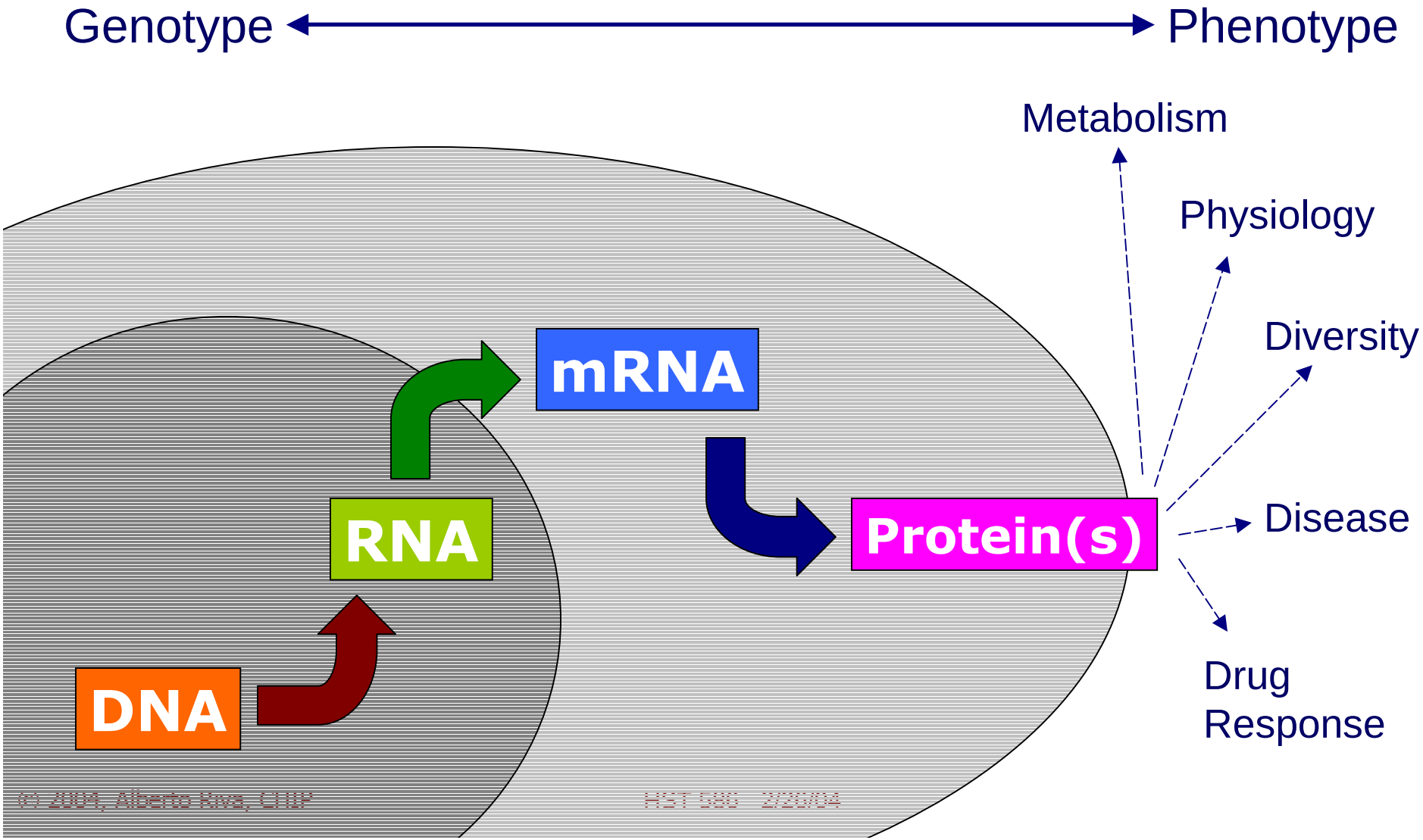
Informational Resources

(Finding your way through the Human Genome)

Alberto Riva, PhD
Children's Hospital Informatics Program
Harvard Medical School

Genomic Medicine
HST 586

The Central Dogma of Molecular Biology



How is information represented?

Where does it come from?

Where is it stored?

How do we find, retrieve and use it?

From genotype to phenotype

- The genotype is "digital".
  - ✓ Each base pair can be precisely represented using one of four symbols (A, T, G, C).
  - ✓ Approximately 3.2 billion base pairs in the human genome.
- The phenotype is "analog".
  - ✓ Proteins are not uniquely determined by their sequence.
  - ✓ Environmental factors are always present.
  - ✓ Most phenotypes are qualitative in nature.

From phenotype to genotype

Our knowledge progresses in the opposite direction:

- First studies of inherited traits: Mendel, 1866;
- (Discovery of DNA: Miescher, 1869);
- Genes are made of DNA: Hershey and Chase, 1952;
- DNA replication mechanism: Watson and Crick, 1953;
- Genetic code decyphering: Nirenberg, 1961-1966;
- Discovery of introns: Sharp and Roberts, 1977;
- Human Genome Project: 1990-2003.

From genotype to phenotype

- "The" human genome is an abstraction.
  - ✓ Single-nucleotide polymorphisms (SNPs), microsatellites (repeats), insertions, deletions, translocations, etc etc…
  - ✓ On the average, one polymorphism every 1,000 bases.
- Phenotypes are generalizations:
  - ✓ Species;
  - ✓ Ethnicity;
  - ✓ Diseases.

Data and Methods

- DNA:
  - ✓ Sequence matching (BLAST, etc.)
  - ✓ Gene prediction (Genscan, etc.)
  - ✓ Homology searches
  - ✓ SNP detection (genotyping)
- RNA:
  - ✓ Alternative splicing, transcriptional rearrangements
  - ✓ Expression analysis (microarrays)
    - ☞ Differential analysis
    - ☞ Clustering

Data and Methods

- Protein:
  - ✓ Prediction of active domains
  - ✓ 3-D structure prediction
  - ✓ Protein homology and conservation
  - ✓ Automated construction and analysis of metabolic / regulatory pathways
- Phenotype:
  - ✓ Population genetics
  - ✓ Association studies
  - ✓ Clinical trials

What is a gene?

- Classical geneticist:

  "Gene = smallest unit of inheritance"

- Medical researcher:

  "Gene = disease-causing trait"

- Molecular biologist:

  "Gene = recipe for one or more proteins"

What is a gene?

- Biochemist:

  "Gene = element in a metabolic network"

- Modern geneticist:

  "Gene = functional locus on a chromosome"

- Bioinformatician:

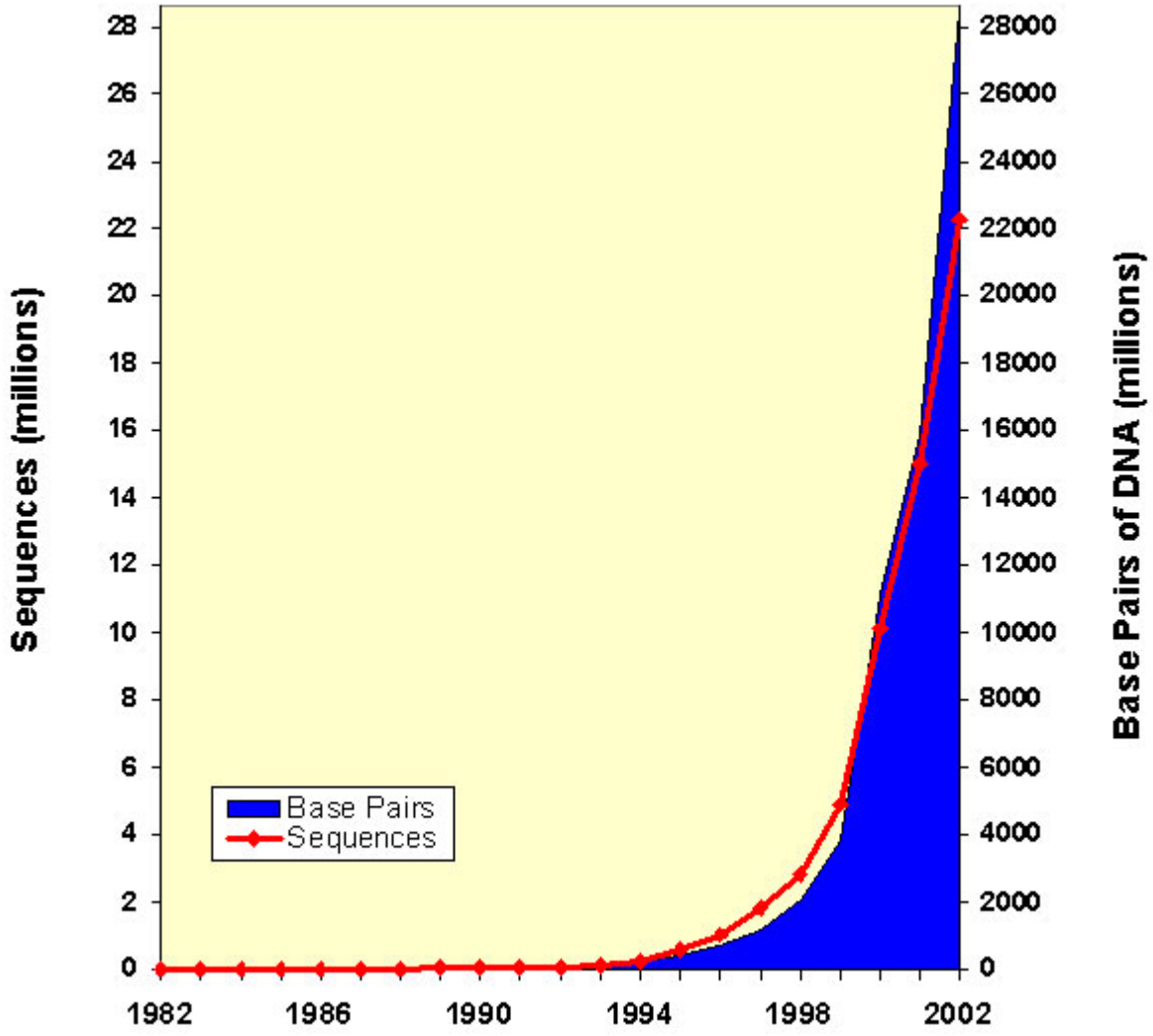
  "Gene = contiguous, characterized DNA sequence"

DNA sequence data

GenBank

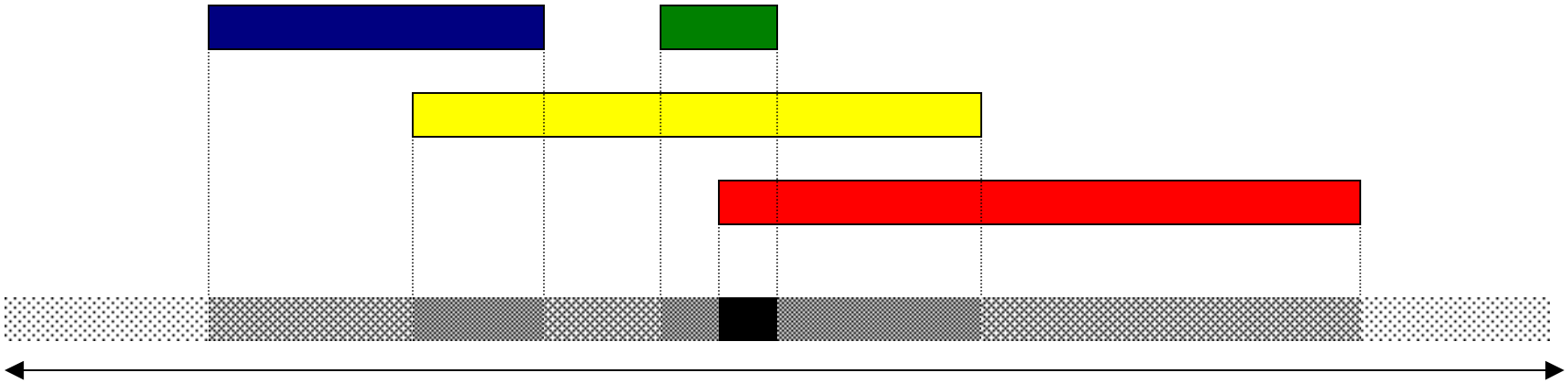
- Largest public repository of sequence data. Accepts direct submissions from researchers.
- As of January 2003:
  - ✓ 22.3 million sequences
  - ✓ 100,000 distinct organisms
  - ✓ 28.5 billion nucleotides
- Foundation of the NCBI cluster: **http://ncbi.nih.gov/Genbank**

Growth of Genbank



Genome assembly

- The assembly process starts with clusters of overlapping sequences, and proceeds in a targeted way to fill the gaps.



- Details available at **http://ncbi.nih.gov/genome/guide/build.html**

Genomes

- Completed or nearing completion:
  - ✓ Viruses: 1,021
  - ✓ Archea: 16
  - ✓ Bacteria: 99
  - ✓ Organelles: 405
  - ✓ Eukariotes: 18
    - ☞ Homo sapiens: 99% finished, 99.99% accuracy
- Data available at: **http://ncbi.nih.gov/Entrez/Genome/main_genomes.html**

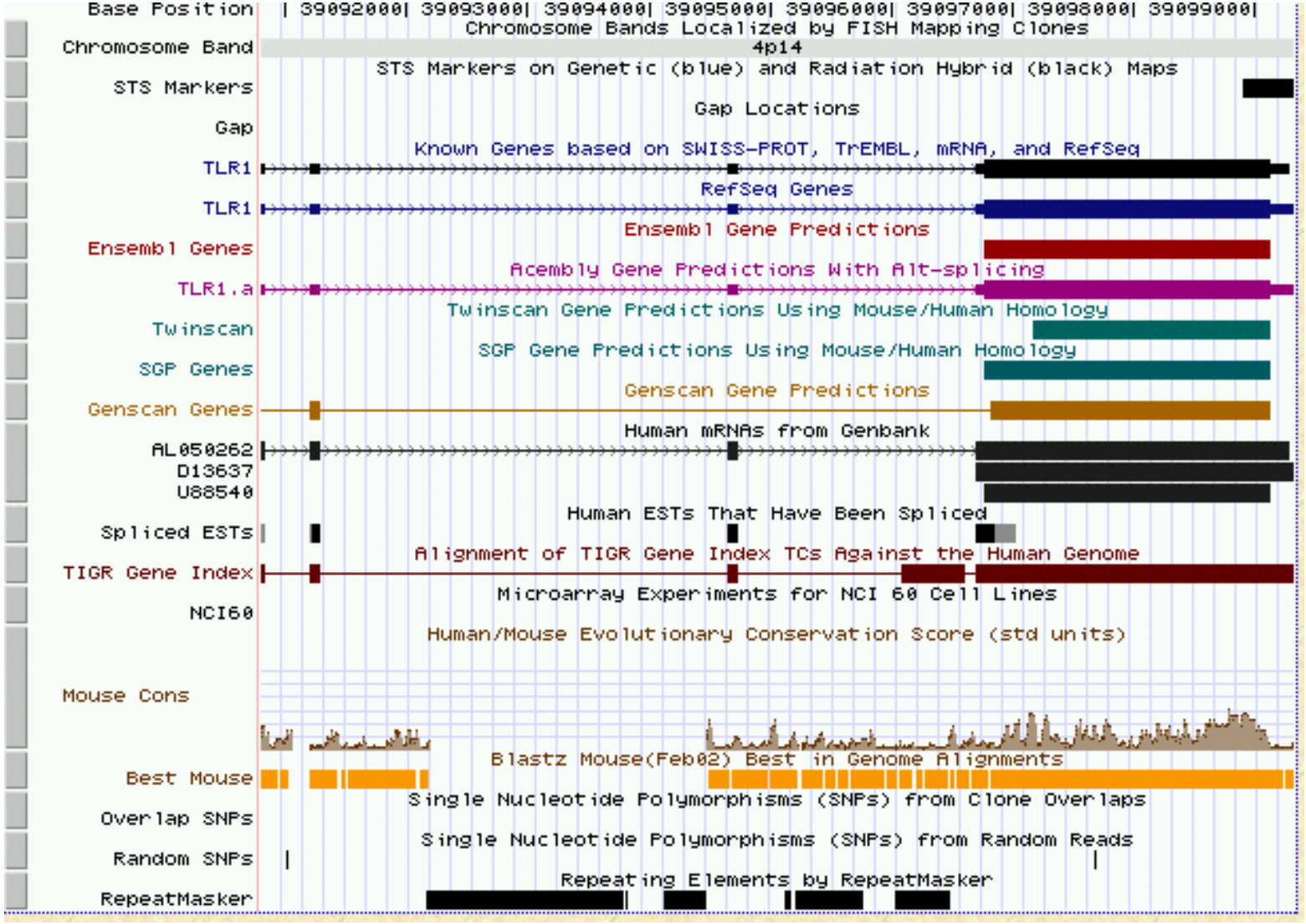
GoldenPath

- Genome browser for human, mouse, rat, chimp, fruitfly, yeast, C. elegans, C. briggsae, SARS.

- Interesting features:
  - ✓ Multi-track graphical view
  - ✓ Can be queried for arbitrary DNA sequences
  - ✓ Uses the NCBI human genome assembly
  - ✓ Provides absolute positional information for genes, markers, mutations, other features.

- Available at **http://genome.ucsc.edu/**

GoldenPath browser



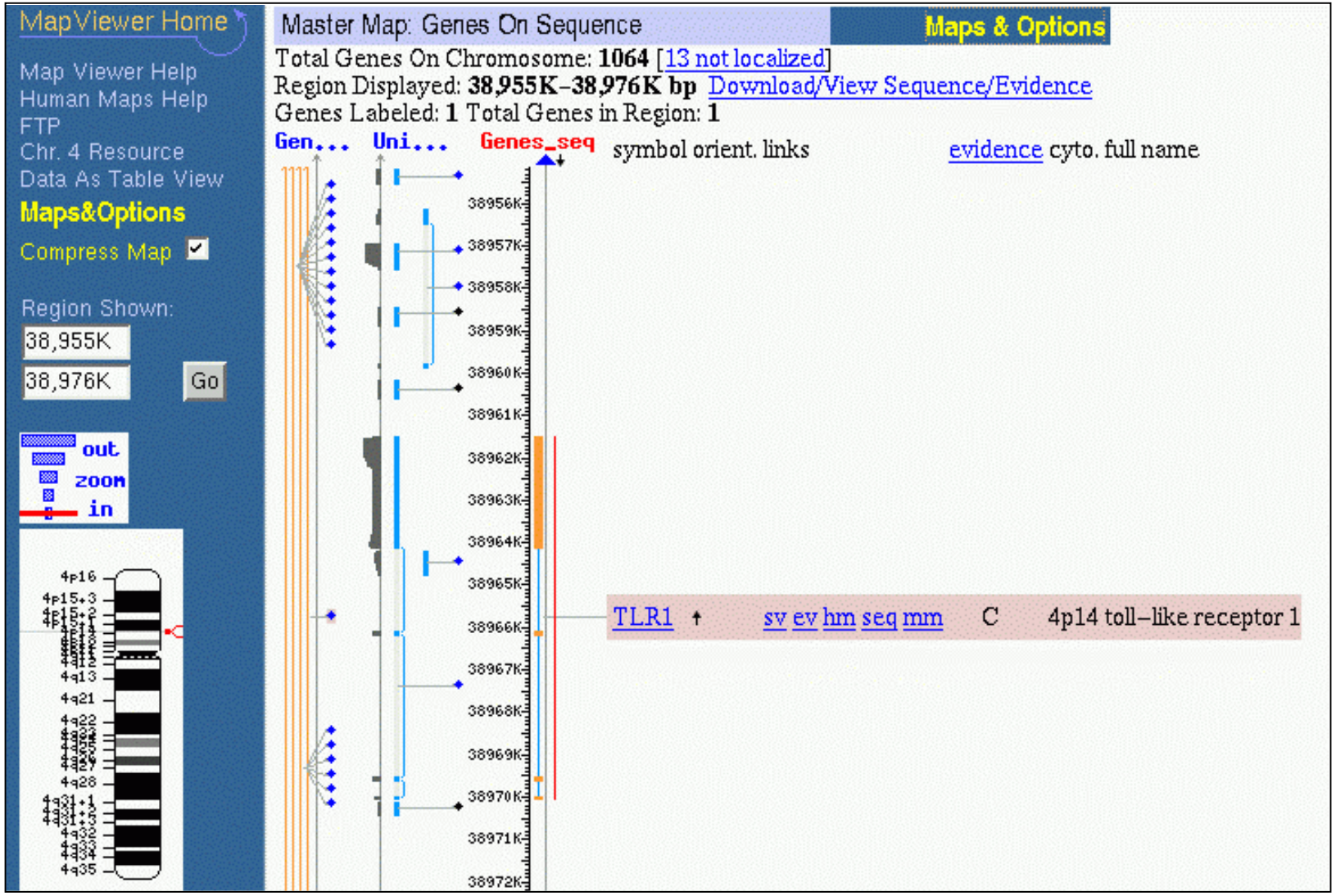
GoldenPath - sequence

```
>hg12_dna range=chr4:39090798-39099316 5'pad=0 3'pad=0 revComp=FALSE strand=? repeatMasking=lower
TACAGACTGCCAAATGGAACAGACAAGCAGGTTGTCTTGGTAAGCAACAC
ATTCTTTTCTTTTGTAAAAGAAAATAATTGTATAGCTAGTTATTAAGTAC
AGAAGTCTCAAAAATCTGTGTAAGTCCTGGGTGTTTTTCTAAGTGGGTTA
TATTTCTGATATGTATATAGTTACTGTGTAGTTTGTACTGGCATTTGTGT
ATCAGTTCTGAGTCCTAAATCAGAGAAAGTCCCCACACTCCTCTGGGAAT
AACACCTCGTGTGTGATTTGCTTATAGGAAATATTTTTGAGTTGGGAAAT
GAATATTTGGGTCACACACGTCTTCTGGTTTTCTTCCAGAGCAGCTGCTA
GTTGTTGATTTTGACAGCATTTCTCTTCACCTAATCCCGCCATTTGTATT
CTCTTCTCAGTGTTAAAGAAAATGAGATATAAGTCAGTTACTCCCGGAGG
CAATGCTGCTGTTCAGCTCTTCTGTTTTTGTGGCCAGGTTTGTTTCAGTT
TTTCTCTCTGGGGCCTTCCCACTAGCTAGTCCTTCTGTTTCTCAGGCCAC
TAACACTGTTCTGGGGAGCTAGACGTGAGGTAGAATTGCAGGTTTTGAAA
ATGTTTCCCAGGCTACATCCAATTTGGTCAAAAGACTTGAAAGTGAATTT
GTTTTATAACAAAGCAAGAGTTCACAAAACCGCATAGAAAGCAAAGCAGA
ACAGTTCTTGAACTGTCTCAGATTCTTTTAACCTTTGGTTAATAGCTATT
TGGTCTTGTGCAGAGAAGGCCCTTAGTAAATATTTACAGGACTAAACTTA
ATGGGCCAGATAGATTGTATGGGTATTGTCCATTAAGACCAGTCAAAGCC
TTGATTTGATGCCTCCAAAGTCTCCAAAAGAAGACAAATTAAATGTATTG
ATTCATTGTATCATATAATGTGAGGCTAAACCTATCATAATAATGAAAAT
TCACAGAAGCTCACCTAGAGGCATTTTACACTTTCAAATTAAAGCATCCT
TTTCTGCTCAGCTTATTTTTTGGATAGTAAAAGAGTTTAAAGGTTCTAAA
ATAGATGAGGCTCAGTCATGGTCATTAAGACGTTGATCAAAAAATTCTTT
GGTCTAGGAAACTGAGACATTGATTTTCATGGATTTAGCAAGTTTGGTAT
ATAAAATTCAAAGATCCTGGAAGTATTAAAAATCAGTTATTCCTGTGTAT
AACATTTTGATTTCTTAATTATAGAAATAAGAATGGTTTCATAAACTGAG
TACTTAAAATTAATGACTTTAAATGAAAGGGCAAGATGGAGGTTAGGCAA
ATAGAAGATATTTGAGTCAATTAATTTAGCTCGACATAAAACTGAAGCTA
TGCTTTCATTAAATGCTtgcgttagcagtggtgaatccatatgggtcgta
gaaacttaattcttgcctcctcagtggaaagaatttgtgtgaggggcata
aggcagagtgagagaccgaggcaagttttagagtaagagagagagtgtca
cgcccagggccaggttccagcccatactgaggtctgaggggaggggtgg
atgagcagatagctgaaagaacactcaggggccgtaggcaggtgaaagg
tgattttattcagcaacagctctcattagcagcttacttacagtagttct
ctcagactgtccgccttgtcctggctgcttagttcggcagcttccacaca
caactgtgcgtccggctctcccttgccctcagggtcagcagcttaactct
ttctctctctgggtacaagcaagccgagctgtgtcctggctcccctcagt
ccatctgcaaagatggacagctttggctctctctctctctctttctctgg
ctgccaatgcacctgtacagcgtcagcagggcaattataccatttacgga
```

NCBI map viewer

- Integrates sequence and map data from a variety of sources. Available maps:
  - ✓ Sequence maps (clone, component, contig, haplotype, gene, STS, transcript, UniGene);
  - ✓ Cytogenetic maps (bands, breakpoints, disease genes);
  - ✓ Linkage maps (deCODE, Genethon, Marshfield);
  - ✓ Radiation hybrid maps;
  - ✓ Human/mouse homology;
- Extremely detailed, complex but flexible.

NCBI viewer


MapViewer Home
Map Viewer Help
Human Maps Help
FTP
Chr. 4 Resource
Data As Table View
Maps&Options
Compress Map
Region Shown:
38,955K
38,976K
Go
out
zoom
in
4p16
4p15.3
4p15.2
4q13
4q21
4q22
4q28
4q31.1
4q32
4q33
4q34
4q35
Master Map: Genes On Sequence
Maps & Options
Total Genes On Chromosome: 1064 [13 not localized]
Region Displayed: 38,955K–38,976K bp Download/View Sequence/Evidence
Genes Labeled: 1 Total Genes in Region: 1
Gen...
Uni...
Genes_seq
symbol orient. links
evidence cyto. full name
38956K
38957K
38958K
38959K
38960K
38961K
38962K
38963K
38964K
38965K
38966K
38967K
38968K
38969K
38970K
38971K
38972K
TLR1
sv ev hm seq mm
C
4p14 toll–like receptor 1

SNPs

- SNPs are the most common form of variation in our genome. SNPs are important as
  - ✓ genomic markers;
  - ✓ causal candidates for diseases;
  - ✓ evolutionary markers.

- dbSNP (**http://ncbi.nih.gov/SNP/**) currently contains over 4,000,000 human SNPs (almost 50% of which are validated).

- TSC (**http://snp.cshl.org/**) offers reliable frequency information on about 100,000 SNPs.

Other SNP resources

- HapMap (**http://www.hapmap.org**), aims at developing a haplotype map of the human genome;
- HGBASE (**http://www.hgbase.com/**), Karolinska Institute (manually curated genotype/phenotype);
- ALFRED (**http://alfred.med.yale.edu/alfred/**), Yale (SNP frequency data in *many* different populations);
- SNPper (**http://snpper.chip.org/**), CHIP.

SNP: rs422951

Position:	chr6:32214410	**Band:**	6p21.32	**Alleles:**	C/T	**Avg Het:**	0.42
dbSNP:	rs422951	**GenBank:**	(unknown)	**Updated:**	01/29/2003	**Validated:**	Y
Gene:	NOTCH4	**Role:**	Exon, Coding sequence	**Relative position:**	3319	**Amino acid change:**	319 A/T

Swiss-Prot domains

Range	Name	Notes
1–1414	Varsplic	MISSING (IN ISOFORM 2).
24–1447	Domain	EXTRACELLULAR (POTENTIAL).
24–2003	Chain	NEUROGENIC LOCUS NOTCH PROTEIN HOMOLOG 4.
314–353	Domain	EGF-LIKE 8, CALCIUM-BINDING (POTENTIAL).
318–332	Disulfid	BY SIMILARITY.

Submitters

dbSNP assay	Submitter	Private ID
ss7859415	DEVINE_LAB	DB_1_92965
ss3177054	WICVAR	WIAF-15800
ss2984851	YUSUKE	IMS-JST006669
ss1954866	KWOK	OVLP-000925-362687
ss1309422	TSC-CSHL	TSC0219404
ss558680	SC_JCM	U89335.1_27304

Frequency

Population	Samples	Major allele	Minor allele
TSC_42_AA	41	**A** (0.72)	**G** (0.28)
TSC_42_C	41	**G** (0.56)	**A** (0.44)

From DNA to genes

LocusLink

- Curated directory of genes from 13 organisms.
- Its central function is "to establish an accurate connection between the defining sequence for a locus and other descriptors for that locus".
- Provides sequence and functional information, links, aliases, phenotypes, homologies, map locations.
- The LocusLink nomenclature is at the basis of several other resources. **http://ncbi.nih.gov/LocusLink/**

UniGene

- UniGene is "an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters".
- Each clusters contains "similar" sequences from multiple forms of the same gene, with related information (tissues, conditions, etc).
- Currently includes 38 organisms. Available at **http://ncbi.nih.gov/UniGene/**

Homologene

- Repository of curated and calculated orthologs.
- 25 organisms, approximately 470,000 putative ortholog pairs.
- Calculated orthologs are based on sequence similarity; similarity score is provided.

Ensembl

- Software system for the automated annotation of genomes. Joint project of EMBL-EBI and the Sanger Institute.
- Currently contains data on 10 organisms (genes, proteins, diseases, SNPs, cross-species analysis, microarray data, etc).
- The EnsMart interface provides powerful data access capabilities and flexible querying of large biological datasets.

Gene regulation

- Gene regulation is an extremely complex mechanism. Our understanding of it is still very limited.

- Gene expression is a function of a very large number of factors, including the following:
  - ✓ Tissue;
  - ✓ Developmental stage;
  - ✓ Time (at widely different resolutions);
  - ✓ External signals;
  - ✓ Expression state of any number of other genes.

Gene regulation

- Transcription factors (TFs) are proteins that bind to the upstream regions of genes, and control their expression and activity.

- TFs interact with the target gene and with each other in a combinatorial fashion. Specific patterns of TFs determine the spatial, temporal, and tissue-dependent expression of the target gene.

Gene regulation

- The reliable detection of TF binding sites (TFBSs) is the first step towards the discovery of the regulatory grammar.
- Commonly used methods are based on pattern matching. Known binding sites are used to train deterministic and probabilistic search methods.
- TRANSFAC (**http://www.gene-regulation.com/**) is the largest database on TFs. It provides information on the factors, their binding sites, their interactions with genes.

Gene expression data

- Gene Expression Omnibus (NCBI) (**http://ncbi.nih.gov/geo/**)
  - ✓ Repository of gene expression and hybridization array data.
  - ✓ 12,000 samples on over 500 platforms.
  - ✓ Very powerful interface.

- Stanford Microarray Database
  - ✓ Data for over 43,000 experiments (6,000 public).

  http://genome-www5.stanford.edu/

- NCI60 (**http://genome-www.stanford.edu/nci60/**)
  - ✓ Gene expression profiles for 60 human cancer cell lines.
  - ✓ Drug activity correlated with gene expression patterns.

Other gene expression data resources

- TREX PGA (565 microarrays from mouse and rat models of sleep, infection, hypertension and pulmonary disease)
  http://pga.tigr.org/data.shtml

- HopGenes PGA, Children's National Medical Center (more than 500 microarrays from many human diseases)
  http://microarray.cnmcresearch.org/pgadatatable.asp

- CardioGenomics PGA (142 microarrays on mouse models of cardiac development and signal transduction)
  http://cardiogenomics.med.harvard.edu/public-data.html

- Human Gene Expression Index (121 microarrays from 19 normal human tissues)
  http://www.hugeindex.org/databases/index.html

From proteins to phenotypes

Protein databases

- The protein world is far more complex than the DNA-RNA world. Proteins interact with each other, combining 3-D and catalyzing chemical reactions.

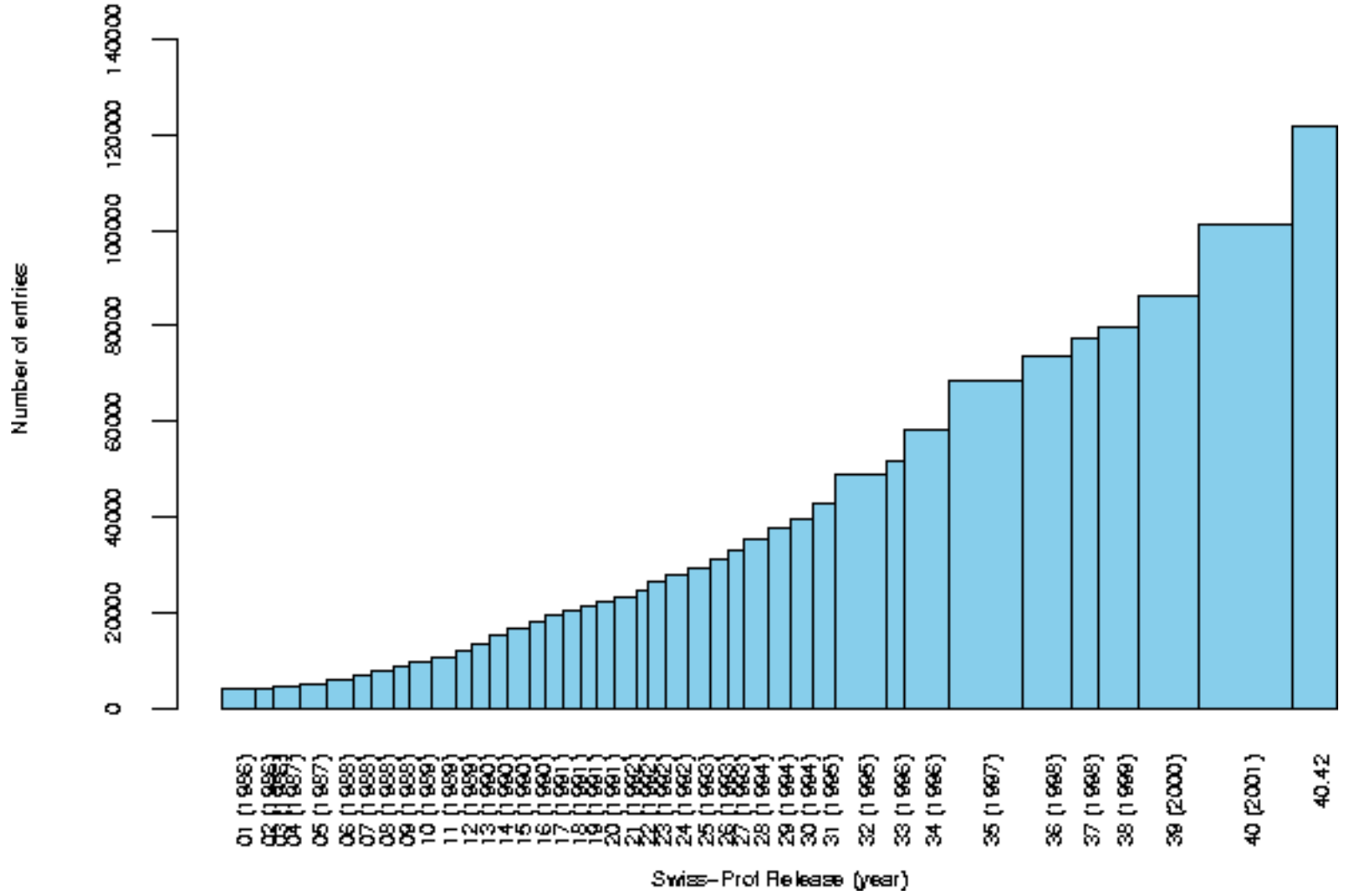
- Protein databases provide information on:
  - ✓ Protein sequence
  - ✓ Known or computed 3-D structure
  - ✓ Known or inferred functional domains

- Protein databases tend to be older, less integrated, less complete. Nomenclature is less standardized.

Swiss-Prot

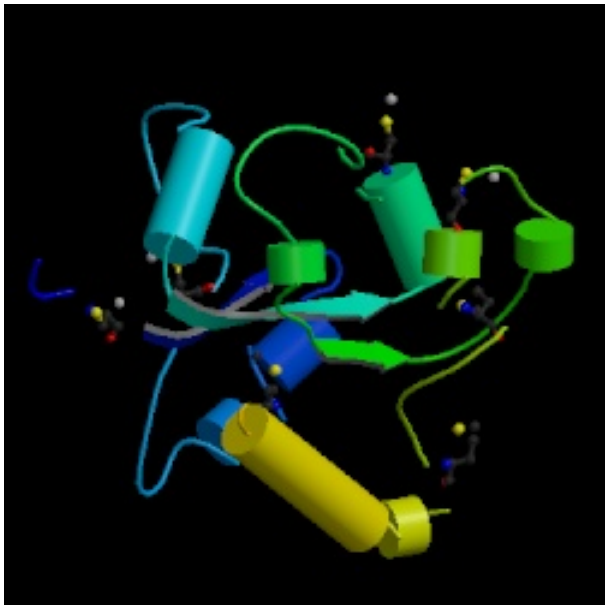
- Current size:
  - ✓ 121,745 sequence entries from 7,752 species.
  - ✓ 9,079 human proteins
- Core data: sequence, references, taxonomic data.
- Annotations: functions, post-translational modifications, domains and sites, secondary and quaternary structure, similarities, diseases, variants.
- **Not** easily linked with LocusLink or Unigene!

Swiss-Prot growth



PDB

- PDB (**http://www.rcsb.org/pdb/**) provides 3-dimensional structure data. Several display options are available.



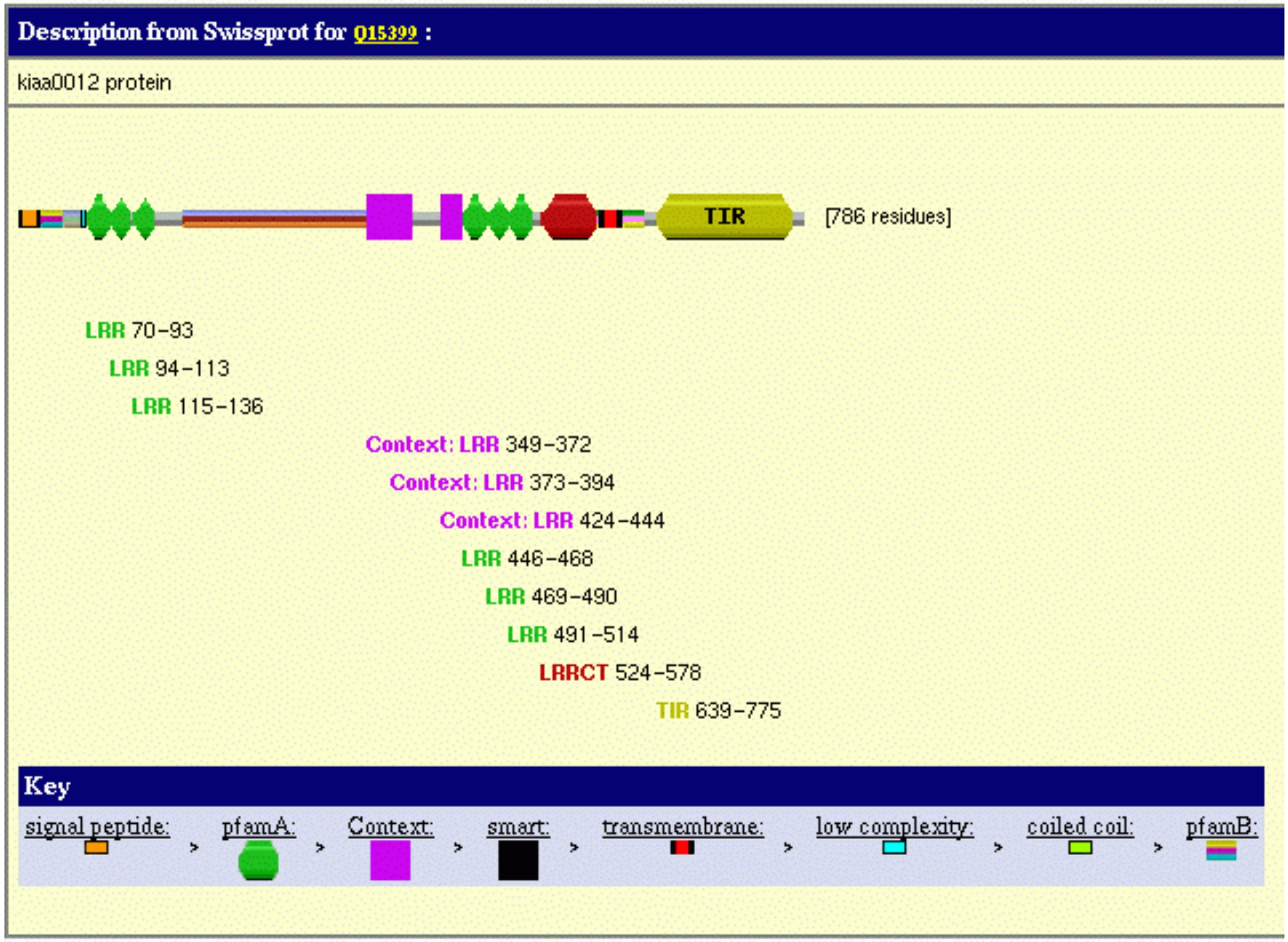
MMDB

- The Molecular Modeling Data Base (NCBI) provides structure data for over 100,000 macromolecules.
- Data originally comes from PDB, but is pre-processed for validation, error correction, format conversion.
- Other NCBI resources:
  - ✓ CDD (Conserved Domain Database)
  - ✓ COG (Clusters of Orthologous Groups)

Pfam

- Pfam (**http://www.sanger.ac.uk/Software/Pfam/**) is a database of protein domains and families, based on alignments (similarity) and Hidden Markov Models.
- Pfam-A (curated) contains 3,700 protein families. Pfam-B contains numerous, smaller families of lower quality.
- Requires Swiss-Prot or TrEMBL identifiers.

Pfam display



Protein interaction

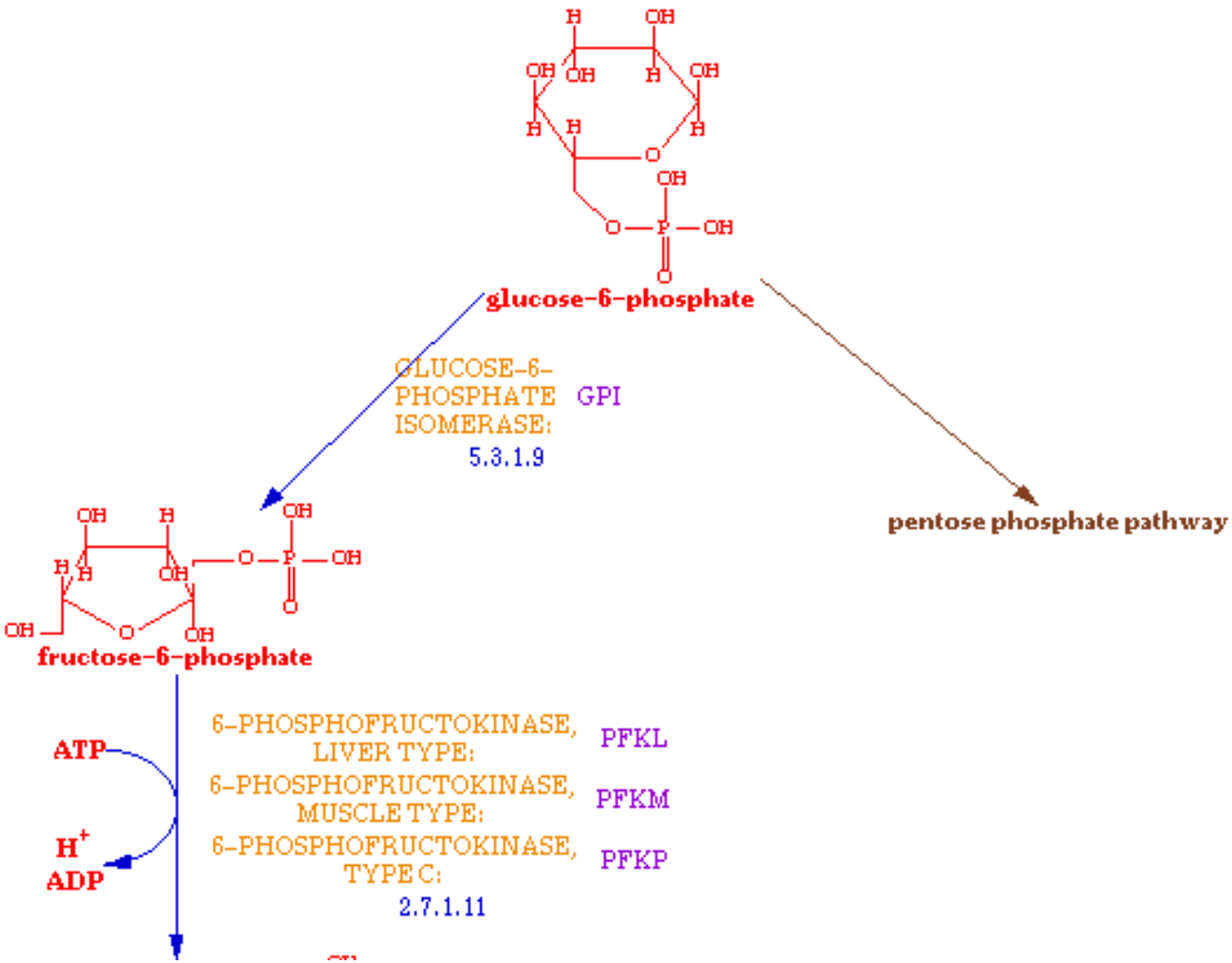
- DIP (**http://dip.doe-mbi.ucla.edu/**) reports experimentally determined interactions between proteins. The data is curated manually and computationally. Graph structure, with proteins as nodes and interactions (described by residue ranges, domains, dissociation constants) as edges.

- BIND (**http://www.bind.ca/**) records Interactions (cellular location, chemical action, kinetics, chemical state, etc), Molecular Complexes (complex-forming interactions) and Pathways (logically connected interactions).

Pathways

- KEGG (**http://www.genome.ad.jp/kegg/**) integrates knowledge on molecular interaction networks, chemical compounds and reactions, genes and proteins. 10,677 pathways with 481,325 genes in 132 organisms.

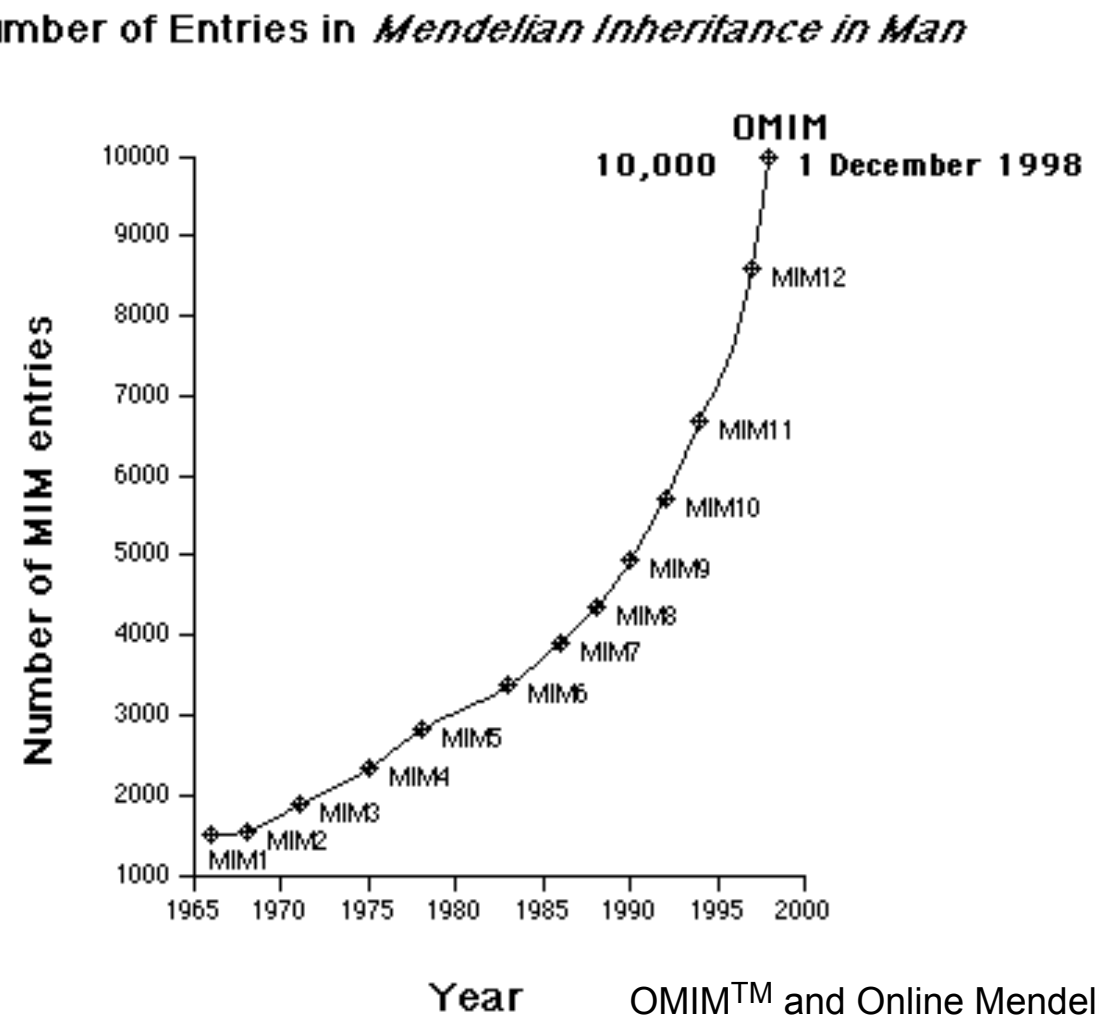
- BioCyc (**http://biocyc.org/**) is a collection of pathway databases for different organisms. 14 species, HumanCyc just released. Pathways are *computationally* derived in most cases.


glucose-6-phosphate
GLUCOSE-6-PHOSPHATE ISOMERASE: GPI
5.3.1.9
pentose phosphate pathway
fructose-6-phosphate
ATP
H+
ADP
6-PHOSPHOFRUCTOKINASE, LIVER TYPE: PFKL
6-PHOSPHOFRUCTOKINASE, MUSCLE TYPE: PFKM
6-PHOSPHOFRUCTOKINASE, TYPE C: PFKP
2.7.1.11

OMIM

- Catalog of human genes and genetic disorders, compiled at Johns Hopkins and hosted by the NCBI.
- Contains textual information, pictures, and reference information.
  - ✓ Description, Cloning, Biochemical features, Gene function, Mapping, Genotype/phenotype correlations, Allelic variants, References.
- 14,218 entries (10,569 genes and 1,229 phenotypes).

OMIM™ growth



PubMed

- Database of citations from the biomedical literature.
- Contains over 12 million entries, dating back to the mid 1960s, from 3,617 journals.
- Provides reference, abstract, links to online resources (full-text, supplementary material).
- Powerful search features. 30 million searches per month.

Speaking the same language: GeneOntology

- GeneOntology (**http://www.geneontology.org/**) is a dynamic controlled vocabulary that can be used to describe biological concepts.

- It is structured on three taxonomies:
  - ✓ Molecular Function
  - ✓ Biological Process
  - ✓ Cellular Component

- 5328 function, 6898 process, and 1174 component terms. Work in progress.

Taxonomy view

GO:0008150 : biological_process (30188)
- GO:0007154 : cell communication (6212)
  - GO:0009605 : response to external stimulus (2726)
    - GO:0009607 : response to biotic stimulus (1544)
      - GO:0006952 : defense response (1206)
        - **GO:0006955 : immune response (971)**
          - GO:0006953 : acute-phase response (9)
          - GO:0030105 : anaphylaxis (2)
          - GO:0019882 : antigen presentation (2)
          - GO:0030333 : antigen processing (16)
          - GO:0045321 : cell activation (63)
          - GO:0006968 : cellular defense response (126)
          - GO:0042107 : cytokine metabolism (30)
          - GO:0006959 : humoral immune response (238)
          - GO:0009682 : induced systemic resistance (0)
          - GO:0045087 : innate immune response (200)
          - GO:0009861 : jasmonic acid/ethylene dependent systemic resistance (25)
          - GO:0045728 : respiratory burst after phagocytosis (0)
          - GO:0009627 : systemic acquired resistance (13)

Conclusions

- We are drowning in data. Our task is to convert it into knowledge.
- Biomedical data covers the whole spectrum of knowledge representation and management techniques.
- Linking, interoperability, data import/export tools are critical. A uniform, stable nomenclature is essential.

Integration of data from multiple sources: SNPper

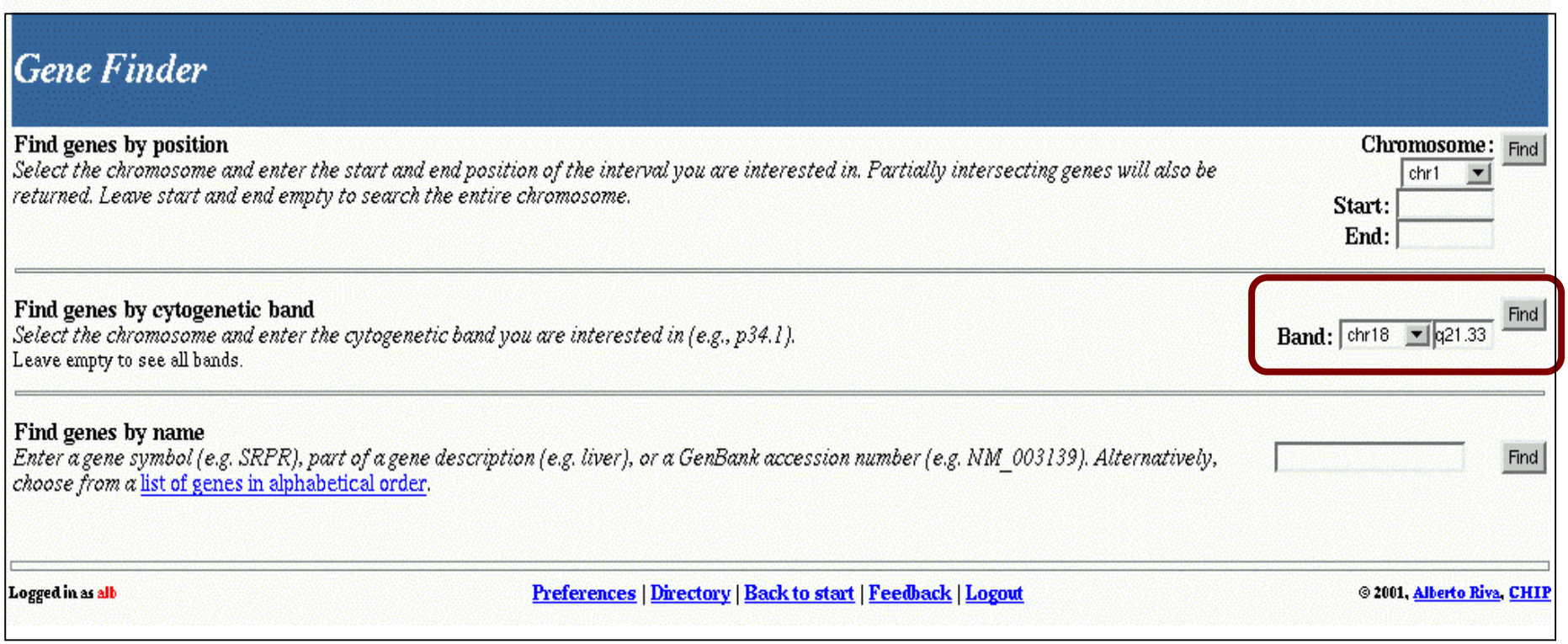
SNPper

- SNPper (**http://snpper.chip.org/**) is a search engine for SNPs. It retrieves known SNPs by name, by position, or by location on one or more genes.
- It can be used to construct sets of SNPs suitable for use in association studies.
- Web-based application written in Common Lisp; relies on a local relational database (mySQL) and on real-time access to GoldenPath. Freely available for academic purposes.

Example

Identify all exonic SNPs in the genes of band 18q21.33. Evaluate their potential significance. Export SNP data and primer design information.

SNPper – Find gene(s)



SNPper – Find gene(s)

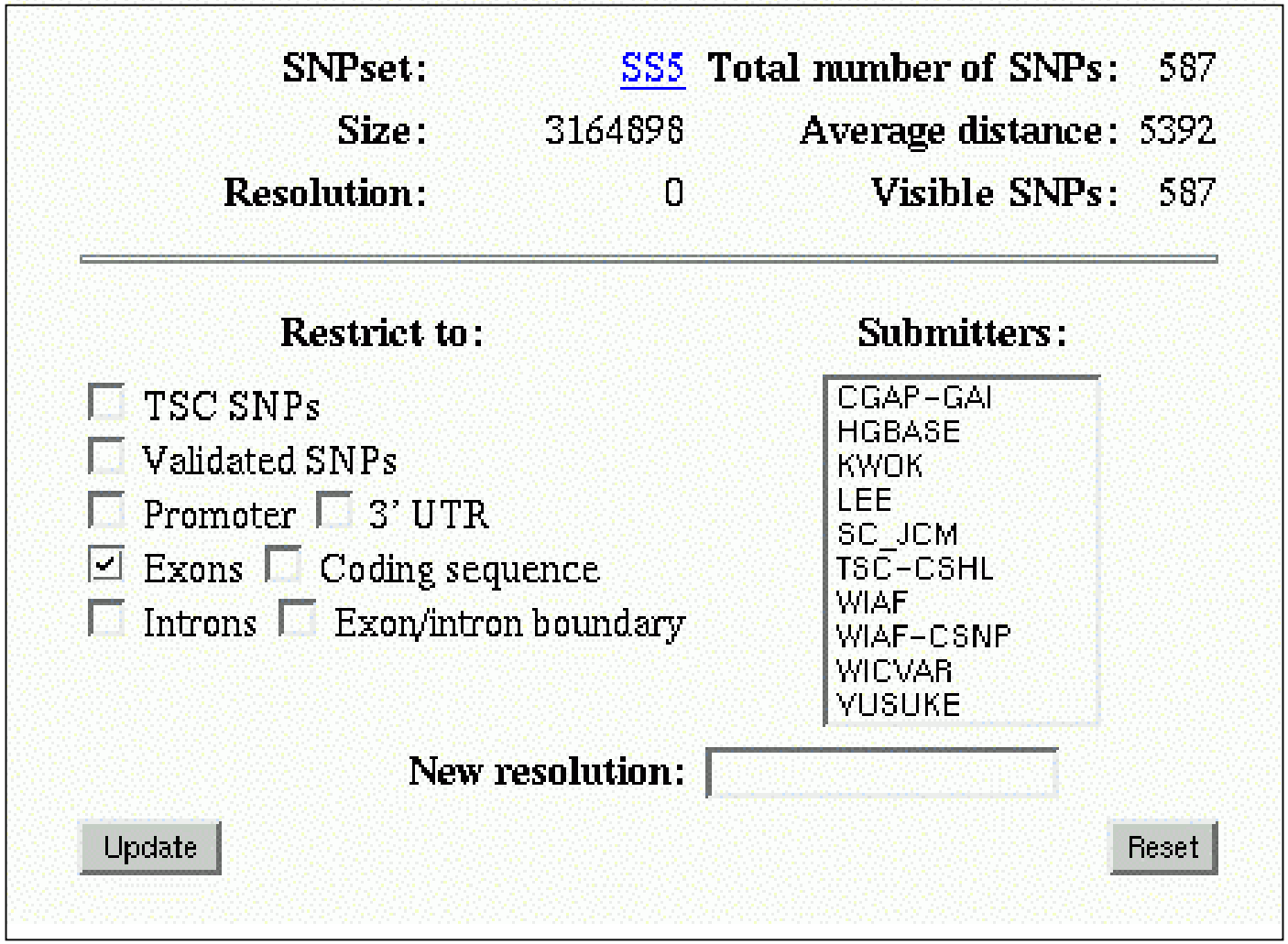
Genes in region:	chr18:62000000–65300000					
SNPS on these genes:	SS5					
Symbol	Chr	Start	End	Size	SNPs	Description
CDH20	chr18	62120522	62185133	64612	147	cadherin 20, type 2 preproprotein
MC4R	chr18	62267103	62268102	1000	8	melanocortin 4 receptor
PIGN	chr18	62969107	63111488	142382	124	phosphatidylinositol glycan, class N
TNFRSF11A	chr18	63245650	63306602	60953	23	tumor necrosis factor receptor superfamily,
MGC13269	chr18	63444076	63445735	1660	16	hypothetical protein MGC13269
FLJ20281	chr18	63494881	63499037	4157	10	hypothetical protein FLJ20281
BCL2	chr18	64495759	64671284	175526	27	B–cell lymphoma protein 2 alpha
BCL2	chr18	64670488	64671399	912	27	B–cell lymphoma protein 2 beta
FVT1	chr18	64683186	64719813	36628	61	follicular lymphoma variant translocation 1
SKD1	chr18	64741777	64775048	33272	22	vacuolar protein sorting factor 4B
SERPINB5	chr18	64829593	64857693	28101	54	serine (or cysteine) proteinase inhibitor, clade
SERPINB12	chr18	64908748	64919600	10853	18	serine (or cysteine) proteinase inhibitor, clade
SERPINB13	chr18	64941257	64949953	8697	15	serine (or cysteine) proteinase inhibitor, clade
SERPINB3	chr18	65007780	65014464	6685	20	serine (or cysteine) proteinase inhibitor, clade
SERPINB11	chr18	65062777	65075983	13207	63	serine (or cysteine) proteinase inhibitor, clade

SNPper - SNPset

SNPset: SS5

Source:	Genes in region chr18:62000000-65300000
Created on:	04/30/2002 11:43:24
SNPs:	587 (avg dist: 5392)
Spacing:	0
Commands:	Save this SNPset Refine this SNPset Export this SNPset XmlXport Get flanking sequences View all 587 SNPs...
(587 SNPs)	

SNPper – Refine SNPset



SNPper – Refine SNPset

SNPset: SS5

Source:	Genes in region chr18:62000000-65300000
Created on:	04/30/2002 11:43:24
SNPs:	64 (avg dist: 49452)
Spacing:	0
Filter:	Exon
Commands:	Save this SNPset Refine this SNPset Export this SNPset XmlXport Get flanking sequences View all 64 SNPs...
(64 SNPs)	

SNPper - SNPset

SNPset: SS5

Source:	Genes in region chr18:62000000–65300000
Created on:	04/30/2002 11:43:24
SNPs:	64 (avg dist: 49452)
Spacing:	0
Filter:	Exon
Commands:	Save this SNPset Refine this SNPset Export this SNPset XmlXport Get flanking sequences Hide SNPs

Name	Position	Gene	Genepos	Role
rs595093	chr18:62133052	CDH20	12518 C/G	Exon, Coding sequence
rs1943330	chr18:62137507	CDH20	16973 A/C	Exon, Coding sequence
rs2282556	chr18:62267395	MC4R	292 C/T	Exon, Coding sequence
rs2229616	chr18:62267410	MC4R	307 A/G	Exon, Coding sequence
rs1016862	chr18:62267609	MC4R	506 G/T	Exon, Coding sequence
rs1053404	chr18:63025676	PIGN	60218 C/T	Exon, Coding sequence
rs2298784	chr18:63034463	PIGN	51431 C/T	Exon, Coding sequence
rs1236159	chr18:63072812	PIGN	13082 A/G	Exon, Coding sequence

SNPper – Gene view

Gene: MC4R

Name:	melanocortin 4 receptor	**XmlXport**	
Sequence:	**Fasta** - **Annotated** - **Protein**	**Strand:**	+
Transcript Position:	chr18:62267103–62268102 (18q21.33)	**Length:**	1000
Coding Sequence Position:	chr18:62267103–62268102	**Length:**	1000

Look up this gene in:

Genbank (mRNA):	NM_005912	**Genbank (prot):**	NP_005903	**Entrez:**	MC4R
LocusLink:	4160	**PubMed:**	MC4R	**OMIM:**	155541
Unigene:	MC4R	**Ensembl:**	MC4R	**SwissProt:**	P32245

Exons:

#	Start		End		Length
1	62267103	(0)	62268102	(999)	1000
XmlXport			**Total:**		1000

Known SNPs:

SNPset: SS3

Source:	**MC4R**
Created on:	**04/30/2002 11:34:04**
SNPs:	**8** (avg dist: **1245**)
Spacing:	**0**
Commands:	Save this SNPset


43,447,605 CCAGTAATCT TTAGAGTACA TCAGAACCAG TTTTCTGATG GCCAATCTGC
43,447,655 TTTTAATTCA CTCTTAGACG TTAGAGAAAT AGGTGTGGTT TCTGCATAGG
43,447,705 GAAAATTCTG AAATTAAAAA TTTAATGGAT CCTAAGTGGA AATAATCTAG
43,447,755 GTAAATAGGA ATTAAATGAA AGAGTATGAG CTACATCTTC AGTATACTTG
43,447,805 GTAGTTTATG AGGTTAGTTT CTCTAATATA GCCAGTTGGT TGATTTCCAC
43,447,855 CTCCAAGGTG TATGAAGTAT GTATTTTTTT AATGACAATT CAGTTTTTGA
43,447,905 GTACCTTGTT ATTTTTGTAT ATTTTCAGCT GCTTGTGAAT TTTCTGAGAC
43,447,955 GGATGTAACA AATACTGAAC ATCATCAACC CAGTAATAAT GATTTGAACA
43,448,005 CCACTGAGAA GCGTGCAGCT GAGAGGCATC CAGAAAAGTA TCAGGGTAGT
43,448,055 TCTGTTTCAA ACTTGCATGT GGAGCCATGT GGCACAAATA CTCATGCCAG
43,448,105 CTCATTACAG CATGAGAACA GCAGTTTATT ACTCACTAAA GACAGAATGA
43,448,155 ATGTAGAAAA GGCTGAATTC TGTAATAAAA GCAAACAGCC TGGCTTAGCA
43,448,205 AGGAGCCAAC ATAACAGATG GGCTGGAAGT AAGGAAACAT GTAATGATAG rs1800063
43,448,255 GCGGACTCCC AGCACAGAAA AAAAGGTAGA TCTGAATGCT GATCCCCTGT
43,448,305 GTGAGAGAAA AGAATGGAAT AAGCAGAAAC TGCCATGCTC AGAGAATCCT rs1799950
43,448,355 AGAGATACTG AAGATGTTCC TTGGAT
43,448,405 AGTTAATGAG TGGTTTTCCA GAAGTG
43,448,455 CACATGATGG GGAGTCTGAA TCAAAT
43,448,505 GTTCTAAATG AGGTAGATGA ATATTC
43,448,555 ACTGGCCAGT GATCCTCATG AGGCTT
43,448,605 ACTCCAAATC AGTAGAGAGT AATATT
43,448,655 TATCGGAAGA AGGCAAGCCT CCCCAACTTA AGCCATGTAA CTGAAAATCT
43,448,705 AATTATAGGA GCATTTGTTA CTGAGCCACA GATAATACAA GAGCGTCCCC
43,448,755 TCACAAATAA ATTAAAGCGT AAAAGGAGAC CTACATCAGG CCTTCATCCT
43,448,805 GAGGATTTTA TCAAGAAAGC AGATTTGGCA GTTCAAAAGA CTCCTGAAAT
43,448,855 GATAAATCAG GGAACTAACC AAACGGAGCA GAATGGTCAA GTGATGAATA
43,448,905 TTACTAATAG TGGTCATGAG AATAAAACAA AAGGTGATTC TATTCAGAAT
43,448,955 GAGAAAAATC CTAACCCAAT AGAATCACTC GAAAAAGAAT CTGCTTTCAA
43,449,005 AACGAAAGCT GAACCTATAA GCAGCAGTAT AAGCAATATG GAACTCGAAT
43,449,055 TAAATATCCA CAATTCAAAA GCACCTAAAA AGAATAGGCT GAGGAGGAAG
43,449,105 TCTTCTACCA GGCATATTCA TGCGCTTGAA CTAGTAGTCA GTAGAAATCT rs1800064
43,449,155 AAGCCCACCT AATTGTACTG AATTGCAAAT TGATAGTTGT TCTAGCAGTG
43,449,205 AAGAGATAAA GAAAAAAAAG TACAACCAAA TGCCAGTCAG GCACAGCAGA
43,449,255 AACCTACAAC TCATGGAAGG TAAAGAACCT GCAACTGGAG CCAAGAAGAG
43,449,305 TAACAAGCCA AATGAACAGA CAAGTAAAAG ACATGACAGC GATACTTTCC rs1799949
rs1799950 Close
chr17:43,448,330 (29630)
Alleles: [A/G]
Amino acid change: Q/R
Validated: N

SNPper – Protein view

Gene:	MC4R (melanocortin 4 receptor)		
Position:	chr18:62267103–62268102	SNPs:	3
View:	Genomic sequence	Stops:	1

```
  1 ATG GTG AAC TCC ACC CAC CGT GGG ATG CAC ACT TCT CTG CAC CTC TGG
  1 Met Val Asn Ser Thr His Arg Gly Met His Thr Ser Leu His Leu Trp

 49 AAC CGC AGC AGT TAC AGA CTG CAC AGC AAT GCC AGT GAG TCC CTT GGA
 17 Asn Arg Ser Ser Tyr Arg Leu His Ser Asn Ala Ser Glu Ser Leu Gly

 97 AAA GGC TAC TCT GAT GGA GGG TGC TAC GAG CAA CTT TTT GTC TCT CCT
 33 Lys Gly Tyr Ser Asp Gly Gly Cys Tyr Glu Gln Leu Phe Val Ser Pro

145 GAG GTG TTT GTG ACT CTG GGT GTC ATC AGC TTG TTG GAG AAT ATC TTA
 49 Glu Val Phe Val Thr Leu Gly Val Ile Ser Leu Leu Glu Asn Ile Leu

193 GTG ATT GTG GCA ATA GCC AAG AAC AAG AAT CTG CAT TCA CCC ATG TAC
 65 Val Ile Val Ala Ile Ala Lys Asn Lys Asn Leu His Ser Pro Met Tyr

241 TTT TTC ATC TGC AGC TTG GCT GTG GCT GAT ATG CTG GTG AGC GTT TCA
 81 Phe Phe Ile Cys Ser Leu Ala Val Ala Asp Met Leu Val Ser Val Ser

289 AAT GGA TCA GAA ACC ATT GTC ATC ACC CTA TTA AAC AGT ACA GAT ACG     rs2282556 rs2229616
 97 Asn Gly Ser Glu Thr Ile Val Ile Thr Leu Leu Asn Ser Thr Asp Thr
       Arg               Ile

337 GAT GCA CAG AGT TTC ACA GTG AAT ATT GAT AAT GTC ATT GAC TCG GTG
113 Asp Ala Gln Ser Phe Thr Val Asn Ile Asp Asn Val Ile Asp Ser Val

385 ATC TGT AGC TCC TTG CTT GCA TCC ATT TGC AGC CTG CTT TCA ATT GCA
129 Ile Cys Ser Ser Leu Leu Ala Ser Ile Cys Ser Leu Leu Ser Ile Ala
```

Swiss-Prot domains		
Pos	**Name**	**Description**
1–43	Domain	EXTRACELLULAR (POTENTIAL).
3–3	Carbohyd	N-LINKED (GLCNAC...) (POTENTIAL).
17–17	Carbohyd	N-LINKED (GLCNAC...) (POTENTIAL).
26–26	Carbohyd	N-LINKED (GLCNAC...) (POTENTIAL).
30–30	Variant	S -> R (IN OBESITY). /FTId=VAR_010704.
37–37	Variant	D -> V (IN OBESITY). /FTId=VAR_010705.
44–69	Transmem	1 (POTENTIAL).
70–81	Domain	CYTOPLASMIC (POTENTIAL).
78–78	Variant	P -> L (IN OBESITY). /FTId=VAR_010706.
82–106	Transmem	2 (POTENTIAL).
103–103	Variant	I -> V. /FTId=VAR_010707.
107–123	Domain	EXTRACELLULAR (POTENTIAL).
112–112	Variant	T -> M (IN OBESITY). /FTId=VAR_010708.
124–145	Transmem	3 (POTENTIAL).
146–165	Domain	CYTOPLASMIC (POTENTIAL).
165–165	Variant	R -> W (IN OBESITY). /FTId=VAR_010709.
166–186	Transmem	4 (POTENTIAL).
169–169	Conflict	I -> S (IN REF. 2).
187–191	Domain	EXTRACELLULAR (POTENTIAL).
192–215	Transmem	5 (POTENTIAL).

SNPper – SNP view

SNP: rs1799950

Position:	chr17:43448330	**Band:**	17q21.33	**Alleles:**	A/G	**Avg Het:**	(unknown)
dbSNP:	rs1799950	**GenBank:**	L78833:1	**Updated:**	01/30/2001	**Validated:**	N
Gene:	BRCA1	**Role:**	Exon, Coding sequence	**Relative position:**	29630	**Amino acid change:**	356 Q/R
Gene:	BRCA1	**Role:**	Exon, Coding sequence	**Relative position:**	29630	**Amino acid change:**	356 Q/R
Gene:	BRCA1	**Role:**	Exon, Coding sequence	**Relative position:**	29630	**Amino acid change:**	356 Q/R
Gene:	BRCA1	**Role:**	Exon, Coding sequence	**Relative position:**	179	**Amino acid change:**	60 Q/R
Gene:	BRCA1	**Role:**	Intron	**Relative position:**	29630	**Amino acid change:**	(none)
Gene:	BRCA1	**Role:**	Exon, Coding sequence	**Relative position:**	29630	**Amino acid change:**	356 Q/R
Gene:	BRCA1	**Role:**	Exon, Coding sequence	**Relative position:**	29630	**Amino acid change:**	356 Q/R
Gene:	BRCA1	**Role:**	Exon, Coding sequence	**Relative position:**	29630	**Amino acid change:**	356 Q/R
Gene:	BRCA1	**Role:**	Exon, Coding sequence	**Relative position:**	29630	**Amino acid change:**	315 Q/R
Gene:	BRCA1	**Role:**	Intron	**Relative position:**	29630	**Amino acid change:**	(none)
Gene:	BRCA1	**Role:**	Intron	**Relative position:**	29630	**Amino acid change:**	(none)
Gene:	BRCA1	**Role:**	Intron	**Relative position:**	29630	**Amino acid change:**	(none)
Gene:	BRCA1	**Role:**	Exon	**Relative position:**	29630	**Amino acid change:**	(none)

Submitters

dbSNP assay	Submitter	Private ID
ss2420002	HGBASE	SNP000002456

SNPper – Export SNPs

SNPset: SS5 **Source:** Genes in region chr18:62000000–65300000
Details: 64 SNPs, created 04/30/2002 11:43:24
Filter: Exon

Choose fields:

- [x] SNP rs#
- [ ] SNP position
- [ ] Band
- [ ] Distance from previous SNP
- [ ] Alleles
- [ ] Gene
- [ ] Role
- [x] Amino acid change
- [x] Amino acid position
- [ ] Flanks
- [ ] Contig
- [ ] Submitters
- [x] Validated

Choose output format:

- ( ) Mail to Alberto.Riva@TCH.Harvard.edu
- (•) View as HTML table

Execute | Reset form

SNPper – Export SNPs

SNPset: SS5 **Source:** Genes in region chr18:62000000–65300000
Details: 64 SNPs, created 04/30/2002 11:43:24
Filter: Exon

SNP rs#	SNP position	Amino acid change	Amino acid position	Validated
rs595093	chr18:62133052	V/V	260	N
rs1943330	chr18:62137507	P/H	328	Y
rs2282556	chr18:62267395	G/R	98	N
rs2229616	chr18:62267410	V/I	103	Y
rs1016862	chr18:62267609	I/S	169	Y
rs1053404	chr18:63025676	L/L	674	N
rs2298784	chr18:63034463	A/A	496	N
rs1236159	chr18:63072812	F/F	207	N
rs1805033	chr18:63245680			N
rs1805034	chr18:63280341	A/V	192	N
rs3017358	chr18:63305708			N
rs2980963	chr18:63305831			N
rs1515682	chr18:63445188	R/G	356	N
rs2674020	chr18:63445519			N
rs659683	chr18:63445653			N
rs1050618	chr18:64499594			N
rs1564483	chr18:64499835			Y
rs2078303	chr18:64499974			N

SNPper – Export SNPs

SNPset: SS5 **Source:** Genes in region chr18:62000000–65300000
Details: 18 SNPs, created 04/30/2002 11:43:24
Filter: Exon, Validated

Choose fields:

- [x] SNP rs#
- [ ] SNP position
- [ ] Band
- [ ] Distance from previous SNP
- [ ] Alleles
- [ ] Gene
- [ ] Role
- [ ] Amino acid change
- [ ] Amino acid position
- [x] Flanks
- [ ] Contig
- [ ] Submitters
- [ ] Validated

Choose output format:

- ( ) Mail to Alberto.Riva@TCH.Harvard.edu
- (•) View as HTML table

Execute | Reset form

SNPper – Export SNPs

SNPset: SS5 **Source:** Genes in region chr18:62000000–65300000
Details: 18 SNPs, created 04/30/2002 11:43:24
Filter: Exon, Validated

SNP rs#	
rs1943330	TATAGTGACGGATTATCAAGAATGCGTGTCCTGGCTTTTCTTCGTAGAACATTACCAGATGAGTGTGTTGGAATCAGCTCCAATTAG
rs2229616	CTGATGGAGGGTGCTACGAGCAACTTTTTGTCTCTCCTGAGGTGTTTGTGACTCTGGGTGTCATCAGCTTGTTGGAGAATATCTTAGT
rs1016862	TGTCATCACCCTATTAAACAGTACAGATACGGATGCACAGAGTTTCACAGTGAATATTGATAATGTCATTGACTCGGTGATCTGTAC
rs1564483	CTCTCCAAAGTCATTTAAAGCCTTGCTTTAAACTCACAGGTGGGCCAAGGCCACACAGCCAACGTGCCATGTGCTACAGCCAAAAT
rs1016860	AGGTTCTGCGGACTTCGGTCTCCTAAAAGCAGGCACTTGTGGCGGCCTGATGCTCTGGGTAACTCTAGCCTTCCTGATGCGGAAGTC
rs6810	CTTCAGAAAACATTTAAGGAGGCTGCCTTTCCCCATGAATGAGAAATGAAACGTTCCCTAAGCATTTGAGTCTAAAGACAAGAAAG
rs1455555	CAGAGTCACTGTCACAGTGGACTAATCCCAGCACCATGGCCAATGCCAAGGTCAAACTCTCCATTCCAAAATTTAAGGTGGAAAAG
rs1020694	AAATTAATAACTTCTGACTGACACAAATCAGTGTTACACTGTTTTAGATTTTTATTGATAATCATGCCATTCTACTCTTCTTTTTTGA
rs6098	ATGTACCAAAAATAAAGTATGTATTAATATAGCATTTCATGATTGTATTCAAAAACTATTACCATGGCTTAAGAACTATCTTGTTTAC
rs6106	CCTTCAGCACCTGCCTTCCATAGCCAACCTCCACTCCCACCCTACCCCAGGTCTCCTAATTTCAATGGGAAGACCATAATTCACCAT
rs6101	CTCCCACCCTACCCCAGGTCTCCTAATTTCAATGGGAAGACCATAATTCACCATTATGCCATGGCTTGTTTGGTATGTATTTTATGTA
rs6105	AGGTCTCCTAATTTCAATGGGAAGACCATAATTCACCATTATGCCATGGCTTGTTTGGTATGTATTTTATGTAGCCTTTGTCATTTTCT
rs6102	ATGGGCATGGAGGACGCCTTCAACAAGGGACGGGCCAATTTCTCAGGGATGTCGGAGAGGAATGACCTGTTTCTTTCTGAAGTGTT
rs6103	GCCAATTTCTCAGGGATGTCGGAGAGGAATGACCTGTTTCTTTCTGAAGTGTTCCACCAAGCCATGGTGGATGTGAATGAGGAGGG
rs6104	GAATGACCTGTTTCTTTCTGAAGTGTTCCACCAAGCCATGGTGGATGTGAATGAGGAGGGCACTGAAGCAGCCGCTGGCACAGGAC
rs12102	TCAGTTTATTTTTATAACATTAACTTTTACTTTGTTATTTATTATTTTATATAATGGTGAGTTTTTAAATTATTGCTCACTGCCTATTTA
rs724558	GAAAACATTTTTATCCACTTTAGGAACATTTTCAACTTTTGTTTGCAATTCCCCAAATAATTTTATTTAGCTTTAGTTTGCTAACATGT

Conclusions (II)

- We need tools to help us make sense of the data we are drowning in.
- Integrating multiple data sources is hard. Differences in nomenclature, semantics, scope. Keeping up with updates is a full-time job.
- Automated, interoperable, autonomous tools will soon become an essential aid to computational biology research.