



Application of Information Sciences to Analysis and Interpretation of Novel Genetic Data

**NIA Alzheimer's Disease Centers
Directors Meeting**

Speaker:

Christopher Dubay

Division of Medical Informatics &
Outcomes Research, OHSU

1



Introduction & Overview

- Speaker Introduction
- Description of the Talk & Bioinformatics
 - An Integrative Information Science
- Illustrative Exercise: Processing a 'Top 50' candidate genes list
- Future goals for Bioinformatics:
 - Systems Biology of Complex Phenotypes
 - The New Medicine

2

Bioinformatics



- General Tools
 - WP, Spreadsheets, Robotics, Instrumentation
- Communications
 - E-Mail, Networks, Internet & World Wide Web
- Databases
 - Storage, Organization
- Analysis Tools
 - Examination & Discovery
- **Informatics Has Changed How Science is Done**

3

Basic Point of Talk



- New Biology ->
- New Genetics ->
- New Medicine ->
- How Information Sciences & Technology can Support and Influence this Process

4

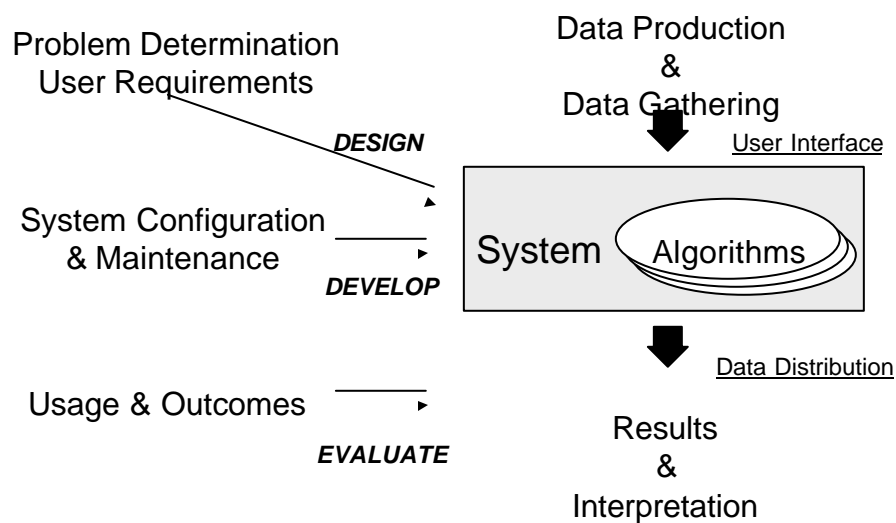
Bioinformatics Skill Set



- Practical Tools
- Cross Cultural Exchange
 - Language of Biomedical Research
 - Language of Informatics
- Solving Scientific Problems using Computers
 - Database Interoperation
 - Process Modeling & Data Visualization
- Bioinformatics is an Information Science

5

Informatics Research



6

Bioinformatics Skill Set



- Practical Tools
- Cross Cultural Exchange
 - Language of Biomedical Research
 - Language of Informatics
- Solving Scientific Problems using Computers
 - Database Interoperation
 - Process Modeling & Data Visualization
- Bioinformatics is an Information Science

7

Bioinformatics Significance



RESEARCH NEWS

Missing Alzheimer's Gene Found

Researchers find the gene that causes Alzheimer's disease in "Volga German" families. It shows a remarkable similarity to another recently discovered Alzheimer's gene.

This recent research on Alzheimer's disease, the memory-impairing disorder that affects 11 million to 20 million people worldwide, has ignited as much in the scientific community as it has in the general public.

Discovery was provocative because it provided a direct link to a characteristic feature of Alzheimer's pathology: Aβ, the cause of a peptide called β-amyloid that is found in the abnormal "senile plaques" that build up in the brains of Alzheimer's patients.

sequence. Ingrid (EST) sequences that DNA sequences known to come from active genes. Watson found an EST with a sequence similar to S182. "That's weird," she said. "Maybe this is the Volga German gene."

After the S182 sequence was published, Watson told Schellenberg about it. "Having seen a million conditions for the Volga German gene (come and won't) occur," Schellenberg readily placed Lave-Labad, in his lab group, ahead and looked. She found that the gene was not only on chromosome 1, but with the very stretch of DNA that she had

pinpointed as the likely site of the Alzheimer's gene. "That was like a sledgehammer to the forehead," says Schellenberg. "It went from being a ho-hum project to ... saying 'oh my God this is the gene.'"

Within a few days, the team sequenced the gene from Volga German family members, with help from David Galas and his col-

leagues. The results show that the protein they found has a similar function. According to Schellenberg, the resemblance "suggests that something about this type of ... protein is very important for the biology of Alzheimer's disease."

As a bonus, the so-called Volga German, who were all descended from a colony of ethnic Germans in the United States, also had a similar gene.

● Discovered in 1993
● Gene isolated in 1994

Family resemblance. Mutations in the amyloid precursor made by the genes S182 and S182/182 cluster around the membrane-spanning regions.

8

Illustrative Exercise



- What do you do with a 'Top 50' list of candidate genes?
 - Nomenclature
 - Similarity:
 - » Coding, non-Coding
 - » Structure & Function
 - » Cross-Species homology
 - Known Variation
 - Published Literature
 - » Systems Biology: pathways & interactions

9

Nomenclature



- Key to being able to find references
 - New and Old references
- Current Central Repository for Gene Names
 - HUMAN Genome Organization - Genew

10

HUGO - Genew

HUGO Gene Nomenclature Committee (HGNC) - Microsoft Internet Explorer

Address: <http://www.gene.ucl.ac.uk/nomenclature/>

HUGO Gene Nomenclature Committee

Home About HGNC Database Guidelines Submissions Downloads Gene Families

Giving unique and meaningful names to every human gene

Commercial Users

Contact Us

Database Links

FAQs

AC (International Union of Pure and Applied Chemistry)

Journal Links

Meetings

Newsletter

Public Interest

Publications

Search Approved Symbols

We have approved symbols for nearly one half of the genes in the human genome and, with an estimated 15,000 more genes to name, we still have plenty to do! Use the Gene database to search for your gene.

Quick Gene Search

Gene Symbol Submission

Obtaining a gene symbol before publication will avoid any possible conflicts with existing symbols and will ensure that your gene is promptly recorded in our database and others. Any information that you provide will be treated in the strictest confidence.

NCBI

BLAST

Protein

GenBank

EMBL

Ensembl

RefSeq

UniProt

GeneCards

Gene Database

<http://www.gene.ucl.ac.uk/nomenclature/genefamils.shtml>

Internet

11

HUGO - Genew

Search Results - Microsoft Internet Explorer

Address: <http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl>

[Back To Search](#)

Returned: 2 of 2 records
where "aliases" does contain "S182|STM2"

Please click on the Approved Gene Symbol to retrieve the complete gene record

Approved Gene Symbol	Approved Gene Name	Location	Sequence Accession IDs	Previous Symbols	Aliases
PSEN1	presenilin 1 (Alzheimer disease 3)	14q24.3	NM_000021	AD3	FAD , S182, PS1
PSEN2	presenilin 2 (Alzheimer disease 4)	1q31-q42	U44572 NM_012486	AD4	AD3L, STM2, PS2

Maintained by noms@ealton.ucl.ac.uk Script last updated: July 2002 [HGNC Homepage](#)

Similarity - Biosequence



- DNA Sequences
 - Get from GenBank
 - Find Genomic Location and Control
 - Gene Family?
- Protein Sequence
 - Function
 - Expression

13

GenBank



The screenshot shows the NCBI Sequence Viewer for the PSEN1 gene (NM_000021). The interface includes a search bar, navigation tabs (Nucleotide, Protein, Genes, Structure, Popset, Taxonomy, OMIM, Maps), and a detailed view of the gene's information.

NCBI Sequence Viewer - Microsoft Internet Explorer
 Address: http://www.ncbi.nlm.nih.gov/entrez/viewer/fg?val=NM_000021

NCBI Nucleotide

Published Nucleotide Protein Genes Structure Popset Taxonomy OMIM Maps

Search: Nucleotide for [Go] [Clear]

Display: default [Save] [Text] [Add to Clipboard] [Get Subsequence] [Details]

☐ L: NM_000021 Homo sapiens presenilin 1 [gi21536454]

LOCUS PSEN1 2763 bp mRNA linear PRI 21-
 DEFINITION Homo sapiens presenilin 1 (Alzheimer disease 3) (PSEN1) tra-
 variant 1-467, mRNA.
 ACCESSION NM_000021
 VERSION NM_000021.2 GI:21536454
 KEYWORDS -
 SOURCE -
 ORGANISM [Homo sapiens](#)
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
 1 (bases 1 to 2763)
 REFERENCE
 AUTHORS Schellenberg, G.P., Bird, T.D., Wjstman, E.H., Orr, H.T., Anderson
 Messing, E., White, J.A., Bonnyycastle, L., Weber, J.L., Alonso, M.E.,
 Potter, H., Boston, L.L. and Martin, G.H.
 TITLE Genetic linkage evidence for a familial Alzheimer's disease locus
 on chromosome 14

Related Unigene
[Map Viewer](#)
[OMIM](#)
[Protein](#)
[PubMed](#)
[SNP](#)
[Taxonomy](#)
[UniSTS](#)
[LinkOut](#)
[Help](#)

14

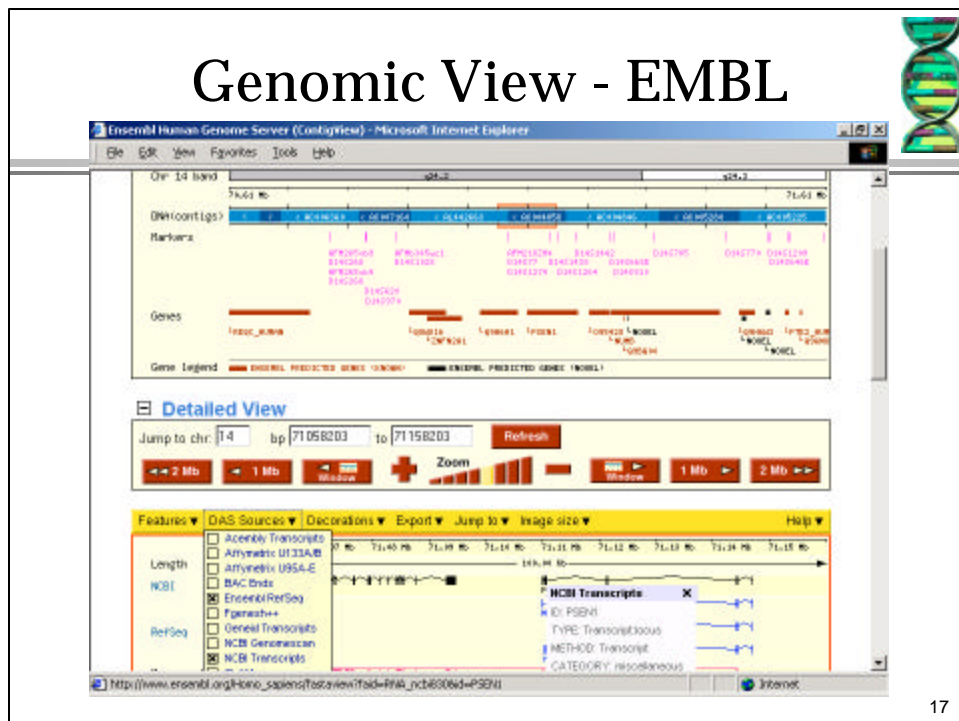
LocusLink

15

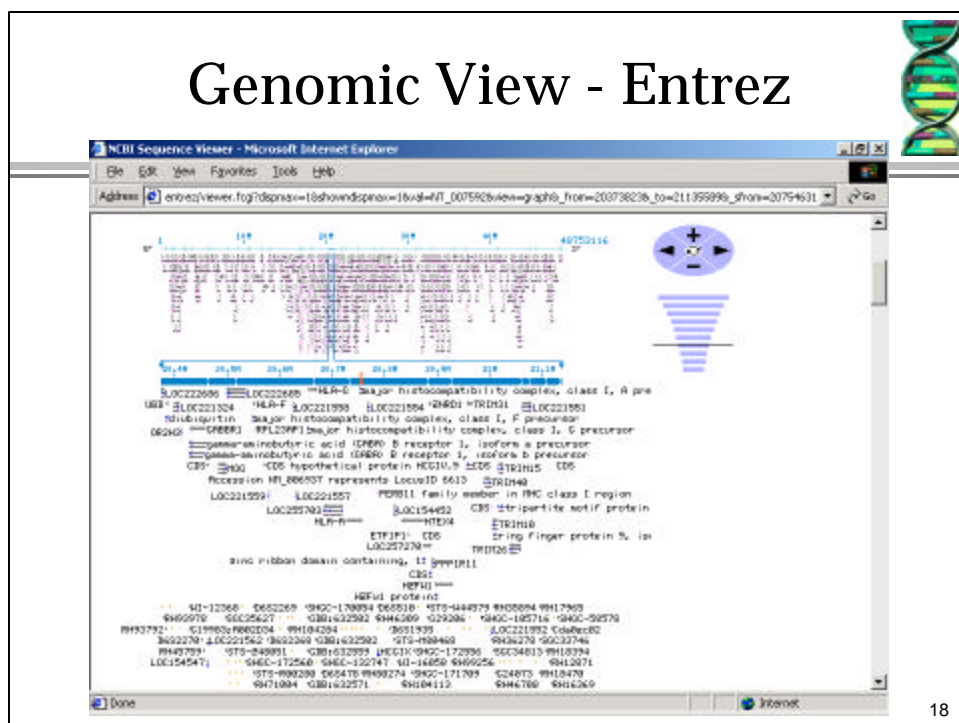
Genomic View - NCBI

16

Genomic View - EMBL



Genomic View - Entrez



Sequence & Results Manager: GCGblank



19

Promoters & Enhancers



- Transcriptional Elements
- For Genes & Genomic Region
 - Enhancers can be distant

20

Transcription Factor Database



TRRD - Microsoft Internet Explorer

Address: http://www.bionet.nrc.ru/ing/gw/tdr/

Gene Express 2.1

HOME DBS RNA PROTEIN GENENETWORK MAP

TRANSCRIPTION REGULATORY REGIONS DATABASE

TRRD

General information
 How to site TRRD?
 TRRD publications
 The latest report on TRRD
 Contact us
 Acknowledgments
 User's guide
 Database scheme
 How to search TRRD?
 How close with other databases
 TRRD Viewer
 FAQ
 What's new?
 How is TRRD updated?
 Standardization of information
 FAQ

TRRD is a unique information resource, accumulating information on structural and functional organization of transcription regulatory regions of eukaryotic genes. Only experimental information is included into TRRD.

[What's new?](#)

ACCESS to TRRD:

[SRS ACCESS](#) [TRRDGENES](#) [TRRDEXP](#) [TRRDSTATS](#) [TRRDFACTORS](#) [TRRDDB](#) [TRRDUNITS](#)

[TRRDViewer](#)
 (Access the TRRD
 TRRD sections, located within functional systems)

General information
[How to site TRRD?](#)
[TRRD publications](#)
[The latest report on TRRD](#)
[TRRDViewer](#)
[Contact us](#)
[Acknowledgments](#)

User's guide
[Database scheme](#)
[How to search TRRD?](#)
[Information with other databases](#)
[TRRD Viewer](#)
[FAQ](#)

How is TRRD updated?
[Standardization of information input](#)
[TRRD progress \(from 1996\)](#)

Current TRRD release
[Information contents](#)
[TRRD statistics](#)

TOOLS

http://www.bionet.nrc.ru/ing/gw/tdr/tdr_viewer.shtml

21

TRRD - Viewer



TRRD Viewer - Microsoft Internet Explorer

Address: http://www.bionet.nrc.ru/ing/gw/tdr/tdr_viewer.shtml

Gene Express 2.1

HOME DBS RNA PROTEIN GENENETWORK MAP

TRANSCRIPTION REGULATORY REGIONS DATABASE

TRRD

General information
 How to site TRRD?
 TRRD publications
 The latest report on TRRD
 Contact us
 Acknowledgments
 User's guide
 Database scheme
 How to search TRRD?
 How close with other databases
 TRRD Viewer
 FAQ
 What's new?
 How is TRRD updated?
 Standardization of information
 FAQ

Visualization of a gene regulatory map is exemplified in Figure 1. Three windows are provided: (1) navigation window, (2) text box with the relevant information and designations, and (3) window with the map of gene regulatory regions.

Figure 1. Visualization of the map of gene regulatory regions by TRRD Viewer exemplified with human cholesterol 7-alpha-hydroxylase gene (CYP7). A hooked arrow indicates the transcription start. Shown on the axis is the distance from the reference point (here, the transcription start). A pop-up text box (yellow rectangle) presents the information on SRE binding site.

Navigation window

Transcription start

Pop-up text box

Designations of transcription factor binding sites

The following options are provided:

22

Similarity - Function



- Group genes of similar function
 - Spatially
 - Temporally
 - Action
- Need for common Vocabulary

23

GO: Gene Ontology



LocusLink Report - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.ncbi.nlm.nih.gov/locuslink/Lod/pt.cgi?w=5663> Go

LocusLink Home

[PSENI Index](#)
[Top of Page](#)
[Nomenclature](#)
[Overview](#)
[Function](#)
[Relationships](#)
[Map](#)
[RefSeq](#)
[GenBank](#)
[Links](#)

[LocusLink](#)
[Collaborators](#)
[Download](#)
[FAQ](#)
[Help](#)
[Statistics](#)

[RefSeq](#)
[About](#)
[Download](#)
[FAQ](#)
[Statistics](#)

Gene Ontology™:

Term	Evidence	Source	Pub
• centromere	E	Proteome	pm
• kinetochore	E	Proteome	pm
• anti-apoptosis	E	Proteome	pm
• membrane fraction	E	Proteome	pm
• chromosome segregation	P	Proteome	pm
• integral plasma membrane protein	P	Proteome	pm
• chromosome organization and biogenesis	P	Proteome	pm
• nuclear outer membrane, integral protein	E	Proteome	pm

Other Ontologies:

Term	Evidence	Source	Pub
• Nuclear	E	Proteome	pm
• Cell death/Apoptosis	E	Proteome	pm
• CNS-specific functions	P	Proteome	pm
• Integral membrane	NR	Proteome	pm
• Unspecified membrane	E	Proteome	pm
• DNA-associated (direct or indirect)	E	Proteome	pm

Relationships ?

Internet

24

Similarity - Orthologs & Paralogs



- Function Clues from:
 - What is known in other organisms
 - Evolution of gene in organisms

25

Jackson Labs: MGI



MGI 2.0 - Genes and Markers Query Results (Details) - Microsoft Internet Explorer

Address: <http://www.informatics.jax.org/searches/marker.cgi?35661>

Type: Gene
 Symbol: Psen1
 Name: presenilin 1
 Chromosome: 12
 cM Position: 37.0
 MGI Accession ID: MGI1202717

Synonyms: S182 / PS1 / presenilin-1 / PS-1 / Aβ3h

Additional Information:

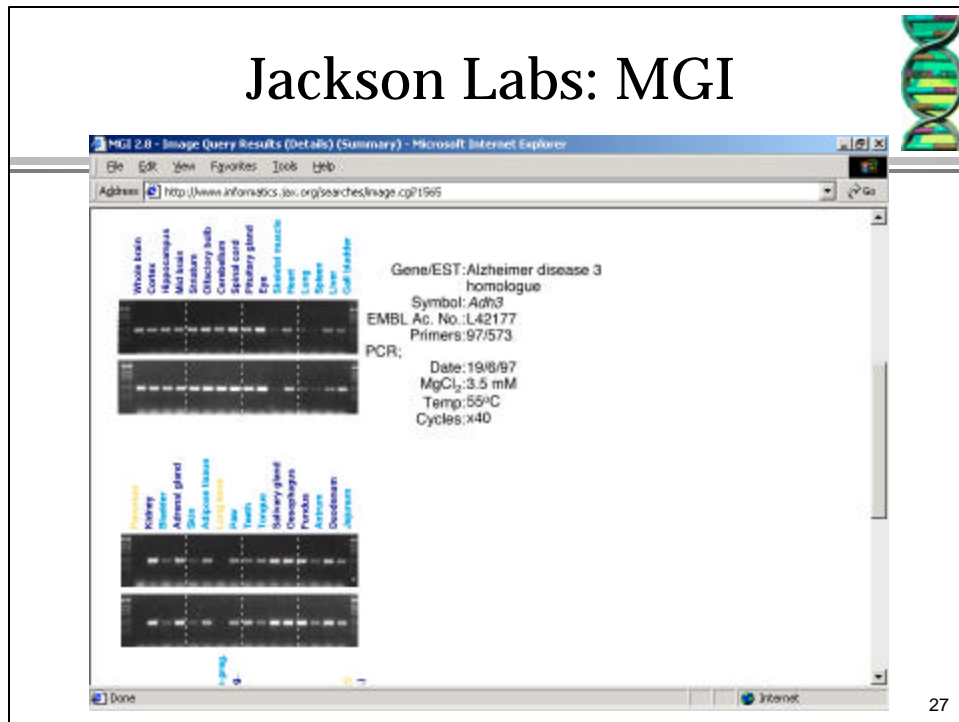
- [Mammalian Homology](#)
- [Marker Mapping Data](#) (3)
- [Phenotype Alleles](#) (8)
- [Phenotype Classifications](#) (19) [new](#)
- [Gene Expression Data](#) (52 results in 3 assays)
- [Gene Expression Literature Index Data](#) (4)
- [Antibodies](#) (2)
- [Molecular Probes and Segments](#) (31)
- [References](#) (78)
- [Mouse Locus Catalog](#)

Gene Classifications: (You can browse the Gene Ontology (GO) Classifications)

<http://www.informatics.jax.org/searches/linkap.cgi?chromosome=12&midpoint=37.0&coverage=1.0&segments=100>

26

Jackson Labs: MGI



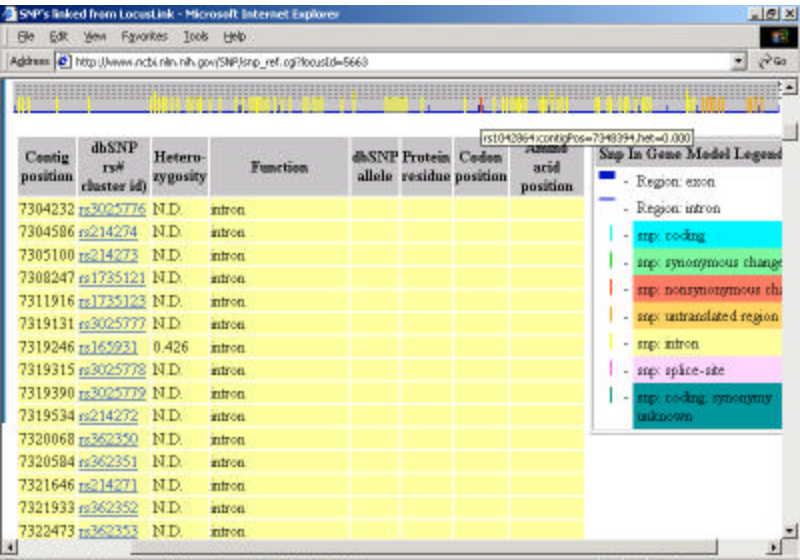
27

Variation

- Today: SNPs
 - Functional & Non-Functional
- Yesterday: RFLPs & SSLPs

28

dbSNP



SNPs linked from LocusLink - Microsoft Internet Explorer

Address: http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?locusId=5660

Contig position	dbSNP rs# cluster id	Heterozygosity	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
7304232	rs3025776	N.D.	intron				
7304586	rs214274	N.D.	intron				
7305100	rs214273	N.D.	intron				
7308247	rs1735121	N.D.	intron				
7311916	rs1735123	N.D.	intron				
7319131	rs3025777	N.D.	intron				
7319246	rs162921	0.426	intron				
7319315	rs3025778	N.D.	intron				
7319390	rs3025779	N.D.	intron				
7319534	rs214272	N.D.	intron				
7320068	rs362350	N.D.	intron				
7320584	rs362351	N.D.	intron				
7321646	rs214271	N.D.	intron				
7321933	rs362352	N.D.	intron				
7322473	rs362353	N.D.	intron				

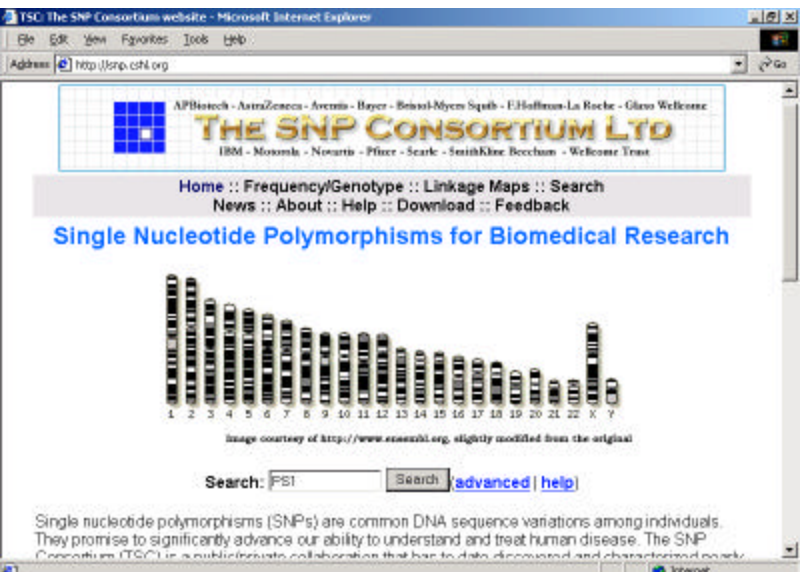
rs1042864:contigPos=7348394,het=0.000

Map In Gene Model Legend

- Region: exon
- Region: intron
- map: coding
- map: synonymous change
- map: nonsynonymous change
- map: untranslated region
- map: intron
- map: splice-site
- map: coding, synonymous unknown

29

SNP Consortium



TSC: The SNP Consortium website - Microsoft Internet Explorer

Address: <http://snp.cshl.org>

APBioscience - AstraZeneca - Aventis - Bayer - Bristol-Myers Squibb - Eli Lilly - Glaxo Wellcome
THE SNP CONSORTIUM LTD
 IBM - Merck - Novartis - Pfizer - Searle - SmithKline Beecham - Wellcome Trust

Home :: Frequency/Genotype :: Linkage Maps :: Search
 News :: About :: Help :: Download :: Feedback

Single Nucleotide Polymorphisms for Biomedical Research

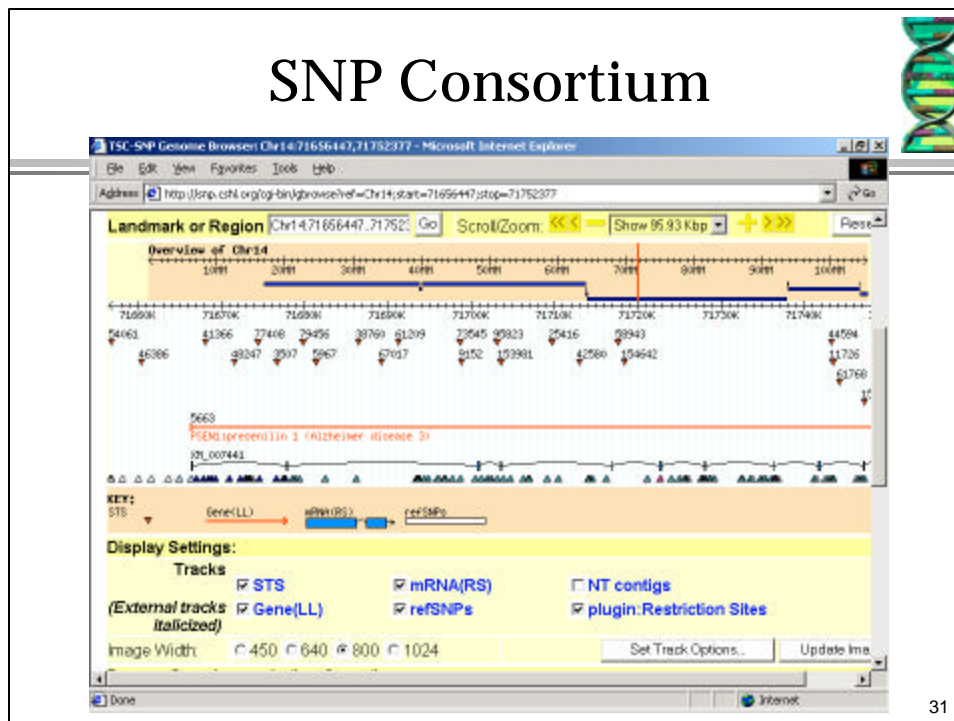
Image courtesy of <http://www.genexl.org>, slightly modified from the original

Search: Search [advanced](#) [help](#)

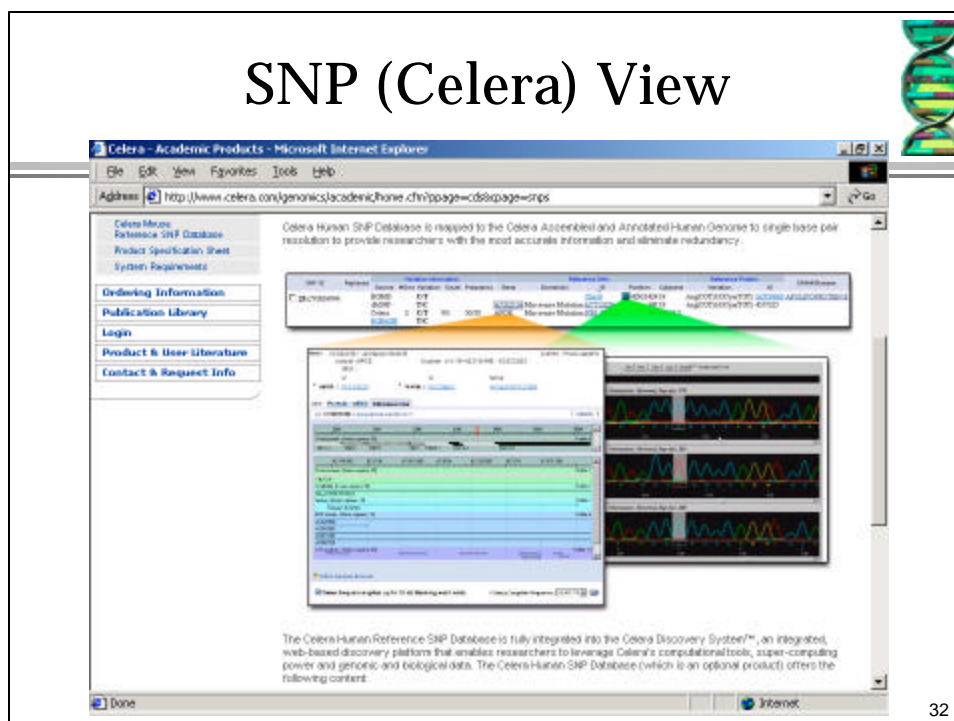
Single nucleotide polymorphisms (SNPs) are common DNA sequence variations among individuals. They promise to significantly advance our ability to understand and treat human disease. The SNP Consortium (TSC) is a public-private collaboration that has to date discovered and characterized nearly

30

SNP Consortium



SNP (Celera) View



Published Literature



- Searching the Medline Abstracts
 - The MESH Vocabulary
- Now: Full Text & Web supplements
- Future: MIAME & other standards
- 3D Protein Structures

33

MedLine via PubMed



34

Systems Biology - Pathways & Interactions



- Use our growing knowledge of:
 - Pathways
 - Gene & Protein Interactions

35

Location: <http://www.genome.ad.jp/kegg/kegg2.html>

KEGG - Table of Contents

[PATHWAY](#) | [GENES](#) | [GENOME](#) | [EXPRESSION](#) | [Form/Draw](#) | [Tools](#) | [SITE](#) | [LIGAND](#)

I. Pathway Information

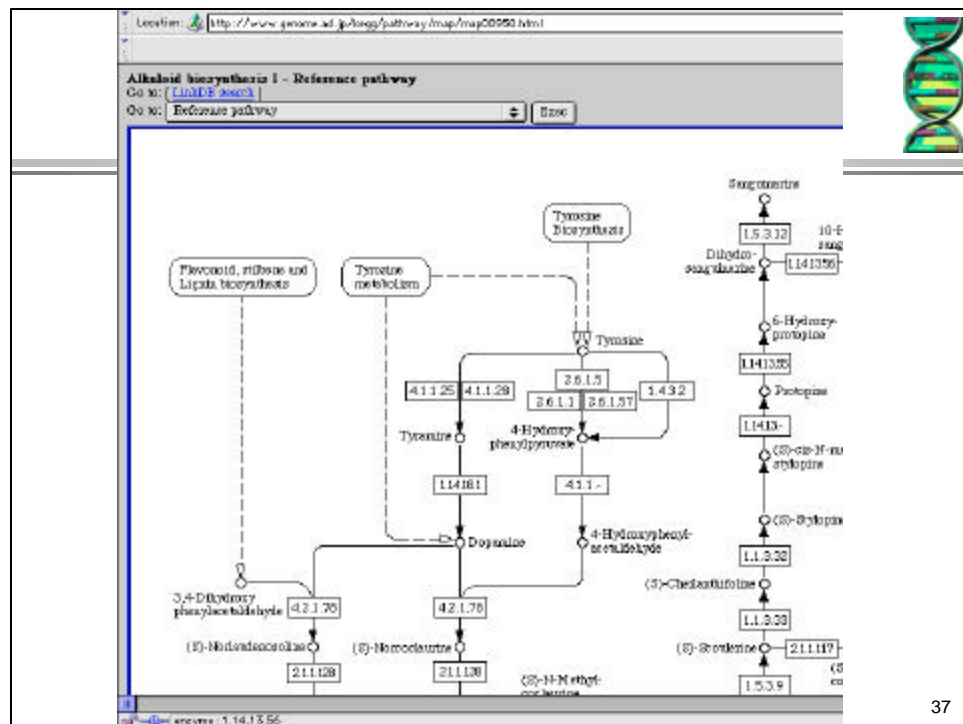
I-1. Pathway Maps and Ortholog Tables


Category	Pathway Map Ortholog Table	Search & Compare	DBGET Search
Pathway	Metabolic pathways	Search objects in pathway maps	PATHWAY
	Regulatory pathways	Color objects in pathway maps	
	Search objects in ortholog tables		
	Search similar and related in pathway maps		
	Search similar orthologs in ortholog tables	Generate possible reaction pathways	

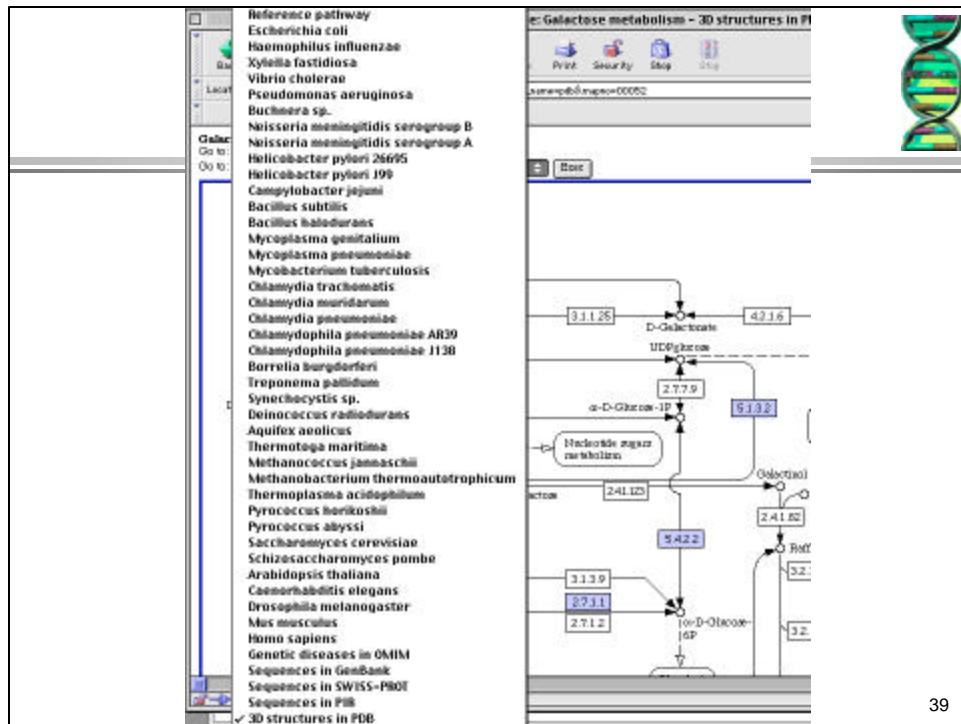
I-2. Disease Catalogs, Cell Catalogs, and Molecule Catalogs

Category	Catalog	DBGET Search
Disease	ICD9 disease classification	OMIM read map OMIM manual map
Organism	Complete genomes in KEGG	Genome projects
Cell	Complete viral genomes	Genome
Gene product	Cell lineage	Gene Ontology GO
Enzyme	Enzyme identifier classification	EC
Enzyme	EC number classification	EC number EC 3D file
Compound	EC number classification	EC 3D file PROSITE motifs
Element	Compound classification	LIGAND
	Periodic table	

36



Location:  http://www.genome.ad.jp/kegg/ortholog/tab00052.html				
Galactose Metabolism				
[P...Pathway map G...Genome map T...Title list]				
Organism	E2.7.1.6	E2.7.7.10	E5.1.3.2	E5.1.3.3
	Galactokinase	Galactose-1P uridylyl- transferase	UDP-glucose 4-epimerase	Aldose 1-epimerase
eco [P G T]	b0757(galK)	b0758(galT)	b0759(galE)	b0756(galM)
hin [P G T]	HI0819	HI0820	HI0351	HI0818
vch [P G T]	VC1595	VC1596	VCA0774	VC1594
pae [P G T]			PA1384 PA4068	
nme [P G T]			NMB0064	NMB0389
nma [P G T]			NMA0190 NMA0203	NMA2099
hpy [P G T]			HP0360	
hpi [P G T]			hpi1020	
cje [P G T]			Cji1131c	
bss [P G T]	galK	galT	galE	
bha [P G T]	BH1107	BH1109	BH1108 BH2891 BH3379 BH3649	BH2755
mge [P G T]			MG118	
mpn [P G T]			A65_orf338	
mtu [P G T]	Rv0620	Rv0618 Rv0619	Rv0501 Rv0536 Rv3468c Rv3634c	



Bioinformatics Future

- Application of our discovered knowledge to health care
- Pharmacogenomics
- Delivering Tools to the Clinician

GeneClinics: Knowledge Base



- Expert-Authored, Up-to-Date
 - Rapid Information Growth
 - Genetics Health Paradigm
- Meet Need of Health Professionals
- OO-DBMS via XML & WWW
- Develop Electronic Peer-Review
- Evaluate Utility & Methods
- Now: GeneSeek -> Integration

41

GeneClinics: Knowledge Base



GeneTests-GeneClinics Home Page - Microsoft Internet Explorer

Address: http://www.genetests.com/server/access

Funded by NIH, HRSA, and DOE

GeneTests • GeneClinics

YOU ARE LOGGED ON.
PLEASE MAKE A SELECTION.

The GeneTests-GeneClinics Web site features:

- [GeneReviews](#)
- The [Laboratory Directory](#)
- The [Clinic Directory](#)
- Expanded [Educational Materials](#)

What's New

New Features

- ▶ [Illustrated Glossary](#)
- ▶ [New Laboratory Directory Search Parameters](#)
- ▶ [Clinic Directory Listings: Updating Online](#)
- ▶ [Use of Aggregate Information: Policy Change](#)

New GeneReviews

New Lab Listings

- ▶ [7 new listings](#)

NEW

Illustrated Glossary accessed from GeneReviews

- Over 225 terms defined
- Over 40 terms illustrated
- New illustrations added regularly

42

GeneReviews: Alzheimer Disease Overview - Microsoft Internet Explorer

Address: <http://ndet.accessiondb.geneReviews.org/ndet/203910key=9ac70CLOF5M2bgr=0for=y0fnc=1MVD3f1ensm=/profiles/alzheimer/index.html>

Home Page | About This Site | **GeneReviews** | Laboratory Directory | Clinic Directory | Educational Materials

[Printable Copy]

Alzheimer Overview

- Summary
- Definition
- Prevalence
- Categories
- Diagnosis
- Genetic Counseling
- Resources
- References
- Review History
- Top of Page

Included Review

- Early-Onset Familial Alzheimer Disease

Go to Search

Disease characteristics. Alzheimer disease (AD) is characterized by adult-onset slowly progressive dementia associated with diffuse cerebral atrophy on neuroimaging studies. It is the most common form of dementia, but less than 5% of families with AD have early-onset **familial AD** (EOFAD), in which symptoms consistently occur before the age of 65 years.

Diagnosis/testing. The diagnosis of Alzheimer disease is based on the histological findings of β -amyloid plaques and intraneuronal neurofibrillary tangles. No accurate clinical diagnostic test for AD exists. A significant association with the $\epsilon 4$ **allele** of apolipoprotein E supports the diagnosis of AD in patients with dementia and increases the risk that asymptomatic individuals will eventually develop AD. ApoE **genotyping**, however, is neither fully specific nor sensitive. Three forms of EOFAD caused by **mutations** in one of three different **genes** (APP, PSEN1, PSEN2) are recognized. A molecular genetic test of the PSEN1 **gene** (chromosomal locus 14q24) is available as clinical laboratory

Click on defined terms (in purple); definition displays here.

<http://www.geneReviews.org/ndet/203910key=9ac70CLOF5M2bgr=0for=y0fnc=1MVD3f1ensm=/profiles/alzheimer/index.html>

43

Questions:



- Are these the best possible interfaces to the data?
- How can Bioinformatics Support Medical Informatics?
- Web links for sites in the talk available at:

<http://medir.ohsu.edu/~bioinf/acdmurls.htm>