COMPARATIVE GENOMICS

-Ms. Rupal Mishra

Introduction

Genomics - Development and application of genetic mapping, sequencing, and computational methods to analyze the genomes of organisms.

Sub-fields of genomics:

1. Structural genomics - genetic and physical mapping of genomes.
2. Functional genomics - analysis of gene function.
3. Comparative genomics - comparison of genomes across species.

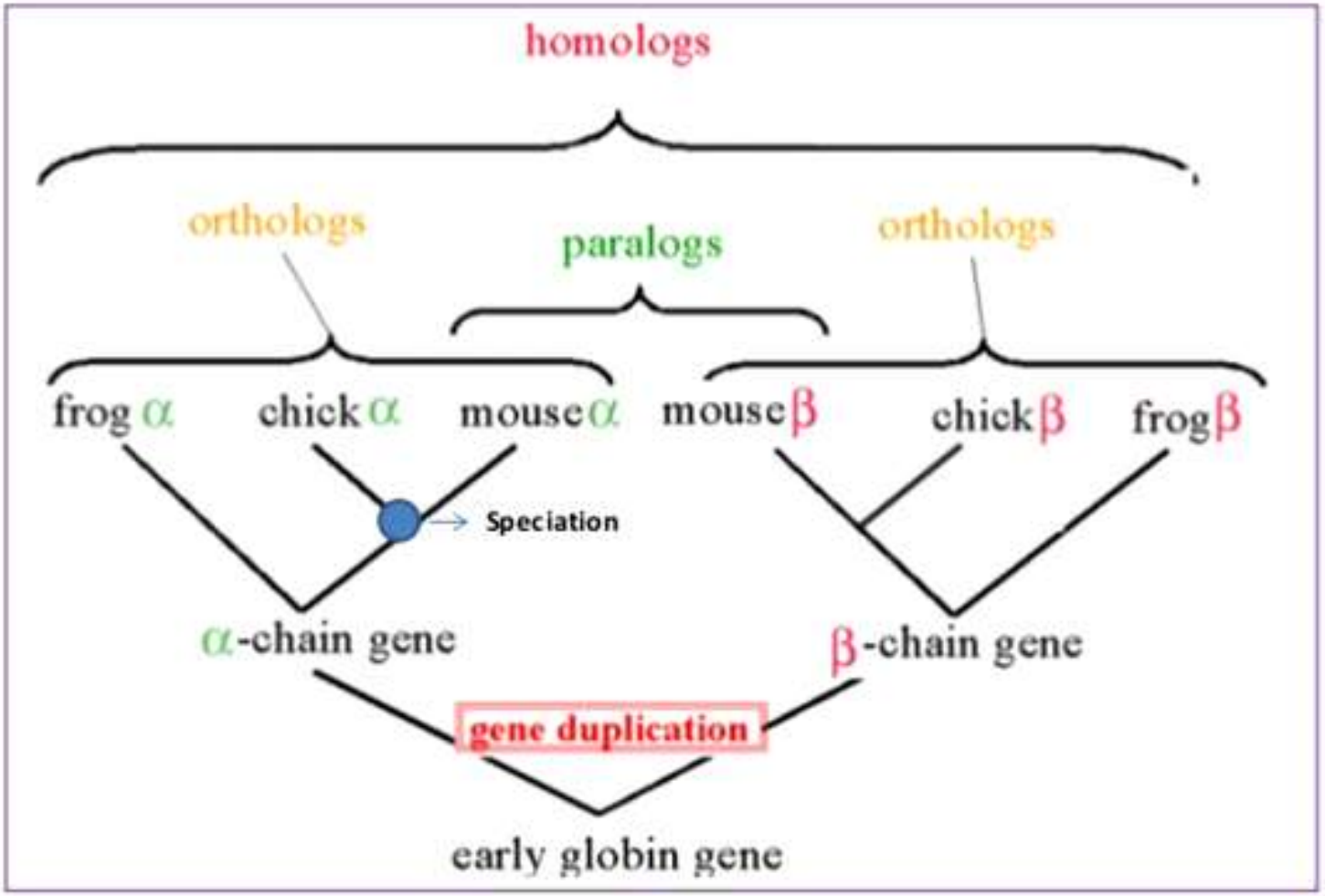
History

- **Comparative genomics** - root in the comparison of virus genomes in the early 1980s.
- **Example**- small RNA viruses infecting animals (picorna viruses) and those infecting plants (cowpea mosaic virus) were compared and turned out to share significant sequence similarity and, in part, the order of their genes.
- In 1986, the **first comparative genomic study at a larger scale** was published, comparing the genomes of varicella-zoster virus and Epstein- Barr virus.

Terms Used

- **Homology** is the relationship of any two characters (such as two proteins that have similar sequences) that have descended, usually through divergence, from a common ancestral character
- **Homologues** are thus components or characters (such as genes/proteins with similar sequences) that can be attributed to a common ancestor of the two organisms during evolution. Homologues can either be orthologues, paralogues, or xenologues
- **Orthologues** are homologues that have evolved from a common ancestral gene by speciation. They usually have similar functions
- **Paralogues** are homologues that are related or produced by duplication within a genome. They often have evolved to perform different functions

Terms Used


homologs
orthologs
paralogs
orthologs
frog α
chick α
mouse α
mouse β
chick β
frog β
Speciation
α-chain gene
β-chain gene
gene duplication
early globin gene

Introduction

- **Comparative genomics** is a field of biological research in which researchers use a variety of tools to compare the complete genome sequences of different species.
- By carefully comparing characteristics that define various organisms, researchers can pinpoint regions of similarity and difference.
- A comparison of gene numbers , gene locations & biological functions of gene in the genomes of different organisms, one objective being to identify groups of genes that play a unique biological role in a particular organism.

Introduction

The comparison helps-

1) to reveal the extent of conservation among genomes, which will provide insights into the mechanism of genome evolution and gene transfer among genomes.
2) to understand the pattern of acquisition of foreign genes through lateral gene transfer.
3) to reveal the core set of genes common among different genomes, which should correspond to the genes that are crucial for survival.

Comparative Genomics

- Comparative genomics has yielded dramatic results. Researchers are increasingly using comparative genomics to explore areas ranging from human development and behavior to metabolism and susceptibility to disease.
- These studies are uncovering new behavioral, neurological and developmental pathways and genes that are shared or related among species.
- Among the results so far some are as follows:
  1) A study discovered that about 60 percent of genes are conserved between fruit flies and humans, meaning that the two organisms appear to share a core set of genes. Two-thirds of human genes known to be involved in cancer have counterparts in the fruit fly.

Comparative Genomics

2) Researchers studying milk production have mapped genes that increase the yield of high-fat milk in cows, resulting in higher production levels and potentially a significant economic impact. This is one of many studies aimed at increasing food production.
3) Comparisons of nearly 50 bird species' genomes revealed a gene network that underlies singing in birds and that may have an important role in human speech and language. The bird researchers also found gene networks responsible for traits such as feathers and beaks.
4) Scientists have found genes that increase muscling in cattle by twofold; they found the same genes in racing dogs, and such results may foster human performance studies.

Comparative Genomics

The subject of comparative genomics has effects on-

- ✓ Evolutionary biology and phylogenetic reconstructions of the tree of life,
- ✓ Drug discovery programs,
- ✓ Function predictions of hypothetical proteins
- ✓ Identification of genes, regulatory motifs and other non-coding DNA motifs
- ✓ Genome flux (a research associate in human genetics, developed methods to extrapolate the amount of DNA that vanished by comparing genome sizes from present day animals to that of their common ancestors)
- ✓ Genome dynamics (*Elucidating the molecular mechanisms and evolutionary processes that shape the structure and function of genes and genomes*)

Components

The main themes of comparative genomics include

I. Whole genome alignment,

II. Comparing gene order between genomes,

III. Constructing minimal genomes,

IV. Lateral gene transfer among genomes

Whole Genome Alignment

- With an ever-increasing number of genome sequences available, it becomes imperative to understand sequence conservation between genomes, which often helps to reveal the presence of conserved functional elements.
- This can be accomplished through direct genome comparison or genome alignment.
- The alignment at the genome level is fundamentally no different from the basic sequence alignment.
- However, alignment of extremely large sequences presents new complexities owing to the sheer size of the sequences.
- Regular alignment programs tend to be error prone and inefficient when dealing with long stretches of DNA containing hundreds or thousands of genes.

Whole Genome Alignment

- Another challenge of genome alignment is effective visualization of alignment results.
- Because it is obviously difficult to sift through and make sense of the extremely large alignments, a graphical representation is a must for interpretation of the result.
- Therefore, specific alignment algorithms are needed to deal with the unique challenges of whole genome alignment.
- Alignment programs for "super-long" DNA sequences are-
  - ✓ MUMmer,
  - ✓ BLASTZ,
  - ✓ LAGAN,
  - ✓ PipMaker,
  - ✓ MAVID,
  - ✓ GenomeVista

Comparing gene order

- **Gene orders** are the permutation of genome arrangement.
- When the order of a number of linked genes is conserved between genomes, it is called **synteny**.
- Gene order is much less conserved compared with gene sequences.
- Gene order conservation is in fact rarely observed among divergent species.
- Computational analysis of genome arrangements is to estimate the number and types of rearrangements that have occurred and also to determine when they occurred.

Comparing gene order

- For comparing gene orders on chromosomes that have undergone rearrangements, lines joining the corresponding genes will intersect
- The greater the amount of rearranging, the greater the number of intersects.
- Genetic analysis has revealed that genes with a related function are frequently found to be clustered at one chromosomal location.
- Tools-
  - ✓ geneCo (https://bigdata.dongguk.edu/geneCo/#/index/main),
  - ✓ Cinteny (http://cinteny.cchmc.org/),
  - ✓ GeneOrder 4.0 (http://binf.gmu.edu:8080/GeneOrder4.0/)

Constructing Minimal Genomes

- One of the goals of genome comparison is to understand what constitutes a minimal genome, which is a minimal set of genes required for maintaining a free-living cellular organism.
- Finding minimal genomes helps provide an understanding of genes constituting key metabolic pathways, which are critical for a cell's survival.
- This analysis involves identification of orthologous genes shared between a number of divergent genomes.

Constructing Minimal Genomes

- The concept of minimal genome arose from the observations that many genes do not appear to be necessary for survival.
- In order to create a new organism a scientist must determine the minimal set of genes required for metabolism and replication.
- This can be achieved by experimental and computational analysis of the biochemical pathways needed to carry out basic metabolism and reproduction
- Program for identifying Minimal Genome is Coregenes

Lateral Gene Transfer

- **Lateral gene transfer** (or horizontal gene transfer) is defined as the exchange of genetic materials between species in a way that is incongruent (incompatible) with commonly accepted vertical evolutionary pathway.
- Lateral gene transfer mainly occurs among prokaryotic organisms when foreign genes are acquired through mechanisms such as transformation (direct uptake of foreign DNA from environment), conjugation (gene uptake through mating behavior), and transduction (gene uptake mediated by infecting viruses).
- The transmission of genes between organisms can occur relatively recently or as a more ancient event.

Lateral Gene Transfer

- If lateral transfer events occurred relatively recently, one would expect to discover traces of the transfer by detecting regions of genomic sequence with unusual properties compared to surrounding regions.
- The basic **tools** for identifying genomic regions that may be a result of lateral gene transfer events using the within-genome approach are-
  - ✓ACT (Artemis Comparison Tool; www.sanger.ac.uk/Software/ACT)
  - ✓Swaap (http://www.bacteriamuseum.org/SWAAP/SwaapPage.htm)
- **Within-Genome Approach** is to identify regions within a genome with unusual compositions. Single or oligonucleotide statistics, such as G–C composition, codon bias, and oligonucleotide frequencies are used.

Methods

- Methods for comparative genomics
    1) Comparative analysis of genome structure
    2) Comparative analysis of coding regions
    3) Comparative analysis of non-coding regions

Comparative analysis of genome structure

- Analysis of the global structure of genomes, such as nucleotide composition, syntenic relationships, and gene ordering offer insight into the similarities and differences between genomes.
- This provide information on the organization and evolution of the genomes, and highlight the unique features of individual genomes.
- The structure of different genomes can be compared at three levels:
  - Overall nucleotide statistics,
  - Genome structure at DNA level,
  - Genome structure at gene level.

Comparative analysis of genome structure

Comparison of overall nucleotide statistics

- Overall nucleotide statistics, such as
  - Genome size,
  - Overall (G+C) content,
  - Regions of different (G+C) content,
  - Genome signature such as codon usage biases,
  - Amino acid usage biases, and the ratio of observed di-nucleotide frequency
  - The expected frequency given random nucleotide distribution
- These all present a global view of the similarities and differences of the genomes.

Comparative analysis of genome structure

Comparison of genome structure at DNA level

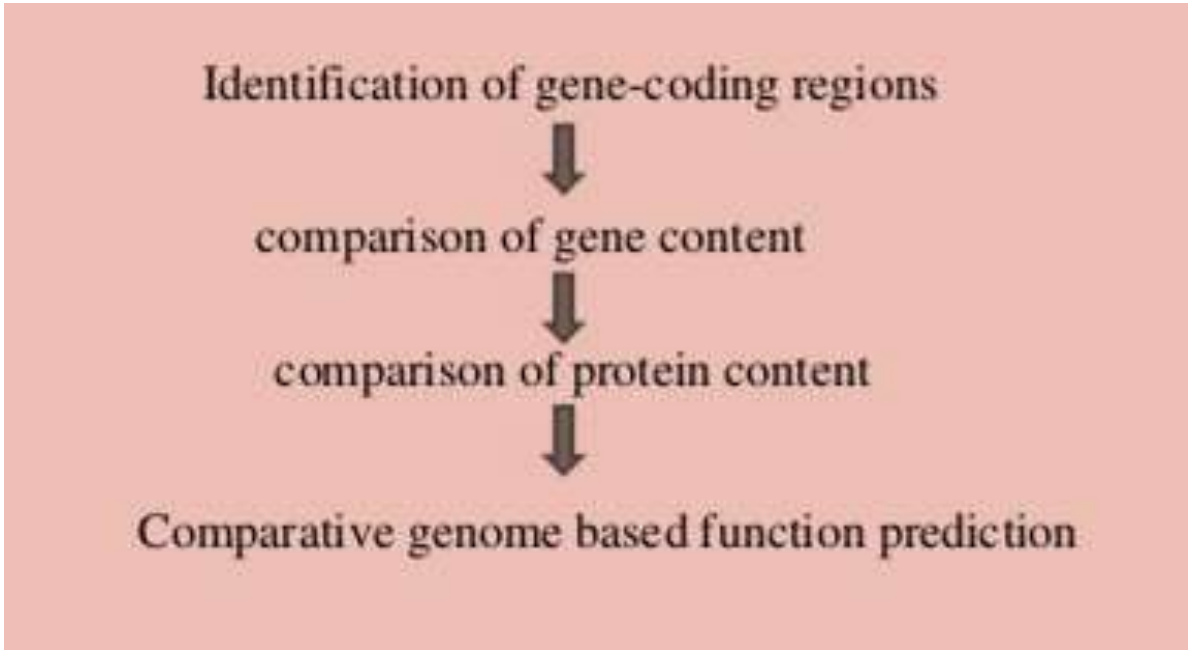
- Chromosomal breakage and exchange of chromosomal fragments are common mode of gene evolution.
- They can be studied by comparing genome structures at DNA level.
  - Identification of conserved synteny and genome rearrangement events
  - Analysis of breakpoints
  - Analysis of content and distribution of DNA repeats

Comparative analysis of genome structure

Comparison of genome structure at gene level

- Chromosomal breakage and exchange of chromosomal fragments cause disruption of gene order
- Therefore gene order correlates with evolutionary distance between genomes

Comparative analysis of coding regions



- Number of algorithms that have been use in comparative genomics to aid function prediction of genes.

Comparative analysis of coding regions

Identification of gene-coding regions

- The analysis and comparison of the coding regions starts with the gene identification algorithm that is used to infer what portions of the genomic sequence actively code for genes.
- There are four basic category for gene identification

Category	Algorithm
1. Based on direct evidence of transcription	EST_GENOME sim4
2. Based on homology with known genes	PROCRUSTES
3. Statistical or ab-initio approaches	Genscan FGENES GeneMark Glimmer
4. Using genome comparison	TwinScan Rosetta

Comparative analysis of coding regions

Comparison of gene content

- After the predicted gene set is generated, it is very interesting and important to compare the content of genes across genomes.
- The first statistics to compare is the estimated total number of genes in a genome, elucidate the similarities and differences between the genomes include percentage of the genome that code for genes, distribution of coding regions across the genome (a.k.a. gene density), average gene length, codon usage
- This is often done using a pairwise sequence comparison tool such as BLASTN or TBLASTX

Comparative analysis of coding regions

Comparison of protein content

- A second level of analysis that can be performed is to compare the set of gene products (protein) between the genomes, which has been termed "comparative proteomics"
- It is important to compare the protein contents in critical pathways and important functional categories across genomes
- Two widely used resources for pathways and functional categories are the KEGG pathway database and the Gene Ontology (GO) hierarchy.
- Interesting statistics to compare include:

  – Level of sequence identity between orthologous pairs across genome – Paralogous pairs within genome, – Number of replicated copies in corresponding paralog families – Functions of the paralogs – Locations of members of paralog families across the genome

Comparative analysis of coding regions

Comparative genomics-based function prediction

- Functional assignment of genes in a non similarity-based manner.
- This rely on the basic premise that genes; that are functionally related, are genes that are closely associated across genomes in some form.
- This include three methods:
  - Co-conservation across genomes
  - Conservation of gene clusters and genomic context across species
  - Physical fusion of functionally linked genes across species (Domain fusion analysis)

Comparative analysis of noncoding regions

- Noncoding regions of the genome gained a lot of attention in recent years because of its predicted role in regulation of transcription, DNA replication, and other biological functions
- This approach is based on the presumption that selective pressure causes regulatory elements to evolve at a slower rate than that of non regulatory sequences in the non coding regions

Genome Alignment Tools

Pairwise Alignment-

- MUMmer
- PipMaker
- VISTA

Multiple Sequence Alignment-

- DIALIGN
- MGA

THANK YOU

MUMmer

-Ms. Rupal Mishra

Introduction

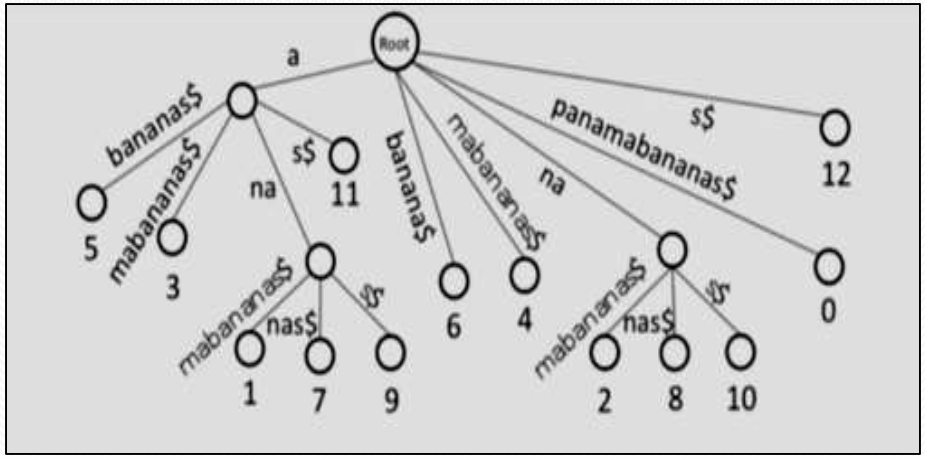
- MUMmer is a system for rapidly aligning large DNA sequences to one another.
- It can align:
  - ✓ whole genomes to other genomes
  - ✓ large genome assemblies to one another
  - ✓ partial (draft) genomes sequences to one another
  - ✓ or (with release 4) a set of reads to a genome.
- It is very fast and easy to run.
- The current version is MUMmer 4.0

Introduction

- MUMmer is an open source software package for the rapid alignment of very large DNA and amino acid sequences.
- MUMmer is a modular and versatile package that relies on a suffix tree data structure for efficient pattern matching.
- Suffix trees are suited for large data sets because they can be constructed and searched in linear time and space.
- This allows mummer to find all 20 base pair maximal exact matches between two ~5 million base pair bacterial genomes in 20 seconds, using 90 MB of RAM, on a typical 1.7 GHz Linux desktop computer.

Suffix tree

- **Suffix tree** [also called **PAT tree** (Pattern Searching Tree) or, in an earlier form, position tree]
- It is a compressed trie (one that has been compacted down to save space) containing all the suffixes of the given text as their keys and positions in the text as their values.
- Suffix trees allow particularly fast implementations of many important operations.



MUMmer

- Uses a seed and extend strategy, as alignment anchors to generate pair-wise alignments.
- **Seed-and-extend strategy** is based on the observation that a good alignment should contain exact or inexact short matches between two sequences.
- It also includes some utilities to handle the alignment output and a primitive plotting tool (mummerplot) that allows the user to convert MUMmer output to gnuplot files for dot and percent identity plots.
- Another graphical utility called MapView is included with the MUMmer distribution and displays sequence alignments to a annotated reference sequence for exon refinement and investigation.

MUMmer

- Advantage of MUMmer is its speed. Its low runtime and memory requirements allow it to be used on most any computer.
- It's efficiency makes it ideal for aligning huge sequences such as completed and draft eukaryotic genomes.
- MUMmer has been successfully used to align the mouse and human genomes, showing it can handle most any input available.
- Because of it's many abilities, inexperienced users may find it difficult to determine the best methods for their application.

MUMmer

- The MUMmer package provides efficient means for comparing an entire genome against another.
- However, until 1999 there were no two genomes of sufficient similarity to compare.
- With the publication of the second strain of *Helicobacter pylori* in 1999, following the publication of the first strain in 1997, the scientific world had its first chance to look at two complete bacterial genomes whose DNA sequences were highly similar.
- The number of pairs of closely-related genomes has exploded in recent years, facilitating many comparative studies.

Program

- The MUMmer pipelines are comprised of three main sections.
  - The **first section** identifies a certain subset of maximal exact matches between the two inputs
  - The **second section** clusters these matches into groups that will likely make good alignment anchors
  - The **third section** extends alignments between these clustered matches to produce the final gapped alignment

Program

These three sections also outline the primary types of programs included in the MUMmer package –

- The **Maximal exact matching section** describes the programs that compute different types maximal exact matches
- The **Clustering section** describes the two different types of clustering algorithms
- **Alignment generators** describes the scripts that combine matching, clustering and extending in order to produce high scoring pair-wise alignments.

Alignment

- With the capability to align the entire human genome to itself, **there is no genome too large for MUMmer**.
- Example:
  - Each human chromosome was used as a reference, and the rest of the genome was used as a query against it.
  - To avoid duplication, we only included chromosomes in the query if they had not already been compared; thus we first used chromosome 1 as a reference, and streamed the other chromosomes against it.
  - Then we used chromosome 2 as a reference, and streamed chromosomes 3–22, X, and Y against that, and so on.

Human vs. Human

The following table gives run times and space requirements for a cross comparison of all human chromosomes.

Chr	Ref length (Mbp)	Suffix time (min)	Qry length (Mbp)	Query time (min)	Total space (Mb)	Suffix space (bytes/bp)
1	221.8	24.6	2617.1	679.5	3702	15.43
2	237.6	27.4	2379.5	625.8	3908	15.43
3	194.8	21.2	2184.7	565.0	3232	15.43
4	188.4	22.4	1996.3	518.0	3121	15.43
5	177.7	18.6	1818.6	461.4	2952	15.43
6	175.8	17.9	1642.8	407.6	2900	15.43
7	153.8	15.7	1489.0	360.1	2550	15.43
8	142.8	14.4	1346.2	322.3	2378	15.43
9	117.0	10.7	1229.2	303.7	1974	15.43
10	131.1	13.2	1098.1	263.3	2195	15.43
11	133.2	13.1	964.9	225.6	2228	15.43
12	129.4	12.5	835.5	195.9	2168	15.43
13	95.2	8.6	740.3	163.6	1633	15.44
14	88.2	7.5	652.1	141.0	1523	15.44
15	83.6	6.8	568.5	122.1	1451	15.44
16	80.9	6.4	487.6	106.3	1409	15.44
17	80.7	6.6	406.9	91.8	1406	15.44
18	74.6	6.3	332.3	78.8	1311	15.44
19	56.4	3.7	275.8	56.1	1026	15.45
20	59.4	4.6	216.4	45.8	1073	15.45
21	33.9	2.1	182.5	33.7	673	15.48
22	33.8	2.0	148.6	26.4	672	15.48
Un	1.4	0.03	147.3	10.0	164	16.96
X	147.3	14.6		4.8	2327	15.57

Table

- **Column 1** indicates the chromosome number ("Un" referring to unmapped contigs).
- **Column 2** shows chromosome length.
- **Column 3** shows the time to construct the suffix tree.
- **Column 4** shows the length of the total genomic DNA searched against the chromosome in column 1.
- **Column 5** the time to stream the query sequence through it.
- **Column 6** shows the maximum amount of computer memory occupied by the program and data.
- **Column 7** shows memory usage for the suffix tree in bytes per base pair.

Installation

- let's suppose you just downloaded the **MUMmer3.0.tar.gz** distribution from the SourceForge site.
- The first step would be to move this file to the desired installation directory and type: **tar -xvzf MUMmer3.0.tar.gz**
- To extract the MUMmer source into a **MUMmer3.0** subdirectory. Switch to this newly created subdirectory and execute: **make check**
- To assure the makefile can identify the necessary utilities. If no error messages appear, the diagnostics were successful and you may continue.
- However, if error messages are displayed, the listed programs are not accessible via your system path. Install the utilities if necessary, add them to your system **PATH** variable, and continue with the MUMmer installation by typing: **make install**
- This will attempt to compile the MUMmer scripts and executables. If the make command issues no errors, the compilation was successful and you are ready to begin using MUMmer.

Running MUMmer

- This example compares a single query sequence to a single reference sequence using mummer, and then uses mummerplot to generate a dot plot representation of the comparison.
- The following input files will be used to demonstrate this example:

  H_pylori26695_Eslice.fasta

  H_pyloriJ99_Eslice.fasta

- Mummer can handle multiple reference and multiple query sequences, however a dotplot of more that two sequences can be confusing.

Running MUMmer

mummer -mum -b -c H_pylori26695_Eslice.fasta H_pyloriJ99_Eslice.fasta > mummer.mums

- This **command will find all** maximal unique matches (-mum) between the reference and query on both the forward and reverse strands (-b) and report all the match positions relative to the forward strand (-c).
- Output is to **stdout**, so we will redirect it into a file named **mummer.mums.**
- A dotplot of all the MUMs between two sequences can reveal their macroscopic similarity.

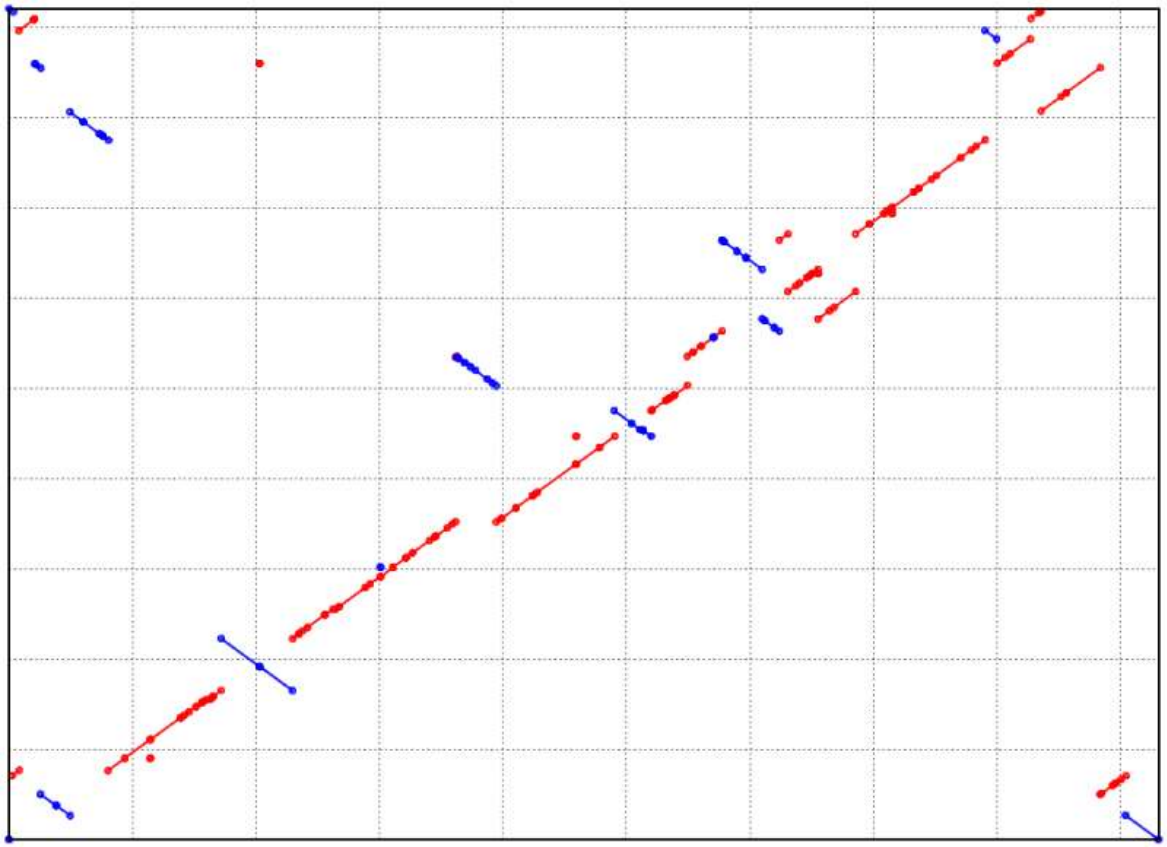
Running MUMmer

mummerplot -x "[0,275287]" -y "[0,265111]" -postscript -p mummer mummer.mums

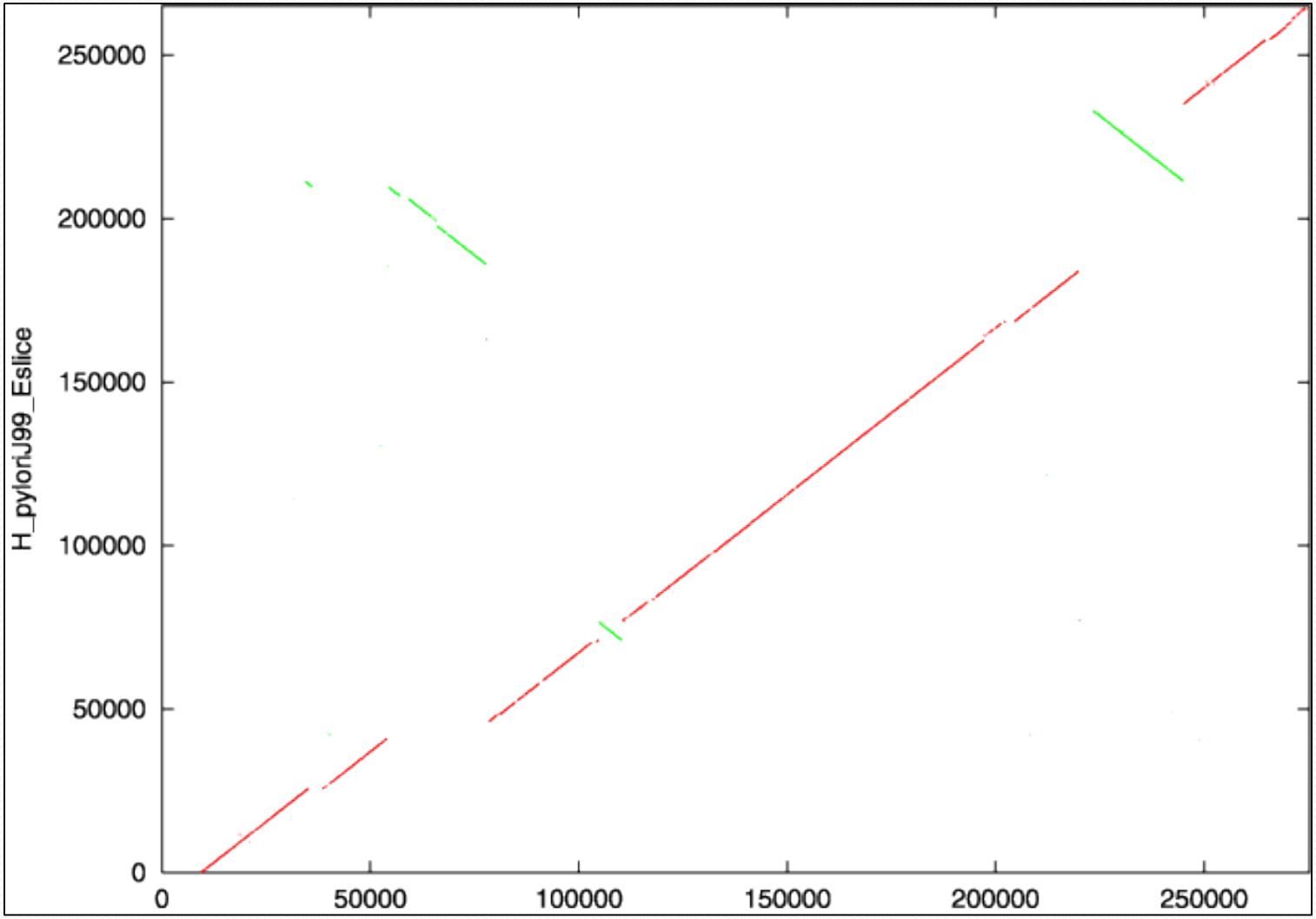
- This command will plot all of the MUMs in the mummer.mums file between the given ranges for the X and Y axes.
- When plotting mummer output, it is necessary to use the lengths of the input sequences to set the plot ranges, otherwise the plot will be automatically scaled around the minimum and maximum data points.

Running MUMmer

- Most image manipulation programs can edit the postscript output, or it can be sent directly to a printer with the lpr command.



Viewing the output



Viewing the output

- The **plot represents the set of all MUMs** between the two input sequences.
- Forward MUMs are plotted as red lines/dots while reverse MUMs are plotted as green lines/dots (blue may be used for reverse matches in newer versions).
- The green segment in the upper left quadrant of the graph shows both an inversion and translocation, as it is of negative slope and inconsistently located relative to the rest of the plot which falls on a line.
- However the green segment in the upper right quadrant of the graph shows only an inversion, as it is of negative slope but is consistent in location with the rest of the plot.

THANK YOU

PipMaker

-Ms. Rupal Mishra

Introduction

- **PipMaker** computes alignments of similar regions in two DNA sequences using the BLASTZ algorithm for comparing and to identify conserved segments.
- The resulting alignments are summarized with a "percent identity plot", or "pip" for short.
- The result also shows a traditional textual form of the alignment.
- http://pipmaker.bx.psu.edu/cgi-bin/pipmaker?basic
- PipMaker generates graphical output as a PDF document.
- PipMaker is appropriate for comparing genomic sequences from any two related species.

Introduction

- To generate a pip, PipMaker requires four user-supplied files.
- First 2 are mandatory files-
  1) The first sequence file is depicted along the horizontal axis.
  2) Finally a second sequence file.
- Next 2 are optional files-
  1) The Mask file of the first sequence (The user generates this file using the RepeatMasker program, available on the web at the Institute for Systems Biology.)
  2) A file of gene and exon positions allows PipMaker to draw the locations of exons and indicate the directionality of genes.

Input to PipMaker

- PipMaker processes the contents of the following **four files**.
- For each of those, you can either paste the data into the multi-line text area or, if your browser supports it, give the filename in the subsequent single-line field.
- **First sequence data file**- A FastA file containing the first sequence, with nucleotides given as capital letters. The maximum length is two million nucleotides.
- For example,

```
>exactly one header line, rest contain ACGT ONLY
ACGTACGTACGT
CGTACGTACGTA
GTACGTACGTAC
TACGTACGTACG
```

Input to PipMaker

- **Second sequence data file**. A FastA file containing the second sequence. The maximum allowable length is 2Mb.
- **First sequence mask file**- The documentation produced by RepeatMasker for the first sequence.
- Inclusion of this file is optional, but it is strongly recommended for mammalian sequences, since otherwise biologically insignificant matches due to repeats will be computed.
- If this file is omitted, then the pip will not include icons showing the positions of interspersed repeat elements.
- **Exons file for the first sequence**- An optional text file providing the positions of transcriptional units in the first sequence.

Input to PipMaker

- The directionality of a gene (< or >), its start and end positions, and name should be on one line, followed by separate lines specifying the start-positions and end-positions of each exon.
- It is also permissible to begin the line with '|', in which case the line indicating the gene's position in the first sequence is drawn without an arrowhead.
- An optional line beginning with a "+" character can indicate the first and last nucleotides of the translated region (including the initiation codon, Met, and the stop codon).
- Blank lines are ignored. Exons must be specified in order of increasing address even if the gene is on the reverse strand (<).

Input to PipMaker

Thus, the Exons file might begin as follows:

- My favorite genomic region

```
> 100 800 Gene 1
100 200

300 400
600 800

< 1000 2000 Second Gene
+ 1100 1900
1000 1200
1800 2000
...
```

PipMaker

PipMaker (instructions) aligns two DNA sequences and returns a percent identity plot of that alignment, together with a traditional textual form of the alignment.

- First sequence (FASTA format):

  or filename (**file must be plain text only**):

  Choose File No file chosen
- Second sequence (FASTA format):

  or filename (**file must be plain text only**):

  Choose File No file chosen
- Your email address:
- Optional features:
  - First sequence mask:

    Choose File No file chosen
  - First sequence exons:

    Choose File No file chosen

Submit Reset

PipMaker

PipMaker (instructions) aligns two DNA sequences and returns a percent identity plot of that alignment, together with a traditional textual form of the alignment.

- First sequence (FASTA format):

```
GGTCTCC
TAGAGCTCCTTACCGGAGT
```

  or filename (**file must be plain text only**):

  Choose File No file chosen

- Second sequence (FASTA format):

```
GCTCCTCCCTCCCTACCTCTGTCTCTGAGTTGCCTGGTAAAGGTGTGTGTCAGTCCTCTTCAT
TCTAGAA
CTAGTGGATCCCCC
```

  or filename (**file must be plain text only**):

  Choose File No file chosen

- Your email address:
- Optional features:
  - First sequence mask:

    Choose File No file chosen
  - First sequence exons:

    Choose File No file chosen

Submit Reset

Submission Details

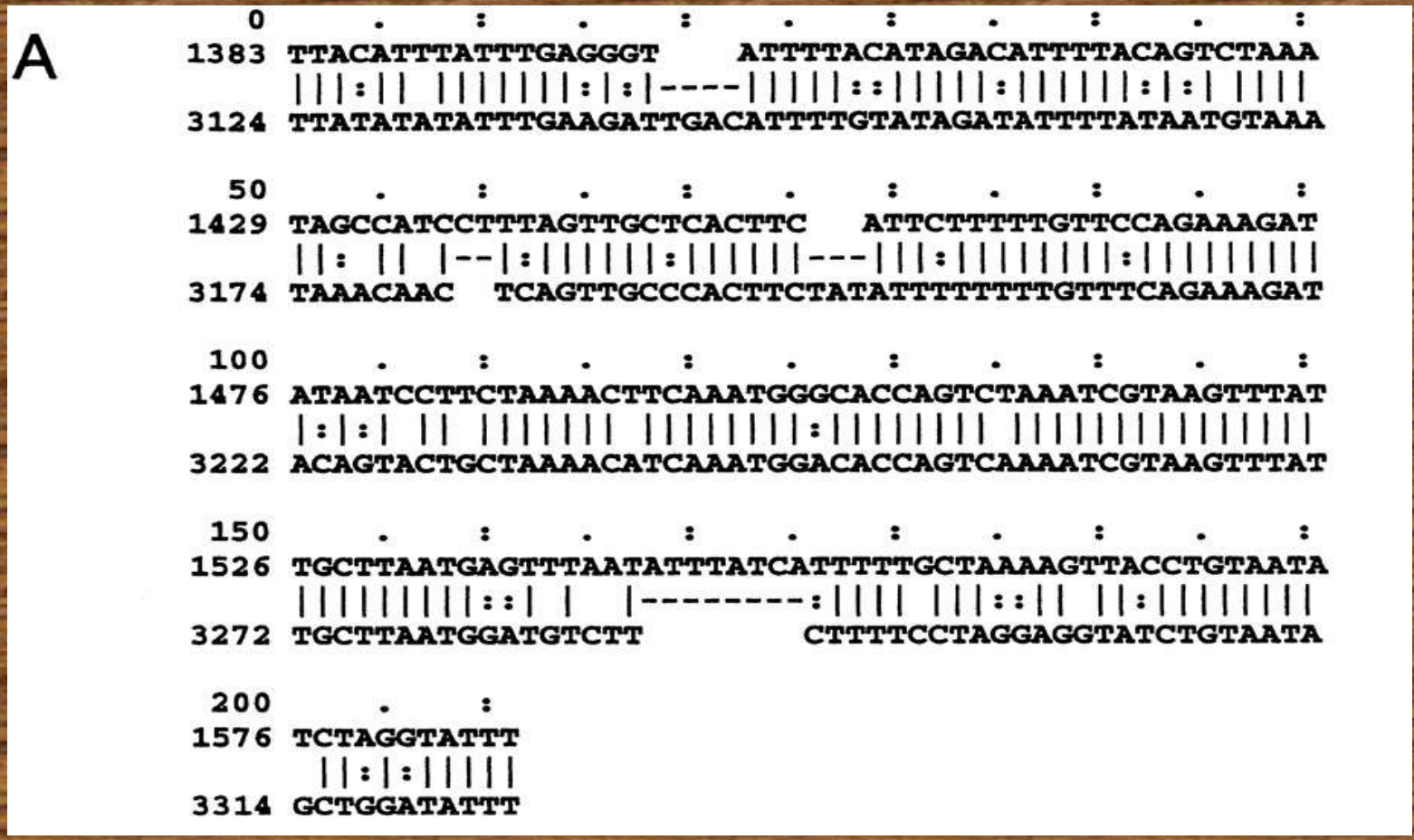
submitted

Expect reply via email.

```
email: rupalmishra555@gmail.com
seq1data: bp 77538: >AC004500.1 Homo sapiens chromosome 5, P1 clone 1076B9 (LBNL H14), complete sequence
seq2data: bp 81837: >AC004775.1 Homo sapiens chromosome 5, P1 clone 1308e5 (LBNL H13), complete sequence
seq1mask: 0 lines
exons: 0 lines
underlay: 0 lines
search strand: both
coverage: show all matches
pip title:
generate pip: yes
generate dotplot: yes
data format: PDF
generate concise text: yes
generate traditional text: yes
generate analysis of exons: no
return raw blastz output: no
```

An illustration of pips

(*A*) An alignment.



An illustration of pips

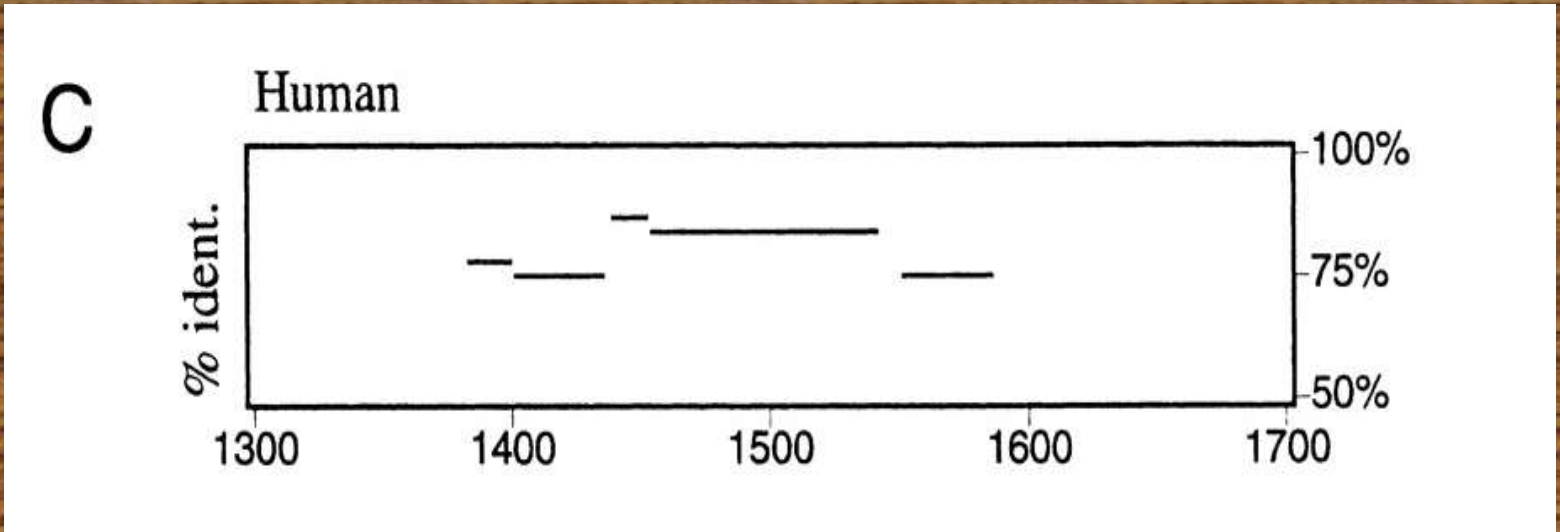
(*B*) Positions and percent identity of gap-free segments within that alignment.

B

human pos.	mouse pos.	length	identity
1383-1400	3124-3141	18 nt	78%
1401-1436	3146-3181	36 nt	75%
1439-1453	3182-3196	15 nt	87%
1454-1542	3200-3288	89 nt	84%
1551-1586	3289-3324	36 nt	75%

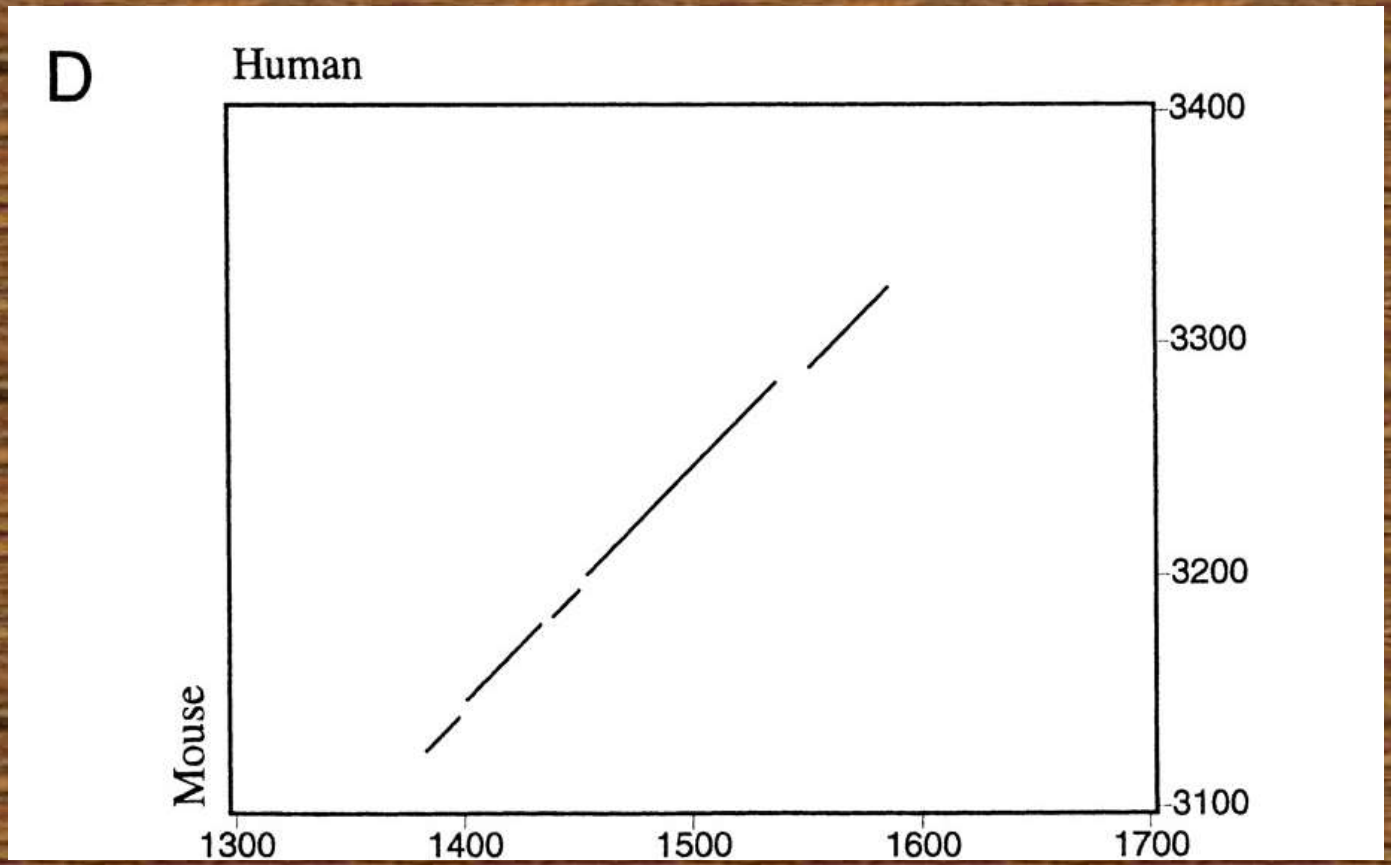
An illustration of pips

(C) The corresponding pip.



An illustration of pips

(*D*) The corresponding dot plot.



PipMaker Output

- The following **three files** are returned as attachments to an email message.
- You will need a MIME-aware (**Multi-purpose Internet Mail Extensions** or **Multimedia Internet Mail Extensions**) email program to read them.

- **The percent identity plot** "pip" is a PDF document.
- At the present time, Acroread has better support for hyperlinks, which PipMaker uses. (Ghostview can also display PostScript.)

PipMaker Output

- **A compressed form of the alignments**- This information is useful for precisely identifying sequence positions corresponding to conserved regions indicated in the pip.
- **The traditional textual form of the alignments**- For example, the following alignment corresponds to the first line of the compressed form shown above

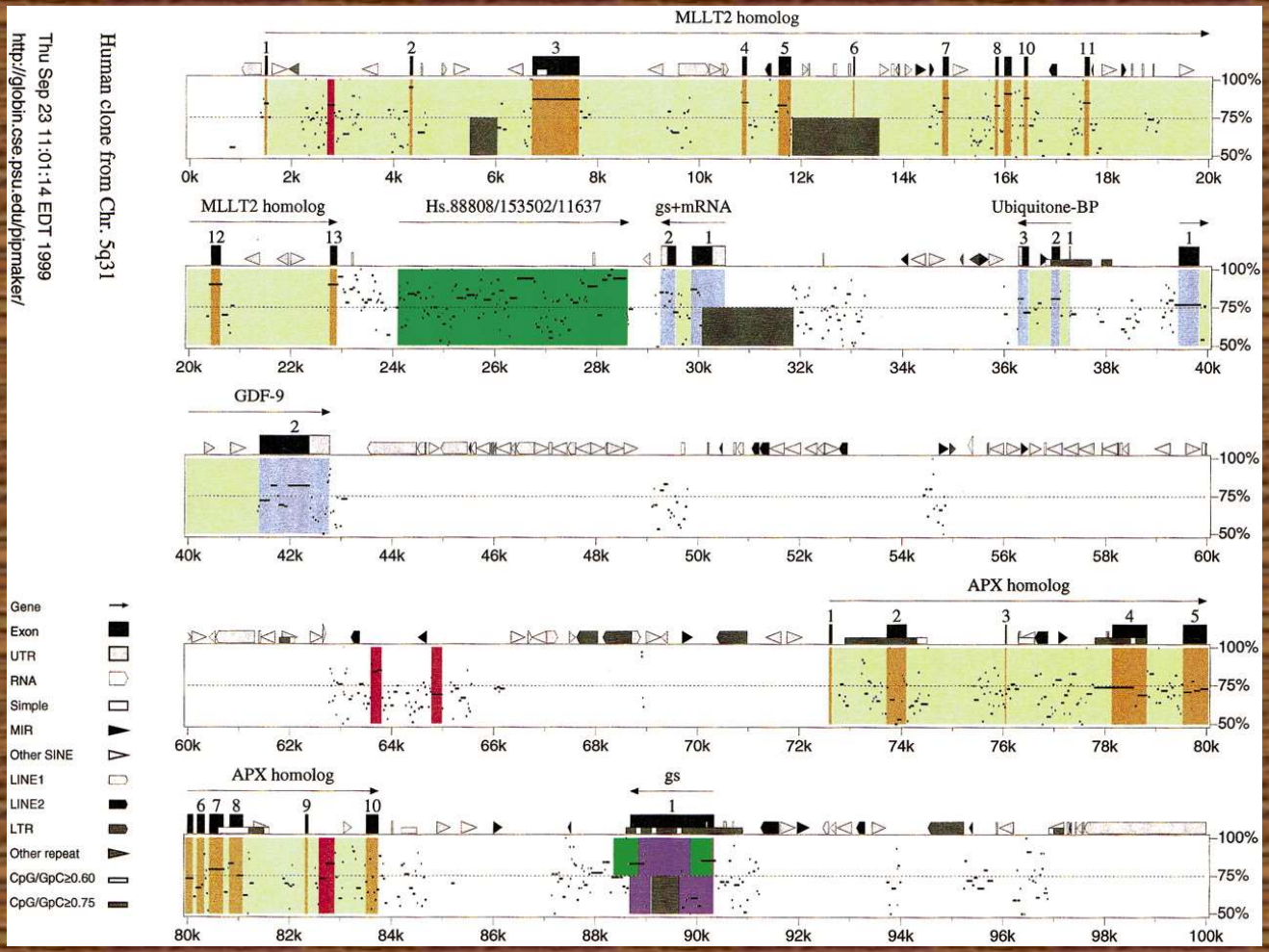
```
   0     .    :    .    :    .    :    .    :    .    :
2870 GCCCAGGCCTGGGCAGCGAGAGGGCCCTGCTCCCCGCTCAAGGCTCCCAG
     ||||||:| ||::||::|:|||| ||||||||:||:|||| ::|::::||
2281 GCCCAGACATGAACAATGGGAGGCCCCTGCTCTCCACTCACAACCTTTAG

  50     .
2920 GACATTC
     :||||||
2331 AACATTC
```

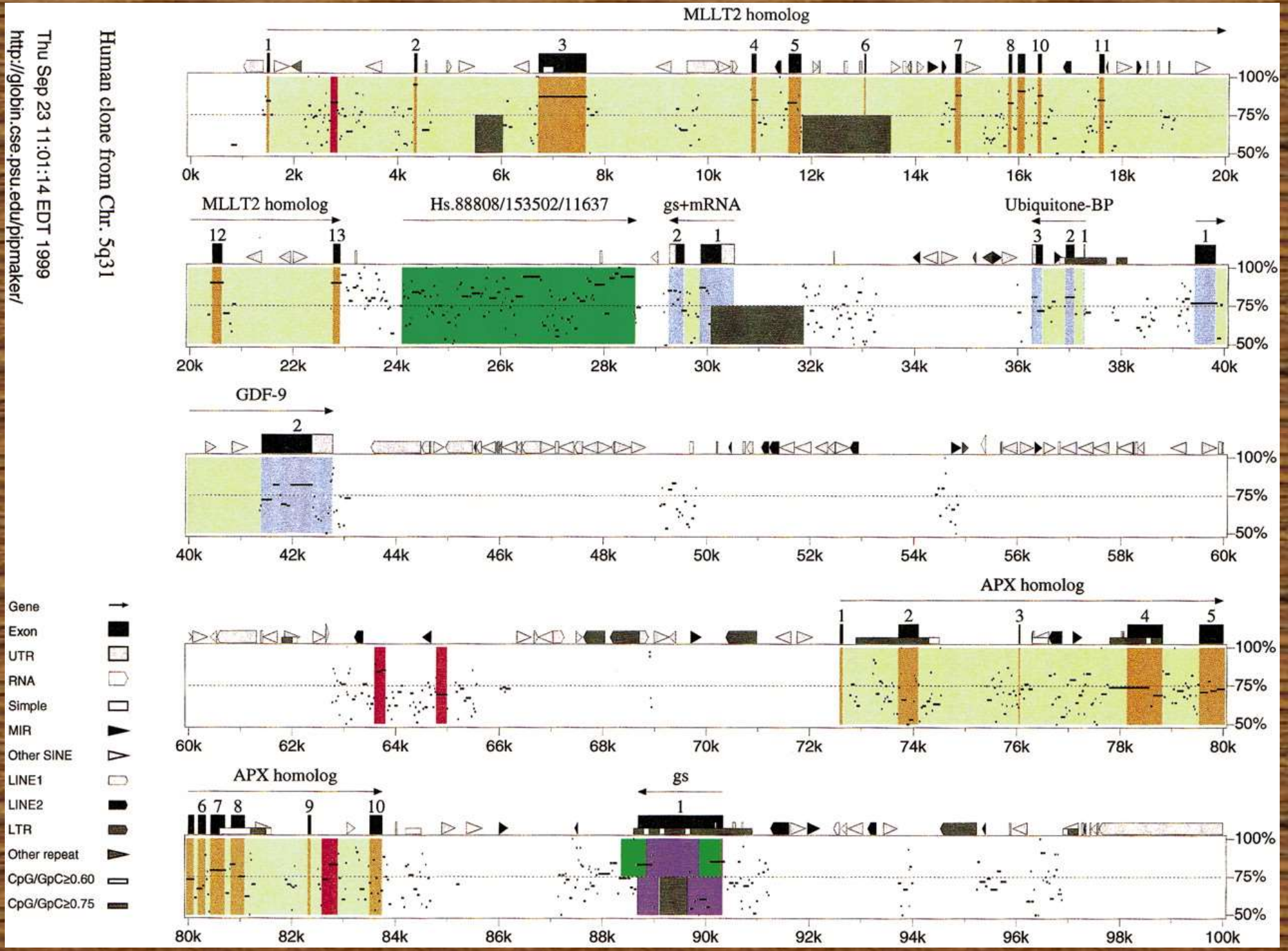
Interpreting a Pip

Icons along the top of the box have the following meanings.

- White pointed boxes are L1 (Line 1) repeats.
- Light gray triangles are SINEs (Short interspersed nuclear elements) other than MIR.
- Black triangles are MIRs (Mammalian-wide interspersed repeats).
- Black pointed boxes are L2 (Line 2) repeats.
- Dark gray triangles and pointed boxes are other kinds of interspersed repeats, such as LTR (long terminal repeat) elements and DNA transposons.
- Short dark gray boxes are CpG islands where the ratio CpG/GpC exceeds 0.75.
- Short white boxes are CpG islands where the ratio CpG/GpC lies between 0.6 and 0.75.

Interpreting a Pip




Human clone from Chr. 5q31
Thu Sep 23 11:01:14 EDT 1999
http://globin.cse.psu.edu/pipmaker/
MLLT2 homolog
MLLT2 homolog
Hs.88808/153502/11637
gs+mRNA
Ubiquitone-BP
GDF-9
APX homolog
APX homolog
gs
Gene
Exon
UTR
RNA
Simple
MIR
Other SINE
LINE1
LINE2
LTR
Other repeat
CpG/GpC≥0.60
CpG/GpC≥0.75
100%
75%
50%
0k
2k
4k
6k
8k
10k
12k
14k
16k
18k
20k
22k
24k
26k
28k
30k
32k
34k
36k
38k
40k
42k
44k
46k
48k
50k
52k
54k
56k
58k
60k
62k
64k
66k
68k
70k
72k
74k
76k
78k
80k
82k
84k
86k
88k
90k
92k
94k
96k
98k
100k


THANK
YOU

GENE ORDER

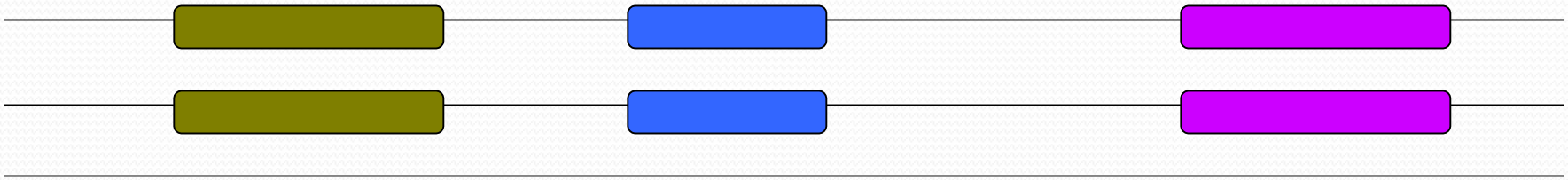
-Ms. Rupal Mishra

Gene Order (Synteny)

- Two species that have recently diverged from a common ancestor might be expected to share a similar set of genes and also similar chromosomes with these genes positioned along the chromosomes in the same order
- You should have heard about sequence polymorphisms, but what about the order of genes
- Two important observations
  1. Order is highly conserved in closely related species but becomes changed by rearrangements over evolutionary time
  2. Groups of genes that have a similar biological function tend to remain localized in a group or cluster

Sequence Divergence

Genomes of 2 closely related organisms



Gene A
ATGCCGGAG

Gene B
TTATATACG

Gene C
TTACGGCA

Evolutionary time ~2.5 M yrs

ATATGCTTAG
Gene A

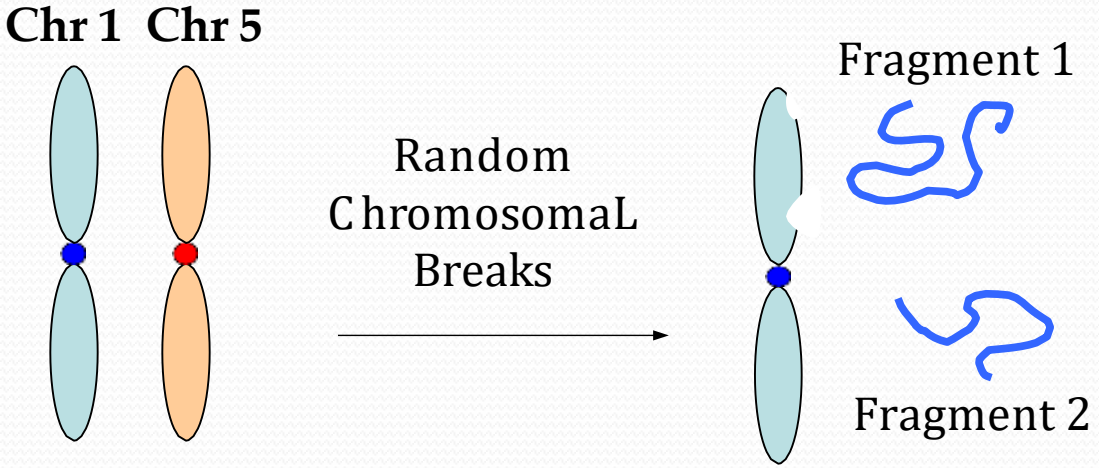
GCGCGCCG
Gene B

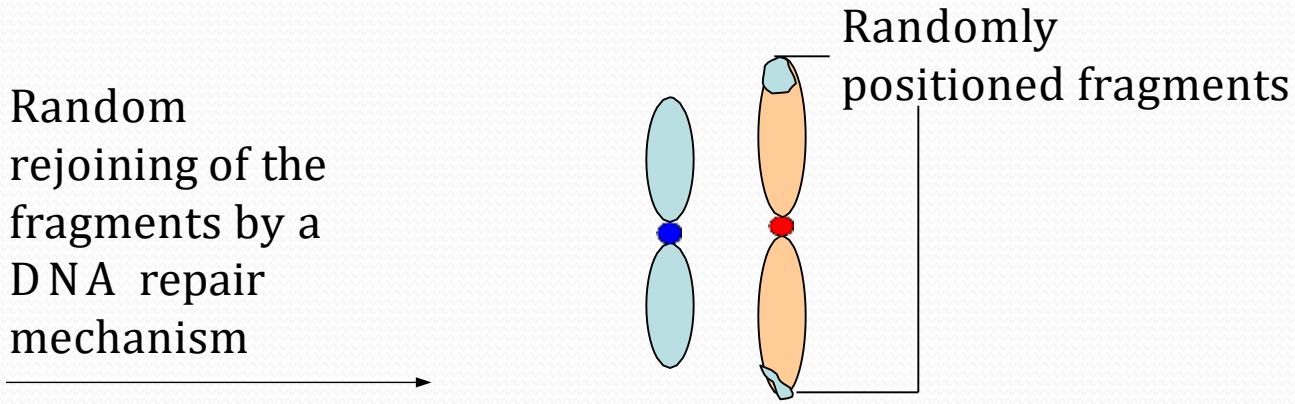
TTATATAT
Gene C

MUTATED BASES

Gene order not predictable

Chromosomal Rearrangements

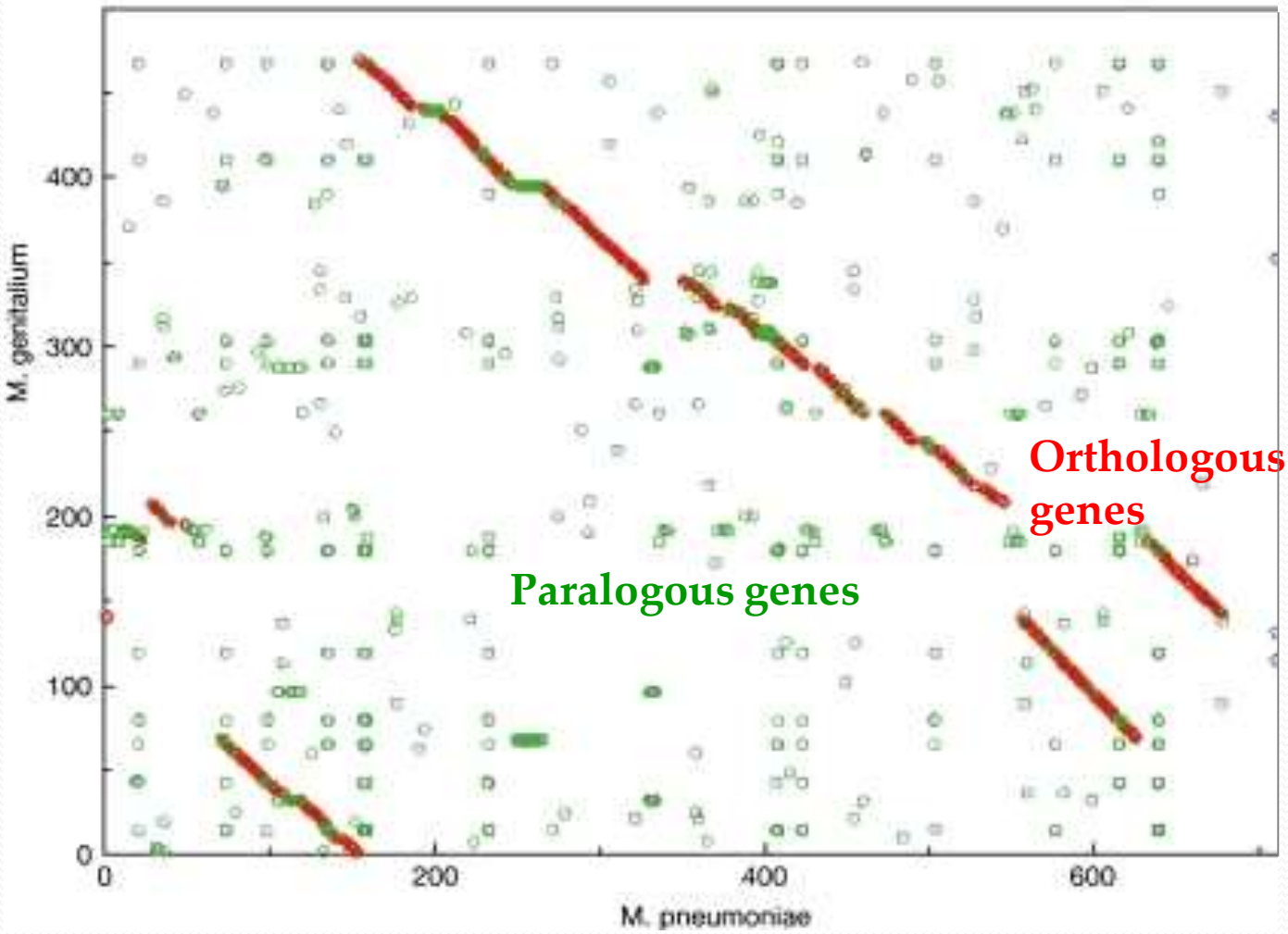




Rearrangements

- Collinearity of gene order is referred to as synteny, and a conserved group of genes in the same order in two genomes as a syntenic group or cluster.
- Collinearity - the relationship between the linear sequence of nucleic acid bases in the DNA of the gene and the sequence of amino acids in the protein encoded by the DNA.
- Rearrangements may be analyzed by comparing the location of orthologs, genes of highly conserved sequence and function in prokaryotic and eukaryotic proteomes from different phylogenetic lineages

Genome Plot


M. genitalium
M. pneumoniae
0
100
200
300
400
0
200
400
600
Orthologous genes
Paralogous genes

Classifying genes to get clear order

A similar plot of orthologous genes in the genomes of the bacterial species *E. coli* and *H. influenzae* appears quite random even though the organisms are only slightly more distant in evolution than the two *Mycoplasma* species.

↓

Classify genes using functional classification scheme

↓

Several genes falling into the same functional category are clustered together on the chromosomes of both of these organisms, and the clusters are in a similar order

Prokaryotic organisms of diverse phylogenetic origin

If gene A has a neighboring gene B

↓

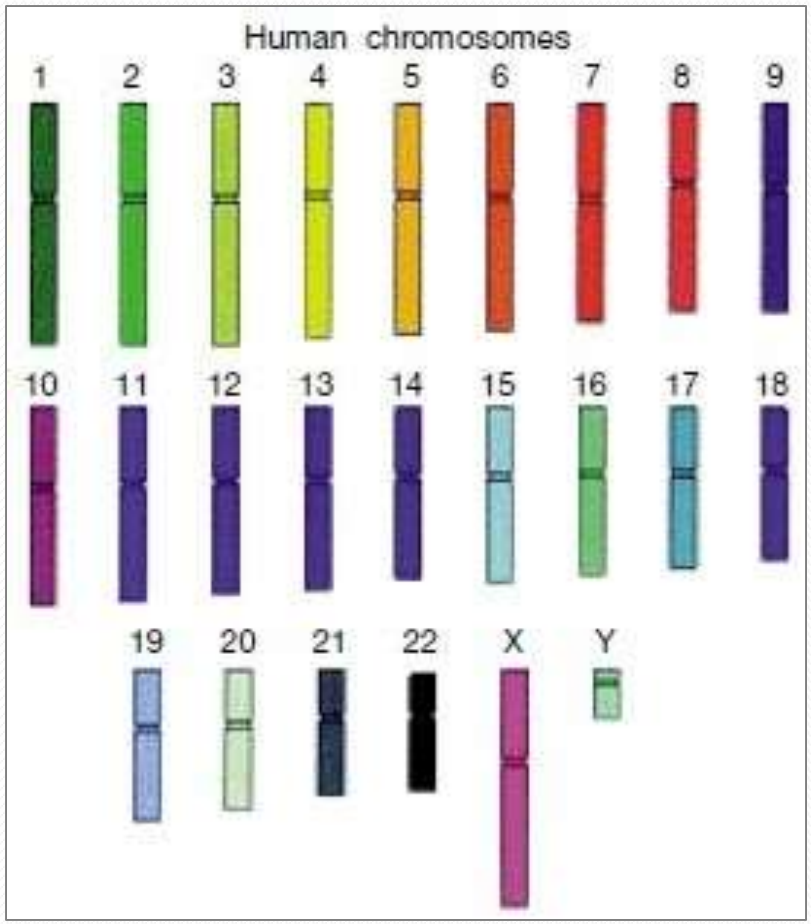
If an ortholog of A occurs in another genome

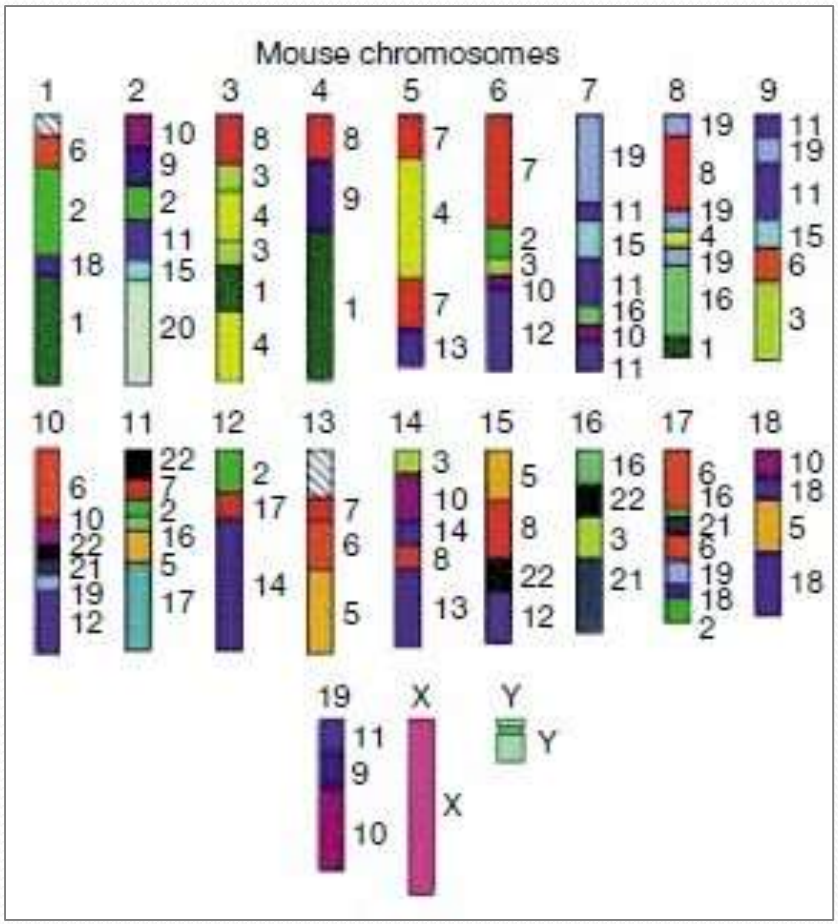
↓

There is an increased probability of an ortholog of B also occurring in the other organism

- However, the B ortholog is less likely to be a neighbor of the a ortholog of the genome of the second species if the two species are more divergent

Eukaryotic Genomes

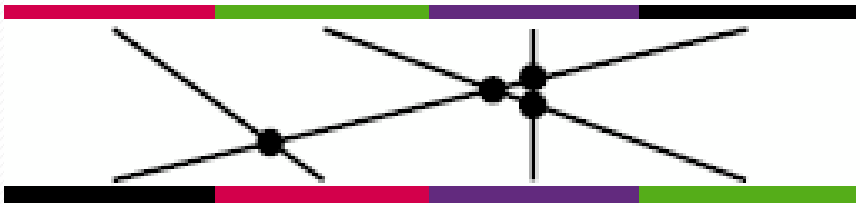




Each chromosome is a mosaic of a similar set of ancestral fragments

Computational analysis of genome arrangements

- To estimate the number and types of rearrangements that have occurred and also to determine when they occurred
- For comparing gene orders on chromosomes that have undergone rearrangements, lines joining the corresponding genes will intersect
- The greater the amount of rearranging, the greater the number of intersects



Clusters of Genes

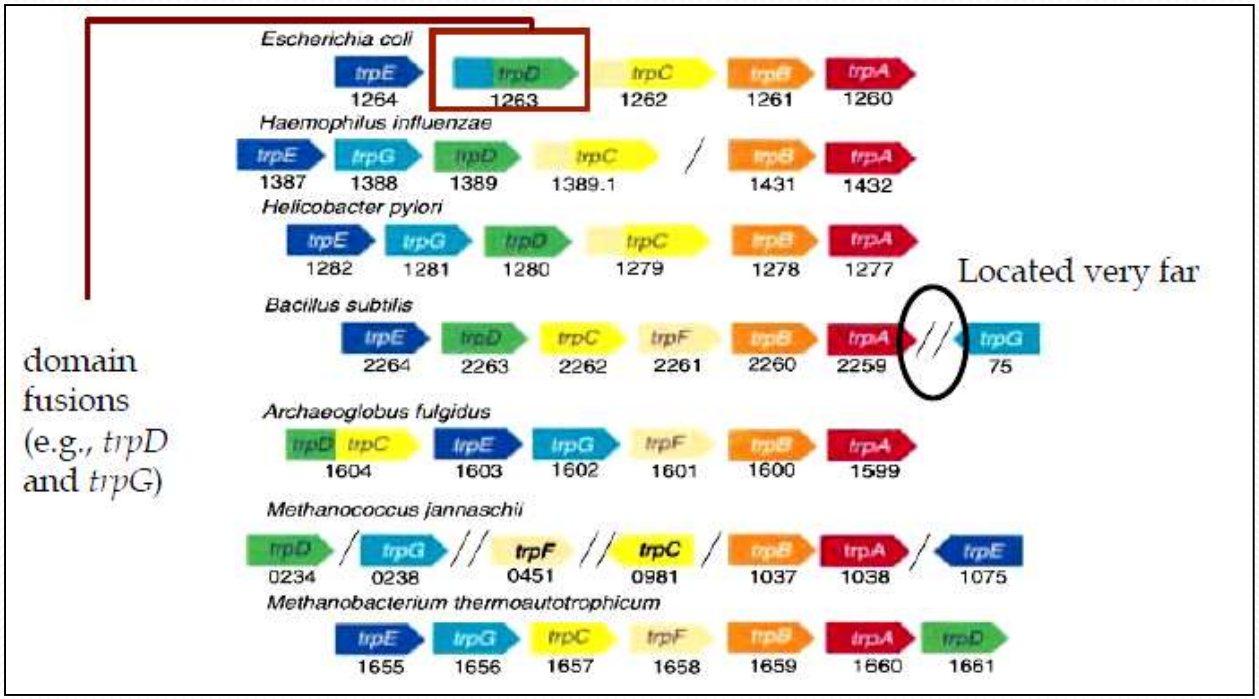
- Genetic analysis has revealed that genes with a related function are frequently found to be clustered at one chromosomal location

- Clustering of related genes presumably provides an evolutionary advantage to a species, but the underlying biological reason is not understood

- **Possibilities:**
  1) There is genetic variation (alleles) within each gene in a cluster of a given species and that only certain allelic combinations of different genes are compatible

  2) Some kind of coordinated translation of the proteins that may aid their folding

Computational analysis of genome arrangements



In bacterial species, genes that act sequentially in a biochemical pathway are frequently found to be adjacent to each other at one chromosomal location. For e.g. *trp* genes are clustered together on the chromosome of *E. coli*

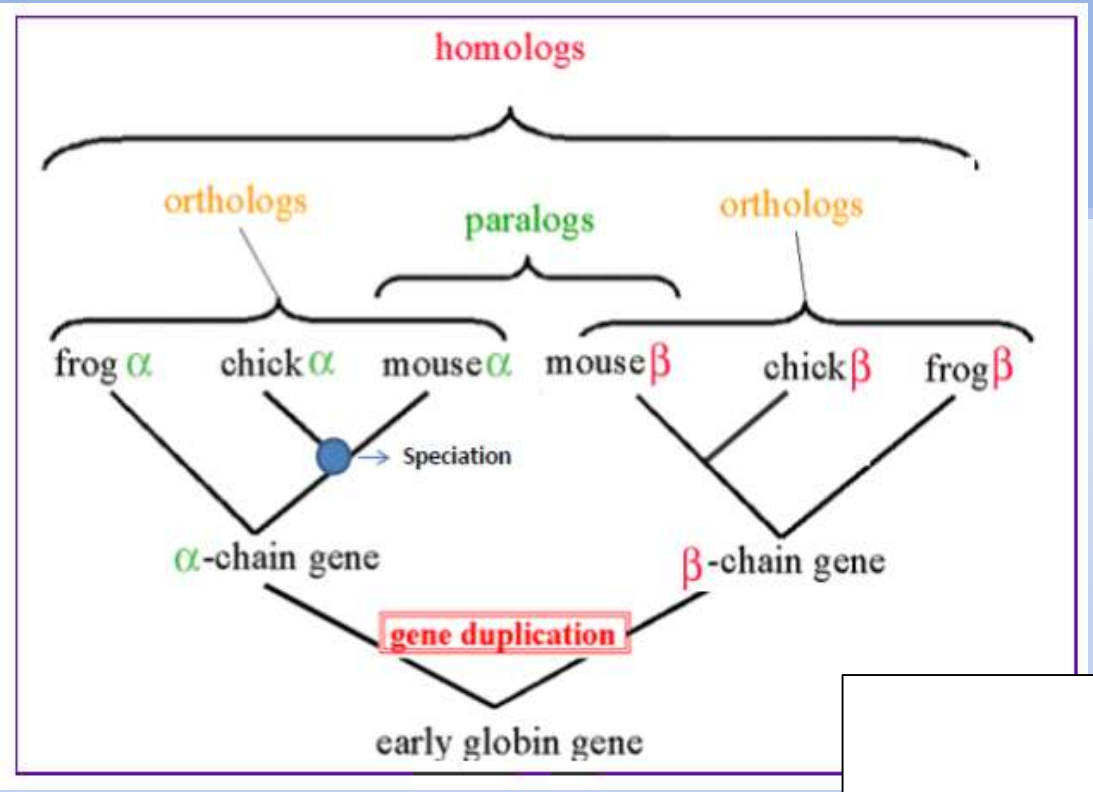
THANK YOU

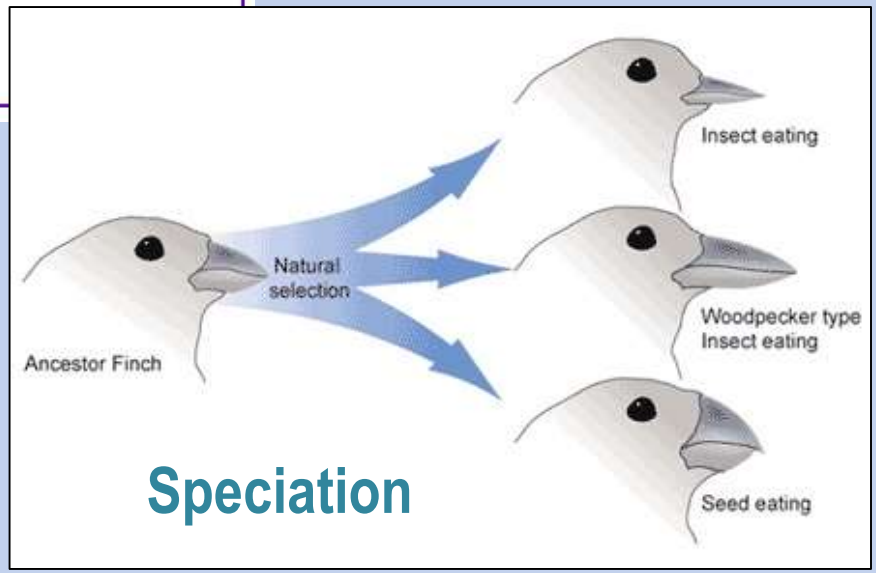
Database of Clusters of Orthologous Genes (COGs)

-Ms. Rupal Mishra

Some Important Terminologies

- **Orthologs** are genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution. Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes.
- **Paralogs** are genes related by duplication within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.
- **Speciation** is the evolutionary process by which populations evolve to become distinct species.


homologs
orthologs
paralogs
orthologs
frog α
chick α
mouse α
mouse β
chick β
frog β
Speciation
α-chain gene
β-chain gene
gene duplication
early globin gene


Insect eating
Natural selection
Woodpecker type Insect eating
Ancestor Finch
Seed eating
Speciation

COGs

- COG - A collection of homologous genes that are useful for study of evolutionary relationships.
- A COG consists of orthologues and paralogues.
- The COG was generated by comparing predicted and known proteins in all completely sequenced microbial genomes to infer sets of orthologs.
- Each COG consists of a group of proteins found to be orthologous across at least three lineages and likely corresponds to an ancient conserved domain.
- Clusters of Orthologous Groups are direct evolutionary counter parts and are considered to be part of an 'ancient conserved domain'.

COGs

- A COG is defined as three or more proteins from the genomes of distant species that are more similar to each other than to any other protein within the individual genome.
- Since the COG database is significantly smaller than the NCBI non-redundant (NR) database, it provides a fast alternative for rapidly describing the functional characteristics of one microbe or a community of microbes.
- Recently, there have been a few successors to the COG db including euKaryotic Orthologous Groups (KOGs) and eggNOG which provide extended analysis of more genomes including eukaryotes.

Applications OF THE COGs

- The most straightforward application of the COGs is for the prediction of functions of individual proteins or protein sets, including those from newly completed genomes.
- COGs can be used to predict the function of homologous proteins in poorly studied species
- Can also be used to track the evolutionary divergence from a common ancestor, hence providing a powerful tool for functional annotation of uncharacterized proteins.

COGs Construction

- The COGs reflect one-to-many and many-to-many orthologous relationships as well as simple one-to-one relationships (hence Orthologous Groups of proteins).
- The original set included the proteins from five bacterial, one archaeal and one eukaryotic genomes and consisted of 720 COGs
- Subsequently, a sixth bacterial genome was added, with the number of COGs increasing to 860 .
- The status of the COG database in 1999 consists of 2091 COGs and includes proteins from 21 complete genomes.

COGs Construction

- COGs have been identified on the basis of an all-against-all sequence comparison of the proteins encoded in complete genomes using the gapped BLAST program after masking low-complexity and predicted coiled-coil regions.
- The COG construction procedure is based on the simple notion that any group of at least three proteins from distant genomes that are more similar to each other than they are to any other proteins from the same genomes are most likely to belong to an orthologous family.

COG construction includes the following steps

1) Perform the all-against-all protein sequence comparison.
2) Detect and collapse obvious paralogs, that is, proteins from the same genome that are more similar to each other than to any proteins from other species.
3) Detect triangles of mutually consistent, genome-specific best hits (BeTs), taking into account the paralogous groups detected at step 2.
4) Merge triangles with a common side to form COGs.
5) A case-by-case analysis of each COG. This analysis serves to eliminate false-positives and to identify groups that contain multi-domain proteins by examining the pictorial representation of the BLAST search outputs
6) Examination of large COGs that include multiple members from all or several of the genomes using phylogenetic trees, cluster analysis and visual inspection of alignments; as a result, some of these groups are split into two or more smaller ones that are included in the final set of COGs.

The COGnitor Program

- New proteins can be assigned to the COGs using the COGnitor program, the principal tool associated with the COGs database.
- COGnitor "BLASTs" the query sequences against all protein sequences encoded in the genomes that are classified in the current release of the COG system.
- To assign proteins to COGs, COGnitor applies the same principle that is embedded in the COG construction procedure, i.e., the consistency of genome-specific Best Hits (BeTs).
- For any given query protein, if the number of BeTs for a particular COG exceeds a predefined cut-off (three by default; the cut-off value can be changed by the user), the query protein is assigned to that COG.

Phyletic Pattern Analysis in COGs

- A phyletic pattern is the pattern of species that are represented in a given COG.
- **Phyletic classifications** are natural classifications that try to identify the evolutionary history of natural groups.
- The COGs show a broad diversity of phyletic patterns i.e., they are represented in all sequenced genomes, whereas COGs present in only three or four species are most abundant.
- This patchy distribution of phyletic patterns probably reflects the major role of horizontal gene transfer and lineage specific gene loss in the evolution of prokaryotes, as well as the rapid evolution of certain genes in specific lineages, which may be linked to functional changes.

Phyletic Pattern Analysis in COGs

- Phyletic patterns are informative not only as indicators of probable evolutionary scenarios but also functionally associated with proteins that have the same phyletic pattern.
- On some occasions, complementary patterns indicate that distinct (sometimes unrelated) proteins are responsible for the same function in different sets of species.
- The COG system includes a simple phyletic pattern search tool that allows the selection of COGs according to species.
- This tool effectively provides the functionality of "differential genome display" (for example, allowing the selection of all COGs that are present in one pair of genomes of interest).

COGs Database

- The Clusters of Orthologous Groups of proteins (COGs) database has been designed as an attempt to classify proteins from completely sequenced genomes on the basis of the orthology concept .
- Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes.
- NCBI provides a COG database that consists of COGs that code for proteins from the genomes of bacteria, archaea and unicellular eukaryotes.
- COGs stands for Clusters of Orthologous Genes also referred to as the Clusters of Orthologous Groups of proteins.
- Available at https://www.ncbi.nlm.nih.gov/research/cog/

COGs Database

- The database was initially created in 1997 followed by several updates, most recently in 2014.
- The current update includes complete genomes of 1,187 bacteria and 122 archaea that map into 1,234 genera.
- The new features include ~250 updated COG annotations with corresponding references and PDB links.
- The current update, substantially expands the scope of the database to include-

COGs	Genomic loci	Taxonomic Categories	Organisms	Protein IDs	COG symbols
4,877	3,456,041	37	1,309	3,213,196	3,821

Current Version

- **The current version of the COGs includes the following new features:**

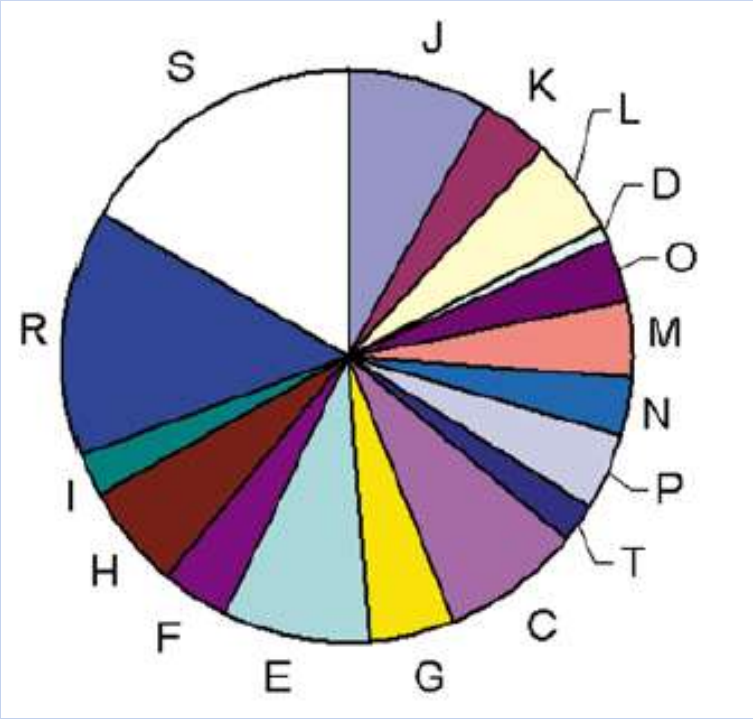
1) The recently deprecated NCBI's gene index (gi) numbers for the encoded proteins are replaced with stable RefSeq or GenBank\ENA\DDBJ coding sequence (CDS) accession numbers.
2) COG annotations are updated for *>200* newly characterized protein families with corresponding references and PDB links, where available.
3) Lists of COGs grouped by pathways and functional systems are added.

Current Version

4) 266 new COGs for proteins involved in CRISPR-Cas immunity, sporulation in Firmicutes and photosynthesis in cyanobacteria are included.
5) The database is made available as a web page.

Functional Category

- The current COG database used is composed of over 4800 COGs.
- While each COG has a specific functional description, it may also have one or more general category letter associations:



One-letter abbreviations for the functional categories

CELLULAR PROCESSES AND SIGNALING

[D] Cell cycle control, cell division, chromosome partitioning

[M] Cell wall/membrane/envelope biogenesis

[N] Cell motility

[O] Post-translational modification, protein turnover, and chaperones

[T] Signal transduction mechanisms

[U] Intracellular trafficking, secretion, and vesicular transport

[V] Defense mechanisms

[W] Extracellular structures

[Y] Nuclear structure

[Z] Cytoskeleton

INFORMATION STORAGE AND PROCESSING

[A] RNA processing and modification

[B] Chromatin structure and dynamics

[J] Translation, ribosomal structure and biogenesis

[K] Transcription

[L] Replication, recombination and repair

METABOLISM

[C] Energy production and conversion

[E] Amino acid transport and metabolism

[F] Nucleotide transport and metabolism

[G] Carbohydrate transport and metabolism

[H] Coenzyme transport and metabolism

[I] Lipid transport and metabolism

[P] Inorganic ion transport and metabolism

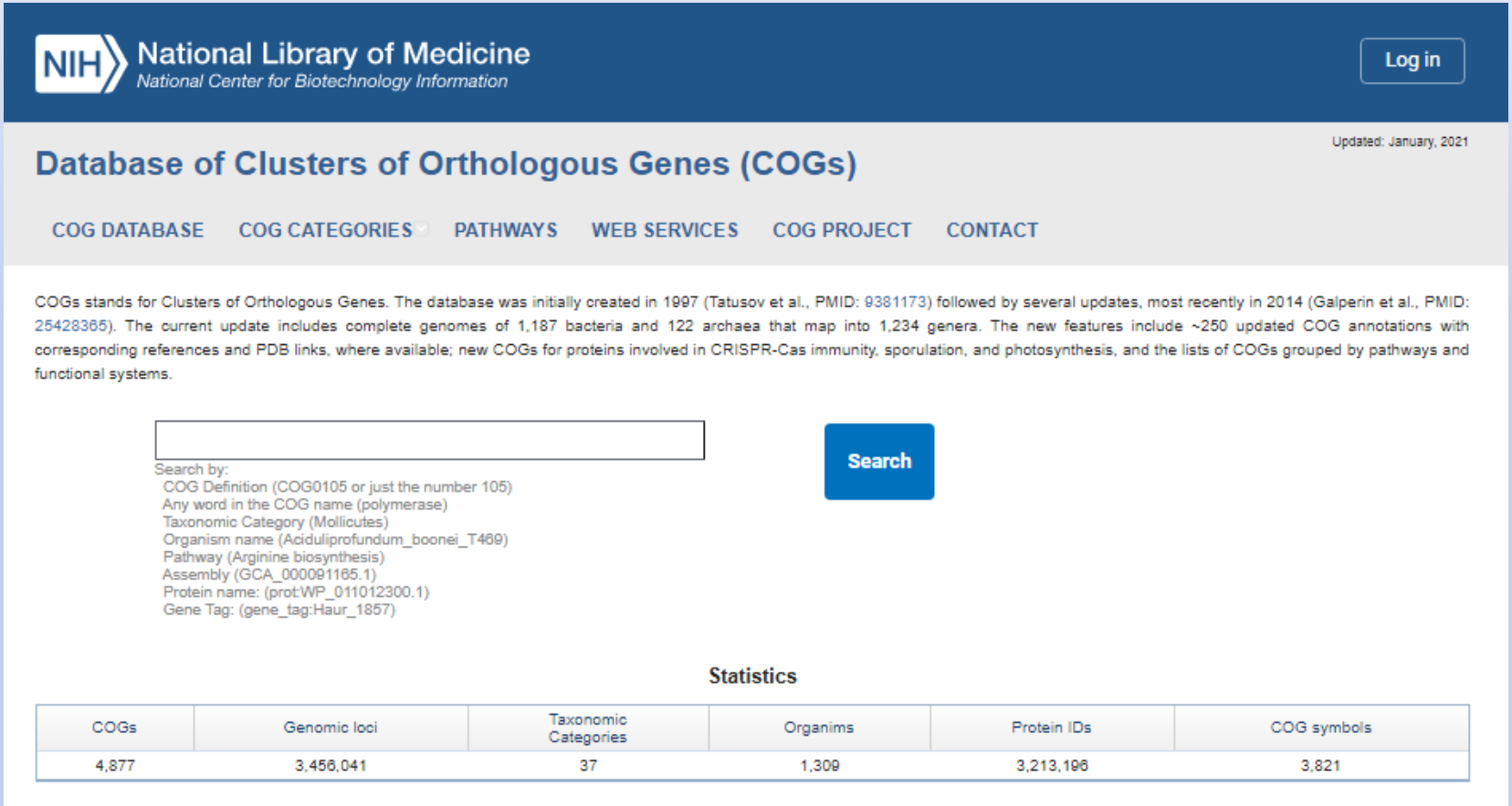
[Q] Secondary metabolites biosynthesis, transport & catabolism

POORLY CHARACTERIZED

[R] General function prediction only

[S] Function unknown

Database


National Library of Medicine
National Center for Biotechnology Information
Log in
Updated: January, 2021
Database of Clusters of Orthologous Genes (COGs)
COG DATABASE
COG CATEGORIES
PATHWAYS
WEB SERVICES
COG PROJECT
CONTACT
COGs stands for Clusters of Orthologous Genes. The database was initially created in 1997 (Tatusov et al., PMID: 9381173) followed by several updates, most recently in 2014 (Galperin et al., PMID: 25428365). The current update includes complete genomes of 1,187 bacteria and 122 archaea that map into 1,234 genera. The new features include ~250 updated COG annotations with corresponding references and PDB links, where available; new COGs for proteins involved in CRISPR-Cas immunity, sporulation, and photosynthesis, and the lists of COGs grouped by pathways and functional systems.
Search
Search by:
COG Definition (COG0105 or just the number 105)
Any word in the COG name (polymerase)
Taxonomic Category (Mollicutes)
Organism name (Aciduliprofundum_boonei_T469)
Pathway (Arginine biosynthesis)
Assembly (GCA_000091165.1)
Protein name: (prot:WP_011012300.1)
Gene Tag: (gene_tag:Haur_1857)
Statistics
COGs
Genomic loci
Taxonomic Categories
Organisms
Protein IDs
COG symbols
4,877
3,456,041
37
1,309
3,213,196
3,821

Categories


COG CATEGORIES
PATHWAYS
WEB SERVICES
COG PROJECT
CONT
J - TRANSLATION, RIBOSOMAL STRUCTURE AND BIOGENESIS
A - RNA PROCESSING AND MODIFICATION
K - TRANSCRIPTION
L - REPLICATION, RECOMBINATION AND REPAIR
B - CHROMATIN STRUCTURE AND DYNAMICS
D - CELL CYCLE CONTROL, CELL DIVISION, CHROMOSOME PARTITIONING
Y - NUCLEAR STRUCTURE
V - DEFENSE MECHANISMS
T - SIGNAL TRANSDUCTION MECHANISMS
M - CELL WALL/MEMBRANE/ENVELOPE BIOGENESIS
N - CELL MOTILITY
Z - CYTOSKELETON
W - EXTRACELLULAR STRUCTURES
U - INTRACELLULAR TRAFFICKING, SECRETION, AND VESICULAR TRANSPORT
O - POSTTRANSLATIONAL MODIFICATION, PROTEIN TURNOVER, CHAPERONES
X - MOBILOME: PROPHAGES, TRANSPOSONS

Pathways

National Library of Medicine
National Center for Biotechnology Information

Log in

Updated: January, 2021

Database of Clusters of Orthologous Genes (COGs)

COG DATABASE | COG CATEGORIES | PATHWAYS | WEB SERVICES | COG PROJECT | CONTACT

Pathways

Download

Pathway	No. COGs
16S rRNA modification	15
23S rRNA modification	12
A/V-type ATP synthase	8
Aminoacyl-tRNA synthetases	26
Archaeal ribosomal proteins	33
Arginine biosynthesis	12
Aromatic amino acid biosynthesis	23
Asparagine biosynthesis	2
Biotin biosynthesis	8
CRISPR-Cas system	46
Cobalamine/B12 biosynthesis	24
Cysteine biosynthesis	8
Entner-Doudoroff pathway	1
Fatty acids biosynthesis	15
FoF1-type ATP synthase	12
Folate biosynthesis	9

Web Services

National Library of Medicine
National Center for Biotechnology Information

Log in

Updated: January, 2021

Database of Clusters of Orthologous Genes (COGs)

COG DATABASE | COG CATEGORIES | PATHWAYS | WEB SERVICES | COG PROJECT | CONTACT

Webs Services

COG

1. Get all COGs:
   Web API: https://www.ncbi.nlm.nih.gov/research/cog/api/cog/
   JSON Format: https://www.ncbi.nlm.nih.gov/research/cog/api/cog/?format=json
2. Filter COGs by gene tag: MK0280
   Web API: https://www.ncbi.nlm.nih.gov/research/cog/api/cog/?gene=MK0280
   JSON Format: https://www.ncbi.nlm.nih.gov/research/cog/api/cog/?gene=MK0280&format=json
3. Filter COGs by COG ID tag: COG0003
   Web API: https://www.ncbi.nlm.nih.gov/research/cog/api/cog/?cog=COG0003
   JSON Format: https://www.ncbi.nlm.nih.gov/research/cog/api/cog/?cog=COG0003&format=json
4. Filter COGs by assembly ID: GCA_000007185.1
   Web API: https://www.ncbi.nlm.nih.gov/research/cog/api/cog/?assembly=GCA_000007185.1
   JSON Format: https://www.ncbi.nlm.nih.gov/research/cog/api/cog/?assembly=GCA_000007185.1&format=json
5. Filter COGs by organism name: Nitrosopumilus_maritimus_SCM1

COG Project

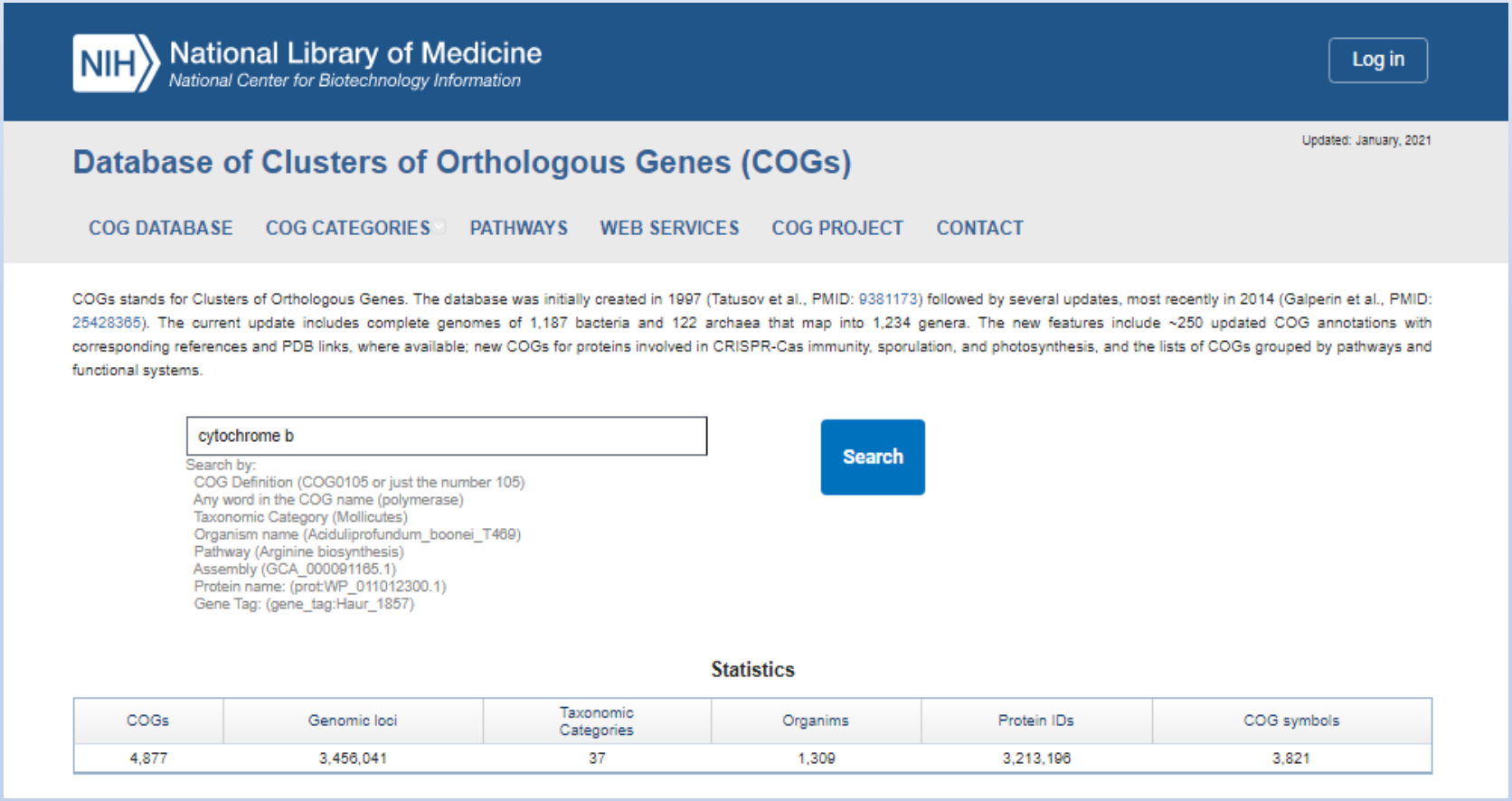
COG links

- 2020 COGs update [Web interface] NEW
- 2020 COGs update [FTP] NEW
- 2003 COGs, 2014 update [Web interface] (superceded by 2020 update)
- 2003 COGs, 2014 update [FTP]
- 2003 COGs, original format [FTP]
- 2003 KOGs, original format [FTP]
- 2003 COGs [FTP]
- arCOGs [FTP]
- NCVOGs [FTP]
- mimiCOGs [FTP]
- 2013 POGs [FTP]
- 2011 POGs, annotated [FTP]
- 2011 POGs, extended [FTP]
- COG software [FTP]

Publications

- COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* 2020 Nov 9; gkaa1018. NEW
- 2019 The COG approach. *Brief Bioinform.* 2019 Jul 19;20(4):1063-1070.
- 2017 ATGC-COGs. *Nucleic Acids Res.* 2017 Jan 4;45(D1):D210-D218.
- 2014 archaeal COGs. *Life* 2015 Mar 10;5(1):818-840.
- 2014 update of 2003 COGs. *Nucleic Acids Res.* 2015 Jan;43:D261-D269.
- mimiCOGs. *Virol J.* 2013 Apr 4;10:106.
- 2013 phage COGs. *J Bacteriol.* 2013 Mar;195(5):941-950.
- 2012 archaeal COGs. *Biol Direct* 2012 Dec 14;7:46.
- Orthologs and BBH. *Genome Biol Evol.* 2012 Jan;4(12):1286-1294.
- 2011 phage COGs. *J Bacteriol.* 2011 Apr;193(8):1806-1814.
- Improved COG algorithm. *Bioinformatics* 2010 Jun 15;26(12):1481-1487.
- NCLDV COGs. *Virol J.* 2009 Dec 17;6:223.
- 2007 archaeal COGs. *Biol Direct* 2007 Nov 27;2:33.
- Lactic acid bacteria COGs. *Proc Natl Acad Sci U.S.A.* 2006 Oct 17;103(42):15611-15616.
- Cyanobacterial COGs. *Proc Natl Acad Sci U.S.A.* 2006 Aug 29;103(35):13126-13131.
- 2003 eukaryotic KOGs. *Genome Biol.* 2004 Jan 15;5(2):R7.
- 2003 database update. *BMC Bioinformatics* 2003 Sep 11;4(1):41
- Original COG paper. *Science* 1997 Oct 24;278(5338):631-7

Example


National Library of Medicine
National Center for Biotechnology Information
Log in
Updated: January, 2021
Database of Clusters of Orthologous Genes (COGs)
COG DATABASE
COG CATEGORIES
PATHWAYS
WEB SERVICES
COG PROJECT
CONTACT
COGs stands for Clusters of Orthologous Genes. The database was initially created in 1997 (Tatusov et al., PMID: 9381173) followed by several updates, most recently in 2014 (Galperin et al., PMID: 25428365). The current update includes complete genomes of 1,187 bacteria and 122 archaea that map into 1,234 genera. The new features include ~250 updated COG annotations with corresponding references and PDB links, where available; new COGs for proteins involved in CRISPR-Cas immunity, sporulation, and photosynthesis, and the lists of COGs grouped by pathways and functional systems.
cytochrome b
Search
Search by:
COG Definition (COG0105 or just the number 105)
Any word in the COG name (polymerase)
Taxonomic Category (Mollicutes)
Organism name (Aciduliprofundum_boonei_T469)
Pathway (Arginine biosynthesis)
Assembly (GCA_000091165.1)
Protein name: (prot:WP_011012300.1)
Gene Tag: (gene_tag:Haur_1857)
Statistics
COGs
Genomic loci
Taxonomic Categories
Organisms
Protein IDs
COG symbols
4,877
3,456,041
37
1,309
3,213,196
3,821

Results

Organism	Protein	COG	Cat	Annotation
677	712	COG0762	O	Cytochrome b6 maturation protein CCB3/Ycf19 and related maturases, YggT family
822	1100	COG1271	C	Cytochrome bd-type quinol oxidase, subunit 1
710	898	COG1290	C	Cytochrome b subunit of the bc complex
795	1032	COG1294	C	Cytochrome bd-type quinol oxidase, subunit 2
152	166	COG1969	C	Ni,Fe-hydrogenase I cytochrome b subunit
647	667	COG2009	C	Succinate dehydrogenase/fumarate reductase, cytochrome b subunit
541	756	COG2193	P	Bacterioferritin (cytochrome b1)
295	405	COG2864	C	Cytochrome b subunit of formate dehydrogenase

Updated: January, 2021

Database of Clusters of Orthologous Genes (COGs)

P - COG2193 - Bacterioferritin (cytochrome b1)

COG symbol: Bfr
PDB entry: 4E6K

Definition to JSON | COGs to JSON

Basic Stats

Organisms: 541/1309
Genes: 758/3456041
Proteins: 756/3213196
Median Protein Length: 160.58

Taxonomy Categories

Click on table rows for more information

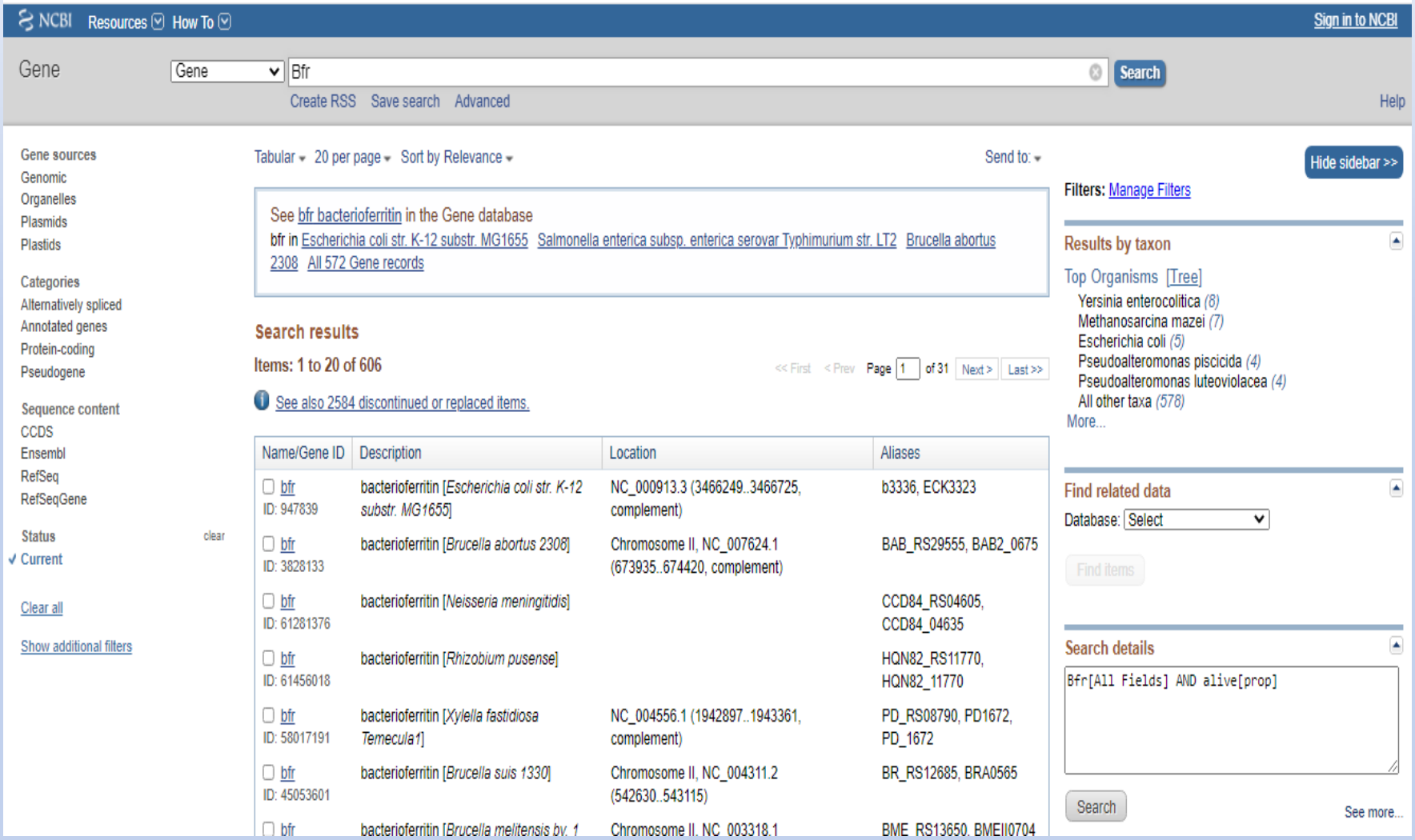
ARCHAEA

- CRENARCHAEOTA [0/25 organisms 0 genes]
- EURYARCHAEOTA [6/79 organisms 6 genes]

BACTERIA

- ACIDOBACTERIA [4/7 organisms 7 genes]
- ACTINOBACTERIA [35/155 organisms 41 genes]
- AQUIFICAE [6/9 organisms 6 genes]
- BACTEROIDETES [2/107 organisms 2 genes]
- CHLAMYDIAE [0/6 organisms 0 genes]
- CHLOROBI [0/5 organisms 0 genes]
- CHLOROFLEXI [8/14 organisms 9 genes]

COG Symbol


NCBI Resources How To
Sign in to NCBI
Gene
Gene
Bfr
Search
Create RSS Save search Advanced
Help
Gene sources
Genomic
Organelles
Plasmids
Plastids
Categories
Alternatively spliced
Annotated genes
Protein-coding
Pseudogene
Sequence content
CCDS
Ensembl
RefSeq
RefSeqGene
Status clear
Current
Clear all
Show additional filters
Tabular 20 per page Sort by Relevance
Send to:
See bfr bacterioferritin in the Gene database
bfr in Escherichia coli str. K-12 substr. MG1655 Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 Brucella abortus 2308 All 572 Gene records
Search results
Items: 1 to 20 of 606
<< First < Prev Page 1 of 31 Next > Last >>
See also 2584 discontinued or replaced items.
Name/Gene ID
Description
Location
Aliases
bfr ID: 947839
bacterioferritin [Escherichia coli str. K-12 substr. MG1655]
NC_000913.3 (3466249..3466725, complement)
b3336, ECK3323
bfr ID: 3828133
bacterioferritin [Brucella abortus 2308]
Chromosome II, NC_007624.1 (673935..674420, complement)
BAB_RS29555, BAB2_0675
bfr ID: 61281376
bacterioferritin [Neisseria meningitidis]
CCD84_RS04605, CCD84_04635
bfr ID: 61456018
bacterioferritin [Rhizobium pusense]
HQN82_RS11770, HQN82_11770
bfr ID: 58017191
bacterioferritin [Xylella fastidiosa Temecula1]
NC_004556.1 (1942897..1943361, complement)
PD_RS08790, PD1672, PD_1672
bfr ID: 45053601
bacterioferritin [Brucella suis 1330]
Chromosome II, NC_004311.2 (542630..543115)
BR_RS12685, BRA0565
bfr
bacterioferritin [Brucella melitensis bv. 1
Chromosome II, NC_003318.1
BME_RS13650, BMEII0704
Hide sidebar >>
Filters: Manage Filters
Results by taxon
Top Organisms [Tree]
Yersinia enterocolitica (8)
Methanosarcina mazei (7)
Escherichia coli (5)
Pseudoalteromonas piscicida (4)
Pseudoalteromonas luteoviolacea (4)
All other taxa (578)
More...
Find related data
Database: Select
Find items
Search details
Bfr[All Fields] AND alive[prop]
Search
See more...

PDB Entry


RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
Documentation
MyPDB
RCSB PDB PROTEIN DATA BANK
178229 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education
Enter search term(s)
Advanced Search | Browse Annotations
Help
PDB-101
EMDataResource
Nucleic Acid Database
Worldwide Protein Data Bank Foundation
Celebrating
YEARS OF Protein Data Bank
Structure Summary
3D View
Annotations
Experiment
Sequence
Genome
Versions
Biological Assembly 1
Display Files
Download Files
4E6K
2.0 A resolution structure of Pseudomonas aeruginosa bacterioferritin (BfrB) in complex with bacterioferritin associated ferredoxin (Bfd)
DOI: 10.2210/pdb4E6K/pdb
Classification: METAL BINDING PROTEIN/ELECTRON TRANSPORT
Organism(s): Pseudomonas aeruginosa PAO1
Expression System: Escherichia coli
Mutation(s): Yes
Deposited: 2012-03-15 Released: 2012-08-01
Deposition Author(s): Lovell, S., Battaile, K.P., Yao, H., Wang, Y., Kumar, R., Ruvinsky, A., Vasker, I., Rivera, M.

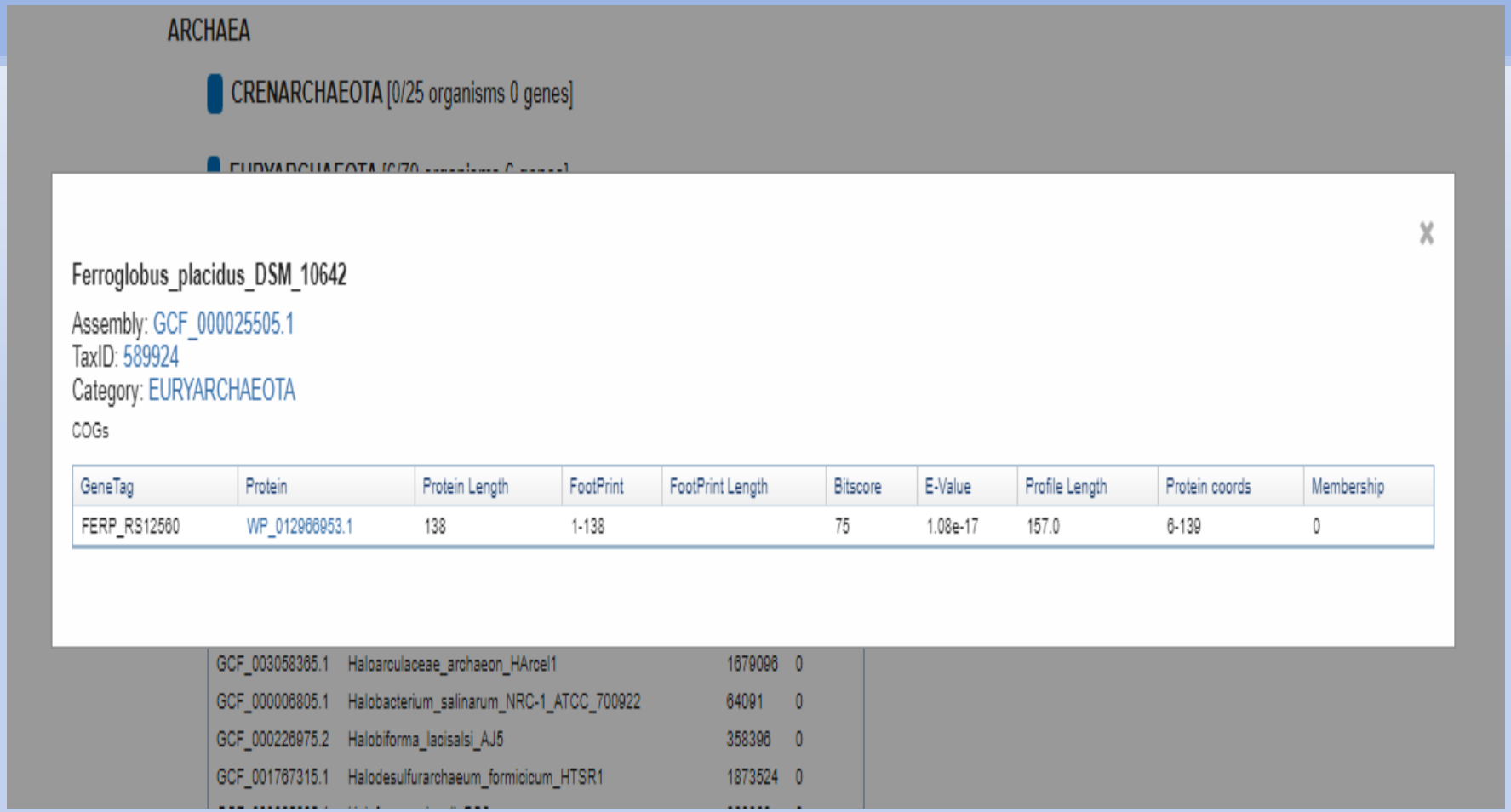
ARCHAEA

+ **CRENARCHAEOTA** [0/25 organisms 0 genes]

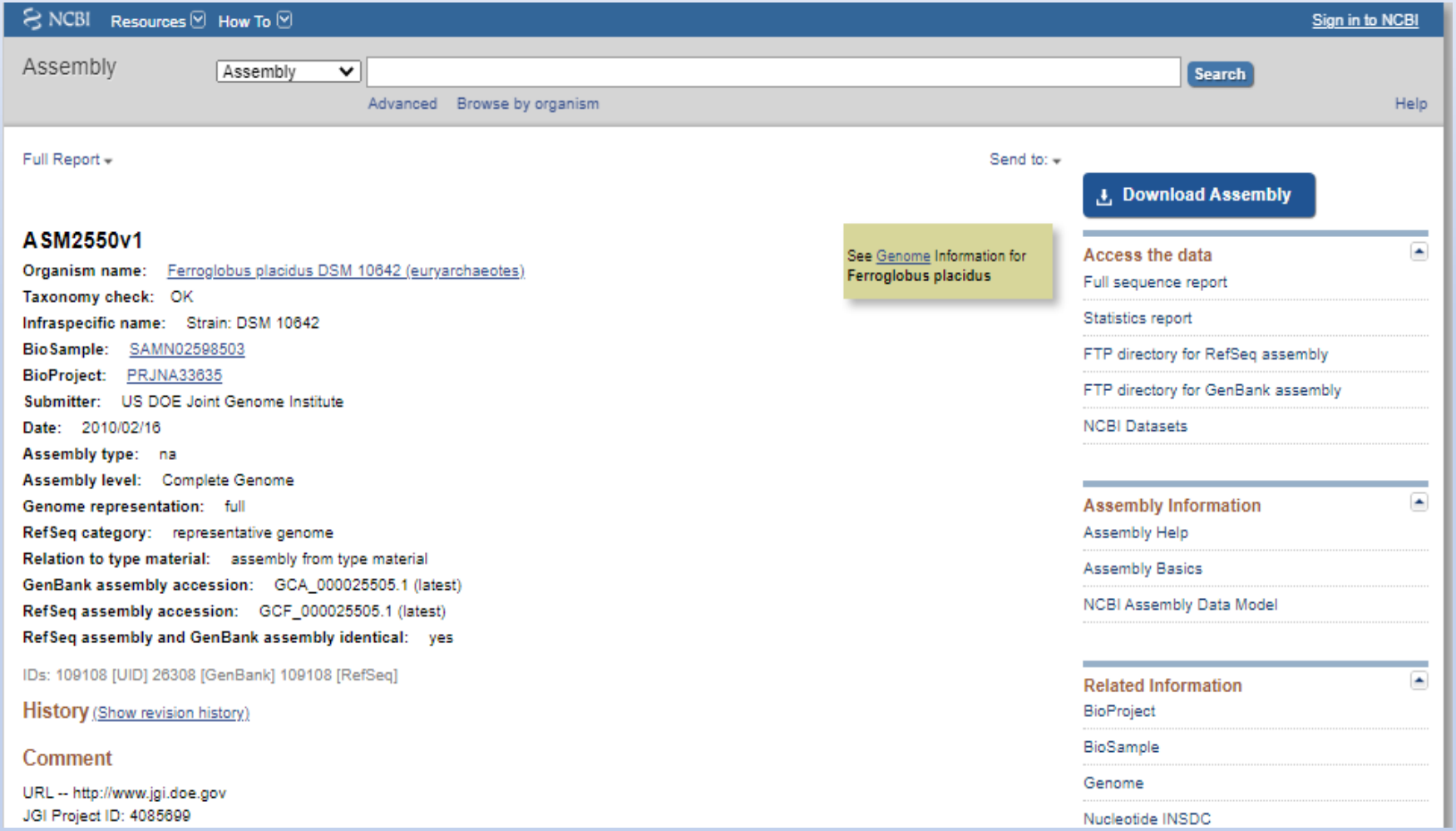
- **EURYARCHAEOTA** [6/79 organisms 6 genes]

Download

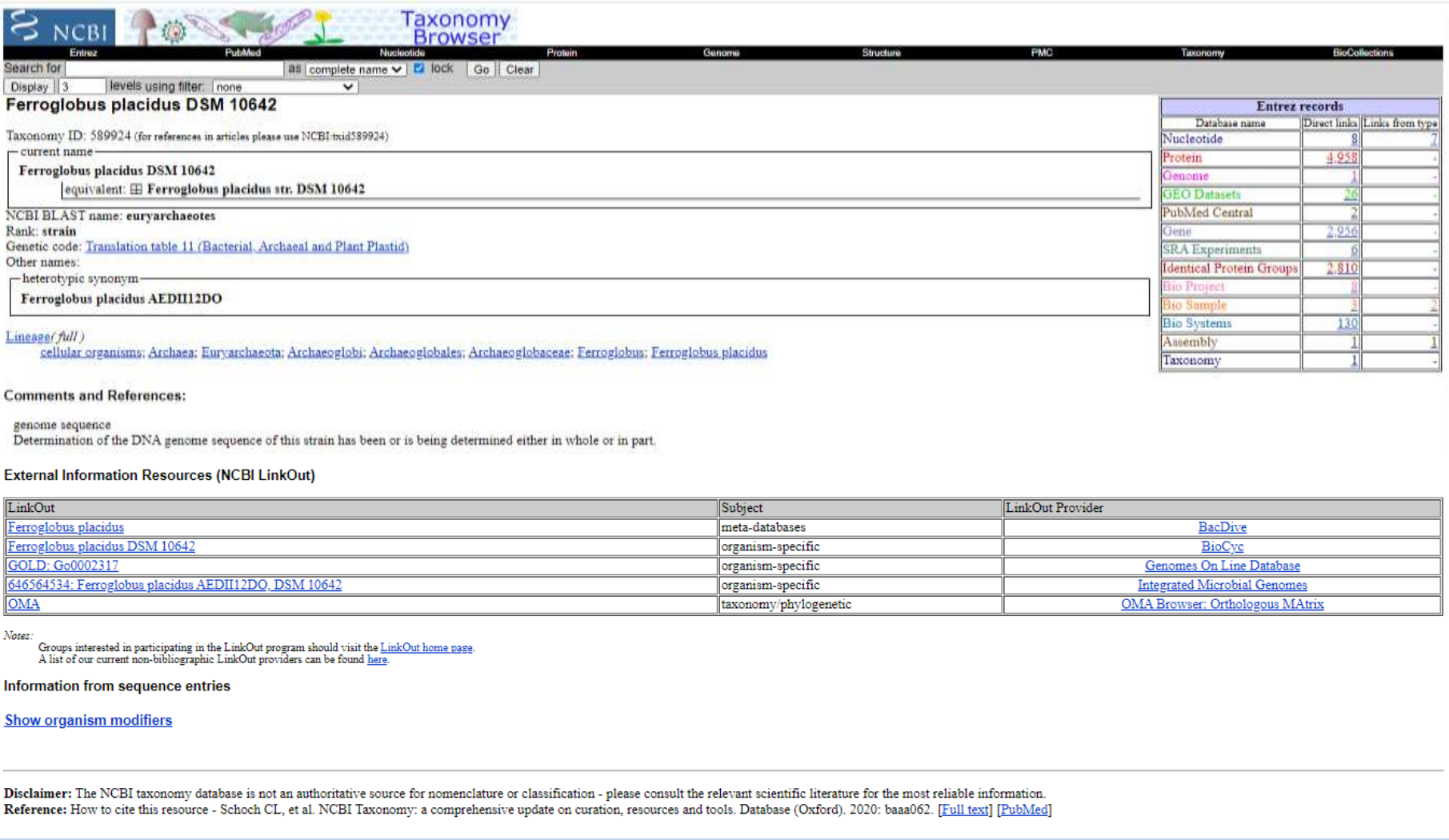
Assembly	Genome	TaxID	No. genes
GCF_000008665.1	Archaeoglobus_fulgidus_DSM_4304	224325	0
GCF_000025505.1	Ferroglobus_placidus_DSM_10642	589924	1
GCF_000789255.1	Geoglobus_acetivorans_SBH6	565033	0
GCF_000025665.1	Aciduliprofundum_boonei_T469	439481	0
GCF_000196895.1	Halalkalicoccus_jeotgali_B3	795797	0
GCF_001011115.1	Halanaeroarchaeum_sulfurireducens_HSR2	1604004	0
GCF_000011085.1	Haloarcula_marismortui_ATCC_43049	272569	0
GCF_003058365.1	Haloarculaceae_archaeon_HArcel1	1679096	0
GCF_000006805.1	Halobacterium_salinarum_NRC-1_ATCC_700922	64091	0
GCF_000226975.2	Halobiforma_lacisalsi_AJ5	358396	0
GCF_001767315.1	Halodesulfurarchaeum_formicicum_HTSR1	1873524	0
GCF_000025685.1	Haloferax_volcanii_DS2	309800	0
GCF_000172995.2	Halogeometricum_borinquense_DSM_11551_PR_3	469382	0
GCF_002788215.1	Halohasta_litchfieldiae_tADL	1073996	0
GCF_000023965.1	Halomicrobium_mukohataei_DSM_12286	485914	0
GCF_002355635.1	Halopenitus_persicus_CBA1233	1048396	0


ARCHAEA
CRENARCHAEOTA [0/25 organisms 0 genes]
Ferroglobus_placidus_DSM_10642
Assembly: GCF_000025505.1
TaxID: 589924
Category: EURYARCHAEOTA
COGs
GeneTag Protein Protein Length FootPrint FootPrint Length Bitscore E-Value Profile Length Protein coords Membership
FERP_RS12560 WP_012966953.1 138 1-138 75 1.08e-17 157.0 6-139 0
GCF_003058365.1 Haloarculaceae_archaeon_HArcel1 1679096 0
GCF_000006805.1 Halobacterium_salinarum_NRC-1_ATCC_700922 64091 0
GCF_000226975.2 Halobiforma_lacisalsi_AJ5 358396 0
GCF_001767315.1 Halodesulfurarchaeum_formicicum_HTSR1 1873524 0

Assembly


NCBI Resources How To Sign in to NCBI
Assembly
Assembly
Search
Advanced Browse by organism Help
Full Report
Send to:
Download Assembly
ASM2550v1
Organism name: Ferroglobus placidus DSM 10642 (euryarchaeotes)
Taxonomy check: OK
Infraspecific name: Strain: DSM 10642
BioSample: SAMN02598503
BioProject: PRJNA33635
Submitter: US DOE Joint Genome Institute
Date: 2010/02/16
Assembly type: na
Assembly level: Complete Genome
Genome representation: full
RefSeq category: representative genome
Relation to type material: assembly from type material
GenBank assembly accession: GCA_000025505.1 (latest)
RefSeq assembly accession: GCF_000025505.1 (latest)
RefSeq assembly and GenBank assembly identical: yes
See Genome Information for Ferroglobus placidus
IDs: 109108 [UID] 26308 [GenBank] 109108 [RefSeq]
History (Show revision history)
Comment
URL -- http://www.jgi.doe.gov
JGI Project ID: 4085699
Access the data
Full sequence report
Statistics report
FTP directory for RefSeq assembly
FTP directory for GenBank assembly
NCBI Datasets
Assembly Information
Assembly Help
Assembly Basics
NCBI Assembly Data Model
Related Information
BioProject
BioSample
Genome
Nucleotide INSDC

Taxonomy



NCBI Taxonomy Browser

Entrez | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | BioCollections

Search for [ ] as [complete name] [lock] [Go] [Clear]

Display [3] levels using filter: [none]

Ferroglobus placidus DSM 10642

Taxonomy ID: 589924 (for references in articles please use NCBI:txid589924)

current name

Ferroglobus placidus DSM 10642

equivalent: ⊞ **Ferroglobus placidus str. DSM 10642**

NCBI BLAST name: **euryarchaeotes**

Rank: **strain**

Genetic code: Translation table 11 (Bacterial, Archaeal and Plant Plastid)

Other names:

heterotypic synonym

Ferroglobus placidus AEDII12DO

Lineage (*full*)

cellular organisms; Archaea; Euryarchaeota; Archaeoglobi; Archaeoglobales; Archaeoglobaceae; Ferroglobus; Ferroglobus placidus

Entrez records

Database name	Direct links	Links from type
Nucleotide	8	7
Protein	4,958	-
Genome	1	-
GEO Datasets	76	-
PubMed Central	2	-
Gene	2,956	-
SRA Experiments	6	-
Identical Protein Groups	2,810	-
Bio Project	8	-
Bio Sample	3	2
Bio Systems	130	-
Assembly	1	1
Taxonomy	1	-

Comments and References:

genome sequence

Determination of the DNA genome sequence of this strain has been or is being determined either in whole or in part.

External Information Resources (NCBI LinkOut)

LinkOut	Subject	LinkOut Provider
Ferroglobus placidus	meta-databases	BacDive
Ferroglobus placidus DSM 10642	organism-specific	BioCyc
GOLD: Go0002317	organism-specific	Genomes On Line Database
646564534: Ferroglobus placidus AEDII12DO, DSM 10642	organism-specific	Integrated Microbial Genomes
OMA	taxonomy/phylogenetic	OMA Browser: Orthologous MAtrix

Notes:

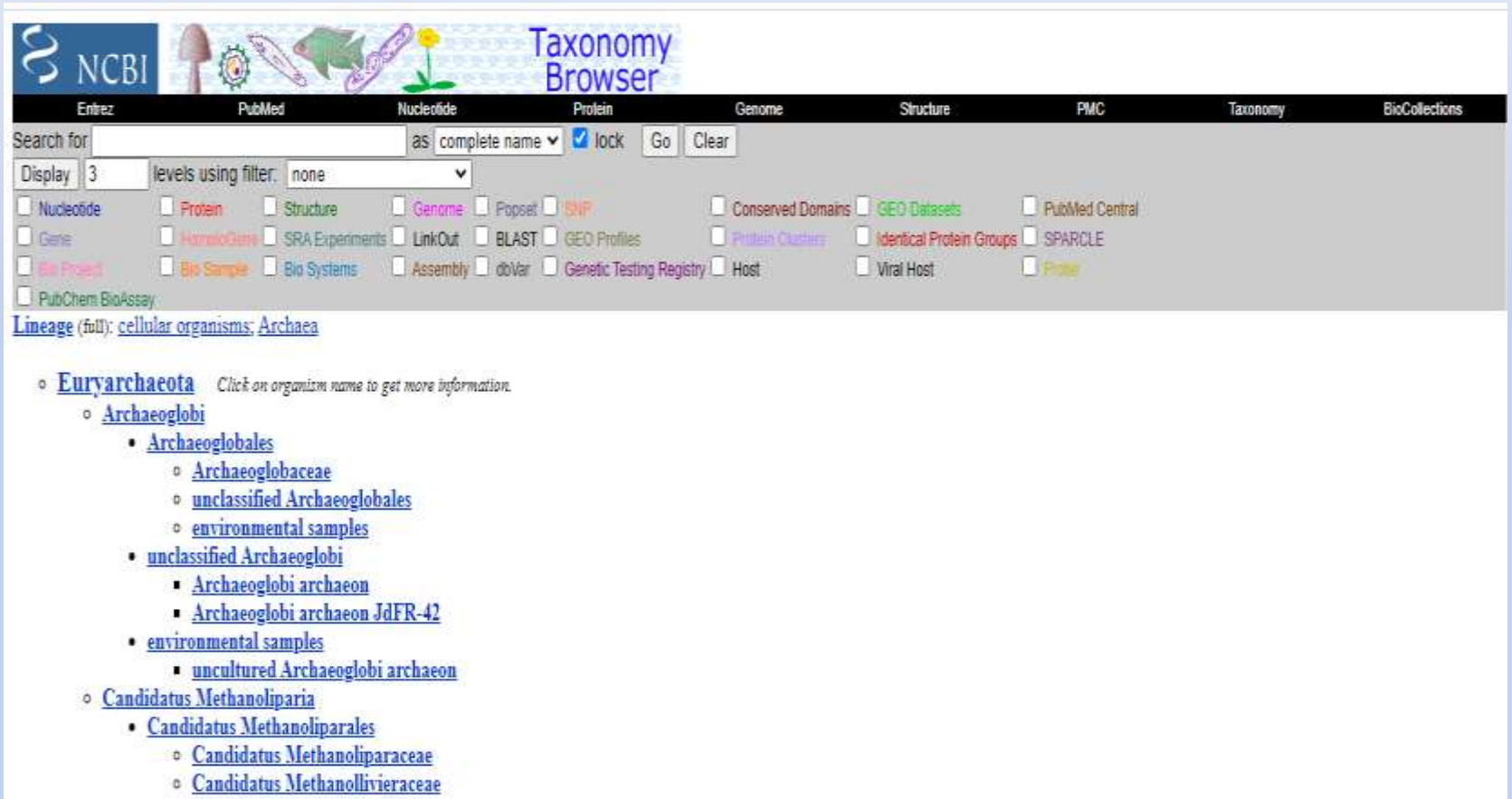
Groups interested in participating in the LinkOut program should visit the LinkOut home page.
A list of our current non-bibliographic LinkOut providers can be found here.

Information from sequence entries

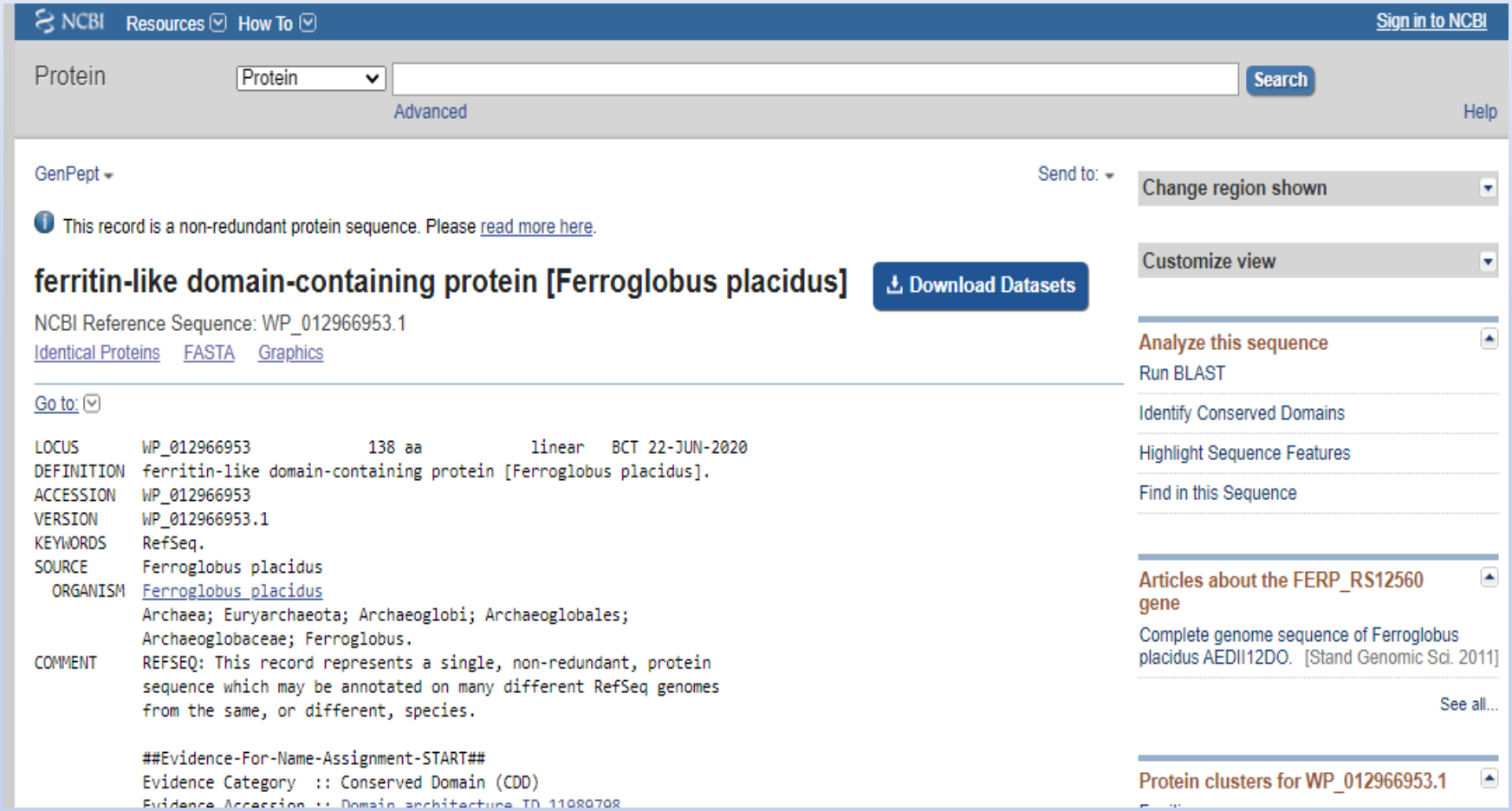
Show organism modifiers

Disclaimer: The NCBI taxonomy database is not an authoritative source for nomenclature or classification - please consult the relevant scientific literature for the most reliable information.
Reference: How to cite this resource - Schoch CL, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford). 2020: baaa062. [Full text] [PubMed]

Category


NCBI
Taxonomy Browser
Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy BioCollections
Search for as complete name lock Go Clear
Display 3 levels using filter: none
Nucleotide Protein Structure Genome Popset SNP Conserved Domains GEO Datasets PubMed Central
Gene HomoloGene SRA Experiments LinkOut BLAST GEO Profiles Protein Clusters Identical Protein Groups SPARCLE
Bio Project Bio Sample Bio Systems Assembly dbVar Genetic Testing Registry Host Viral Host Probe
PubChem BioAssay
Lineage (full): cellular organisms; Archaea
Euryarchaeota Click on organism name to get more information.
Archaeoglobi
Archaeoglobales
Archaeoglobaceae
unclassified Archaeoglobales
environmental samples
unclassified Archaeoglobi
Archaeoglobi archaeon
Archaeoglobi archaeon JdFR-42
environmental samples
uncultured Archaeoglobi archaeon
Candidatus Methanoliparia
Candidatus Methanoliparales
Candidatus Methanoliparaceae
Candidatus Methanollivieraceae

Cog Protein ID

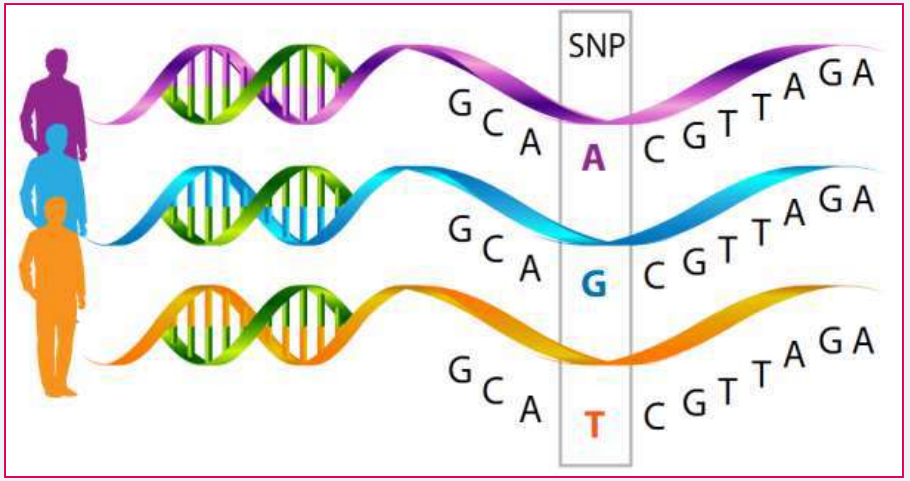


THANK YOU

-Ms. Rupal Mishra

Introduction

- **Polymorphism** is a generic term that means 'many shapes'. It is the ability to appear in different form .
- A **Single Nucleotide Polymorphism (SNP)** is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome differs between members of a species (or between paired chromosomes in an individual).
- Single nucleotide polymorphisms or SNP (pronounced “snips”), are the most common type of genetic variation among peoples.



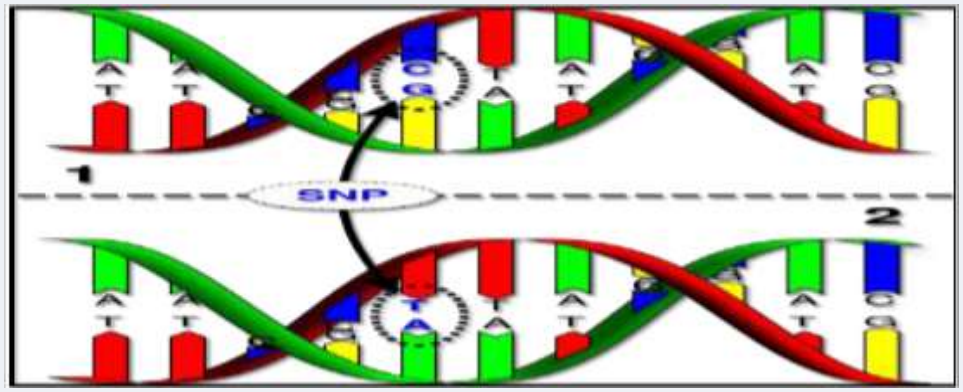
SNP

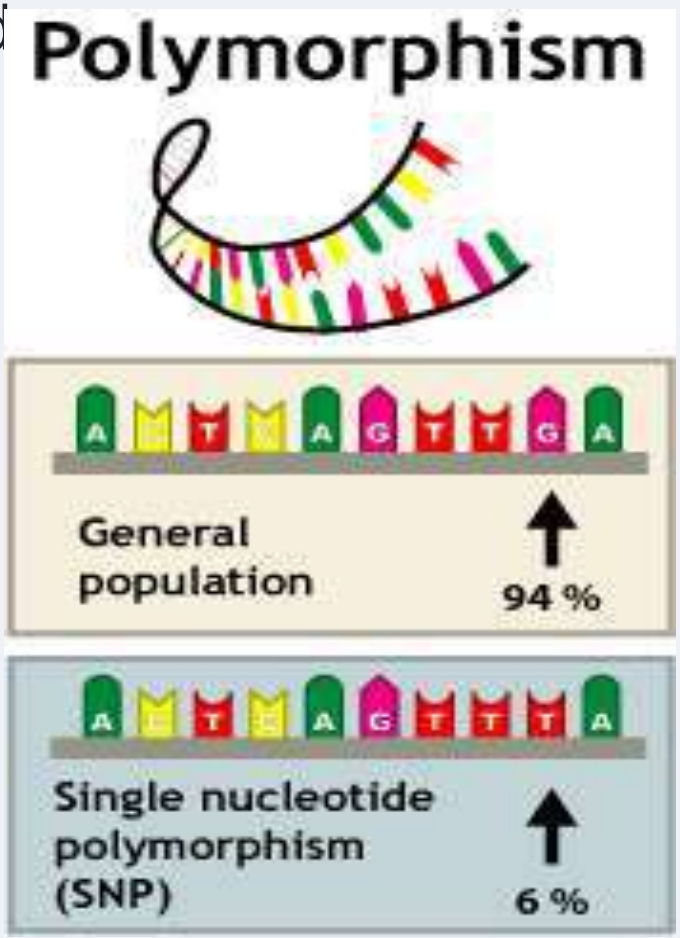
- Each SNP represents a difference in a single DNA building block, called a nucleotide .
- For a variation to be considered a SNP, it must occur in at least 1% of the population.
- For **example**, two sequenced DNA fragments from different individuals, **AAGCCTA** to **AAGCTTA,** contain a difference in a single nucleotide .
- They occur once in every 300~600 nucleosides on average , which means 10 million SNPs in the human genome .
- Most commonly these variations are found in DNA between genes .

SNP

- The **nucleotide on SNP locus** is called
  - a major allele
  - a minor allele

94% ---- A C T T A G C T G - G : major allele
6% ---- A C T T A G C T T – T : minor allele





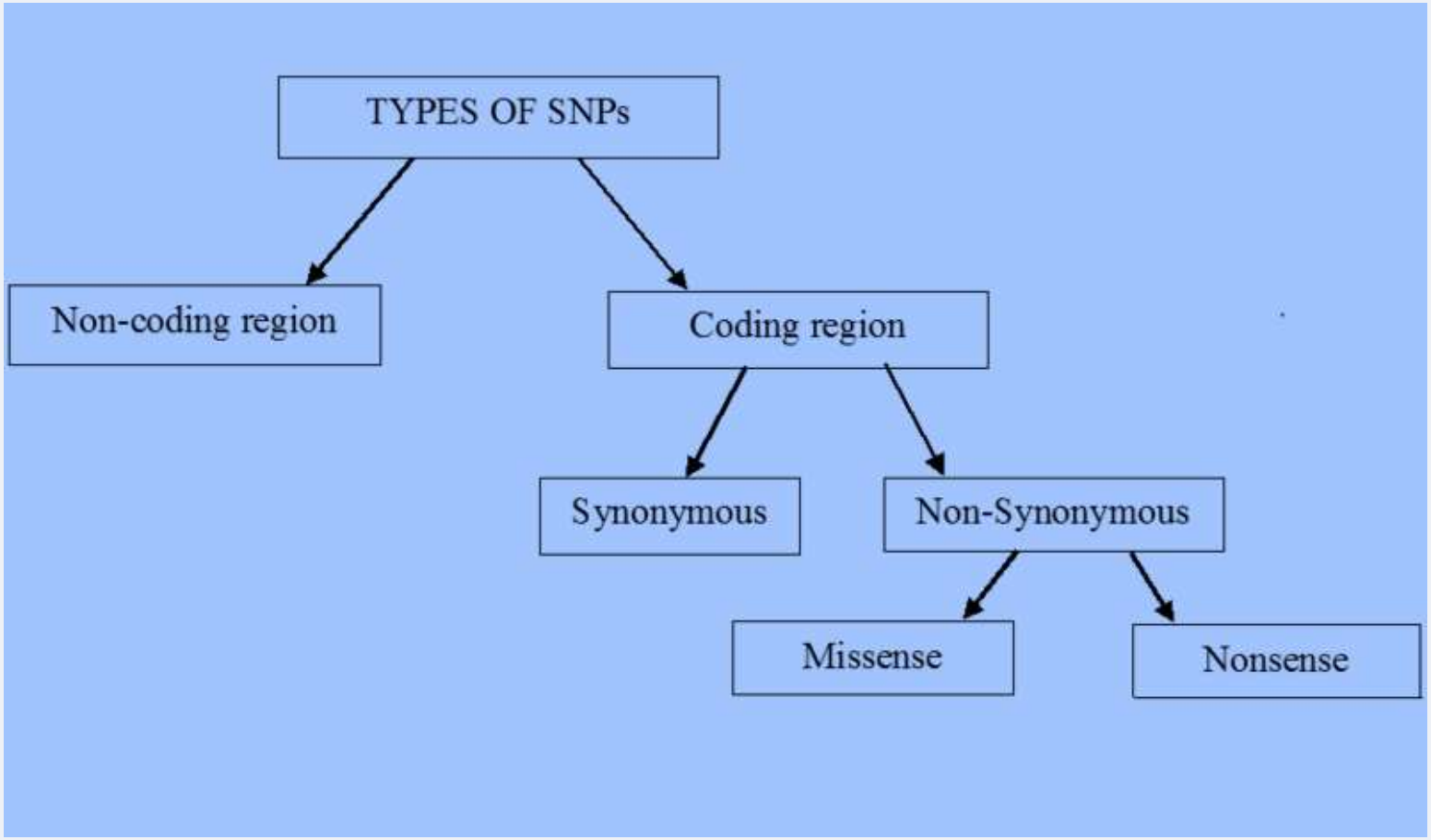
Characteristics Of SNP

- In **human beings**, 99.9 percent bases are same. Remaining 0.1 percent makes a person unique.
- Different attributes / characteristics / traits ~> How a person looks, diseases he or she develops.
- These **variations can be**:
  Harmless (change in phenotype)
  Harmful (diabetes, cancer, heart disease, and hemophilia)
- The abundance of SNPs and the ease with which they can be measured make these genetic variations significant.
- SNPs close to particular gene acts as a marker for that gene.
- SNPs in coding regions may alter the protein structure made by that coding region.

Types of SNPs

1) **Non-coding region -** A segment of DNA that does comprise a gene and does not code for a *protein* .
2) **Coding region -** Regions of DNA/RNA sequence that code for proteins
   a) **SYNONYMOUS** - A SNP in which both forms lead to the same polypeptide sequence is called synonymous (sometimes called silent mutations)
   b) **NON-SYNONYMOUS -** If a different polypeptide sequence is produced they are non synonymous. A non synonymous change may either be **missense or nonsense.**
      i. a missense change results in a different amino acid .
      ii. a nonsense change results in a premature *stop codon* .

Types of SNPs



SNP Application

- Genetic variation
- Diagnostics
- Risk profiling
- Human genetic study
- Personalized medicines
- Genetic marker

dbSNP

- **dbSNP** is world's largest database for nucleotide variations, and is part of NCBI.
- dbSNP **contains** human single nucleotide variations, microsatellites, and small-scale insertions and deletions along with publication, population frequency, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations.
- dbSNP was first **created and released** to public in September of 1998.

Terms

- **Microsatellite** is a tract of repetitive DNA in which certain DNA motifs (ranging in length from one to six or more base pairs) are repeated, typically 5–50 times. Microsatellites occur at thousands of locations within an organism's genome.
- **Molecular consequence** is a calculation of the effect of the sequence change, reported per transcript. ClinVar calculates the predicted molecular consequence
- **ClinVar** is a public archive with free access to reports on the relationships between human variations and phenotypes, with supporting evidence.
- **Flanking sequence** - A DNA sequence located adjacent to a gene, either upstream from its 5'-end or downstream from its 3'-end. Flanking regions of the gene are often found to be of importance in determining the pattern and level of expression of the gene.

Overview of dbSNP

- dbSNP is a database that includes entries submitted by public laboratories and private organizations for a large number of organisms.
- Each submission includes information about the actual nucleotide variation and the 5' and 3' flanking sequences.
- It may also include other information such as genotype and frequency information.
- Each submitted entry is assigned a unique ID number that begins with letters "ss" (submitted SNP).

Overview of dbSNP

- If a number of submitted SNP entries align to the same position on the genome assembly, then they are also reported as a part of a group called the Reference SNP cluster (refSNP), which is assigned a new ID number that begins with letters “rs”.
- This procedure is performed periodically in the process that results in a new database “build”.
- The current build number and the build history can be obtained from http://www.ncbi.nlm.nih.gov/projects/SNP/buildhistory.cgi.

Overview of dbSNP

- **dbSNP** is a public-domain archive for a broad collection of simple genetic polymorphisms.
- This **collection of polymorphisms** includes-
  1) Single-base nucleotide substitutions (also known as single nucleotide polymorphisms or SNPs)
  2) Small-scale multi-base deletions or insertions (also called deletion insertion polymorphisms or DIPs)
  3) Retroposable element insertions and microsatellite repeat variations (also called short tandem repeats or STRs)

Overview of dbSNP

- The **dbSNP** has been designed to support submissions and research into a broad range of biological problems.
- These include
  1) Physical mapping
  2) Functional analysis
  3) Pharmacogenomics
  4) Association studies
  5) Evolutionary studies

Physical Mapping

- In the physical mapping of nucleotide sequences, variations are used as positional markers.
- When mapped to a unique location in a genome, variation markers work with the same logic as Sequence Tagged Sites (STSs) [short DNA sequence that has a single occurrence in the genome and whose location and base sequence are known.] or framework microsatellite markers.
- The position of a variation is defined by its unique flanking sequence, and hence, variations can serve as stable landmarks in the genome, even if the variation is fixed for one allele in a sample.

Functional Analysis

- Variations that occur in functional regions of genes or in conserved non-coding regions might cause significant changes in the complement of transcribed sequences.
- This can lead to changes in protein expression that can affect aspects of the phenotype such as metabolism or cell signaling.
- dbSNP notes possible functional implications of DNA sequence variations in terms of how the variation alters mRNA transcripts.

Association Studies

- The associations between variations and complex genetic traits are more ambiguous than simple, single-gene mutations that lead to a phenotypic change.
- When multiple genes are involved in a trait, then the identification of the genetic causes of the trait requires the identification of the chromosomal segment combinations, that carry the putative gene variants.

Evolutionary Studies

- The variations in dbSNP currently represent an uneven but large sampling of genome diversity starting from few entries.
- The human data in dbSNP include submissions from-
  - ✓ The SNP Consortium
  - ✓ Variations mined from genome sequence as part of the human genome project
  - ✓ Individual lab contributions of variations in specific genes
  - ✓ mRNAs
  - ✓ ESTs or genomic regions

Searching dbSNP

The SNP database can be queried from the dbSNP homepage by using the six basic dbSNP search options.

1) **Entrez SNP** - dbSNP is a part of the Entrez integrated information retrieval system and may be searched using either qualifiers (aliases) or a combination of 25 different search fields. A complete list of the qualifiers and search fields can be found on the Entrez SNP site

Field full name	Field aliases	Description	Search term values and rules	Example
All Fields	ALL, *	Search all searchable (indexed) fields	Asterisk (*) in the search term is not interpreted as a wildcard	SNV AND pathogenic
Base Position	POSITION, SNPPOS	Chromosome base position on GRCh38 (current)	A natural number representing the SNP's start coordinate on its chromosome on the latest assembly (ie. GRCh38). Most useful when search in combination with the CHR field.	19956018[POSITION] AND 8[CHR]
Base Position Previous	POSITION_GRCH37, CHRPOS_PREV_ASSM	Chromosome base position on GRCh37 (previous)	A natural number representing the SNP's start coordinate on its chromosome on the previous assembly (ie. GRCh37). Most useful when search in combination with the CHR field.	19813529[POSITION_GRCH37] AND 8[CHR]
Chromosome	CHR, CHRNUM	Chromosomes	One of 1-22, X, Y, MT	7[CHR]
Clinical Significance	CLIN	Variations with defined clinical effects or significances	16 search term values, defined for a relatively small subset of SNPs.	"likely pathogenic"[CLIN]
Filter	FILT, FLTR, SUBSET, SB, FIL	Limits the records returned	A variety of filters is available, including functional, positional, source, etc.	get all dbSNP records "all[sb]" or subsets "splice 5 snp"[Filter]
Function Class	FXN, Function_class, FUNC, FUNCTION, FUNCTION_CLASS	Function class	21 function classes are defined	"frameshift"[Function Class]
Gene Name	GENE, GENE_SYMBOL	Entrez Gene symbol	Corresponds to the Official Symbol field in the Entrez Gene resource	MAPK1[GENE]
Gene ID	GENE_ID	Entrez Gene UID	The numeric ID referencing the Entrez Gene ID	5594[GENE_ID]
Global Minor Allele Frequency	GMAF	Minor Allele Frequency derived from global population (i.e., 1000G); can also be study-wide MAF that is not from global population	Most useful when entered as a range, as in the example	(0.0[GMAF] : 0.01[GMAF])
Project or Submitter Handle	HAN, PROJECT	Submitter Handle or Project Name	Submitter lab or project name including 1000Genomes, GnomAD, and DebNick	1000genomes[Submitter Handle] or 1000genomes[PROJECT]
Reference SNP ID	RS, SNPID	Clustered SNP ID (rs)	The numeric ID must be prefixed with "rs". Also retrieves SNPs that have been merged into the specified SNP.	rs328[RS]
SNP Class	SCLS, SNPCLASS	SNP class	Possible values are: "del", "delins", "ins", "mnv", and "snv".	del[SNPCLASS]
Submitter SNP ID	SS, SSNUM	The ID assigned to each report of a SNP at submission time	Must be prefixed with "ss". Note that the query still returns Reference SNPs rather than Submitter SNPs.	ss329[SS]
Validation Status	VALI, VALIDATE, VALIDATION	Validation status	Possible values are: "by cluster" or "by frequency"	"by cluster"[Validation Status]

Searching dbSNP

2) **Single Record (Search by ID Number) Query in dbSNP** - Use this query module to select SNPs based on dbSNP record identifiers. These include reference SNP (refSNP) cluster ID numbers (rs#), submitted SNP Accession numbers (ss#), local (or submitter) IDs, Genbank accession numbers, and STS accession numbers.

3) **SNP Submission Information Queries** - Use this module to construct a query that will select SNPs based on submission records by laboratory (submitter), new data, the methods used to assay for variation populations of interest, and publication information.

Searching dbSNP

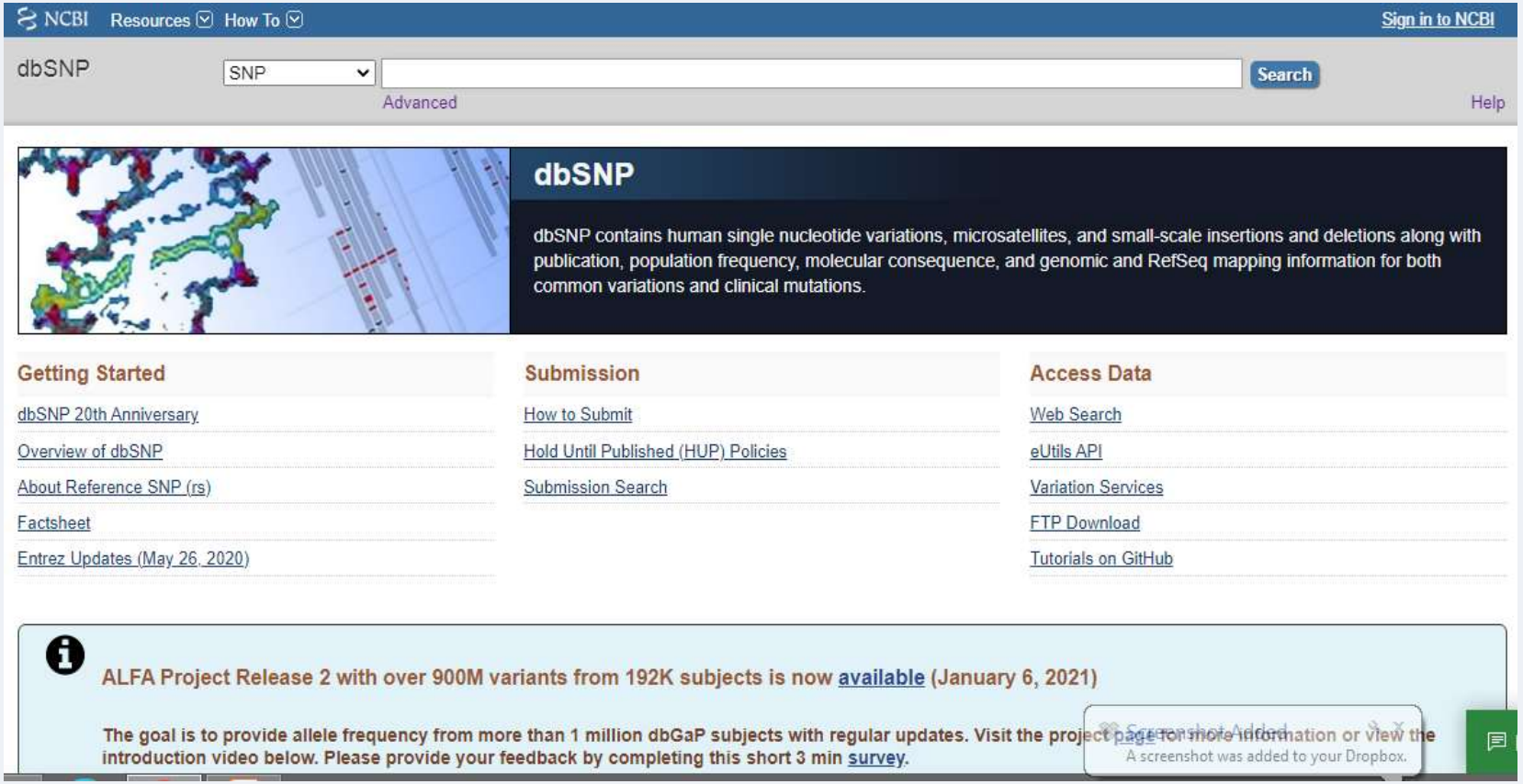
4) **dbSNP Batch Query** - Use sets of variation IDs (including RefSNP (rs) IDs, Submitted SNP (ss) IDs, and Local SNP IDs) collected from other queries to generate a variety of SNP reports.

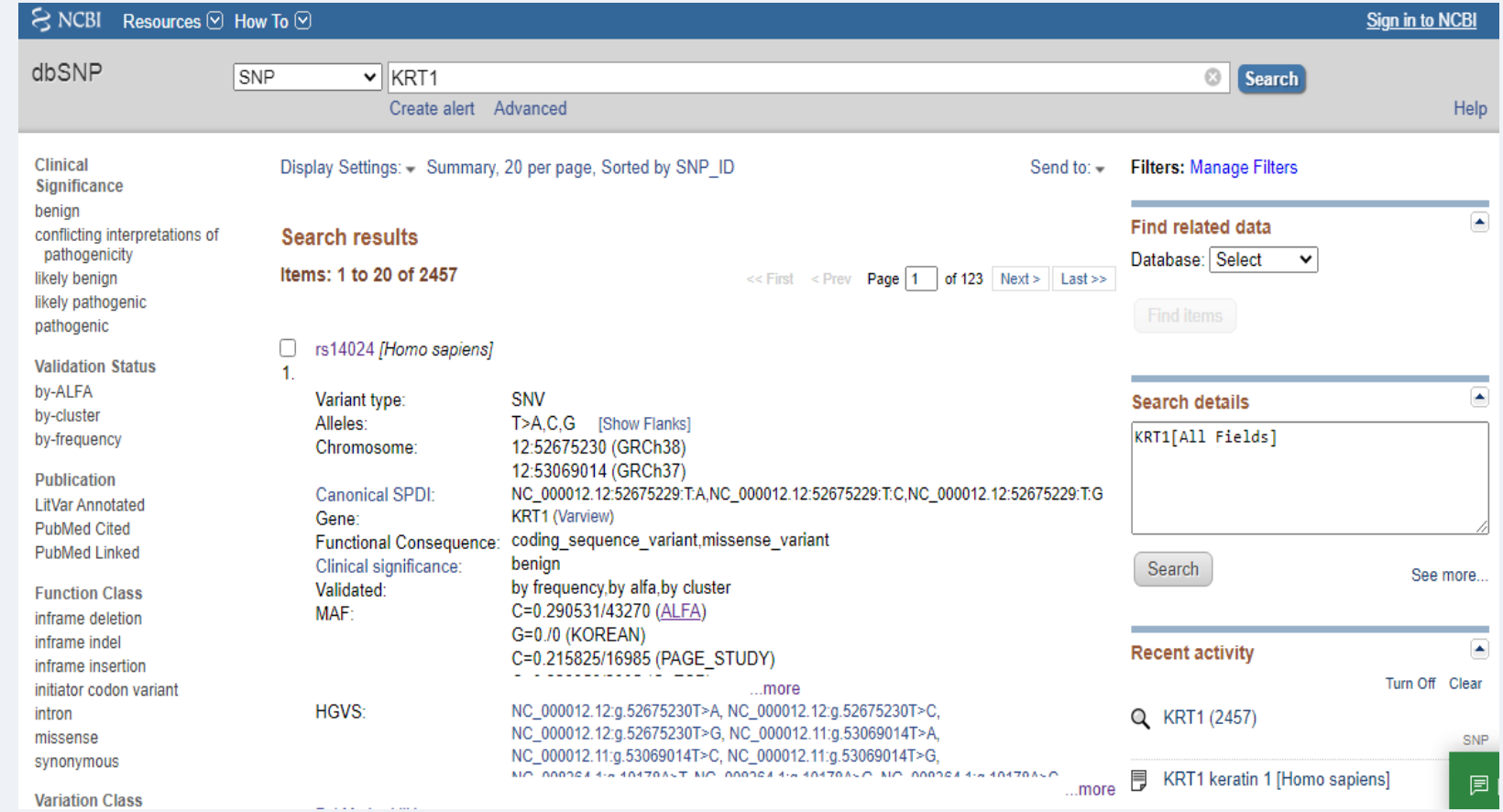
5) **Locus Information Query** - This search was originally accomplished by LocusLink, which has now been replaced by Entrez Gene. Entrez Gene is the successor to LocusLink and has two major differences that differentiate it from Locus Link: Entrez Gene is greater in scope (more of the genomes represented by NCBI Reference Sequences or RefSeqs) and Entrez Gene has been integrated for indexing and query in NCBI's Entrez system.

Searching dbSNP

6) **Between-Markers Positional Query** - Use this query approach if you are interested in retrieving variations that have been mapped to a specific region of the genome bounded by two STS markers. Other map-based queries are available through the NCBI Map Viewer tool.

dbSNP


NCBI Resources How To
Sign in to NCBI
dbSNP
SNP
Search
Advanced
Help
dbSNP
dbSNP contains human single nucleotide variations, microsatellites, and small-scale insertions and deletions along with publication, population frequency, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations.
Getting Started
dbSNP 20th Anniversary
Overview of dbSNP
About Reference SNP (rs)
Factsheet
Entrez Updates (May 26, 2020)
Submission
How to Submit
Hold Until Published (HUP) Policies
Submission Search
Access Data
Web Search
eUtils API
Variation Services
FTP Download
Tutorials on GitHub
ALFA Project Release 2 with over 900M variants from 192K subjects is now available (January 6, 2021)
The goal is to provide allele frequency from more than 1 million dbGaP subjects with regular updates. Visit the project page for more information or view the introduction video below. Please provide your feedback by completing this short 3 min survey.
Screenshot Added
A screenshot was added to your Dropbox.


NCBI Resources How To
Sign in to NCBI
dbSNP
SNP
KRT1
Search
Create alert Advanced
Help
Clinical Significance
benign
conflicting interpretations of pathogenicity
likely benign
likely pathogenic
pathogenic
Validation Status
by-ALFA
by-cluster
by-frequency
Publication
LitVar Annotated
PubMed Cited
PubMed Linked
Function Class
inframe deletion
inframe indel
inframe insertion
initiator codon variant
intron
missense
synonymous
Variation Class
Display Settings: Summary, 20 per page, Sorted by SNP_ID
Send to:
Filters: Manage Filters
Search results
Items: 1 to 20 of 2457
<< First < Prev Page 1 of 123 Next > Last >>
rs14024 [Homo sapiens]
1.
Variant type: SNV
Alleles: T>A,C,G [Show Flanks]
Chromosome: 12:52675230 (GRCh38)
12:53069014 (GRCh37)
Canonical SPDI: NC_000012.12:52675229:T:A,NC_000012.12:52675229:T:C,NC_000012.12:52675229:T:G
Gene: KRT1 (Varview)
Functional Consequence: coding_sequence_variant,missense_variant
Clinical significance: benign
Validated: by frequency,by alfa,by cluster
MAF: C=0.290531/43270 (ALFA)
G=0./0 (KOREAN)
C=0.215825/16985 (PAGE_STUDY)
...more
HGVS: NC_000012.12:g.52675230T>A, NC_000012.12:g.52675230T>C,
NC_000012.12:g.52675230T>G, NC_000012.11:g.53069014T>A,
NC_000012.11:g.53069014T>C, NC_000012.11:g.53069014T>G,
...more
Find related data
Database: Select
Find items
Search details
KRT1[All Fields]
Search
See more...
Recent activity
Turn Off Clear
KRT1 (2457)
SNP
KRT1 keratin 1 [Homo sapiens]

National Library of Medicine
National Center for Biotechnology Information

Log in

dbSNP Short Genetic Variations

Search for terms | Search

Examples: rs268, BRCA1 and more | Advanced search

Welcome to the Reference SNP (rs) Report

All alleles are reported in the Forward orientation. Click on the Variant Details tab for details on Genomic Placement, Gene, and Amino Acid changes. HGVS names are in the HGVS tab.

Reference SNP (rs) Report

Download

Switch to classic site

rs14024

Current Build 154
Released April 21, 2020

Organism	*Homo sapiens*	**Clinical Significance**	Reported in ClinVar
Position	chr12:52675230 (GRCh38.p12)	**Gene : Consequence**	KRT1 : Missense Variant
Alleles	T>A / T>C / T>G	**Publications**	3 citations LitVar 3
Variation Type	SNV Single Nucleotide Variation		
Frequency	C=0.309242 (77483/250558, GnomAD_exome) C=0.290531 (43270/148934, ALFA Project) C=0.249849 (31373/125568, TOPMED) (+ 22 more)	**Genomic View**	See rs on genome

FEEDBACK

Variant Details

Clinical Significance

Frequency

HGVS

Submissions

History

Publications

Flanks

FEEDBACK

Genomic Placements

Sequence name	Change
GRCh37.p13 chr 12	NC_000012.11:g.53069014T>A
GRCh37.p13 chr 12	NC_000012.11:g.53069014T>C
GRCh37.p13 chr 12	NC_000012.11:g.53069014T>G
GRCh38.p12 chr 12	NC_000012.12:g.52675230T>A
GRCh38.p12 chr 12	NC_000012.12:g.52675230T>C
GRCh38.p12 chr 12	NC_000012.12:g.52675230T>G
KRT1 RefSeqGene	NG_008364.1:g.10178A>T
KRT1 RefSeqGene	NG_008364.1:g.10178A>G
KRT1 RefSeqGene	NG_008364.1:g.10178A>C
KRT1 RefSeqGene	NG_008364.2:g.10178A>T
KRT1 RefSeqGene	NG_008364.2:g.10178A>G
KRT1 RefSeqGene	NG_008364.2:g.10178A>C

Variant Details

Clinical Significance

Frequency

HGVS

Submissions

History

Publications

Flanks

Allele: C (allele ID: 333512)

ClinVar Accession	Disease Names	Clinical Significance
RCV000317034.1	Bullous ichthyosiform erythroderma	Benign
RCV000373990.1	Nonepidermolytic palmoplantar keratoderma	Benign

Variant Details

Clinical Significance

Frequency

HGVS

Submissions

History

Publications

Flanks

ALFA Allele Frequency (New)

The ALFA project provide aggregate allele frequency from dbGaP. More information is available on the project page including descriptions, data access, and terms of use.

Release Version: 20201027095038

Search:

Population	Group	Sample Size	Ref Allele	Alt Allele
Total	Global	305340	T=0.698228	C=0.301772
European	Sub	256902	T=0.696530	C=0.303470
African	Sub	11296	T=0.92502	C=0.07498
African Others	Sub	388	T=0.979	C=0.021
African American	Sub	10908	T=0.92308	C=0.07692
Asian	Sub	6862	T=0.4129	C=0.5871
East Asian	Sub	4908	T=0.3775	C=0.6225
Other Asian	Sub	1954	T=0.5015	C=0.4985
Latin American 1	Sub	1394	T=0.7798	C=0.2202
Latin American 2	Sub	6636	T=0.7263	C=0.2737

Variant Details

Clinical Significance

Frequency

HGVS

Submissions

History

Publications

Flanks

Search:

Placement	T=	A	C	G
GRCh37.p13 chr 12	NC_000012.11:g.53069014=	NC_000012.11:g.53069014T>A	NC_000012.11:g.53069014T>C	NC_000012.11:g.53069014T>G
GRCh38.p12 chr 12	NC_000012.12:g.52675230=	NC_000012.12:g.52675230T>A	NC_000012.12:g.52675230T>C	NC_000012.12:g.52675230T>G
keratin, type II cytoskeletal 1	NP_006112.3:p.Lys633=	NP_006112.3:p.Lys633Met	NP_006112.3:p.Lys633Arg	NP_006112.3:p.Lys633Thr
KRT1 RefSeqGene	NG_008364.1:g.10178=	NG_008364.1:g.10178A>T	NG_008364.1:g.10178A>G	NG_008364.1:g.10178A>C
KRT1 RefSeqGene	NG_008364.2:g.10178=	NG_008364.2:g.10178A>T	NG_008364.2:g.10178A>G	NG_008364.2:g.10178A>C
KRT1 transcript	NM_006121.4:c.1898=	NM_006121.4:c.1898A>T	NM_006121.4:c.1898A>G	NM_006121.4:c.1898A>C
KRT1 transcript	NM_006121.3:c.1898=	NM_006121.3:c.1898A>T	NM_006121.3:c.1898A>G	NM_006121.3:c.1898A>C

- Variant Details
- Clinical Significance
- Frequency
- HGVS
- **Submissions**
- History
- Publications
- Flanks

121 SubSNP, 24 Frequency, 2 ClinVar submissions

Search:

No	Submitter	Submission ID	Date (Build)
15	1000GENOMES	ss111950983	Jan 25, 2009 (130)
22	1000GENOMES	ss235937980	Jul 15, 2010 (132)
23	1000GENOMES	ss242496942	Jul 15, 2010 (132)
32	1000GENOMES	ss491042957	May 04, 2012 (137)
47	1000GENOMES	ss1345045211	Aug 21, 2014 (142)
122	1000Genomes	NC_000012.11 - 53069014	Oct 12, 2018 (152)
144	A Vietnamese Genetic Variation Database	NC_000012.11 - 53069014	Jul 13, 2019 (153)
			Mar 10, 2006

- Variant Details
- Clinical Significance
- Frequency
- HGVS
- Submissions
- **History**
- Publications
- Flanks

Search:

Associated ID	History Updated (Build)
rs1050879	Jan 18, 2001 (92)
rs3191224	Jul 03, 2002 (106)
rs17855875	Mar 10, 2006 (126)
rs3825223	Oct 08, 2002 (108)
rs52799434	Sep 21, 2007 (128)

Added to this RefSNP Cluster:

Search:

Submission IDs	Observation SPDI	Canonical SPDI	Source RSIDs
104906, ss89155861, ss111950983, ss160348043, ss175130482, ss281399285, ss286562940, ss291252176, ss479958468, ss491665760, ss1397634004, ss1599212742, ss1713328300, ss2635036721, ss3642957150	NC_000012.10:51355280:T:C	NC_000012.12:52675229:T:C	(self)
34099638, ss3926922244	NC_000012.11:53069013:T:A	NC_000012.12:52675229:T:A	
57795142, 32092108, 22710598, 1220339, 78560, 2966841, 157733440, 9069711, 1189690, 14320783, 34099638, 446662,	NC_000012.11:53069013:T:C	NC_000012.12:52675229:T:C	(self)

Variant Details
Clinical Significance
Frequency
HGVS
Submissions
History
Publications
Flanks

3 citations for rs14024

Search:

PMID	Title	Author	Year	Journal
31067553	Significance of Cytokeratin-1 Single-Nucleotide Polymorphism and Protein Level in Susceptibility to Vocal Leukoplakia and Laryngeal Squamous Cell Carcinoma.	Yang Y et al.	2019	ORL; journal for oto-rhino-laryngology and its related specialties
29028840	Polymorphism of keratin 1 associates with systemic lupus erythematosus and systemic sclerosis in a south Chinese population.	Luo W et al.	2017	PloS one
17668073	In vitro human keratinocyte migration rates are associated with SNPs in the KRT1 interval.	Tao H et al.	2007	PloS one

View All in PubMed

- Variant Details
- Clinical Significance
- Frequency
- HGVS
- Submissions
- History
- Publications
- **Flanks**

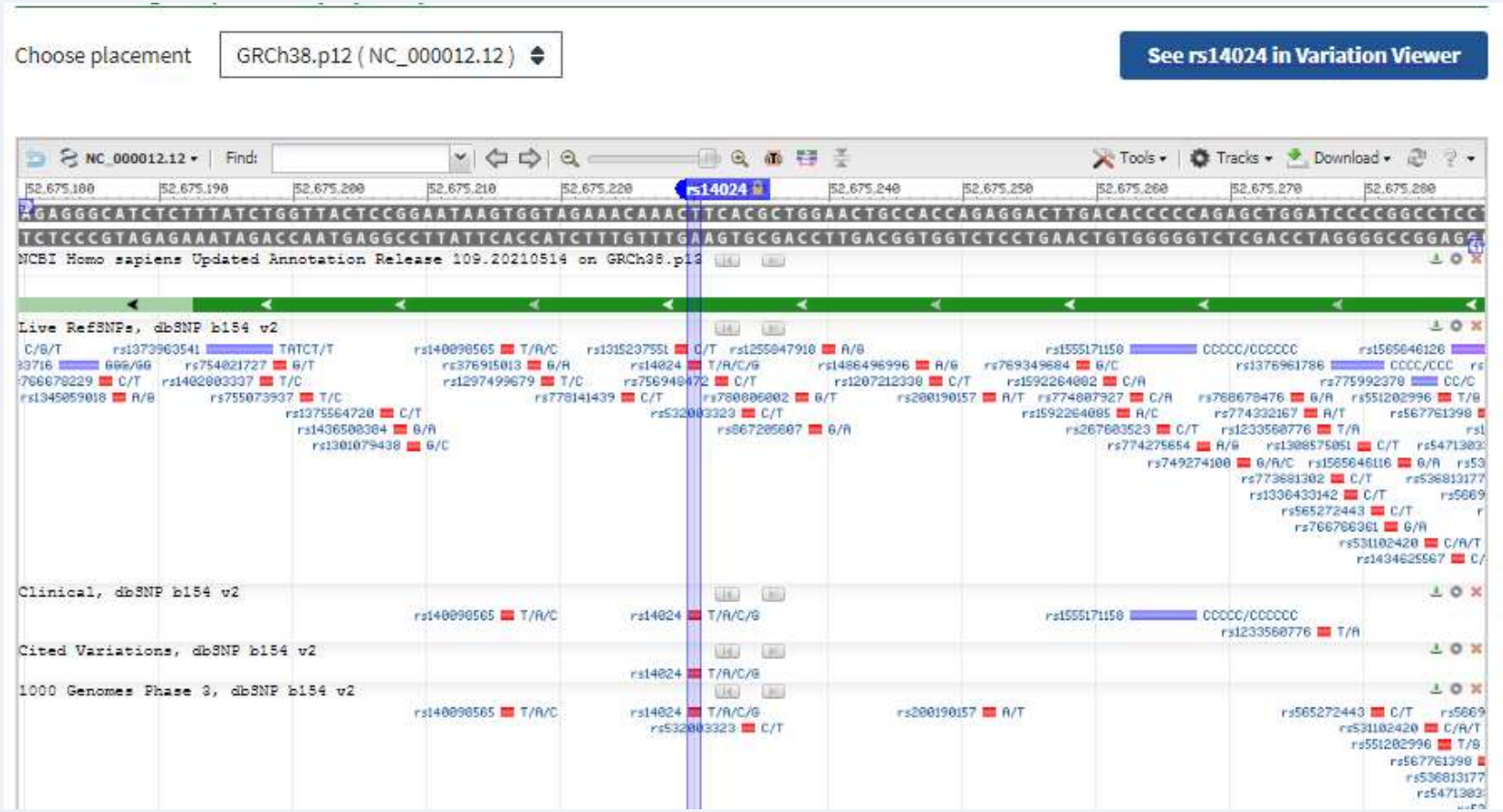
Genome context: GRCh38.p12 (NC_000012.12)

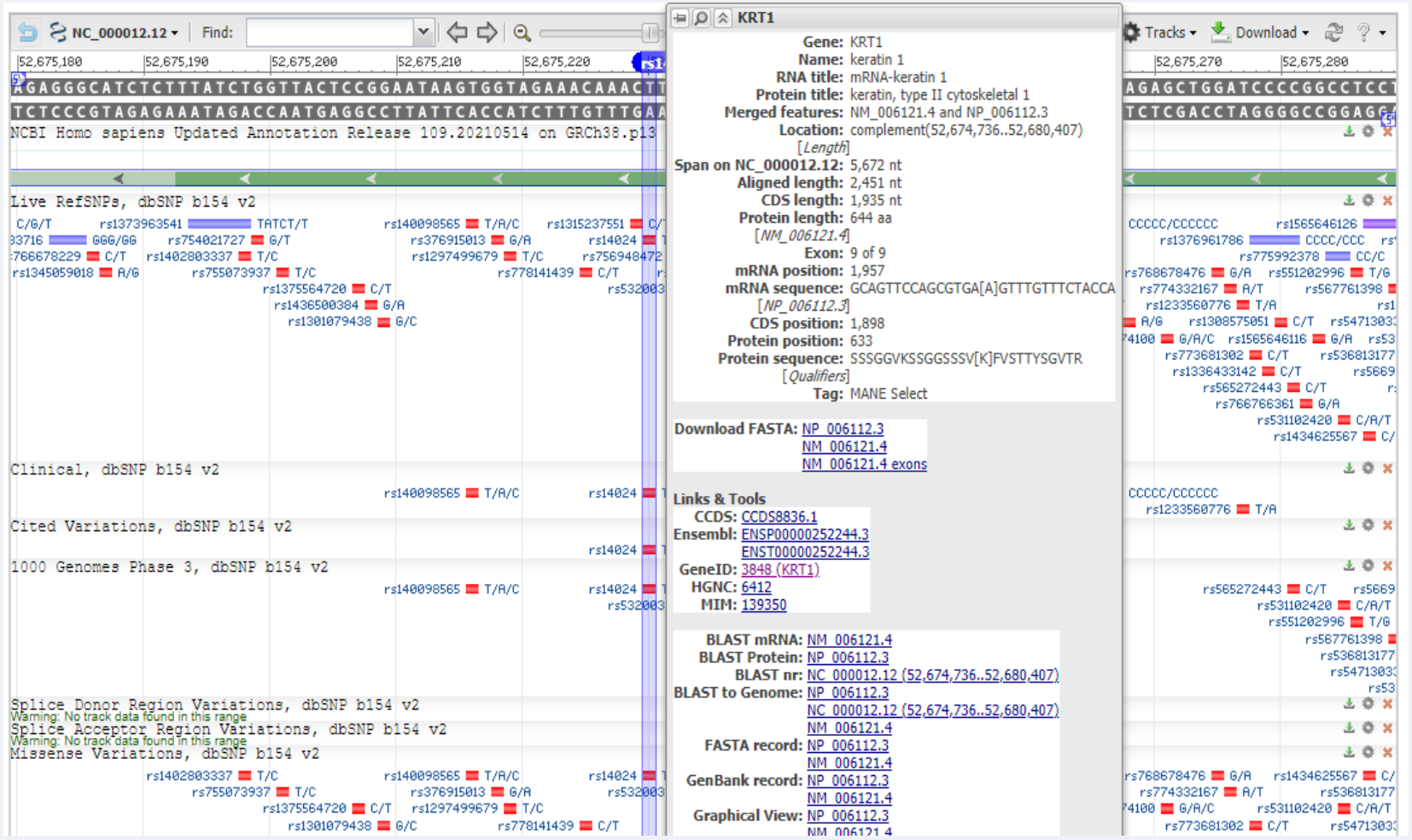
Select flank length: 25 nt

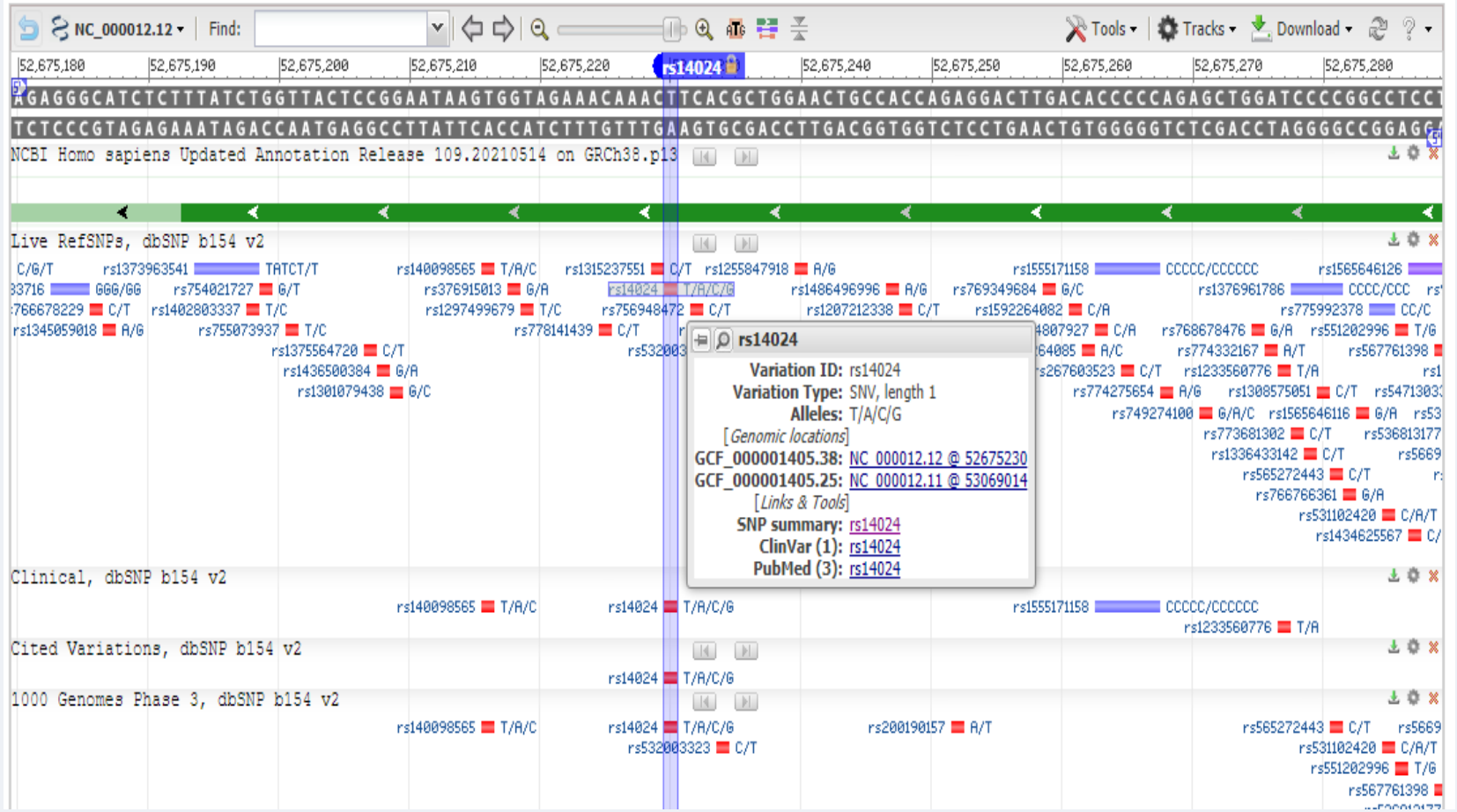
Retrieve

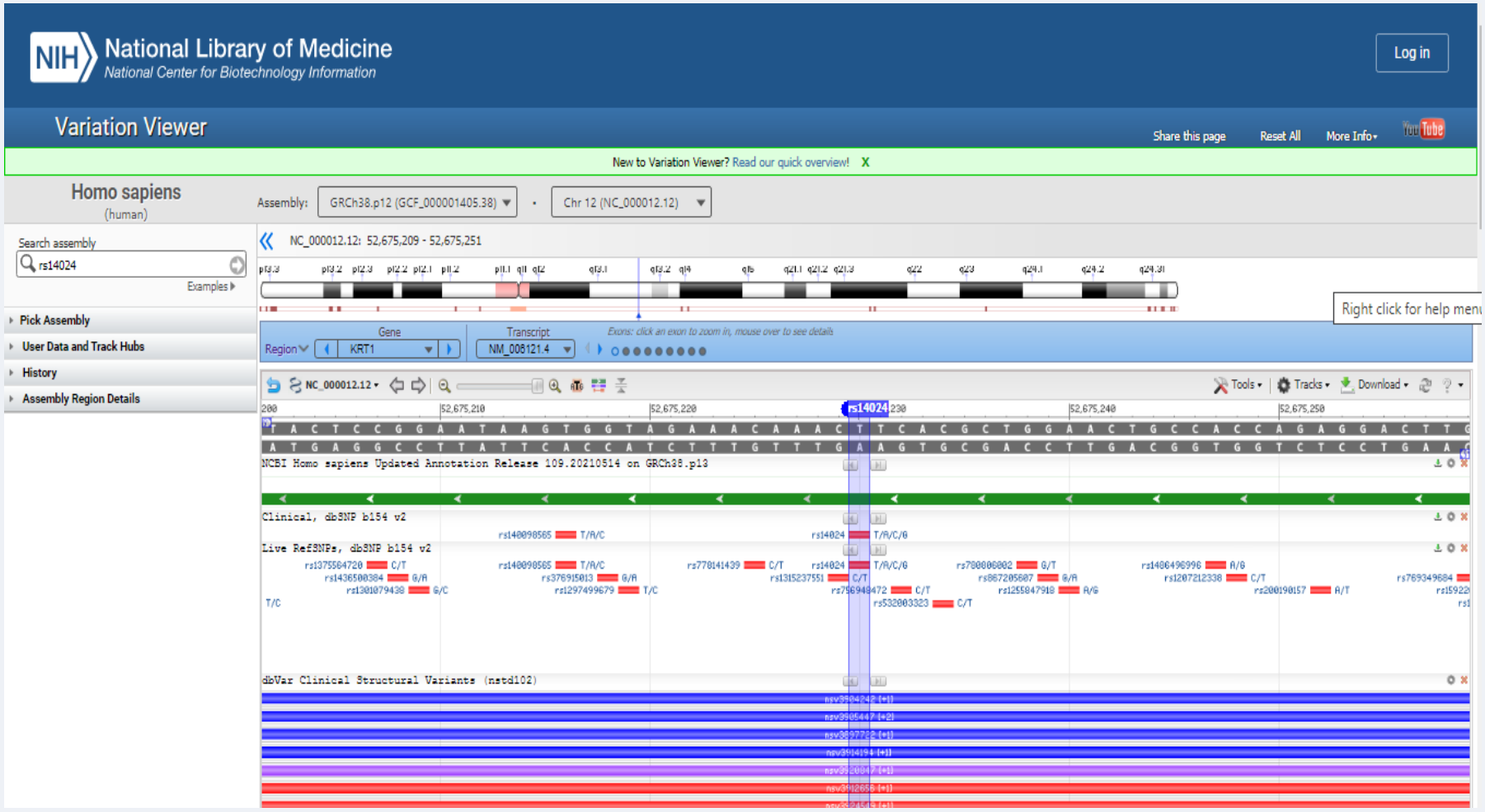
Gene: KRT1, keratin 1 (minus strand)

Molecule type	Change	Amino acid[Codon]	SO Term
keratin, type II cytoskeletal 1	NP_006112.3:p.Lys633Met	K (Lys) > M (Met)	Missense Variant
keratin, type II cytoskeletal 1	NP_006112.3:p.Lys633Arg	K (Lys) > R (Arg)	Missense Variant
keratin, type II cytoskeletal 1	NP_006112.3:p.Lys633Thr	K (Lys) > T (Thr)	Missense Variant
KRT1 transcript	NM_006121.4:c.1898A>T	K [AAG] > M [ATG]	Coding Sequence Variant
KRT1 transcript	NM_006121.4:c.1898A>G	K [AAG] > R [AGG]	Coding Sequence Variant
KRT1 transcript	NM_006121.4:c.1898A>C	K [AAG] > T [ACG]	Coding Sequence Variant


Choose placement
GRCh38.p12 (NC_000012.12)
See rs14024 in Variation Viewer
NC_000012.12
Find:
Tools
Tracks
Download
52,675,180
52,675,190
52,675,200
52,675,210
52,675,220
rs14024
52,675,240
52,675,250
52,675,260
52,675,270
52,675,280
NCBI Homo sapiens Updated Annotation Release 109.20210514 on GRCh38.p13
Live RefSNPs, dbSNP b154 v2
Clinical, dbSNP b154 v2
Cited Variations, dbSNP b154 v2
1000 Genomes Phase 3, dbSNP b154 v2


NC_000012.12
Find:
52,675,180
52,675,190
52,675,200
52,675,210
52,675,220
NCBI Homo sapiens Updated Annotation Release 109.20210514 on GRCh38.p13
Live RefSNPs, dbSNP b154 v2
Clinical, dbSNP b154 v2
Cited Variations, dbSNP b154 v2
1000 Genomes Phase 3, dbSNP b154 v2
Splice Donor Region Variations, dbSNP b154 v2
Warning: No track data found in this range
Splice Acceptor Region Variations, dbSNP b154 v2
Warning: No track data found in this range
Missense Variations, dbSNP b154 v2
KRT1
Gene: KRT1
Name: keratin 1
RNA title: mRNA-keratin 1
Protein title: keratin, type II cytoskeletal 1
Merged features: NM_006121.4 and NP_006112.3
Location: complement(52,674,736..52,680,407)
[Length]
Span on NC_000012.12: 5,672 nt
Aligned length: 2,451 nt
CDS length: 1,935 nt
Protein length: 644 aa
[NM_006121.4]
Exon: 9 of 9
mRNA position: 1,957
mRNA sequence: GCAGTTCCAGCGTGA[A]GTTTGTTTCTACCA
[NP_006112.3]
CDS position: 1,898
Protein position: 633
Protein sequence: SSSGGVKSSGGSSSV[K]FVSTTYSGVTR
[Qualifiers]
Tag: MANE Select
Download FASTA: NP_006112.3
NM_006121.4
NM_006121.4 exons
Links & Tools
CCDS: CCDS8836.1
Ensembl: ENSP00000252244.3
ENST00000252244.3
GeneID: 3848 (KRT1)
HGNC: 6412
MIM: 139350
BLAST mRNA: NM_006121.4
BLAST Protein: NP_006112.3
BLAST nr: NC_000012.12 (52,674,736..52,680,407)
BLAST to Genome: NP_006112.3
NC_000012.12 (52,674,736..52,680,407)
NM_006121.4
FASTA record: NP_006112.3
NM_006121.4
GenBank record: NP_006112.3
NM_006121.4
Graphical View: NP_006112.3
Tracks
Download
52,675,270
52,675,280


NC_000012.12
Find:
Tools
Tracks
Download
52,675,180
52,675,190
52,675,200
52,675,210
52,675,220
rs14024
52,675,240
52,675,250
52,675,260
52,675,270
52,675,280
AGAGGGCATCTCTTTATCTGGTTACTCCGGAATAAGTGGTAGAAACAAACTTCACGCTGGAACTGCCACCAGAGGACTTGACACCCCCAGAGCTGGATCCCCGGCCTCCT
TCTCCCGTAGAGAAATAGACCAATGAGGCCTTATTCACCATCTTTGTTTGAAGTGCGACCTTGACGGTGGTCTCCTGAACTGTGGGGGTCTCGACCTAGGGGCCGGAGG
NCBI Homo sapiens Updated Annotation Release 109.20210514 on GRCh38.p13
Live RefSNPs, dbSNP b154 v2
rs14024
Variation ID: rs14024
Variation Type: SNV, length 1
Alleles: T/A/C/G
[Genomic locations]
GCF_000001405.38: NC_000012.12 @ 52675230
GCF_000001405.25: NC_000012.11 @ 53069014
[Links & Tools]
SNP summary: rs14024
ClinVar (1): rs14024
PubMed (3): rs14024
Clinical, dbSNP b154 v2
Cited Variations, dbSNP b154 v2
1000 Genomes Phase 3, dbSNP b154 v2


National Library of Medicine
National Center for Biotechnology Information
Log in
Variation Viewer
Share this page
Reset All
More Info
New to Variation Viewer? Read our quick overview! X
Homo sapiens
(human)
Assembly:
GRCh38.p12 (GCF_000001405.38)
Chr 12 (NC_000012.12)
Search assembly
rs14024
Examples
Pick Assembly
User Data and Track Hubs
History
Assembly Region Details
NC_000012.12: 52,675,209 - 52,675,251
Region
Gene
KRT1
Transcript
NM_006121.4
Exons: click an exon to zoom in, mouse over to see details
Right click for help menu
Tools
Tracks
Download
rs14024
52,675,210
52,675,220
52,675,230
52,675,240
52,675,250
NCBI Homo sapiens Updated Annotation Release 109.20210514 on GRCh38.p13
Clinical, dbSNP b154 v2
Live RefSNPs, dbSNP b154 v2
dbVar Clinical Structural Variants (nstd102)

5,200 52,675,210 52,675,220 52,675,230 52,675,240 52,675,250

NC_000012.12: 53M..53M (58 nt) Tracks shown: 7/669

Variation Data

Filter by

Source database
- dbSNP (14)
- dbVar (0)

In ClinVar
- Yes (1)
- No (13)

Most severe clinical significance

Variant type

Molecular consequence

1000 Genomes MAF

GO_ESP MAF

ExAC MAF

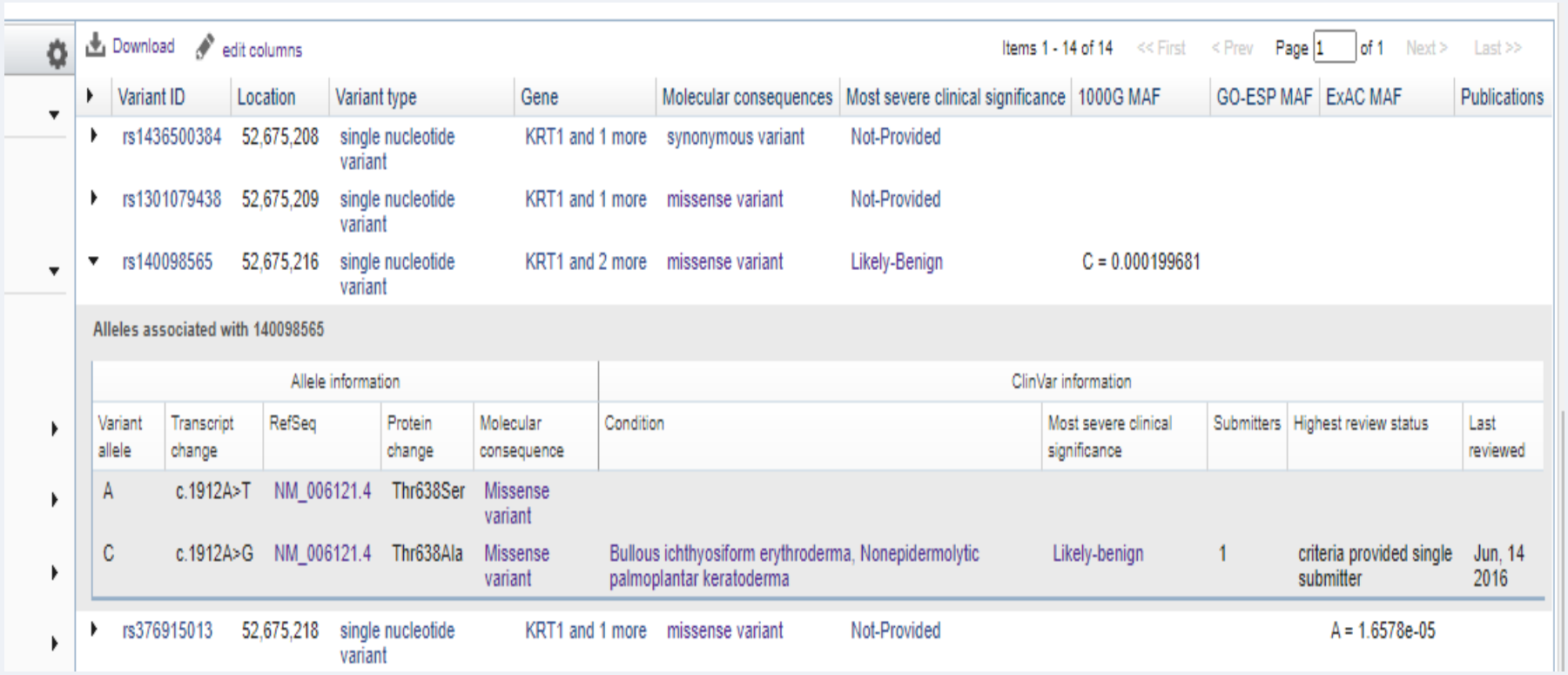
Has publications
- Yes (0)
- No (14)

Download edit columns

Items 1 - 14 of 14 << First < Prev Page 1 of 1 Next > Last >>

Variant ID	Location	Variant type	Gene	Molecular consequences	Most severe clinical significance	1000G MAF	GO-ESP MAF	ExAC MAF	Publications
rs1436500384	52,675,208	single nucleotide variant	KRT1 and 1 more	synonymous variant	Not-Provided				
rs1301079438	52,675,209	single nucleotide variant	KRT1 and 1 more	missense variant	Not-Provided				
rs140098565	52,675,216	single nucleotide variant	KRT1 and 2 more	missense variant	Likely-Benign	C = 0.000199681			
rs376915013	52,675,218	single nucleotide variant	KRT1 and 1 more	missense variant	Not-Provided			A = 1.6578e-05	
rs1297499679	52,675,219	single nucleotide variant	KRT1 and 1 more	missense variant	Not-Provided				
rs778141439	52,675,225	single nucleotide variant	KRT1 and 1 more	missense variant	Not-Provided			T = 8.29256e-06	
rs1315237551	52,675,229	single nucleotide variant	KRT1 and 1 more	synonymous variant	Not-Provided				
rs756948472	52,675,232	single nucleotide variant	KRT1 and 1 more	synonymous variant	Not-Provided			T = 8.31158e-06	
rs532003323	52,675,234	single nucleotide variant	KRT1 and 1 more	missense variant	Not-Provided	T = 0.000199681			
rs780806002	52,675,238	single nucleotide variant	KRT1 and 1 more	synonymous variant	Not-Provided			T = 8.31504e-06	
rs867205607	52,675,239	single nucleotide variant	KRT1 and 1 more	missense variant	Not-Provided				
rs1255847918	52,675,240	single nucleotide variant	KRT1 and 1 more	missense variant	Not-Provided				
rs1486496996	52,675,247	single nucleotide variant	KRT1 and 1 more	synonymous variant	Not-Provided				
rs1207212338	52,675,248	single nucleotide variant	KRT1 and 1 more	missense variant	Not-Provided				

Items 1 - 14 of 14 << First < Prev Page 1 of 1 Next > Last >>



Download edit columns

Items 1 - 14 of 14 << First < Prev Page 1 of 1 Next > Last >>

Variant ID	Location	Variant type	Gene	Molecular consequences	Most severe clinical significance	1000G MAF	GO-ESP MAF	ExAC MAF	Publications
rs1436500384	52,675,208	single nucleotide variant	KRT1 and 1 more	synonymous variant	Not-Provided				
rs1301079438	52,675,209	single nucleotide variant	KRT1 and 1 more	missense variant	Not-Provided				
rs140098565	52,675,216	single nucleotide variant	KRT1 and 2 more	missense variant	Likely-Benign	C = 0.000199681			

Alleles associated with 140098565

Allele information					ClinVar information				
Variant allele	Transcript change	RefSeq	Protein change	Molecular consequence	Condition	Most severe clinical significance	Submitters	Highest review status	Last reviewed
A	c.1912A>T	NM_006121.4	Thr638Ser	Missense variant					
C	c.1912A>G	NM_006121.4	Thr638Ala	Missense variant	Bullous ichthyosiform erythroderma, Nonepidermolytic palmoplantar keratoderma	Likely-benign	1	criteria provided single submitter	Jun, 14 2016

rs376915013	52,675,218	single nucleotide variant	KRT1 and 1 more	missense variant	Not-Provided			A = 1.6578e-05	

THANK YOU