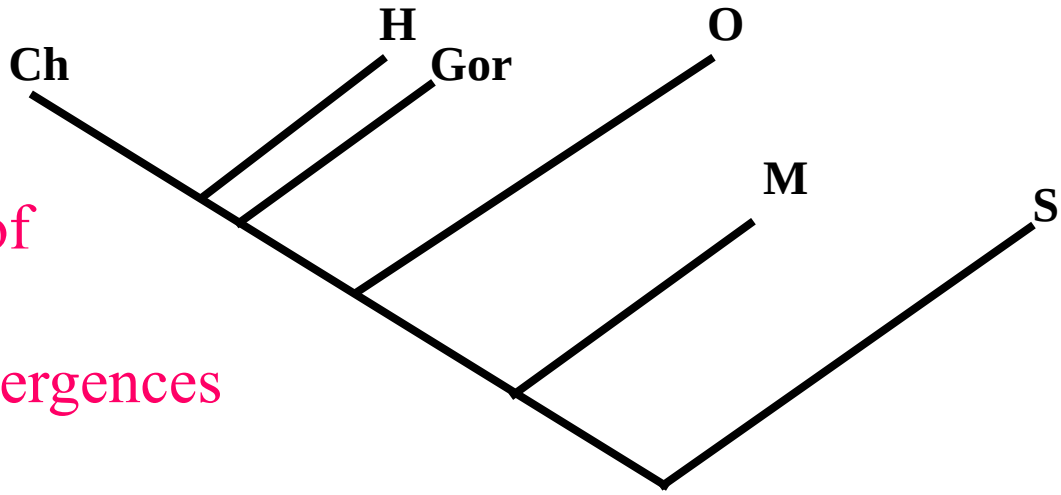
Character Differences → Branch Lengths and Order

Characters

```
Chimp      ..AGCTAAAGGGTCAGGGGAAGGGCA..
Gorilla    ..AGCATAGGGGTCAGGGGAAAGGCT..
Human      ..AGCAAAAGGGTCAGGGGAAGGGGA..
Macaque    ..AGCTCATCGGTAAGGAGAAAGGAT..
Orangutan  ..AGCCCATCGGTCAGGAGAAAGGAT..
Squirrel   ..AGCGGACCGGTAAGGAGAAAGGAC..
```

Phylogenetic Tree

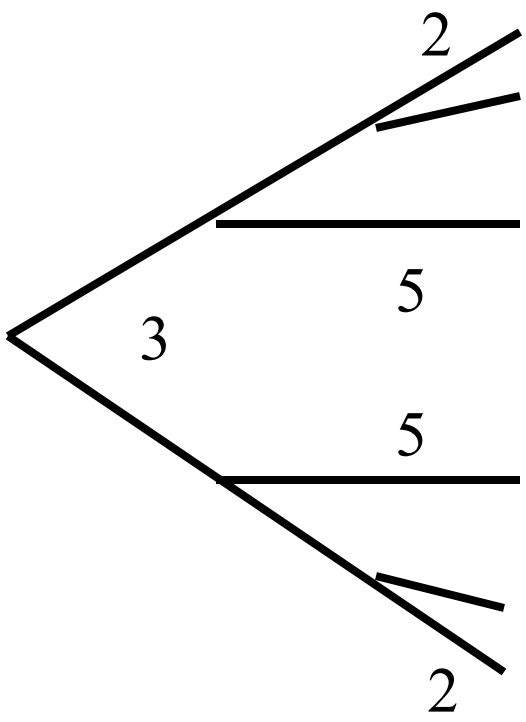
- Length of a branch is proportional to number of differences
- Tree shows order of divergences



Branch length represents the extent of change - expected number of substitutions per nucleotide site.

The longer the branch the more change

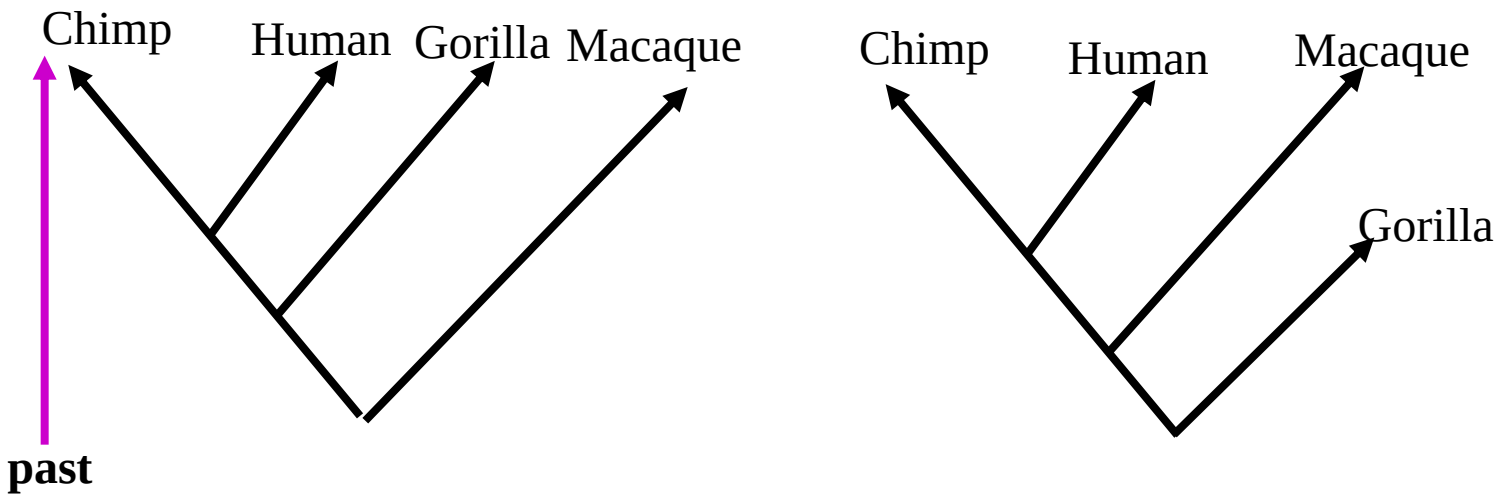
Sequence Distances → Branch Lengths and Order




```
Chimp       ..AGCTAAAGGGTCAGGGGAAGGGCA..
Human       ..AGCAAAAGGGTCAGGGGAAGGGGA..

Gorilla     ..AGCATAGGGGTCAGGGGAAAGGCT..


Squirrel    ..AGCGGACCGGTAAGGAGAAAGGAC..

Macaque     ..AGCTCATCGGTAAGGAGAAAGGAT..
Orangutan   ..AGCCCATCGGTCAGGAGAAAGGAT..
```

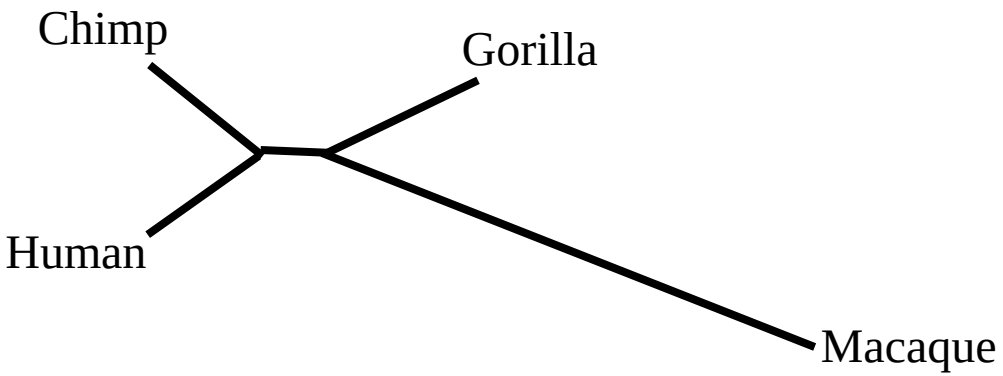
Rooted Versus Unrooted Trees

Rooted Trees



These two rooted trees are different
But both have the same Unrooted tree:

Unrooted Tree



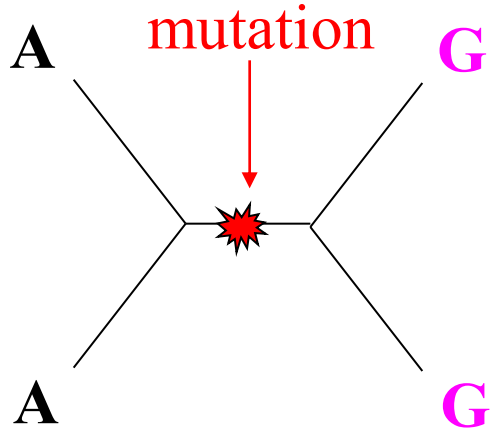
The direction of diversification is specified in a rooted tree but not in an unrooted tree

The Principle of Parsimony

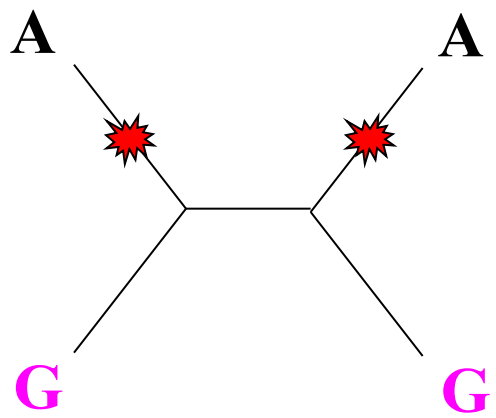
- Out of the possible trees, the one with the fewest required "mutation events" is best.

- A mutation event is a letter substitution at a particular point on the tree; sequences on one side of the event have letter X, and on the other side have letter Y.

Parsimony:
Prefer Tree with Least Mutation Events



Requires only one mutation event



Requires at least two mutation events

#mutations depends on tree

Methods

- Maximum parsimony -Heuristic
  - least complex explanation for an observation
- Maximum likelihood
  - Statistical estimation of assumption
- Bayesian inference
  - Statistical interference in which evidence or observations are used.
- Phenetics
  - Numerical taxonomy
- Neighbour –Joining
  - bottom-up clustering method

BLAST

BLAST Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

New Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool. Go

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

SDSPB | Science Data Sharing Platform for Bioinformation All Databases Please enter keywo

Resources | Subject Databases | Services | About Us

LSBI BLAST

- BLAST Assembled Genomes
- Basic BLAST
- Specialized BLAST
- My BLAST
- Recent Results
- Databases Explanation

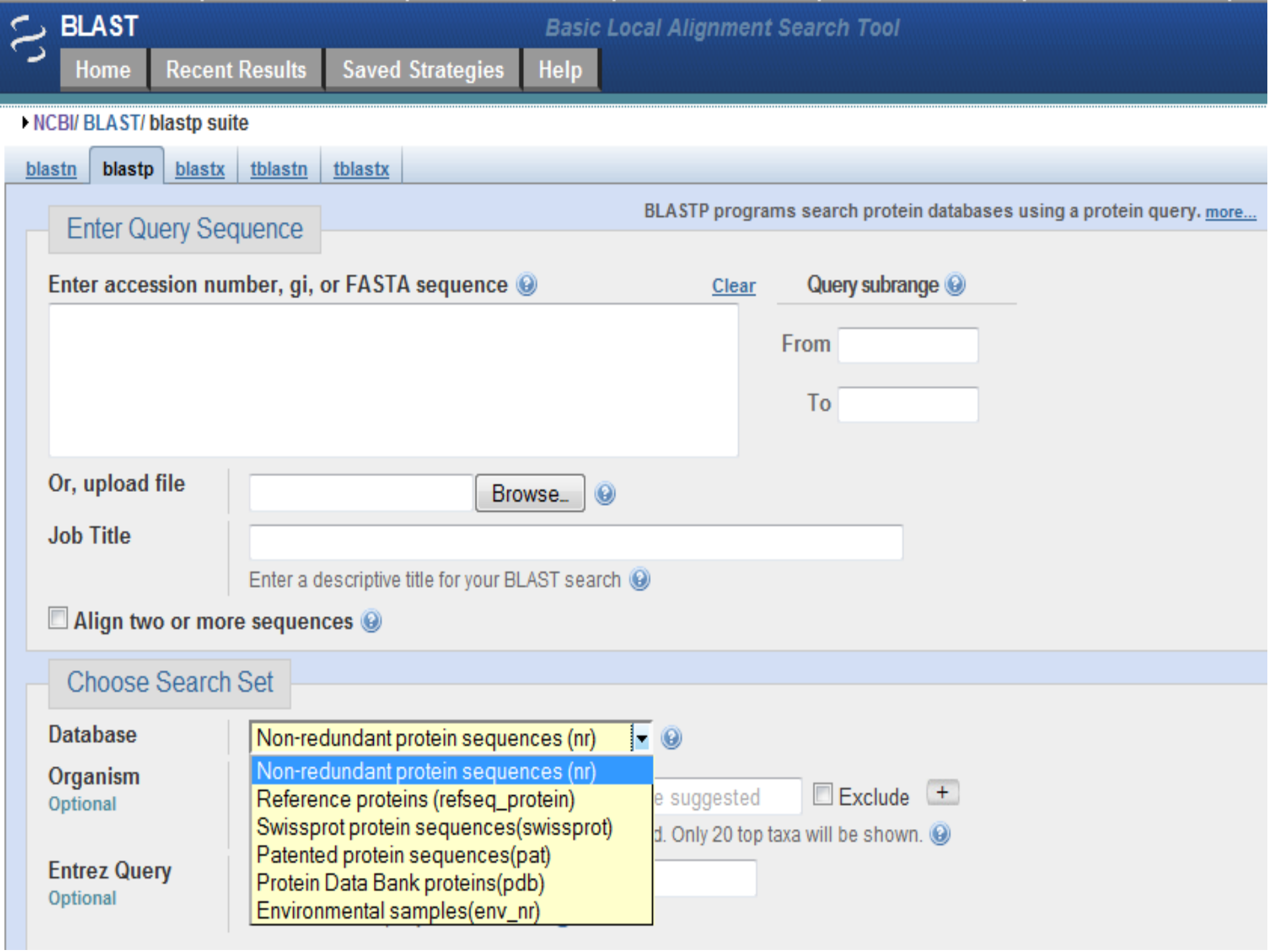
Related Links

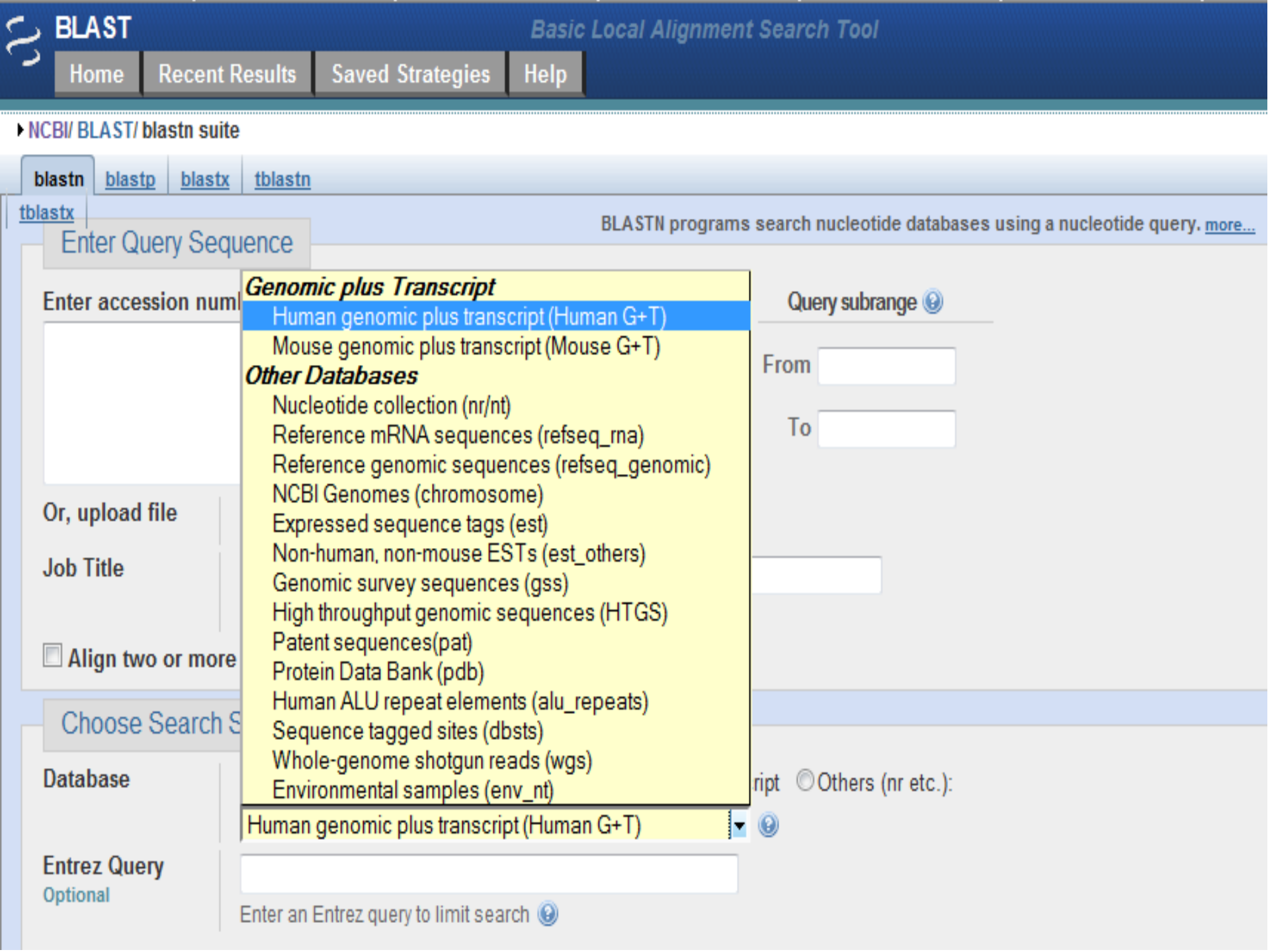
- NCBI Blast

BLAST Assembled Genomes

Choose a species genome to search

1: Baker's yeast
2: Leptospira interrogans
3: Rattus
4: 流感病毒
5: Schistosoma mansoni
6: Vector
7: Hepatitis Virus
8: Leptospira
9: Human immunodeficiency virus
10: Homo sapiens


BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastp suite
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein databases using a protein query. more...
Enter Query Sequence
Enter accession number, gi, or FASTA sequence
Clear
Query subrange
From
To
Or, upload file
Browse...
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Non-redundant protein sequences (nr)
Non-redundant protein sequences (nr)
Reference proteins (refseq_protein)
Swissprot protein sequences(swissprot)
Patented protein sequences(pat)
Protein Data Bank proteins(pdb)
Environmental samples(env_nr)
Organism
Optional
suggested
Exclude
d. Only 20 top taxa will be shown.
Entrez Query
Optional


BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastn suite
blastn
blastp
blastx
tblastn
tblastx
BLASTN programs search nucleotide databases using a nucleotide query. more...
Enter Query Sequence
Enter accession numl
Or, upload file
Job Title
Align two or more
Genomic plus Transcript
Human genomic plus transcript (Human G+T)
Mouse genomic plus transcript (Mouse G+T)
Other Databases
Nucleotide collection (nr/nt)
Reference mRNA sequences (refseq_rna)
Reference genomic sequences (refseq_genomic)
NCBI Genomes (chromosome)
Expressed sequence tags (est)
Non-human, non-mouse ESTs (est_others)
Genomic survey sequences (gss)
High throughput genomic sequences (HTGS)
Patent sequences(pat)
Protein Data Bank (pdb)
Human ALU repeat elements (alu_repeats)
Sequence tagged sites (dbsts)
Whole-genome shotgun reads (wgs)
Environmental samples (env_nt)
Query subrange
From
To
Choose Search S
Database
ript
Others (nr etc.):
Human genomic plus transcript (Human G+T)
Entrez Query
Optional
Enter an Entrez query to limit search

NCBI/ BLAST/ blastn suite

blastn | blastp | blastx | tblastn | tblastx

BLASTN programs search nucleotide databases using a r

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

Query subrange

From

To

Or, upload file Browse...

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Human genomic + transcript | Mouse genomic + transcript | Others (nr etc.):

Human genomic plus transcript (Human G+T)

Entrez Query
Optional

Enter an Entrez query to limit search

Program Selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Edit and Resubmit | Save Search Strategies | ▷Formatting options | ▷Download

EST_Clone_AI2

Query ID lcl|8485
Description EST_Clone_AI2
Molecule type nucleic acid
Query Length 611

Database Name nr
Description All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
Program BLASTN 2.2.22+ ▷Citation

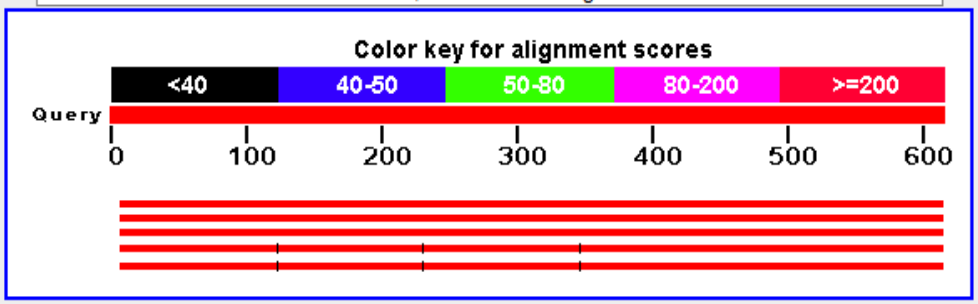
Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results]

▼ Graphic Summary

▼ Graphic Summary

Distribution of 11 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments



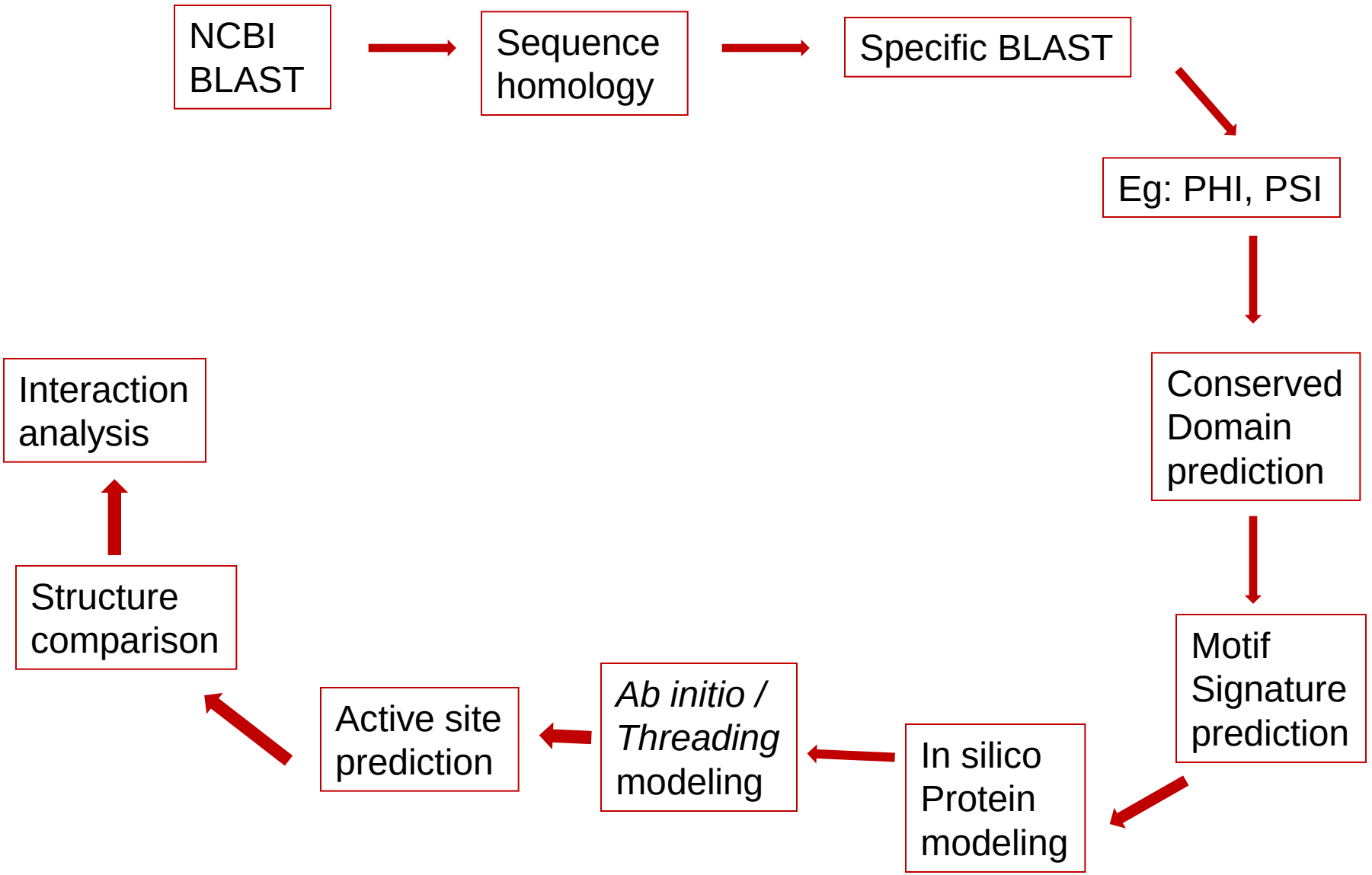
▼ Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

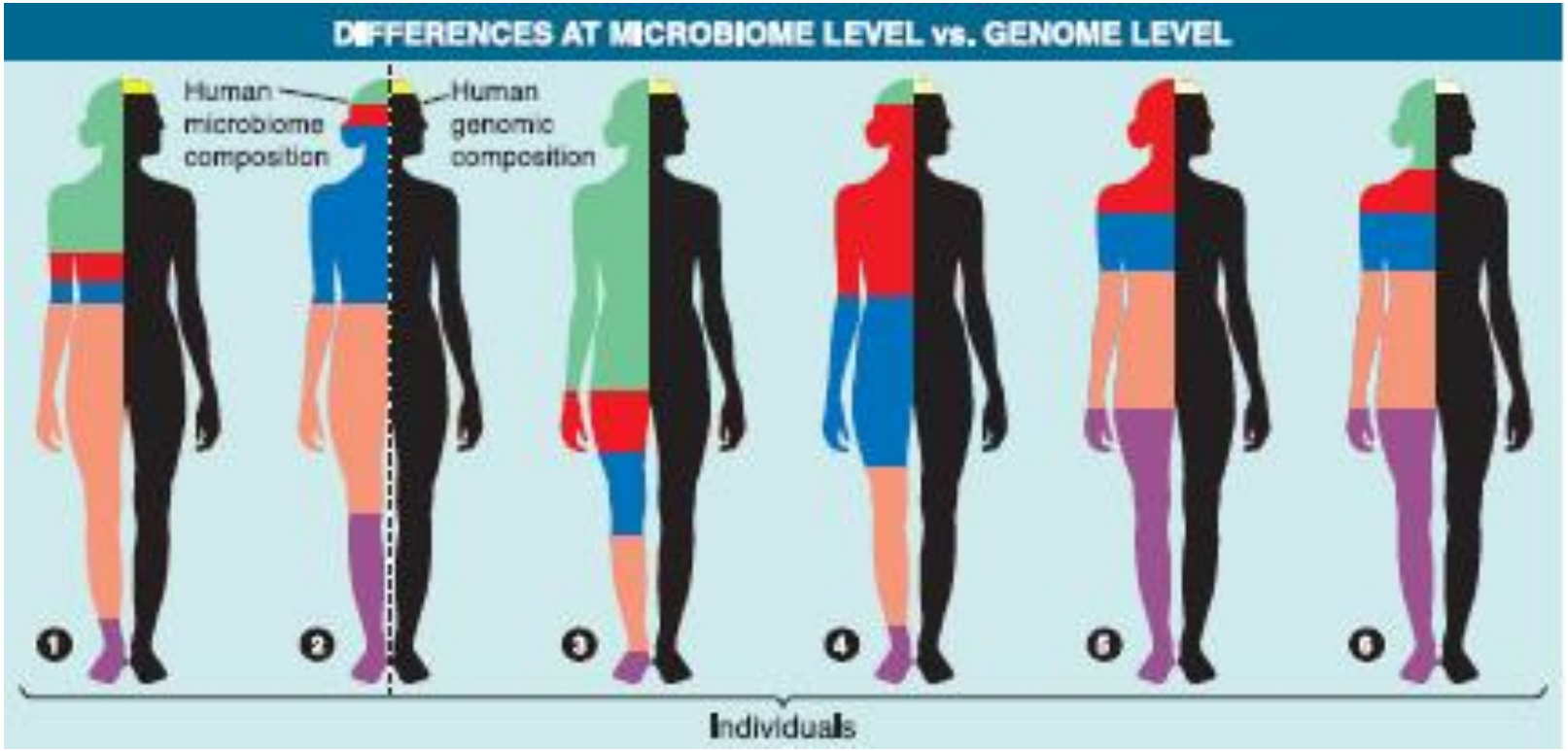
Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
Transcripts							
NM_138715.2	Homo sapiens macrophage scavenger receptor 1 (MSR1), transcript	1114	1114	98%	0.0	100%	G M
NM_138716.2	Homo sapiens macrophage scavenger receptor 1 (MSR1), transcript	1114	1114	98%	0.0	100%	G M
NM_002445.3	Homo sapiens macrophage scavenger receptor 1 (MSR1), transcript	1114	1114	98%	0.0	100%	G M
Genomic sequences [show first]							
NT_167187.1	Homo sapiens chromosome 8 genomic contig, GRCh37 reference pr	499	1127	98%	1e-138	100%	

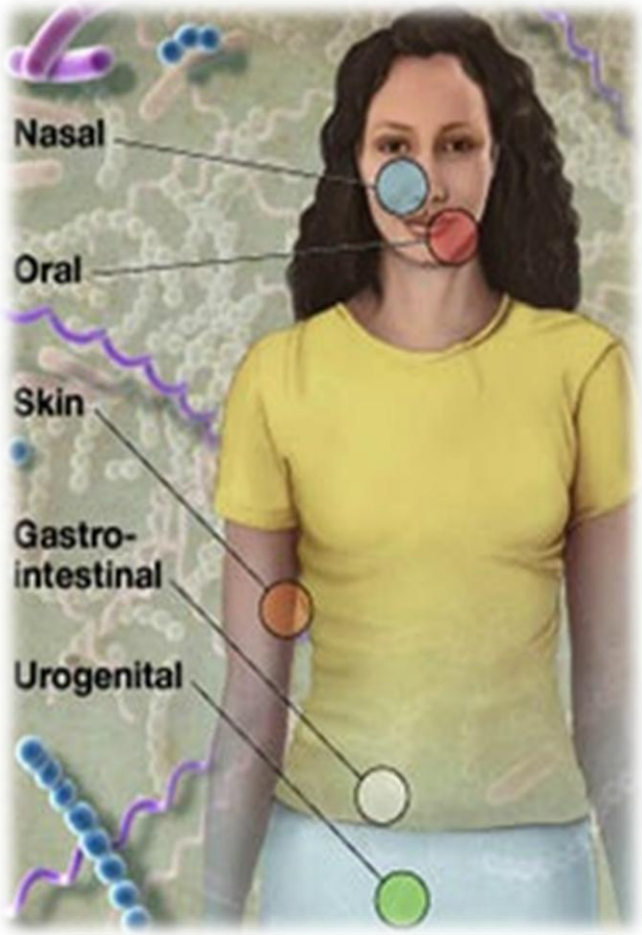
Data Integration

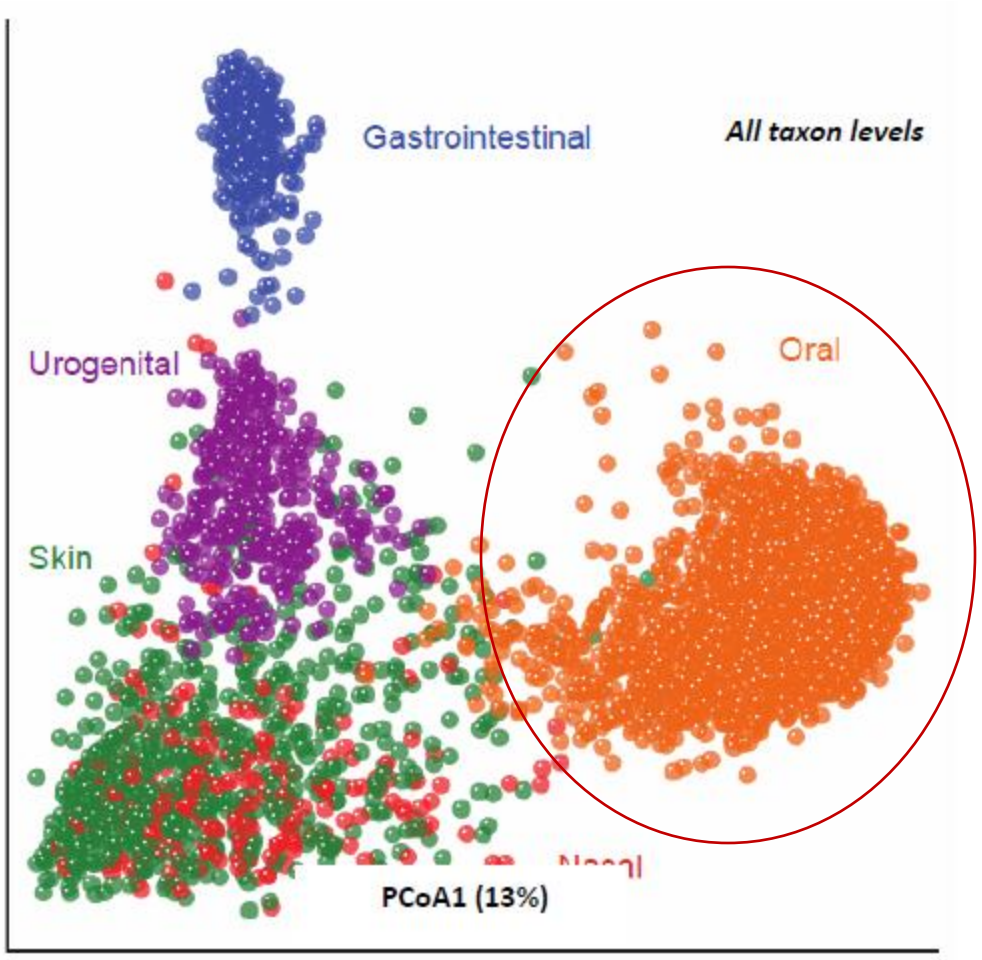


The Human Microbiome



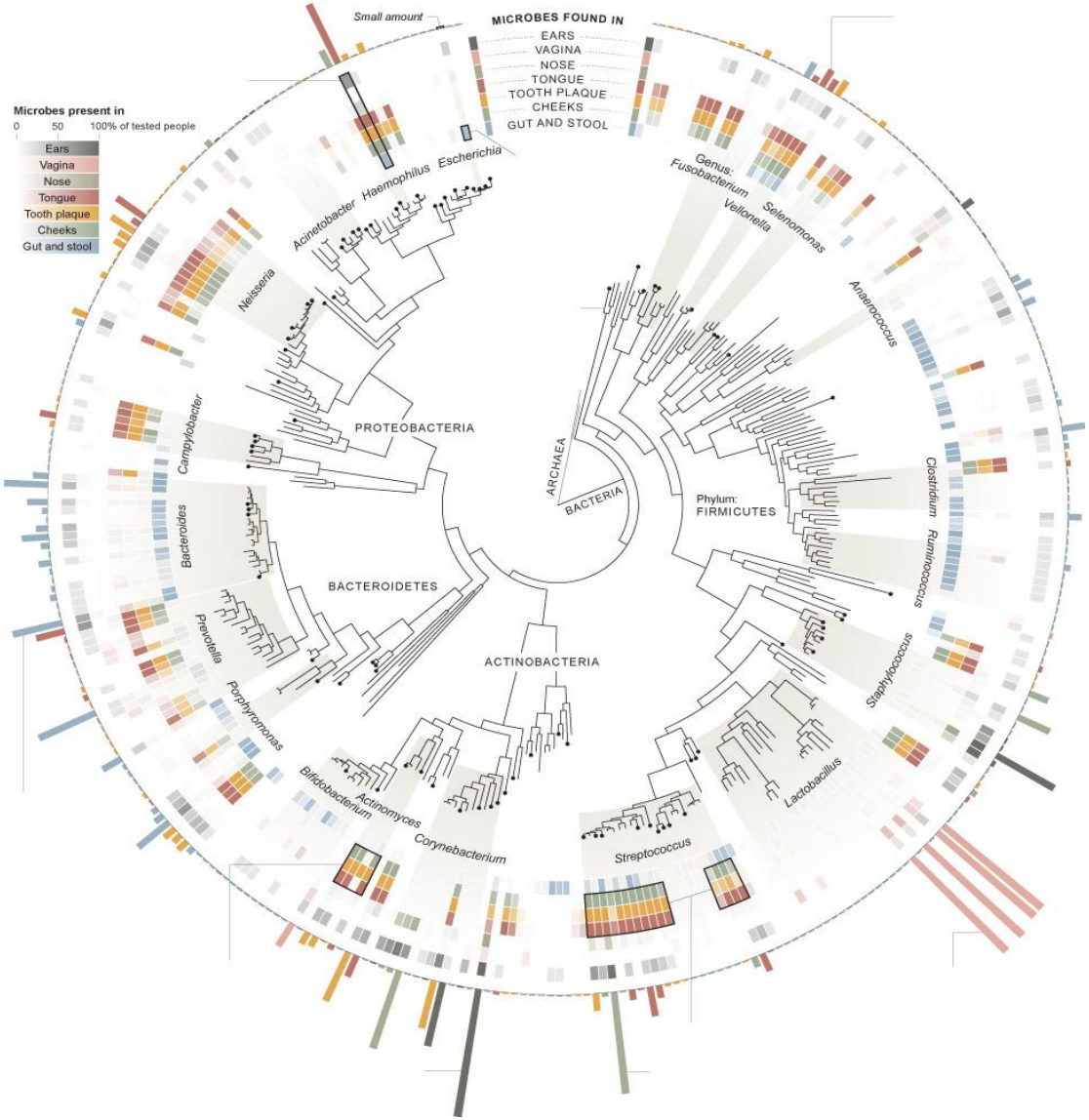
Host associated bacterial communities

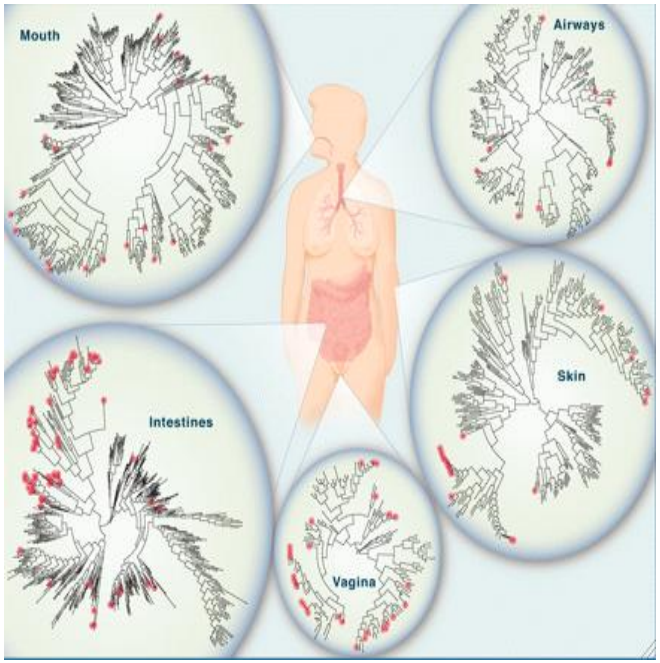




HGP consortium, 2012

Complex combination of organisms living in the body

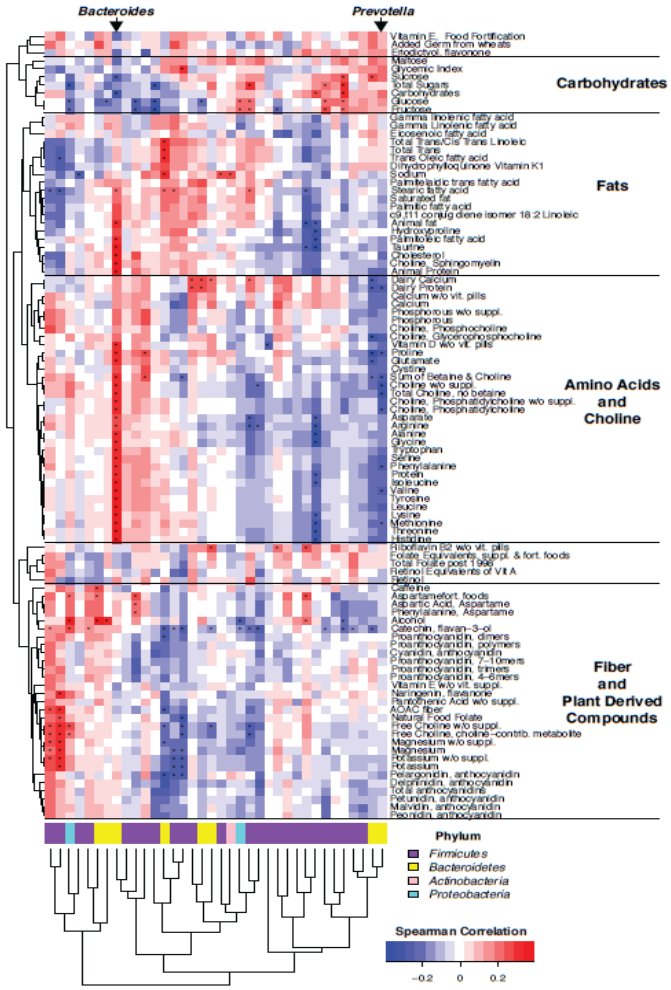




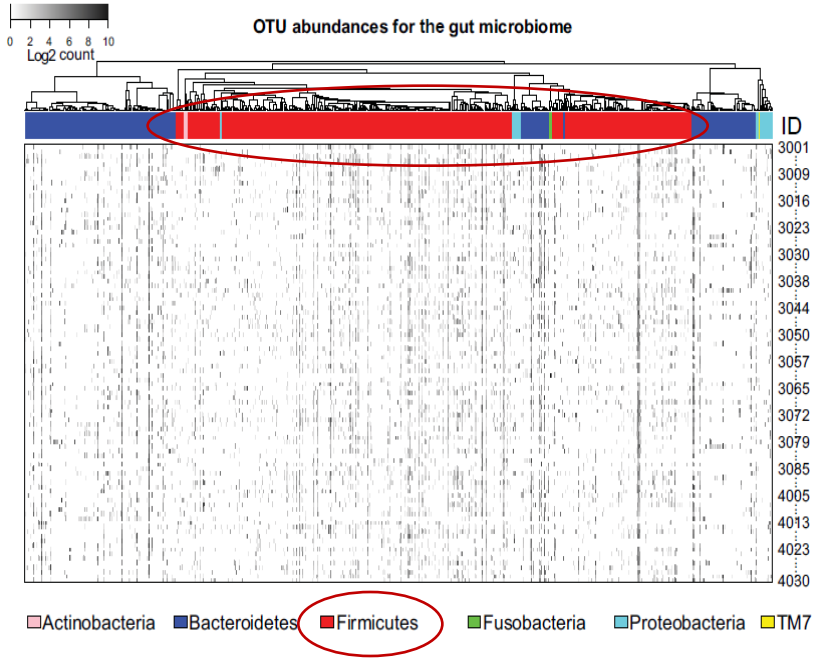
Tongue
Tooth plaque
Gut

Mason et al 2013

Gut Microbiome Project



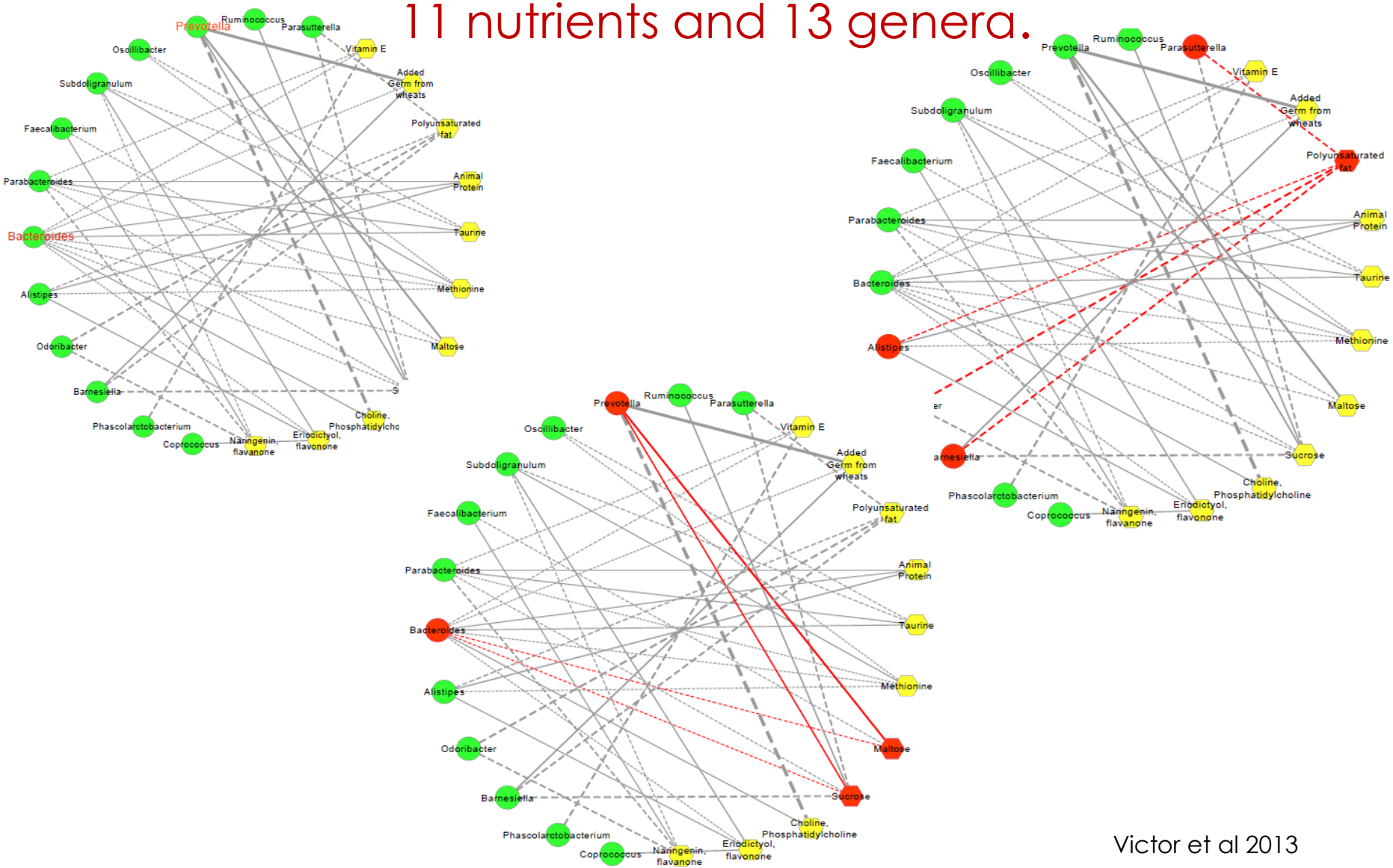
Wu, Chen *et. al* (Science 2011).



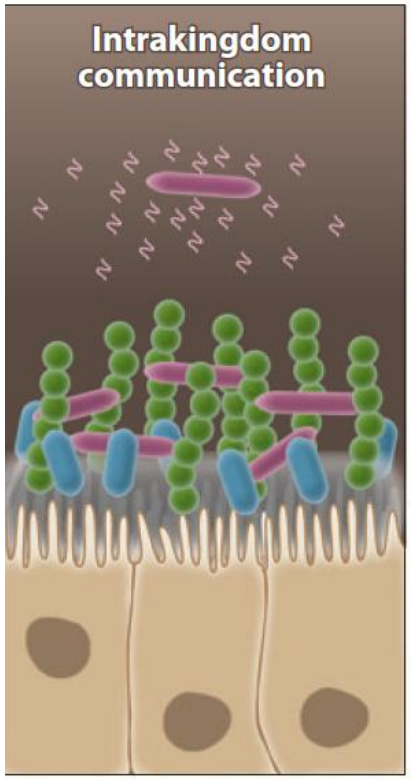
Major human associated Phyla

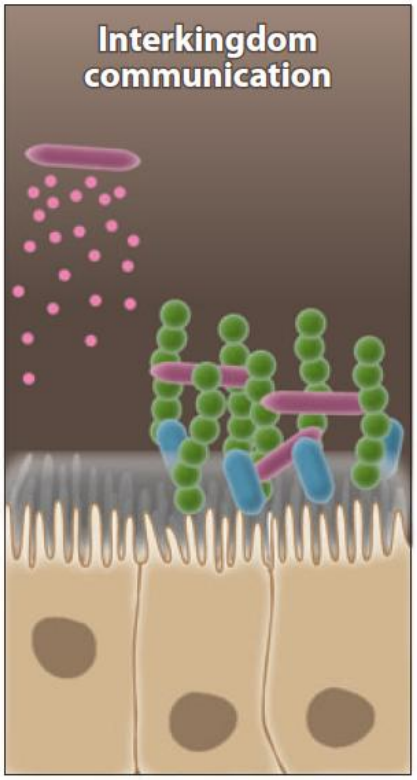
- Actinobacteria
- Firmicutes
- Proteobacteria
- Bacteroidetes

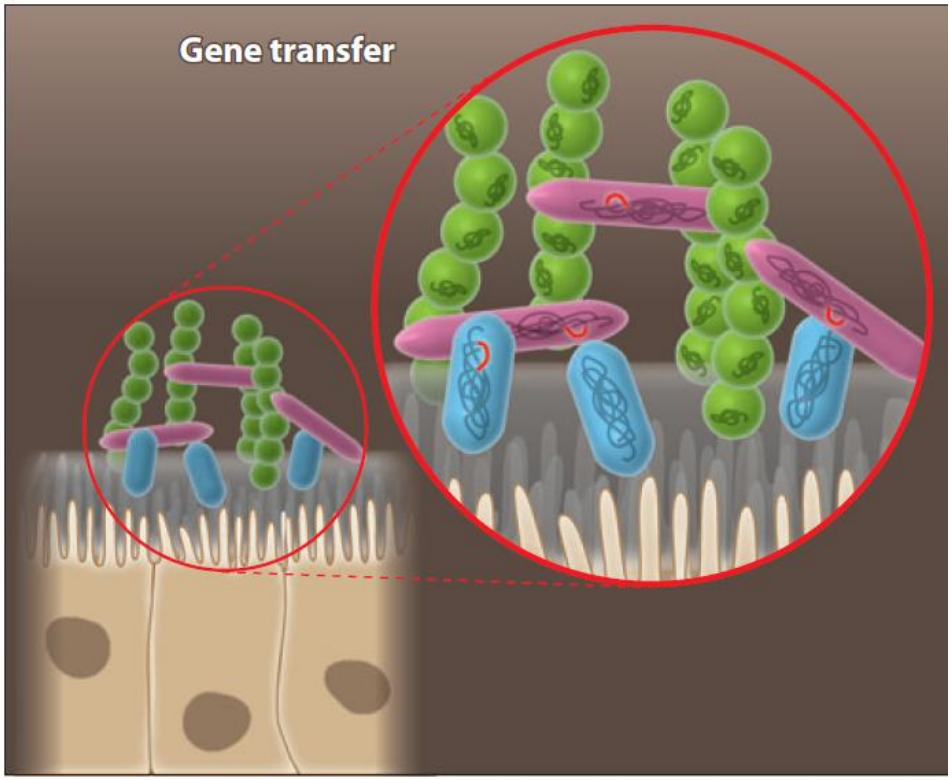
30 common genera, 118 nutrients
11 nutrients and 13 genera.



Victor et al 2013







- Regulation of virulence factors
- Supplementation of nutrients
- Modulation of community composition by bacteria–derived antimicrobials

- Modulation of immune response
- Regulation of transcription factors
- Supplementation of nutrients

- Transfer of antibiotic resistance–encoding genes
- Transfer of virulence factor–encoding genes
- Transfer of metabolism–related genes

Kathryan etal 2012

Human Oral Microbiome

- Falls 2nd to Gut Microbiome (but nearly as equal)
- 700 species of bacteria
- 68% yet-culturable (76% in gut)

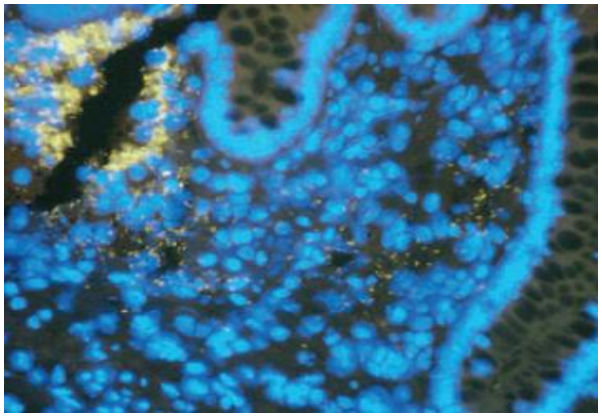
Periodontal infections

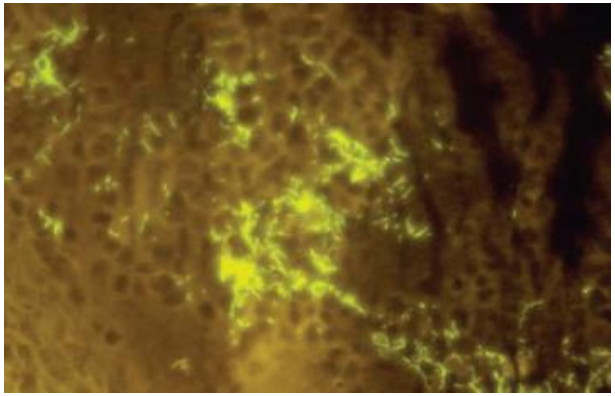
- 30 genera represented in disease conditions
- Red complex bacteria.
  - *Porphyromonas gingivalis,*
  - *Tannerella forsythia*
  - *Treponema denticola*

- Species present in low quantities orchestrate inflammatory periodontal diseases through commensal microbiota.
- *Megasphaera parvimonas, Desulfobulbus* and *Filifactor alocis*.

Microbial dysbiosis

- Inflammatory bowel syndrome
- Ulcerative colitis
- Chron's disease
  - *Faecalibacterium prausnitzii*
- *Appendicitis*
  - *Fusobacterium necrophorum*
  - *Fusobacterium nuclaetum*





Intestinal Microbiome

- Type -2 diabetes
  - Firmicutes & Clostridia reduced
- Obesity
- Helicobacter pylori
  - Removal causes reduced gastric ulcers but increase esophageal cancer

Intestinal Microbiome

Phenotypes
Genotypes
Enterotypes

ARTICLE

doi:10.1038/nature09944

Enterotypes of the human gut microbiome

Manimozhiyan Arumugam[1]*, Jeroen Raes[1,2]*, Eric Pelletier[3,4,5], Denis Le Paslier[3,4,5], Takuji Yamada[1], Daniel R. Mende[1], Gabriel R. Fernandes[1,6], Julien Tap[1,7], Thomas Bruls[3,4,5], Jean-Michel Batto[7], Marcelo Bertalan[8], Natalia Borruel[9], Francesc Casellas[9], Leyden Fernandez[10], Laurent Gautier[8], Torben Hansen[11,12], Masahira Hattori[13], Tetsuya Hayashi[14], Michiel Kleerebezem[15], Ken Kurokawa[16], Marion Leclerc[7], Florence Levenez[7], Chaysavanh Manichanh[9], H. Bjørn Nielsen[8], Trine Nielsen[11], Nicolas Pons[7], Julie Poulain[3], Junjie Qin[17], Thomas Sicheritz-Ponten[8,18], Sebastian Tims[15], David Torrents[10,19], Edgardo Ugarte[3], Erwin G. Zoetendal[15], Jun Wang[17,20], Francisco Guarner[9], Oluf Pedersen[11,21,22,23], Willem M. de Vos[15,24], Søren Brunak[8], Joel Doré[7], MetaHIT Consortium†, Jean Weissenbach[3,4,5], S. Dusko Ehrlich[7] & Peer Bork[1,25]

Our knowledge of species and functional composition of the human gut microbiome is rapidly increasing, but it is still based on very few cohorts and little is known about variation across the world. By combining 22 newly sequenced faecal metagenomes of individuals from four countries with previously published data sets, here we identify three robust clusters (referred to as enterotypes hereafter) that are not nation or continent specific. We also confirmed the enterotypes in two published, larger cohorts, indicating that intestinal microbiota variation is generally stratified, not continuous. This indicates further the existence of a limited number of well-balanced host–microbial symbiotic states that might respond differently to diet and drug intake. The enterotypes are mostly driven by species composition, but abundant molecular functions are not necessarily provided by abundant species, highlighting the importance of a functional analysis to understand microbial communities. Although individual host properties such as body mass index, age, or gender cannot explain the observed enterotypes, data-driven marker genes or functional modules can be identified for each of these host properties. For example, twelve genes significantly correlate with age and three functional modules with the body mass index, hinting at a diagnostic potential of microbial markers.

- Conversion
  - Symbionts to pathobionts
- Excessive bacterial numbers and diversity
  - Contribute to infection
- Reduction in number
  - diminishes resistance to infection
- Fluctuations
  - enhanced predisposition to disease

Gum Disease: Hidden Bacteria Cause Heart Attack and Stroke

Gum disease is caused by bacteria that can enter the bloodstream undetected by the immune system, form blood clots, and cause heart disease and stroke.

Scientific research has already demonstrated a strong connection between gum disease and increased risk of stroke and heart attack. Bacteria from the gums invade the coronary arteries. New research has discovered that bacteria in the gums can enter the bloodstream and make their way to the heart undetected by the body's immune system.

Harvard Health Publications
HARVARD MEDICAL SCHOOL
Trusted advice for a healthier life

Bacteria In Strokes & Heart Disease

Reflections On The 'Cure' Of A Paralyzed Stroke Victim

6-5-9

Heart disease and oral health: role of oral bacteria in heart plaque

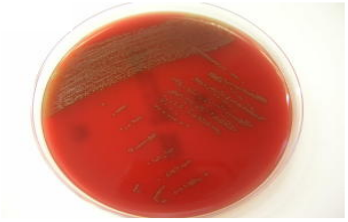
Periodontal Disease and Heart Health

Brushing and flossing may actually save your life.

Science News

Newly Identified Oral Bacterium Linked to Heart Disease and Meningitis

ScienceDaily (Feb. 20, 2012) — A novel bacterium, thought to be a common inhabitant of the oral cavity, has the potential to cause serious disease if it enters the bloodstream, according to a study in the *International Journal of Systematic and Evolutionary Microbiology*. Its identification will allow scientists to work out how it causes disease and evaluate the risk that it poses.

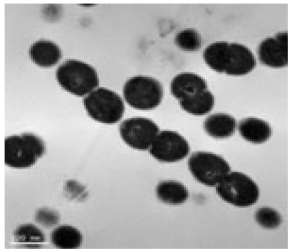


The bacterium was identified by researchers at the Institute of Medical Microbiology of the University of Zurich and has been

Scientists Decode Genome Of Oral Pathogen

Virginia Commonwealth University researchers have decoded the genome of a bacteria normally present in the healthy human mouth that can cause a deadly heart infection if it enters the bloodstream.

The finding enables scientists to better understand the organism, Streptococcus sanguinis, and develop new strategies for treatment and infection prevention.[**Right:** Transmission electron micrograph of S. sanguinis.]



S. sanguinis, a type of bacteria that is naturally present in the mouth, is among a variety of microorganisms responsible for the formation of dental plaque. In general, S. sanguinis is harmless. However, if it enters the bloodstream, possibly through a minor cut or wound in the mouth, it can cause bacterial endocarditis, a serious and often lethal infection of the heart.

Individuals with preexisting heart problems are at an increased risk of developing bacterial endocarditis. The infection may result in impaired heart function and complications such as heart attack and stroke. Typically, before dental surgery, such patients are given high dose antibiotics to prevent infection.

Science News

... from universities, journals, and other research organizations

The More Oral Bacteria, The Higher The Risk Of Heart Attack, Study Shows

ScienceDaily (Apr. 2, 2009) — Several studies have suggested there is a connection between organisms that cause gum disease, known scientifically as periodontal disease, and the development of heart disease, but few studies have tested this theory.

Bacteria's Role in Heart Disease Discovered

January 02 2008

ORAL BACTERIA & YOUR HEART

PREVENT ORAL BACTERIA FROM THREATENING YOUR HEART

CALABASAS, CA (Press Release) – Jun 24, 2010 – Some people have a higher risk for heart attacks because of a combination of factors; their genetic make-up, their weight, their diet and their level of exercise.

But the owner of a leading dental organization located in the heart of this affluent community is working hard to alert the public about another heart disease-causing factor to watch out for;

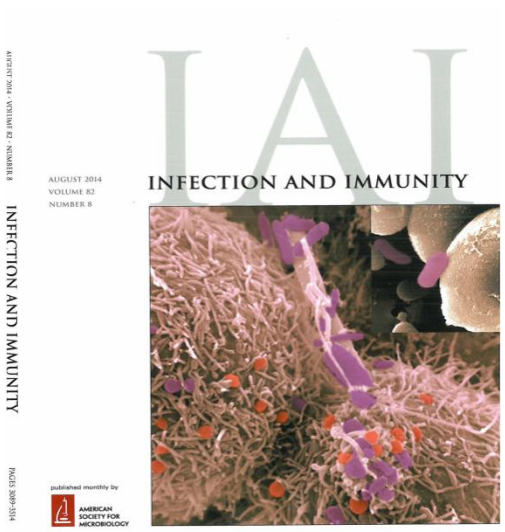
Oral Bacteria that causes gum disease can also damage the heart.

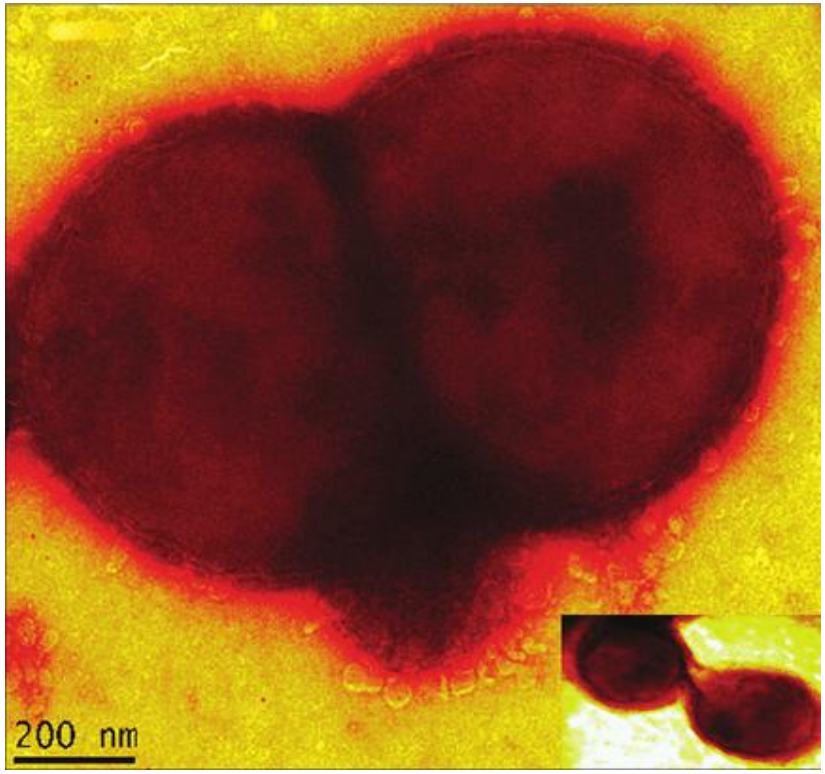
Journal of Cardiovascular Disease Research
* One World, One Mission, Your Outstanding Research Work *

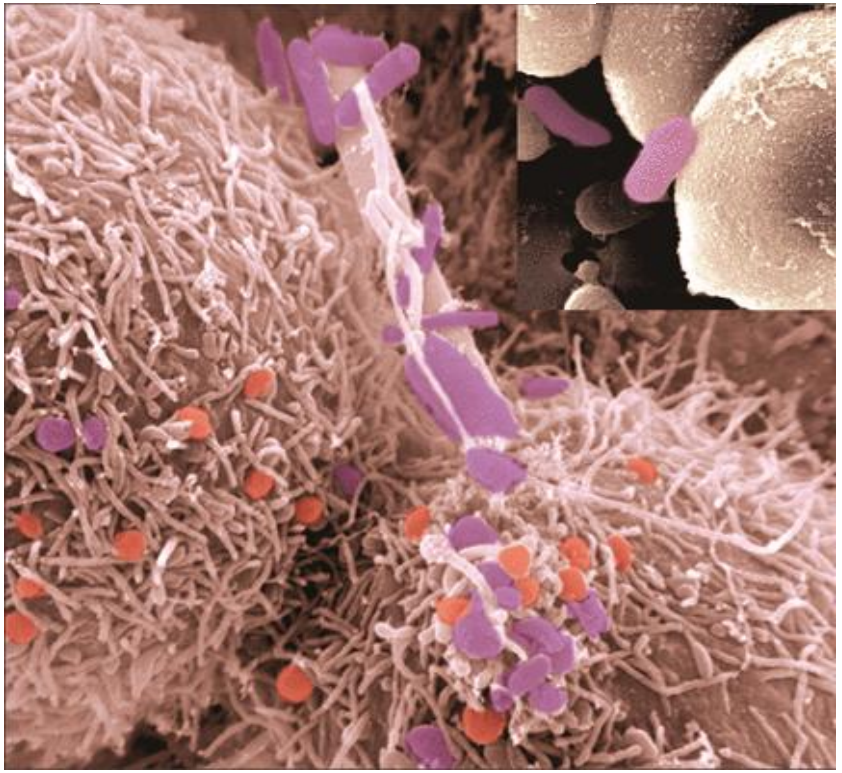
Year : 2010 | **Volume** : 1 | **Issue** : 3 | **Page** : 161

Heart failure and oral bacteria: How could be prevented?

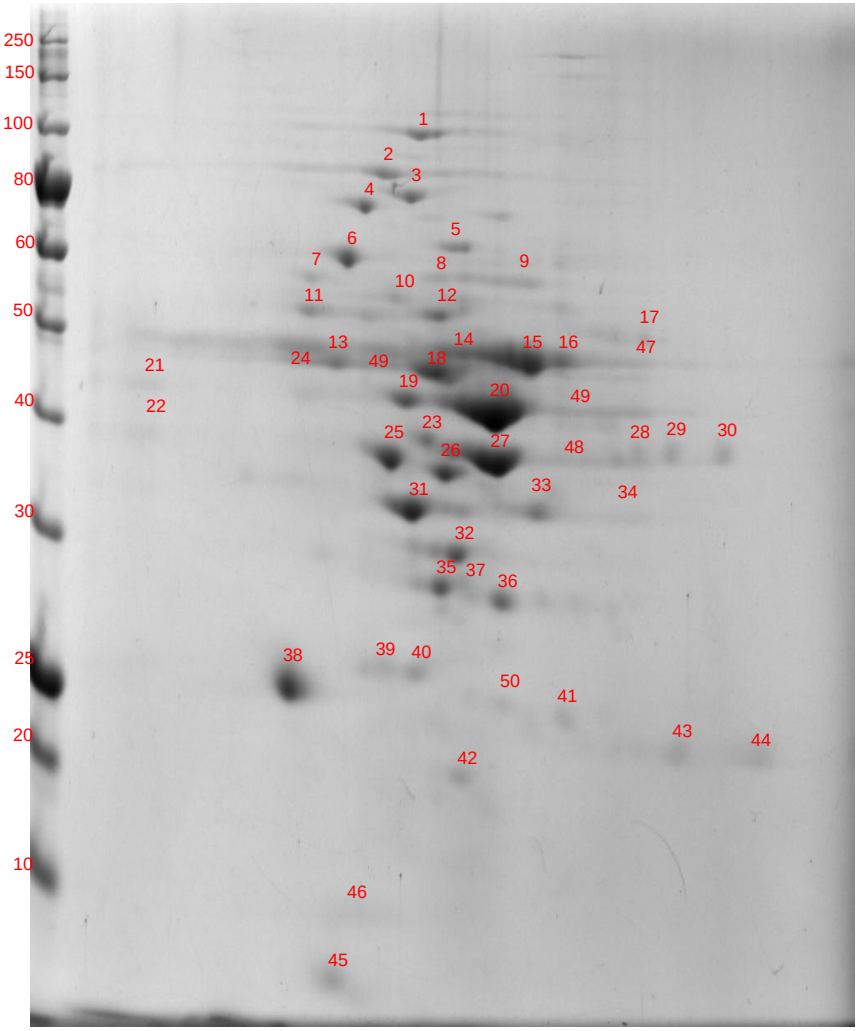

IAI
INFECTION AND IMMUNITY

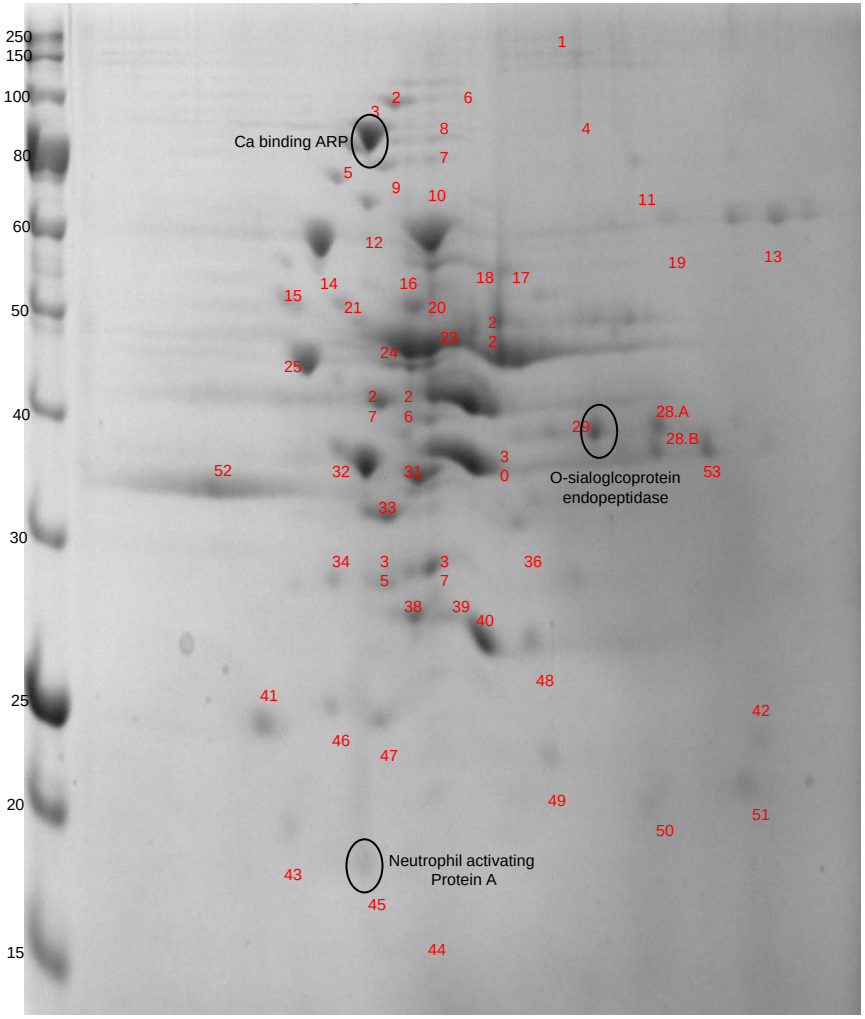

IAI
INFECTION AND IMMUNITY
AUGUST 2014
VOLUME 82
NUMBER 8
published monthly by
AMERICAN SOCIETY FOR MICROBIOLOGY

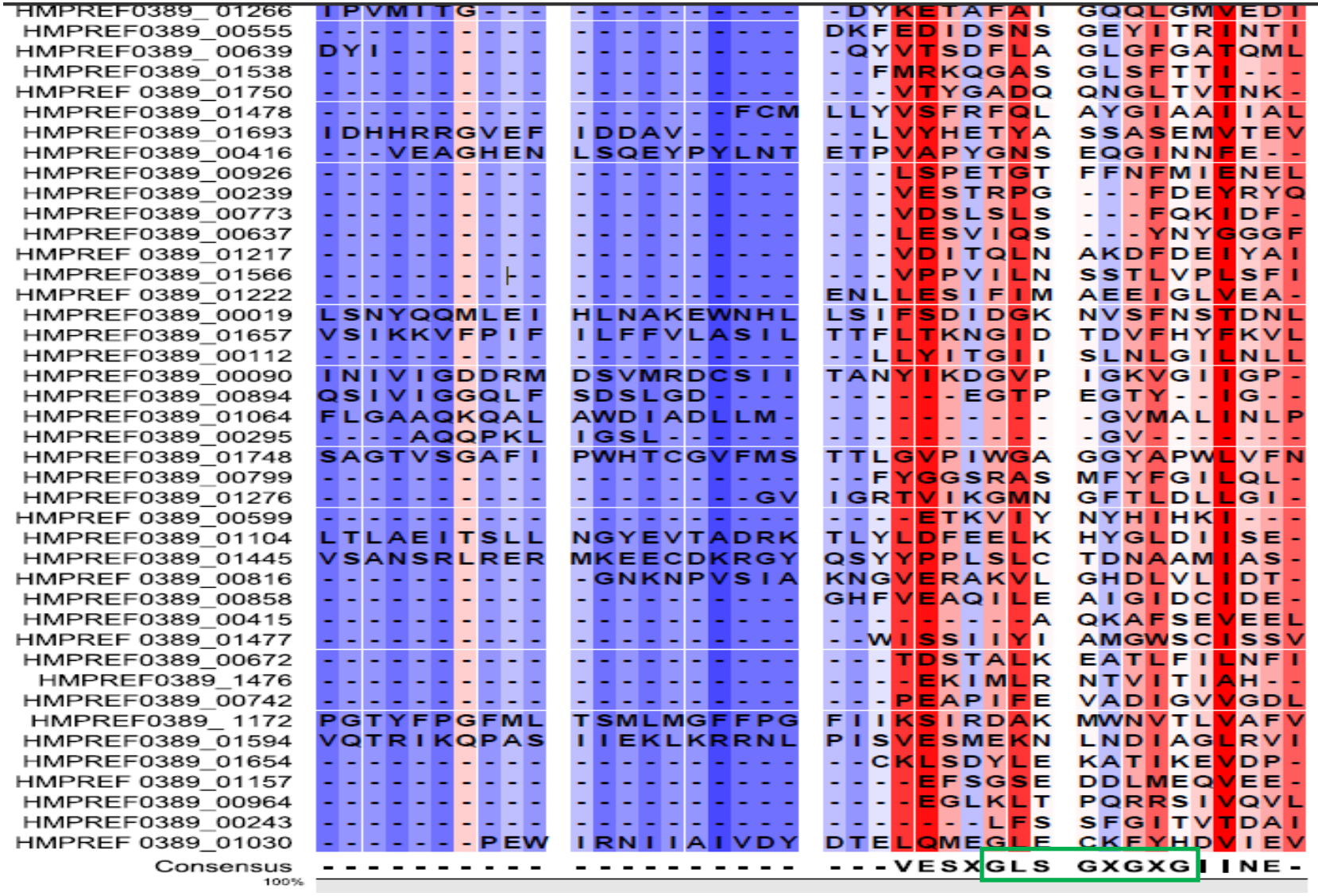

200 nm



2D-PAGE of the membrane fraction of *F. alocis* –ATCC-35896 & D62D strains.

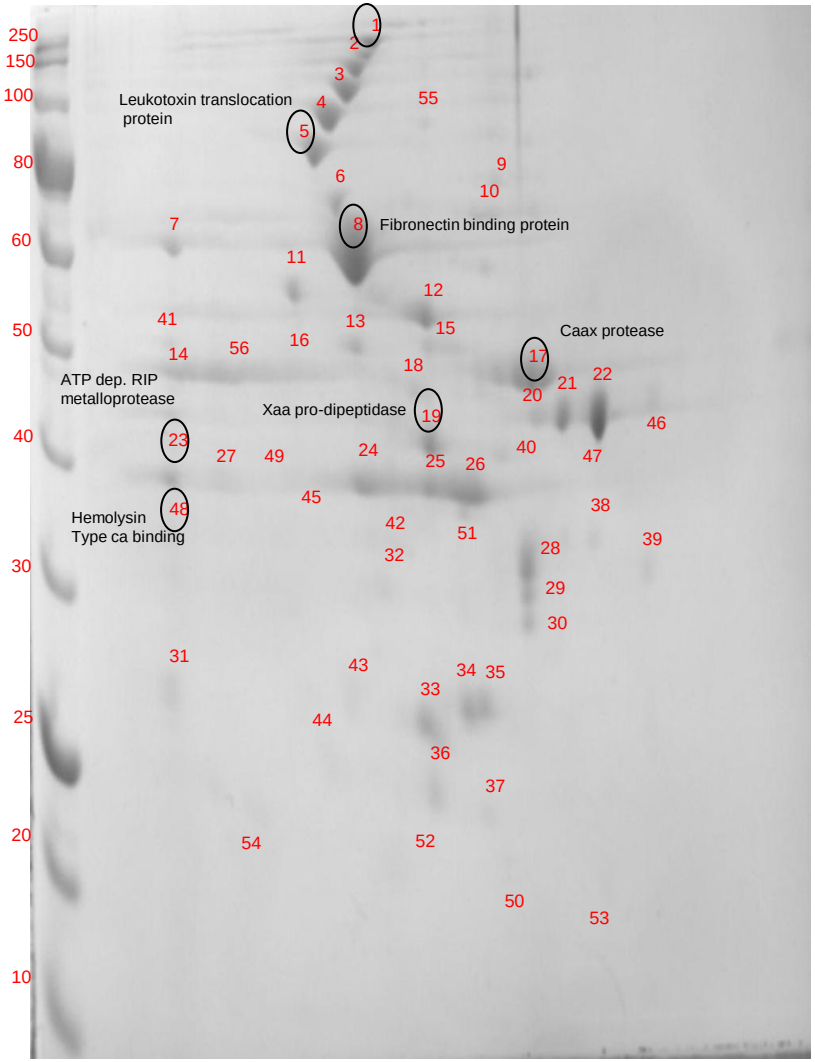


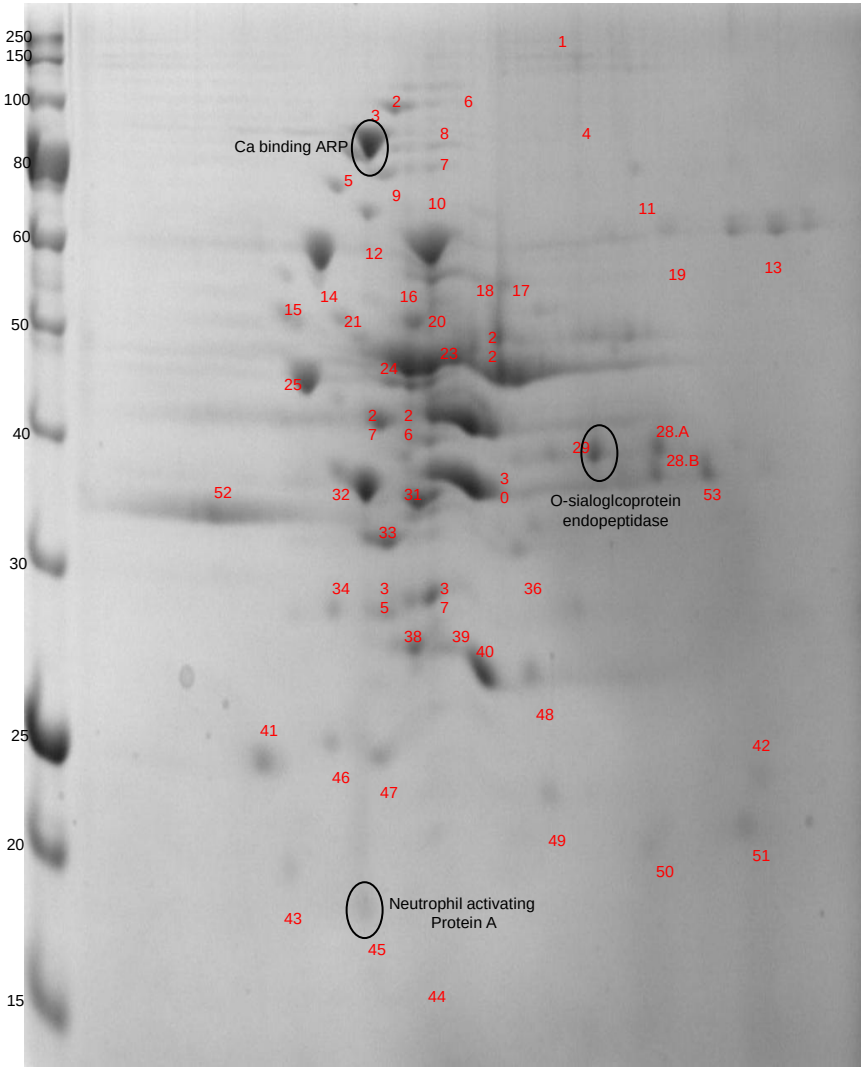




The membrane proteins of *F. alocis* clinical strains contain a C terminal Glycine rich domain with a concensus GxxGxGxGx motif

2D-PAGE of the cytosolic fraction of *F. alocis* –ATCC-35896 & D62D strains.





SWISS-2DPAGE

Two-dimensional polyacrylamide gel electrophoresis database

SWISS-2DPAGE

Search by

[accession number]
[description, ID or gene]
[author names]
[spot ID / serial number]
[identification methods]
[pI / Mw range]
[combined fields]

Maps

[experimental info]
[protein list]
[graphical interface]

Select Remote Interfaces

[All Interfaces]
World-2DPAGE Portal
World-2DPAGE Repository

Exclude local DBs
has only effect if a remote interface is selected

SWISS-2DPAGE contains data on proteins identified on various 2-D PAGE and SDS-PAGE reference maps. You can locate these proteins on the 2-D PAGE maps or display the region of a 2-D PAGE map where one might expect to find a protein from UniProtKB/Swiss-Prot [More details / References / Linking to SWISS-2DPAGE / Commercial users].

Release 19.00, 23rd of May 2011, and updates up to the 9th of November 2011 (containing 1265 entries in 36 reference maps from human, mouse, *Arabidopsis thaliana*, *Dictyostelium discoideum*, *Escherichia coli*, *Saccharomyces cerevisiae*, and *Staphylococcus aureus (N315)*).

Access to SWISS-2DPAGE

- [How to use this interface]
- by description (any word in the ID, DE, GN and KW lines)
- by accession number (AC lines)
- by clicking on a spot: select one of our 2-D PAGE or SDS-PAGE reference maps, click on a spot and then get the corresponding information from the SWISS-2DPAGE database.
- by author (RA lines)
- by spot serial number (2D and 1D lines)
- by experimental pI/Mw range
- by experimental identification methods
- by full text search
- retrieve all the protein entries identified on a given reference map
- user defined / complex queries (SRS like)

SWISS-2DPAGE documents

- [Facts and statistics]
- User manual
- Release notes (September 26, 2006)
- FAQ (Frequently Asked Questions about SWISS-2DPAGE)
- **Protocols:**
  - Technical information about 2-D PAGE (IPG's, silver staining, protocols, etc)
- **Figure captions of SWISS-2DPAGE maps available from publications:**
  - Human CSF, ELC, HEPG2, HEPG2SP, LIVER, LYMPHOMA, PLASMA, PLATELET, RBC, U937, CEC, KIDNEY.
  - *Dictyostelium discoideum*, *Escherichia coli*, *Saccharomyces cerevisiae*.

Services

- Downloading SWISS-2DPAGE by FTP
- Proteomics Core Facility at University of Geneva - Get your 2-D Gels performed according to Swiss standards

Software

- ImageMaster / Melanie - Software package for 2-D PAGE analysis
- Make2D-DB package ver. 3.10 new - A package preparing the data and the programs necessary to build a federated 2-DE database on one's own web site.

Gateways to other 2-D PAGE related servers and services

Proteome 2D-PAGE Database - Home

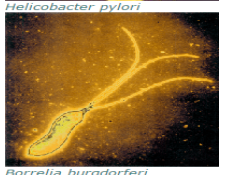
Home - Organisms - Info - Team - References - Search - Mass Peaklists - Gel Gallery - Statistics

Retrieve descriptive information about the proteins identified on 2-D PAGE maps, representing proteomes of the following organisms:

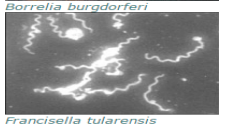
- *Bacillus amyloliquefaciens*
- *Bacillus anthracis*
- *Bartonella henselae*
- *Bifidobacterium longum*
- *Borrelia garinii*
- *Chlamydia pneumoniae*
- *Chlamydophila pneumoniae*
- *Francisella tularensis*
- *Helicobacter pylori*
- *Homo sapiens*
- *Leishmania mexicana*
- *Mus musculus*
- *Mycobacterium bovis*
- *Mycobacterium tuberculosis*
- *Mycoplasma pneumoniae*
- *Paracoccus denitrificans*
- *Rattus norvegicus*
- *Shigella flexneri 2a*

Mycobacteria engulfed by a macrophage

Helicobacter pylori

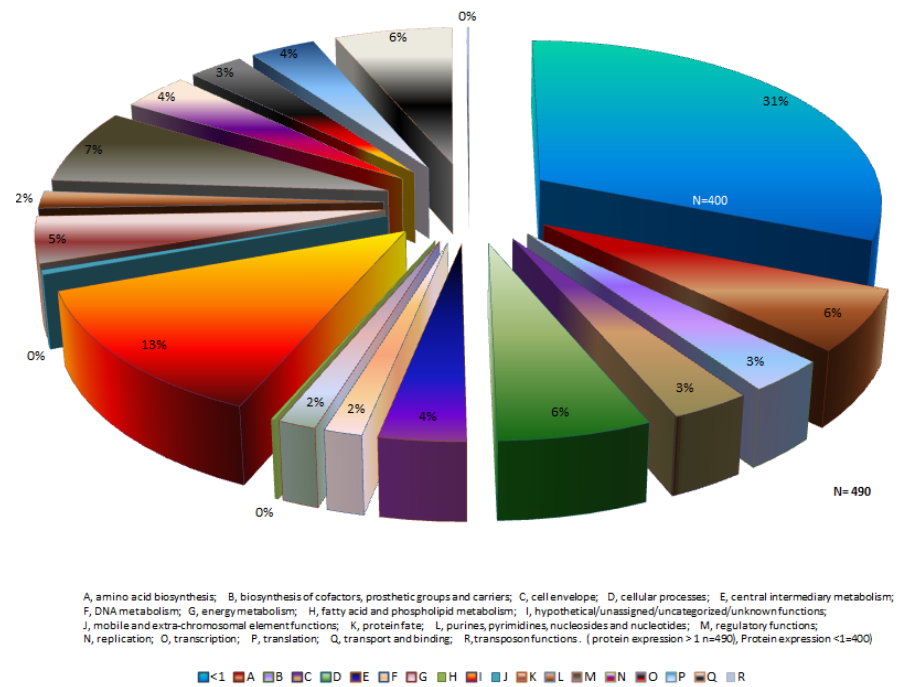


Borrelia burgdorferi



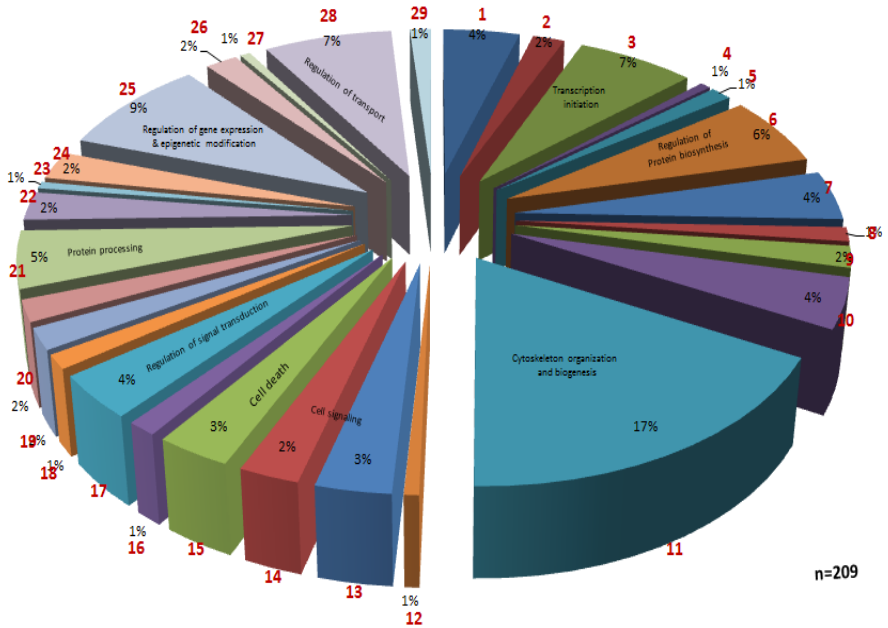
Francisella tularensis

Global Proteome variation



A, amino acid biosynthesis; B, biosynthesis of cofactors, prosthetic groups and carriers; C, cell envelope; D, cellular processes; E, central intermediary metabolism; F, DNA metabolism; G, energy metabolism; H, fatty acid and phospholipid metabolism; I, hypothetical/unassigned/uncategorized/unknown functions; J, mobile and extra-chromosomal element functions; K, protein fate; L, purines, pyrimidines, nucleosides and nucleotides; M, regulatory functions; N, replication; O, transcription; P, translation; Q, transport and binding; R, transposon functions. (protein expression > 1 n=490), Protein expression <1=400)

<1 A B C D E F G H I J K L M N O P Q R



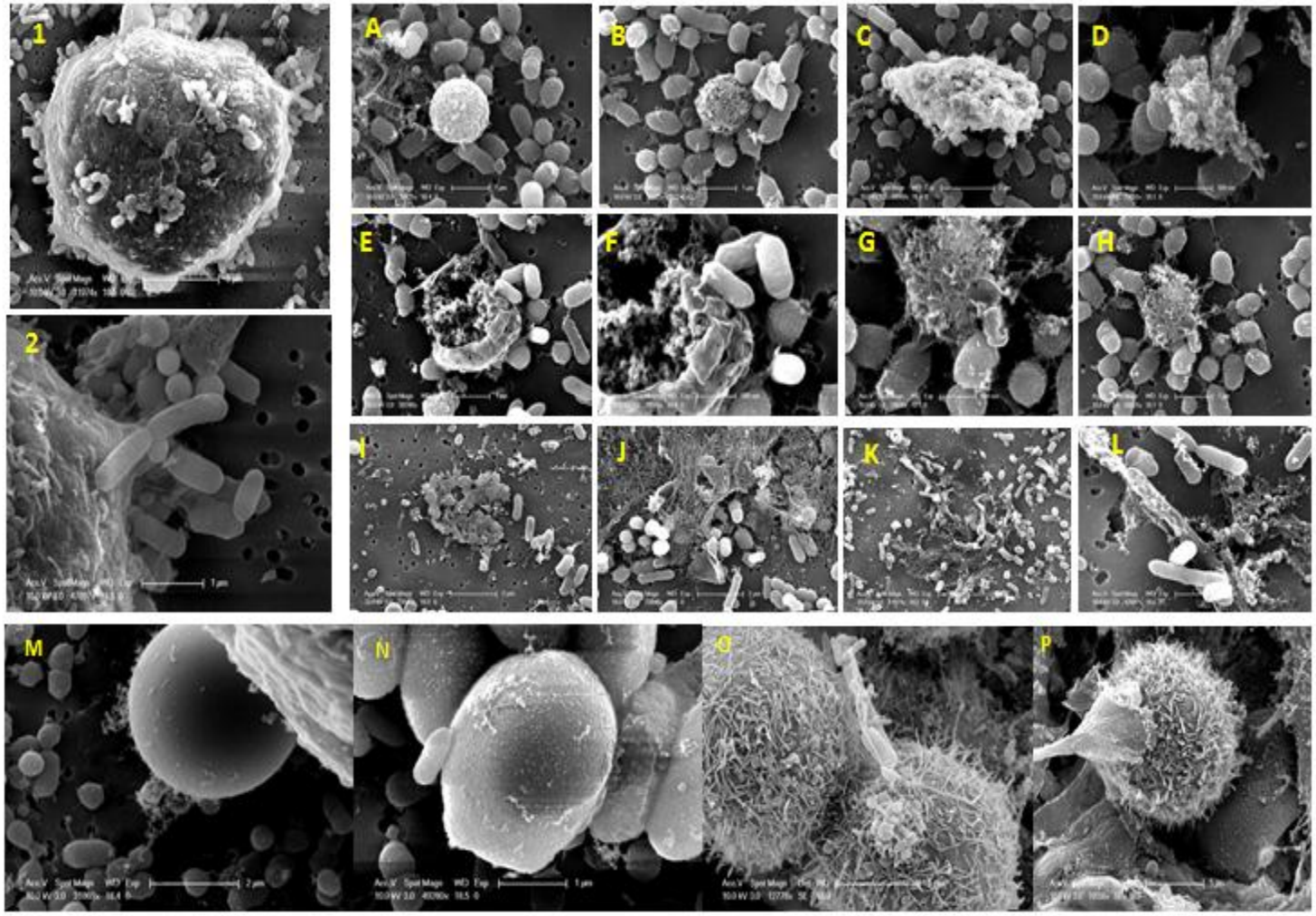
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29

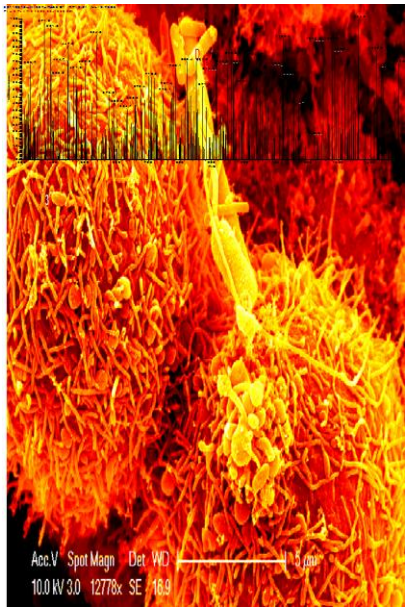
Pathogen

Host

Wilson Aruni et.al, 2014 | Infection and Immunity

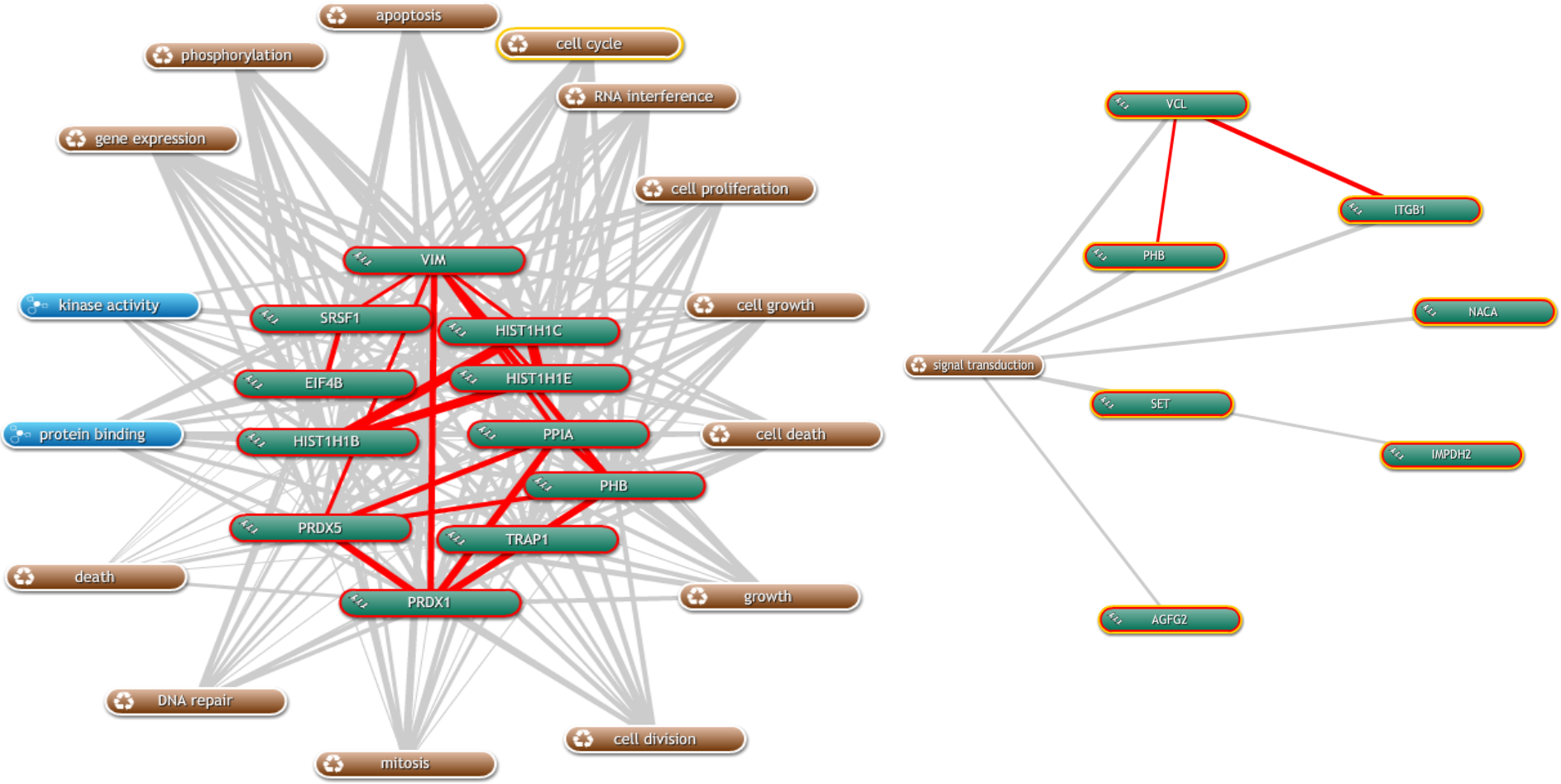
Host cell Apoptosis – co-infection





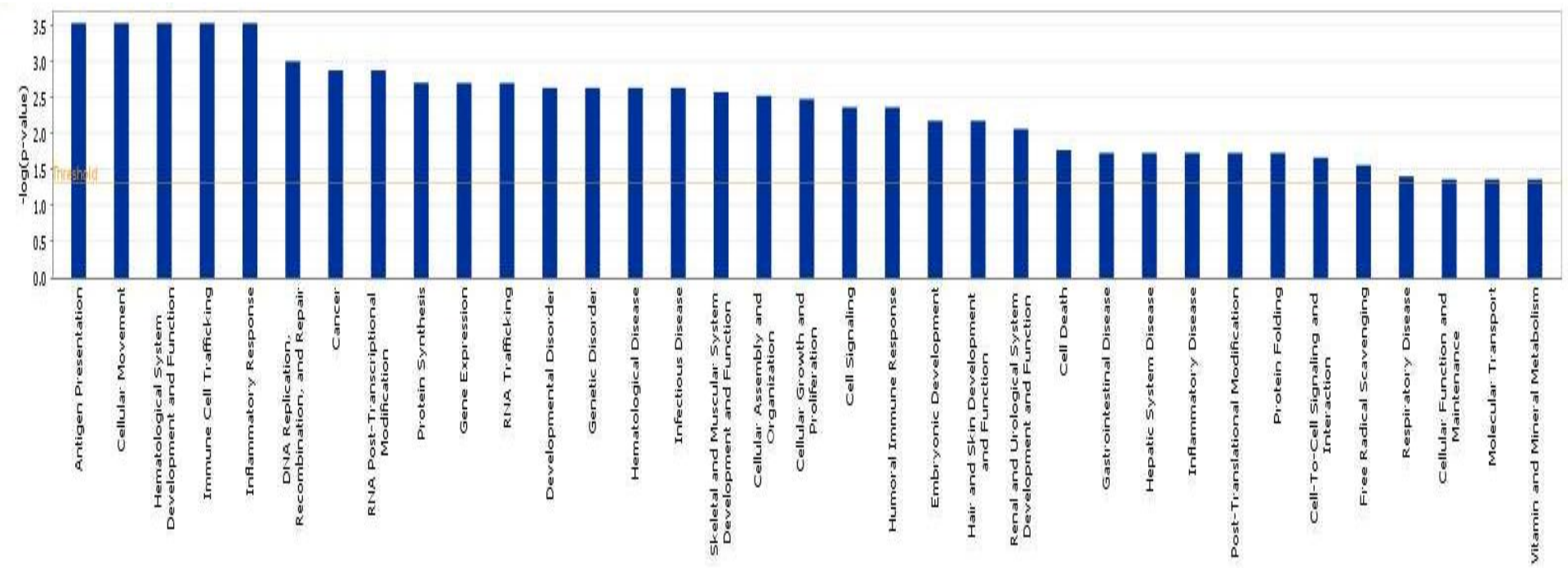
Wilson Aruni et.al, 2014 | Infection and Immunity

Gene network mining of down regulated proteins during co-infection of *Filifactor alocis* with *P. gingivalis*



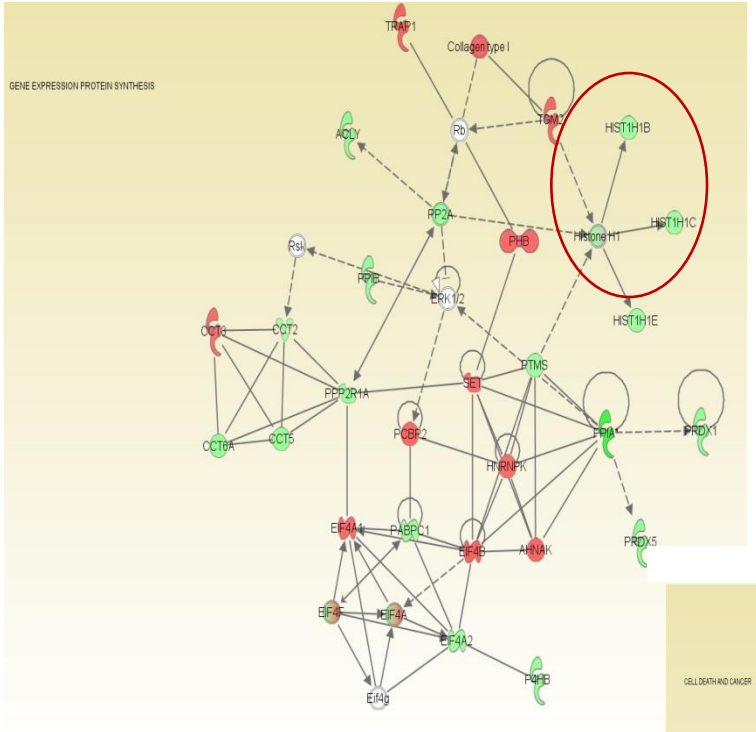
Wilson Aruni et.al, 2014 | Infection and Immunity

Eukaryotic proteins up regulated during *F. alocis* co culture with *P. gingivalis* Vs monoculture



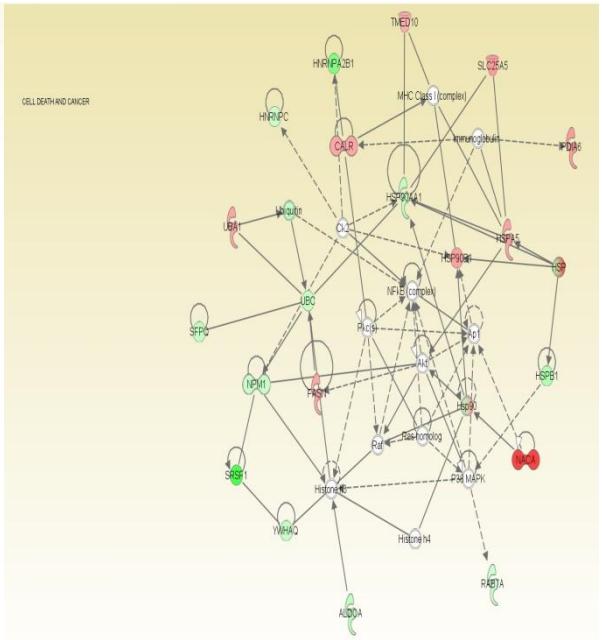
Wilson Aruni et.al, 2014 | Infection and Immunity

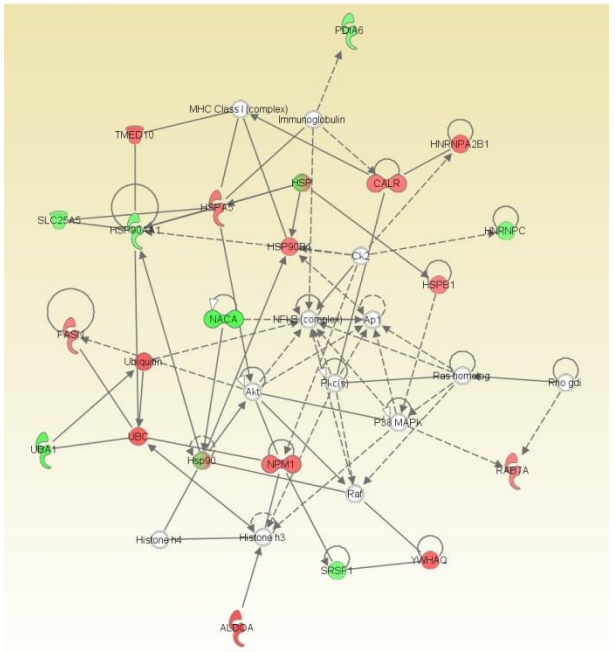
Regulation of host pathways



Gene expression

Cell death and cancer

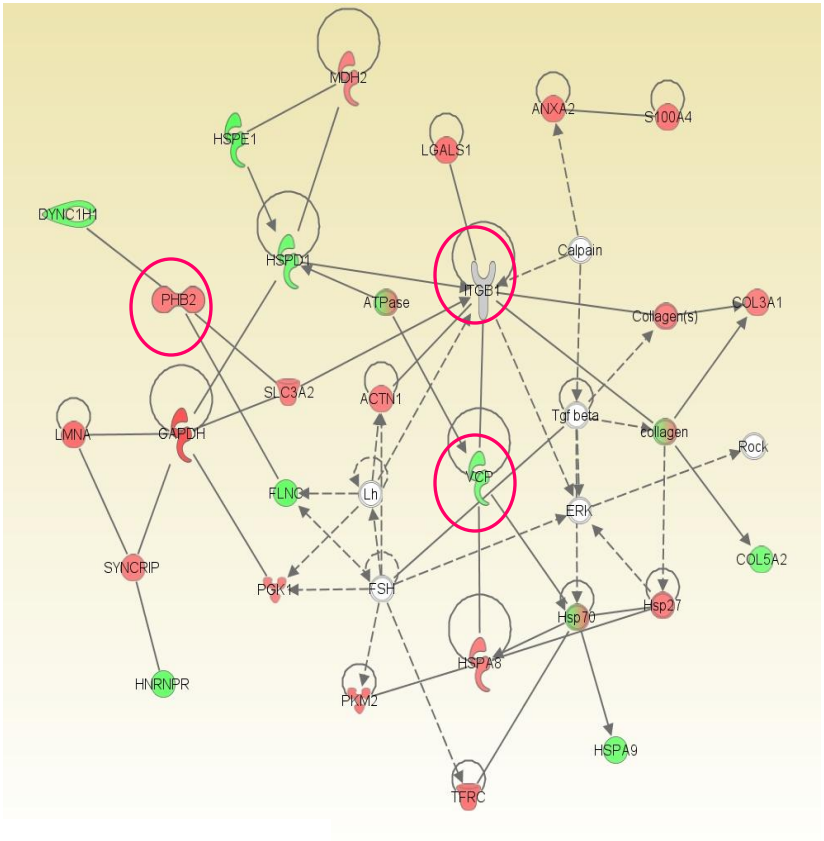




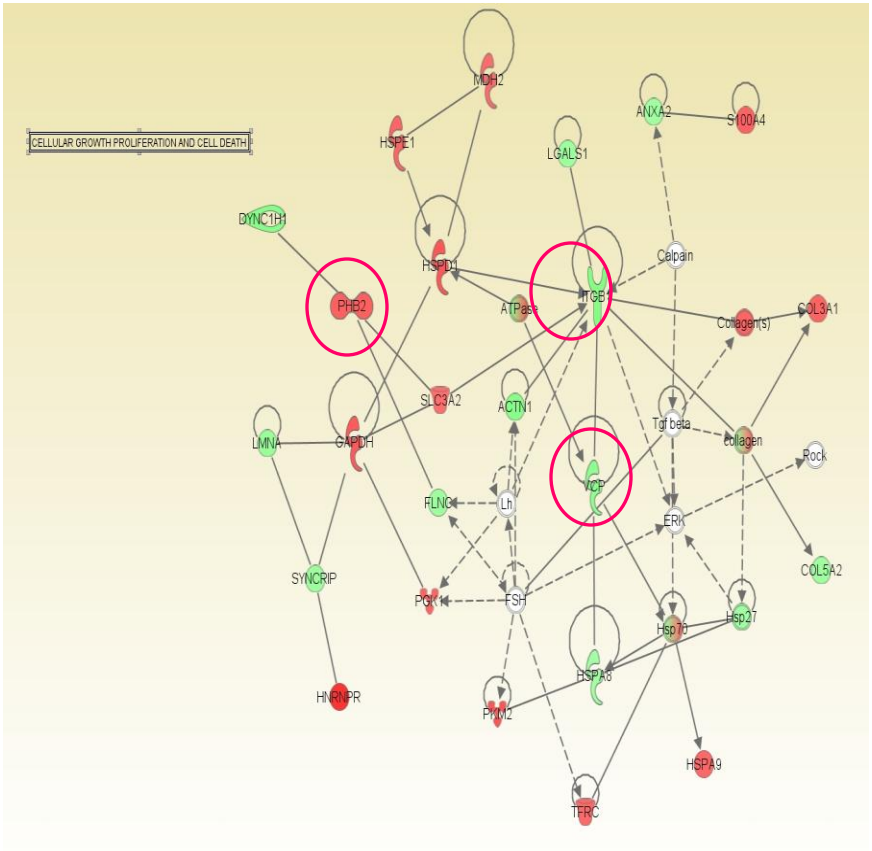
Wilson Aruni et.al, 2014 Infection and Immunity

Figure – S4

Co-infection of *Filifactor alocis* with *P. gingivalis* show upregulation of proteins involved in cell growth and proliferation pathway



Monoculture

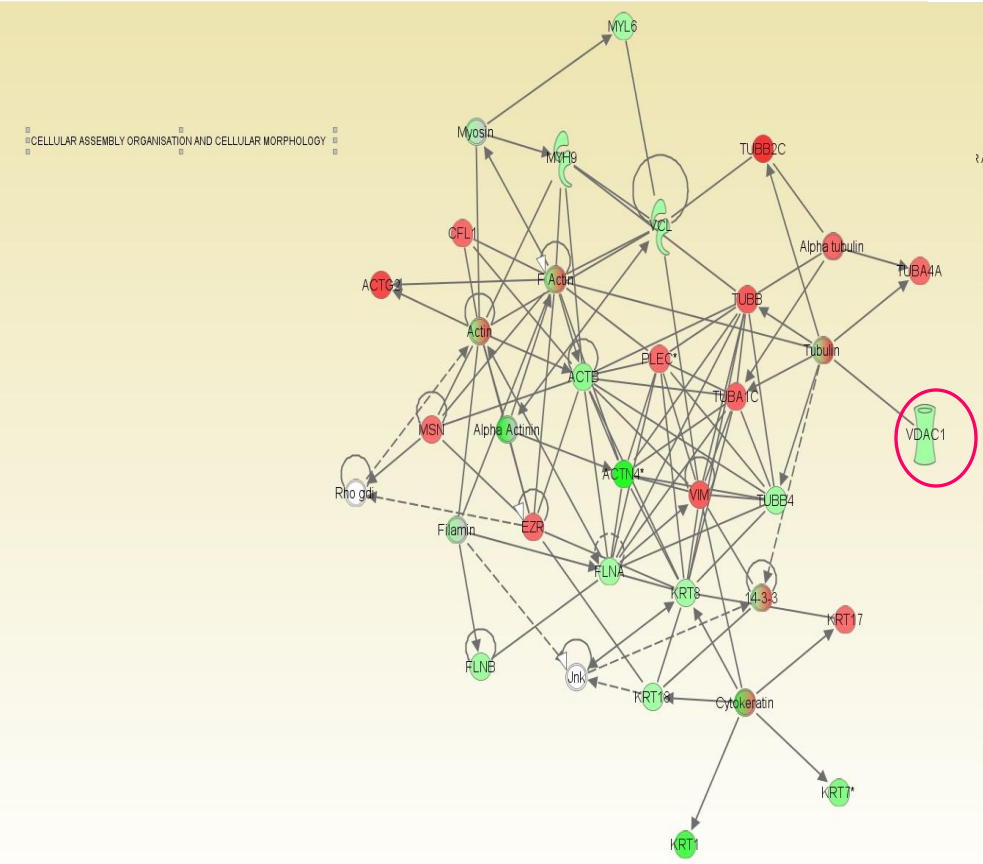


Co-culture

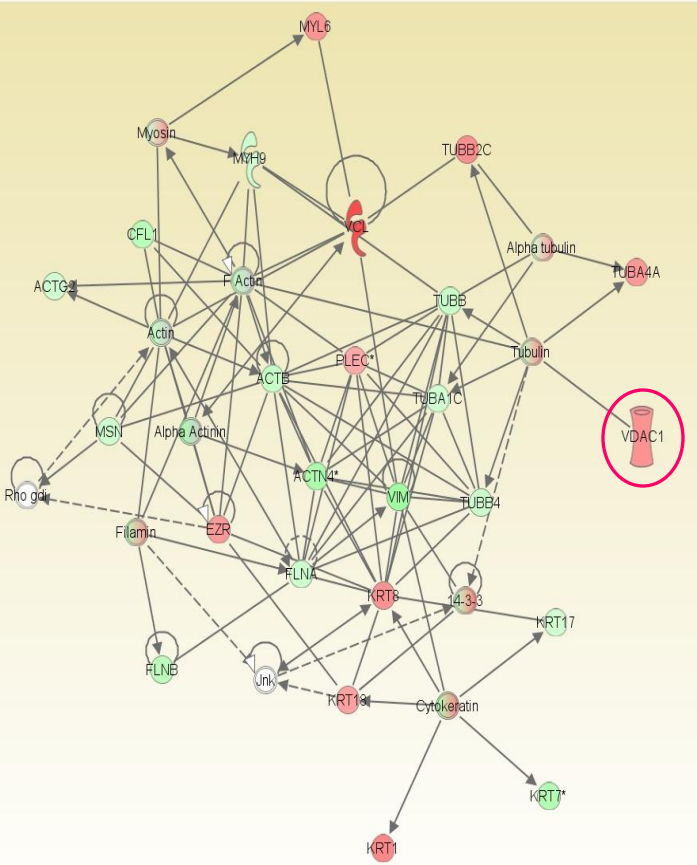
Analysis of the host proteome data showed downregulation of proteins involved in cytoskeleton integrity and ubiquitin dependent protein degradation. Proteins involved in negative regulation of cell proliferation were upregulated.

Wilson Aruni et.al, 2014 Infection and Immunity

Co-infection of *Filifactor alocis* with *P. gingivalis* show modulation of proteins involved in cellular assembly and organization pathways.



Monoculture

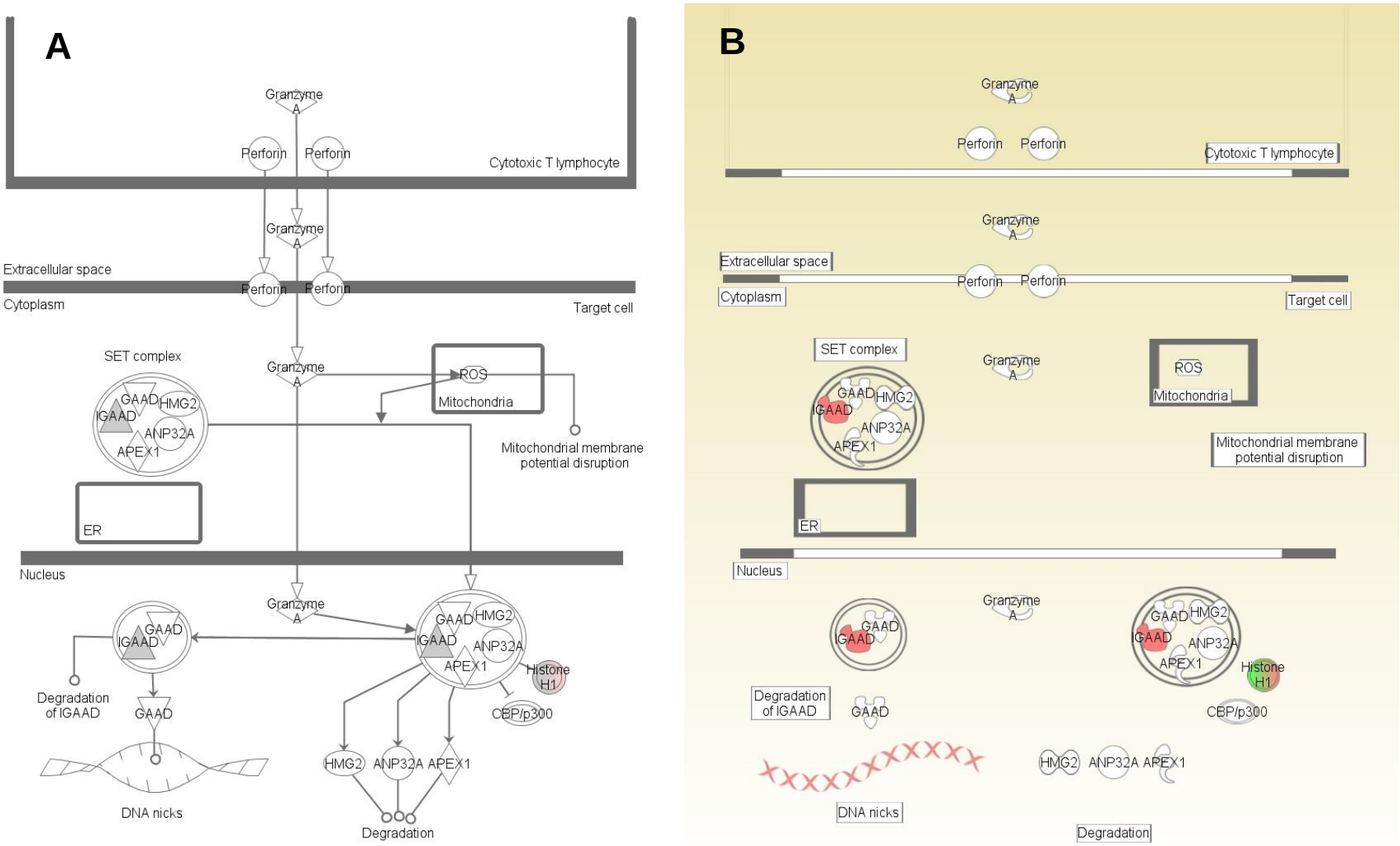


Co-culture

Analysis of the host proteome data showed modulation of many cytoskeleton proteins such as VCL, VDAC1 and downregulation of many proteins involved in actin pathway.

Wilson Aruni et.al, 2014 | Infection and Immunity

Grandzyme mediated apoptic signaling pathway



Coinfection of *F. alocis* with *P. gingivalis* showed upregulation of Grandzyme activating Dnase (GAAD) and downregulation of Histone proteins.

Wilson Aruni et.al, 2014 Infection and Immunity

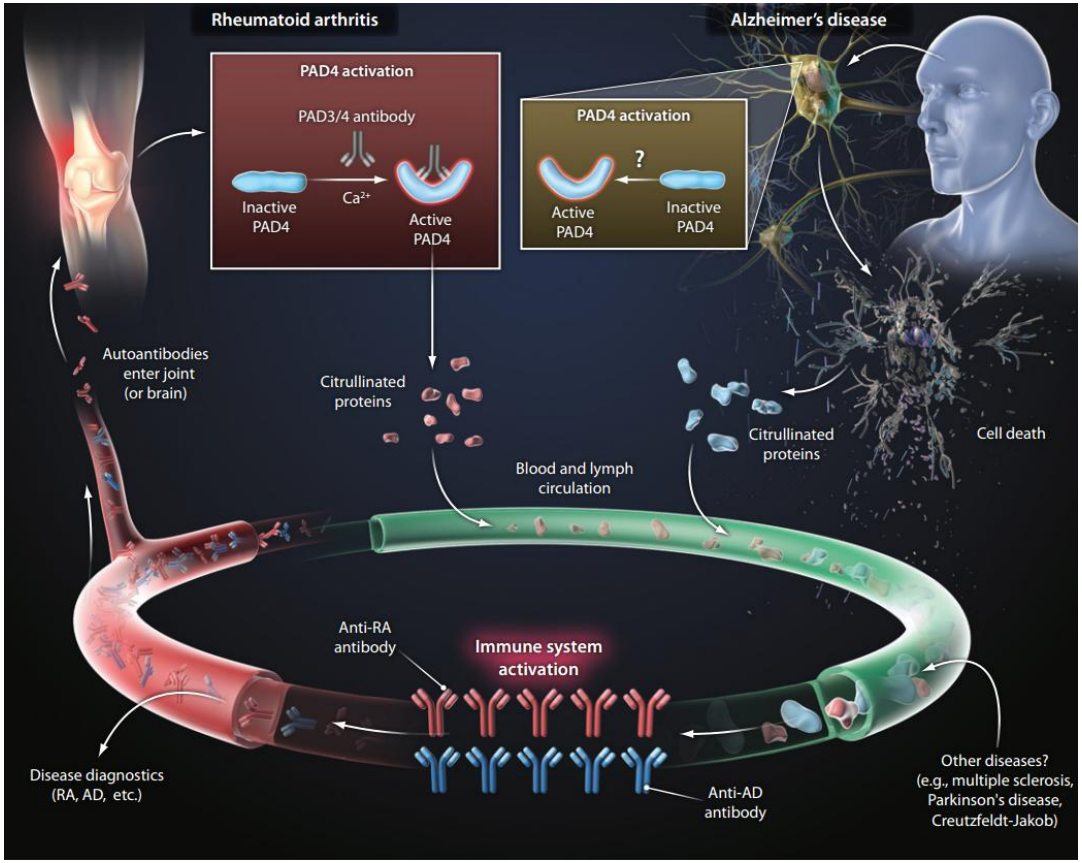
Citrulination

- Arginine deiminase
- Peptidyl arginine deiminase

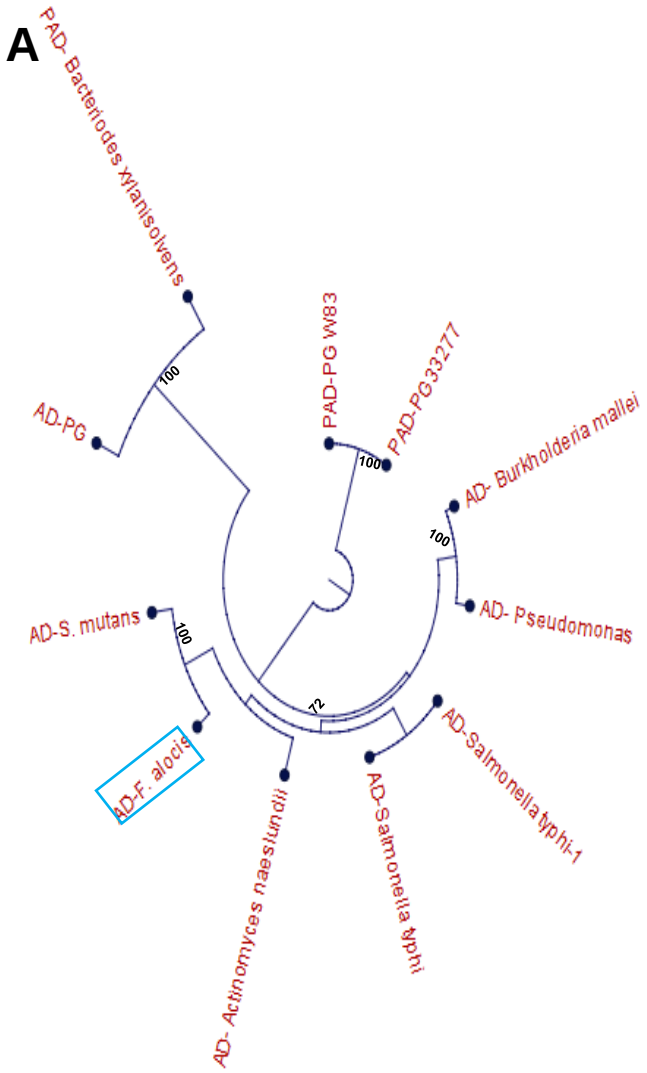
- Peptidyl arginine deiminase

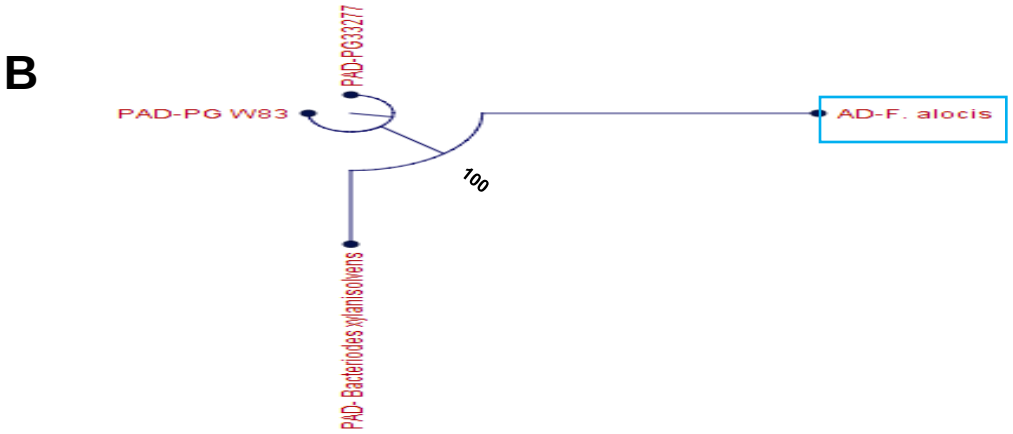
Homologous to humans

- P. gingivalis
- F. alocis



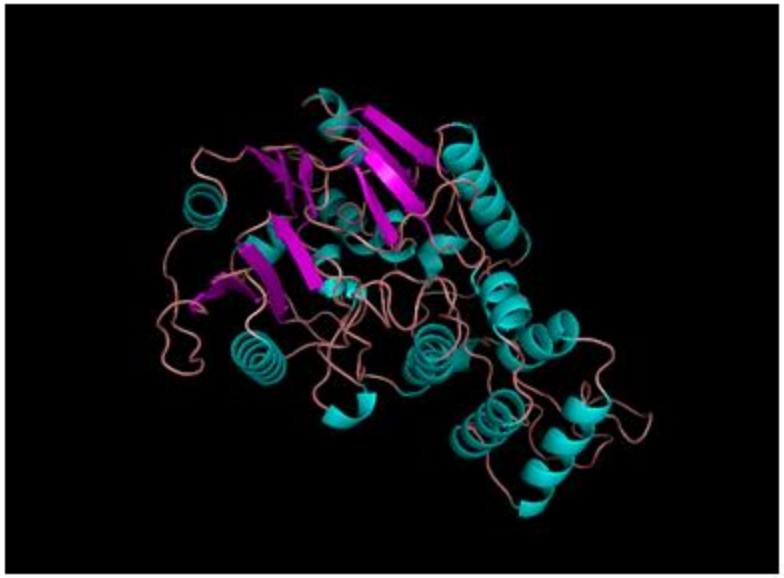
Molecular relatedness of *F. alocis* arginine deiminase



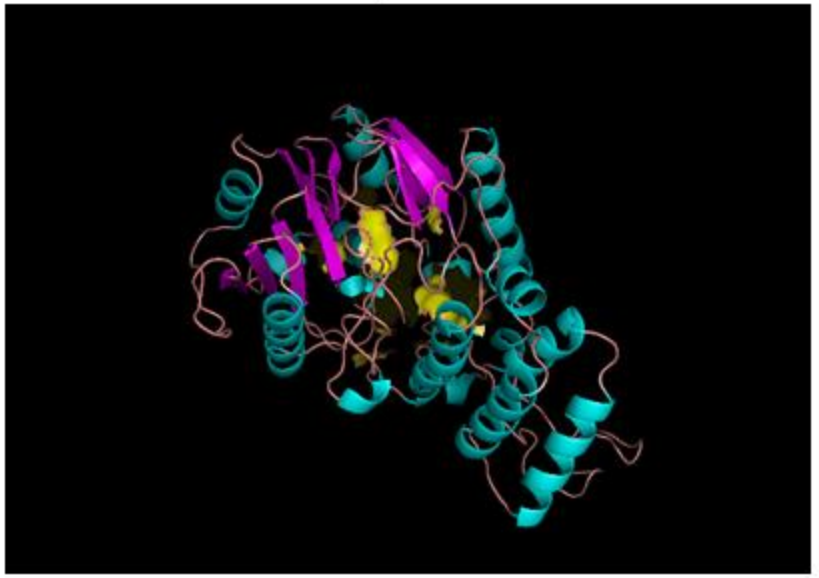


In silico modeling of *F. alocis* arginine deiminase

model



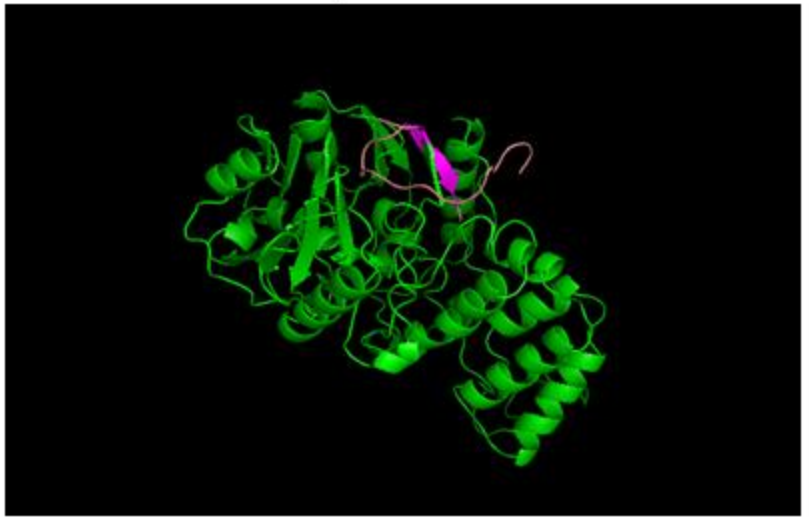
Binding site



158-164,272-276

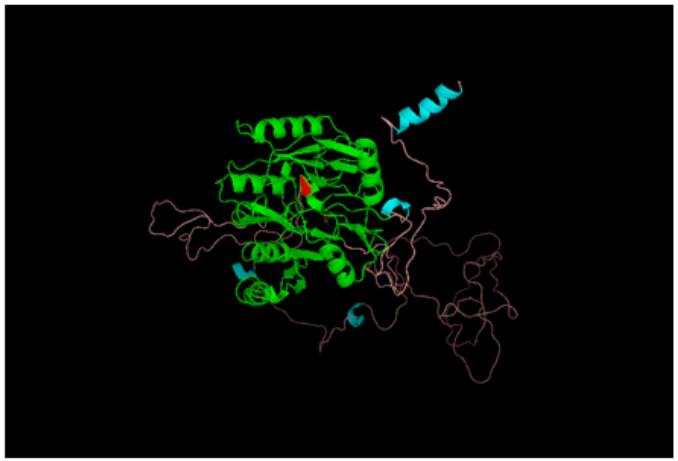
Active site

H-273,C-400



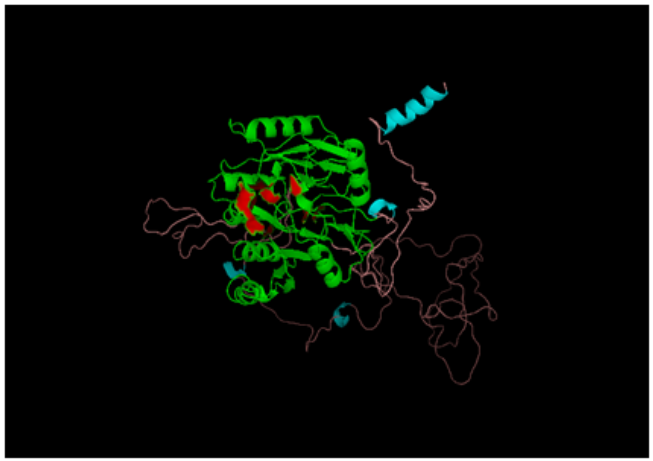
Amidinotransferase domain -green

Active site

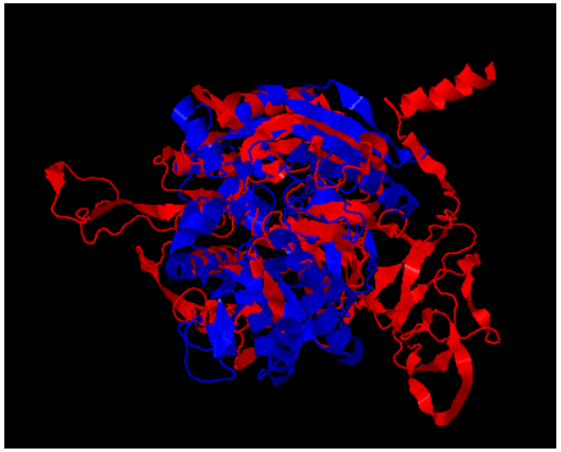


130,236-238

Binding site



127-130,236-238,344-351



Blue –*F. alocis* (AD)(410 amino acids) ,
Red – *P. gingivalis* (PAD) (556 amino acids)

Aligned length= 263, RMSD= 3.89, aligned= 0.125 TM-score= 0.53220
(> 0.5 means the structures share the same SCOP/CATH fold).

```
--RRN------------LQVYSEIGRLRTVLLHRPGREIENVTPDLLERLLFDDIPFLEIAQQEHDAFAQILRDNGVEVLYLEELAAESLTTD
  ...             ..::::::::::::::::                          ::::::::::::::: ::::::::..
IEEMQRAFQETNPPAGPVRAIAEYERSAAVLVRYP----------------------------FGIPMELIKELAKN-DKVITIVAS--------

-LNMVIKMMEGIEKAELPAFEKTTLSDMVDSEYPFVIDPMPNLYFTRDPFATIGS------GISLNHMRTVTRNRETIFAKYIFLYHPRFKDQEIPRWYDRIEKTAIEGGDELILSKDVLAIGISE-RTDA-ASIEKLAKKVFDINESFKKILALHIPNKRAFMHLDTVFTMVDY-DKFTIHP------E-IEGPLTVYS
.. :. ::                          : : ::::::::::::     ::::::::..:: :::::::: ::.:: :  ::: ::  :  ::::::::::::::::::::.: .... :..:::::::... .:::::::::::::..:::::::::::: :::::::     : :::::.
NFI-IA-KT----------------------------D-S-YWTRDYTGWFAMYDTNKVGLVDFIYNRPRPN-DDEFPKYEA-QYLGI-E--MFG-MK--L--KQTGGNYMTDGYGSAVQSHIAYTENSSLSQAQVNQMKKDY-LGITHHDVVQDPNGEYINHVDCWGKYLAPNKILIRKVPDNHPQHQALEDMA---

--E-IEGPLTVYSLTKGENGTIKIEKESKELEQVLAEALGLEKVTLLRCGGGNPIAGAREQWNDGANTLAIAPGEVVVYAR-----NHVTNKLLQEAGI--KLHIMPSSELSR-G-RGGPRCMSMPLWREEI---
  : :::::.                  .:::::::....:.:::::::::      : ::  : :::::::::. :::::::     ::::::::::::  :::::::::::: : ::::::::::::...
PQHQALEDMA------------------AYFAAQTCAWGTKYEVYRALA------T-NE--Q-PYTNSLILNN-RVFVPVNGPASVDNDALNVYKTAMPGYEIIGVKGASGTPWLGTDALHCRTHEVADKGYLYI
```

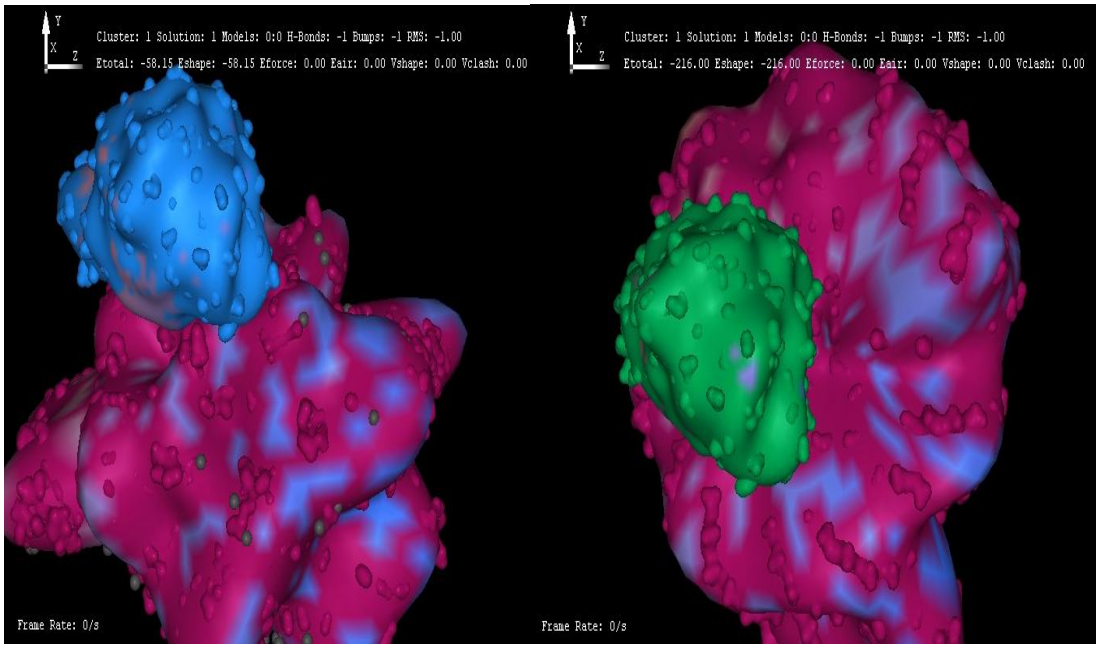
(":" denotes aligned residue pairs of d < 5.0 A, "." denotes other aligned residues)

Comparison of protein Structure between *F. alocis* – Arginine deiminase *P. gingivalis* –Peptidyl arginine deiminase

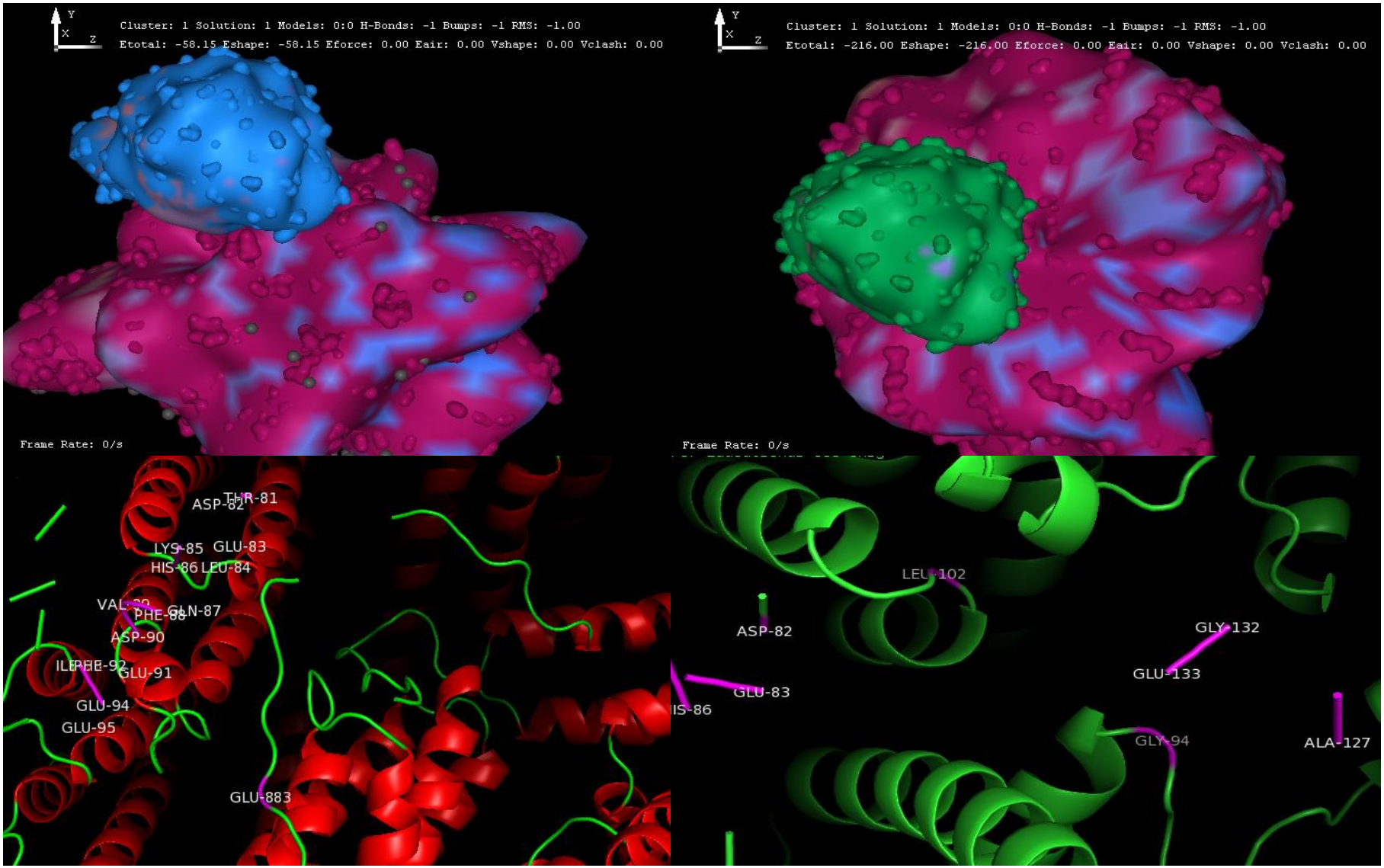
Global proteome analysis

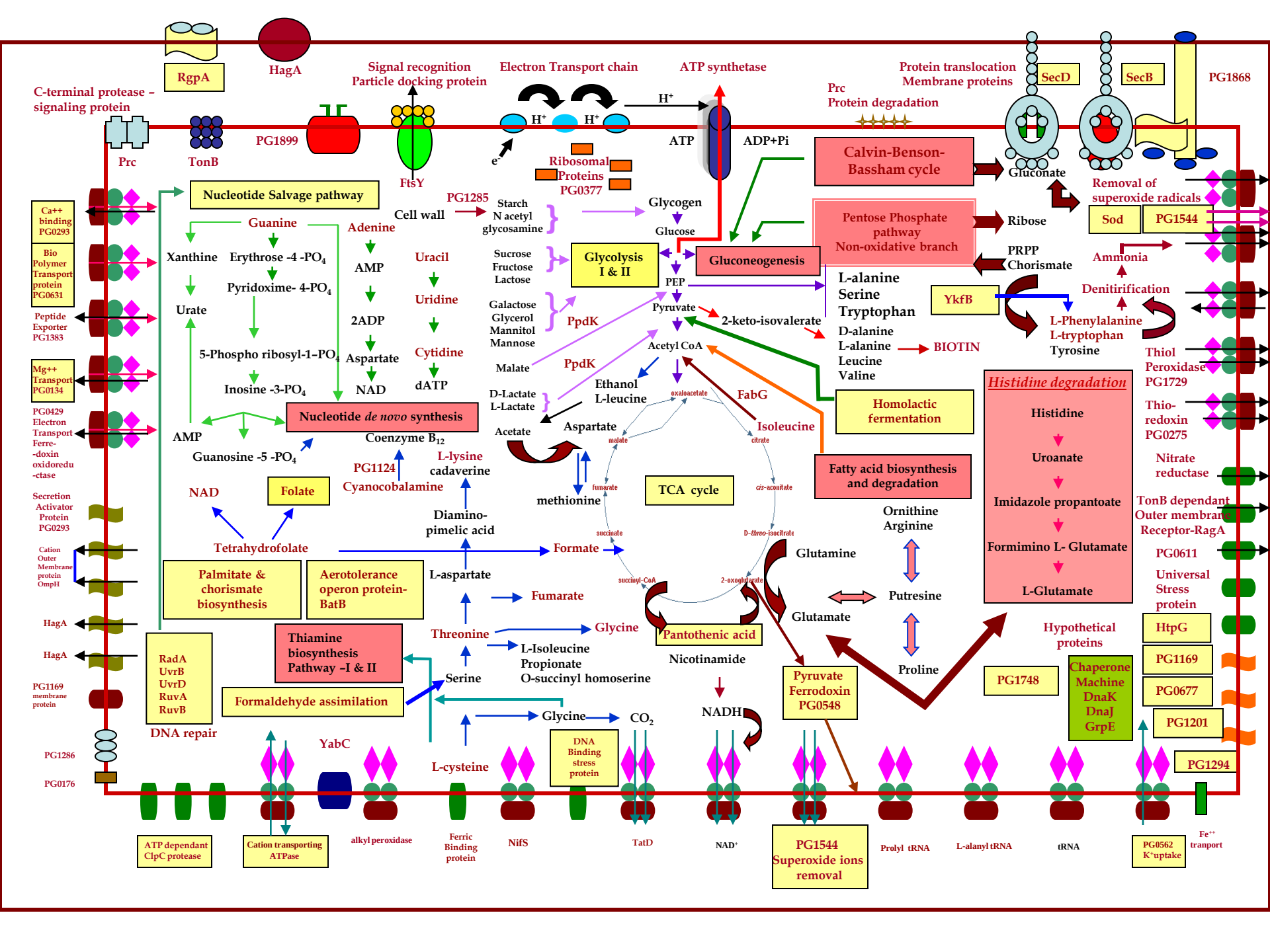
Orbitrap analysis using Isobaric mass tagging technique and confirmation of pathogen protein Interaction using docking

Annotation	Proteins	Fold change
IPI00306959.10	Keratin, type II cytoskeleton 7	1.046
IPI00554648.3	Keratin, type II cytoskeleton 8	1.335
IPI00450768.7	Keratin, type II cytoskeleton 17	0.999
IPI00784347.2	Keratin, type II cytoskeleton 18	1.150
IPI00418471.6	Vimentin	0.463
IPI00291175.7	Isoform 1 of Vinculin	2.189
IPI00007752.1	Tubulin beta – 2C chain	1.383
IPI00218343.4	Tubulin aplha – 6 chain	0.980
IPI00023598.2	Tubulin beta – 4 chain	0.860
IPI00329801.12	Annexin A5	1.226
IPI00455315.4	Annexin A2	1.302
IPI00218918.5	Annexin A1	1.214
IPI00021439.1	Actin, cytoplasmic 1	0.758
IPI00025416.3	Actin, gamma enteric smooth muscle	0.988
IPI00013808.1	Alpha – actin -4	0.966
IPI00013508.5	Alpha – actin - 1	0.875
IPI00465248.5	Isoform alpha – enolase	0.761
IPI00739099.1	Collagen alpha- 2(V) chain precursor	1.218
IPI00644576.1	Filamin A, alpha	0.887
IPI00216694.3	Plastin 3	1.061
IPI00455315.4	Annexin A2	1.302
IPI00398806.2	H2A histone family, membrane V isoform 5	2.509
IPI00453473.6	Histone H4	1.394
IPI00171611.7	Histone H3.2	1.454

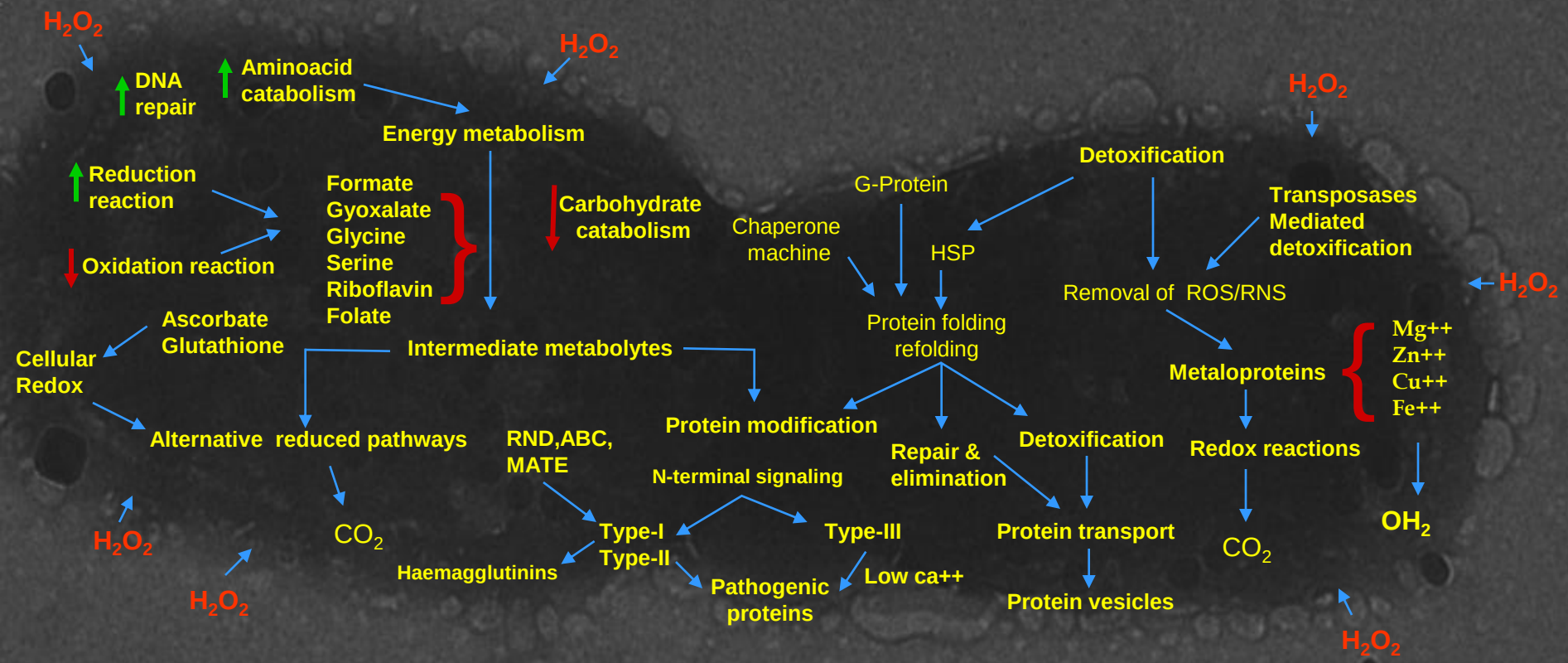


Lingand receptor docking showing similar interacting motifs among the two proteins




RgpA
HagA
C-terminal protease – signaling protein
Signal recognition Particle docking protein
Electron Transport chain
ATP synthetase
Protein translocation Membrane proteins
SecD
SecB
PG1868
Prc
Protein degradation
Prc
TonB
PG1899
FtsY
H+
e-
ATP
ADP+Pi
Ribosomal Proteins PG0377
Calvin-Benson-Bassham cycle
Gluconate
Removal of superoxide radicals
Sod
PG1544
Ca++ binding PG0293
Bio Polymer Transport protein PG0631
Peptide Exporter PG1383
Mg++ Transport PG0134
PG0429 Electron Transport Ferre--doxin oxidoredu-ctase
Secretion Activator Protein PG0293
Cation Outer Membrane protein OmpH
HagA
HagA
PG1169 membrane protein
PG1286
PG0176
Nucleotide Salvage pathway
Guanine
Xanthine
Urate
Erythrose -4 -PO4
Pyridoxime- 4-PO4
5-Phospho ribosyl-1-PO4
Inosine -3-PO4
AMP
Guanosine -5 -PO4
Adenine
AMP
2ADP
Aspartate
NAD
Uracil
Uridine
Cytidine
dATP
Nucleotide de novo synthesis
Coenzyme B12
PG1124
Cyanocobalamine
L-lysine
cadaverine
Diamino-pimelic acid
NAD
Folate
Tetrahydrofolate
Palmitate & chorismate biosynthesis
Aerotolerance operon protein-BatB
RadA
UvrB
UvrD
RuvA
RuvB
DNA repair
Thiamine biosynthesis Pathway –I & II
Formaldehyde assimilation
PG1285
Cell wall
Starch
N acetyl glycosamine
Glycogen
Glucose
Sucrose
Fructose
Lactose
Glycolysis I & II
Galactose
Glycerol
Mannitol
Mannose
PpdK
Malate
PpdK
D-Lactate
L-Lactate
Acetate
Ethanol
L-leucine
Aspartate
methionine
Gluconeogenesis
PEP
Pyruvate
Acetyl CoA
2-keto-isovalerate
FabG
Isoleucine
Pentose Phosphate pathway Non-oxidative branch
Ribose
PRPP
Chorismate
L-alanine
Serine
Tryptophan
D-alanine
L-alanine
Leucine
Valine
BIOTIN
YkfB
Ammonia
Denitirification
L-Phenylalanine
L-tryptophan
Tyrosine
Homolactic fermentation
Fatty acid biosynthesis and degradation
TCA cycle
oxaloacetate
citrate
cis-aconitate
D-threo-isocitrate
2-oxoglutarate
succinyl-CoA
succinate
fumarate
malate
Histidine degradation
Histidine
Uroanate
Imidazole propantoate
Formimino L- Glutamate
L-Glutamate
Ornithine
Arginine
Glutamine
Glutamate
Putresine
Proline
Thiol Peroxidase PG1729
Thio-redoxin PG0275
Nitrate reductase
TonB dependant Outer membrane Receptor-RagA
PG0611
Universal Stress protein
HtpG
PG1169
PG0677
PG1201
PG1294
Hypothetical proteins
PG1748
Chaperone Machine DnaK DnaJ GrpE
L-cysteine
Serine
Threonine
L-aspartate
Fumarate
Glycine
L-Isoleucine
Propionate
O-succinyl homoserine
Glycine
CO2
NADH
NAD+
Pantothenic acid
Nicotinamide
Pyruvate Ferrodoxin PG0548
YabC
DNA Binding stress protein
PG1544 Superoxide ions removal
ATP dependant ClpC protease
Cation transporting ATPase
alkyl peroxidase
Ferric Binding protein
NifS
TatD
Prolyl tRNA
L-alanyl tRNA
tRNA
PG0562 K+uptake
Fe++ tranport

Oxidative stress resistance strategies *in Prphyromonas gingivalis*



Data Integration to study Host –Pathogen interaction

Dry lab - Wet lab concept