SNP Resources: Finding SNPs Databases and Data Extraction

Mark J. Rieder, PhD
Robert J. Livingston, PhD

NIEHS Variation Workshop
January 30-31, 2005

Genotype - Phenotype Studies

Typical Approach:

I have candidate gene/region and samples ready to study.
Tell me what SNPs to genotype.

Other questions:

How do I know I have *all* the SNPs?
What is the validation/quality of the SNPs that are known?
Are these SNPs informative in my population/sample?

Whatdo I need to know for selecting the best SNPs?
How do I pick the bestSNPs?

What information do I need to characterize a SNP for genotyping?

Minimal SNP information forgenotyping/characterization

- What is the SNP? Flanking sequence and alleles.
  - ✓ FASTA format

  >snp_name
  ACCGAGTAGCCAG
  [A/G]
  ACTGGGATAGAAC

- dbSNP reference SNP # (rs #)

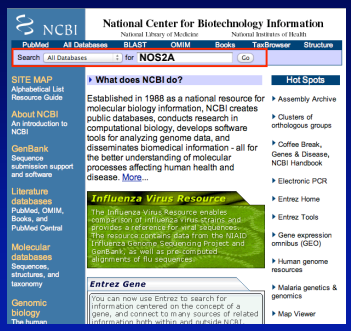
- Where is the SNP mapped? Exon, promoter, UTR, etc
  - ✓ picture of gene with mapped to the gene structure.
- How was it discovered? Method
- What assurances do you have that it is real? Validated how?
- What population African, European, etc?
- What is the allele frequency of each SNP? Common (>10%), rare
- Are other SNPs associated redundant? Genotyping data!

Finding SNPs: Databases and Extraction

How do I find and download SNP data for analysis/genotyping?

1. **Entrez Gene**
   - **-dbSNP**
   - **Entrez SNP**
2. **HapMap Genome Browser**
3. **NIEHS Environmental Genome Project (EGP)**
   Candidate gene website
4. **NIEHS web applications and other tools**
   GeneSNPs, PolyDoms, TraFac, PolyPhen, ECR Browser, GVS

NCBI - Database Resource



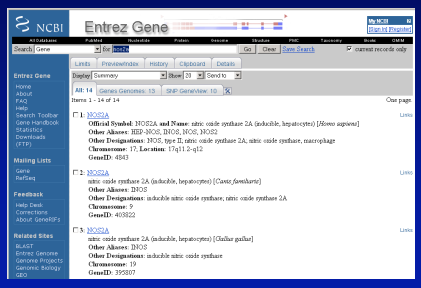
www.ncbi.nlm.nih.gov

Finding SNPs: Where do I start?

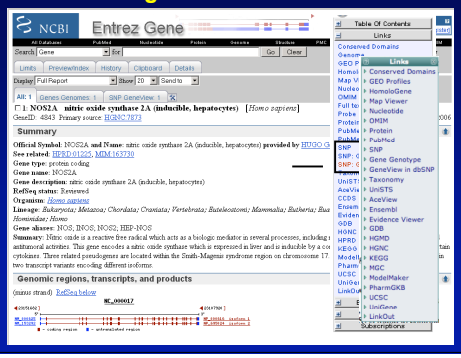
http://www.ncbi.nlm.nih.gov/gquery

Finding SNPs: Where do I start?

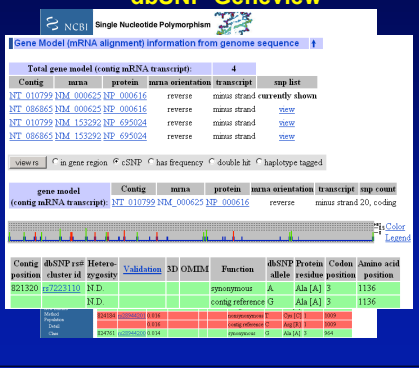
NCBI - Entrez Gene



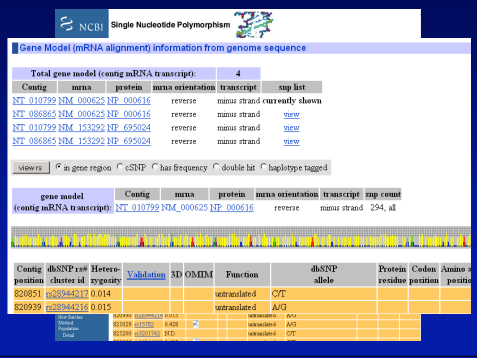
Finding SNPs: Entrez Gene



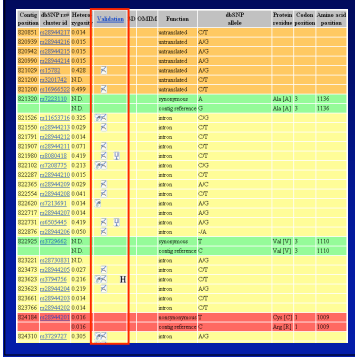
dbSNP Geneview



dbSNP Geneview



Finding SNPs: dbSNP validation

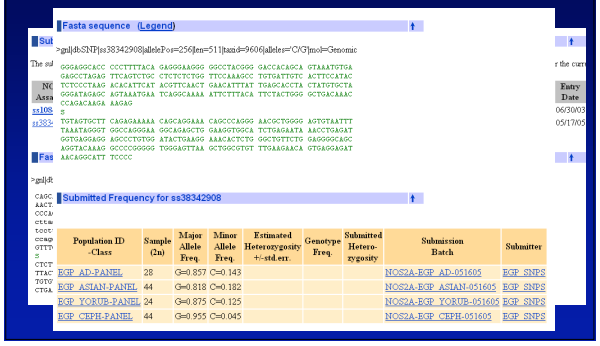


Validation status description

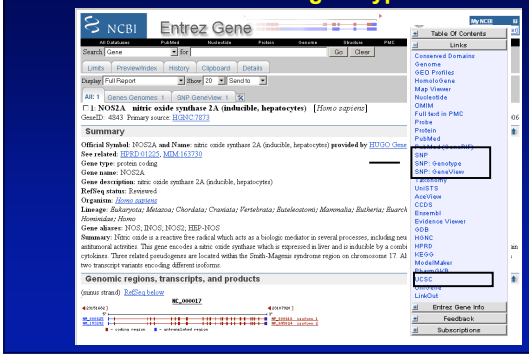
- validated by multiple, independent submissions to the refSNP cluster
- validated by frequency or genotype data: minor alleles observed in at least two chromosomes.
- validated by submitter confirmation
- all alleles have been observed in at least two chromosomes apiece (by Zhih-Zaillele)

H HapMap Verified

Finding SNPs: dbSNP database

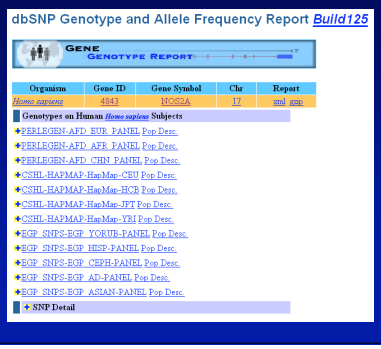


Entrez SNP - dbSNP genotype retrieval

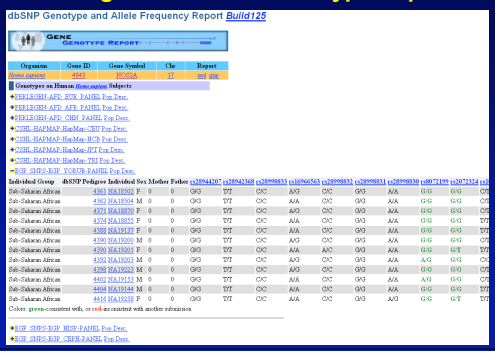


Finding SNPs - Gene Genotype Report

Finding SNPs - Gene Genotype Report



Finding SNPs - Gene Genotype Report



Minimal SNP information for genotyping/characterization

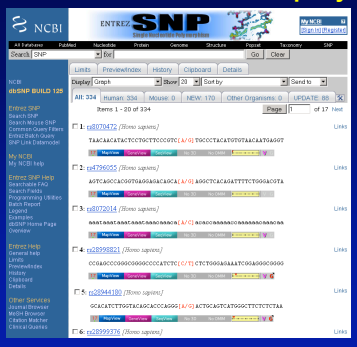
- What is the SNP? Flanking sequence and alleles.
  - ✓ FASTA format

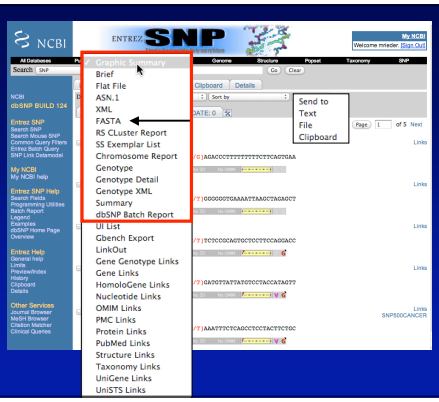
  >snp_name
  ACCGAGTAGCCAG
  [A/G]
  ACTGGGATAGAAC

  dbSNP - data is there
- dbSNP reference SNP # (rs #)
- Where is the SNP mapped? Exon, promoter, UTR, etc
  - ✓ picture of gene with mapped to the gene structure.
- How was it discovered? Method
- What assurances do you have that it is real? Validated how?
- What population African, European, etc?
- What is the allele frequency of each SNP? Common (>10%), rare
- Are other SNPs associated redundant? Genotyping data!

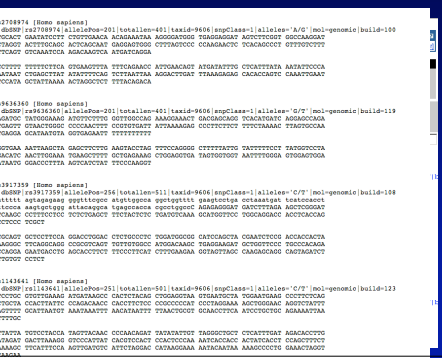
Entrez Gene Entry - Entrez SNP

Entrez SNP - direct dbSNP querying



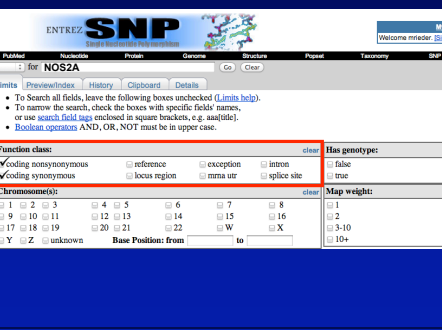


Entrez SNP - Parseable Multi-SNP Reports

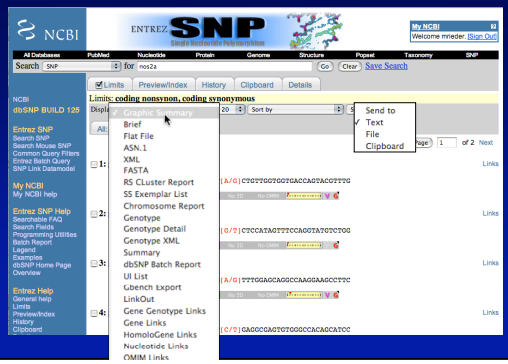


Entrez SNP - Parseable Multi-SNP reports

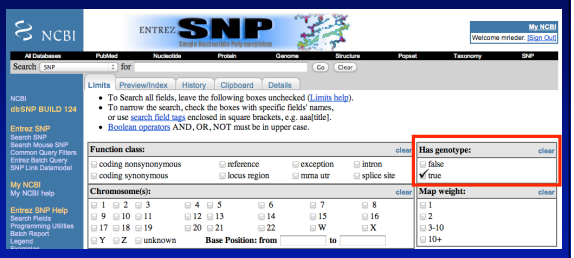
Entrez SNP - Search Limiting Capabilities

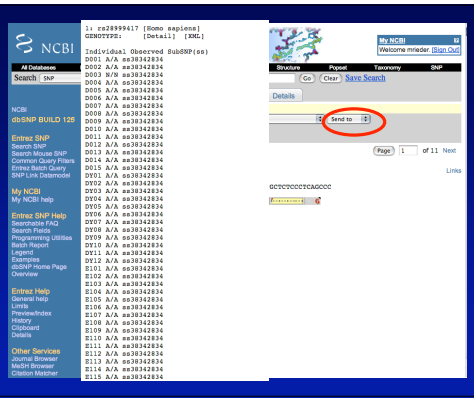


Entrez SNP - Search Limits

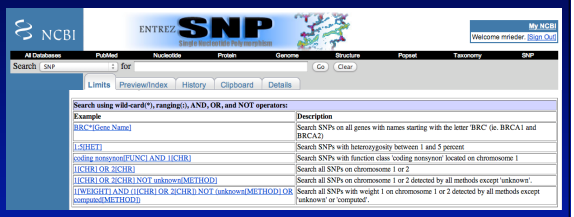


Entrez SNP - Search Limiting Capabilities





Entrez SNP - Query Term Capabilities



Entrez SNP - Search Terms Fields

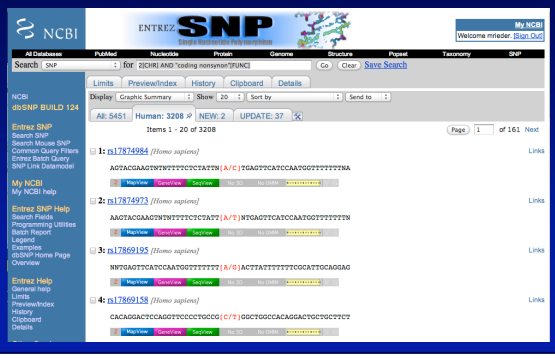
Search Field	Qualifier	Type	Description
Allele	[ALLELE],[VARIATION], [VARI]	IUPAC	[Observed allele(s) Example: N[ALLELE]
Chromosome	[CHR]	Textnum	Mapped chromosome number Available values [1-22,W,Z, and Un (unknown)] Example: 2[CHR] or X[CHR]
Base Position	[CHRPOS],[BPOS]	Integer*	Mapped chromosome position; use in conjunction with chromosome field [CHR] Example: 7[CHR] AND 88556398:88580839[CHRPOS]
Create Build ID	[CREATE_BUILD],[CBID]	Integer*	SNP create build ID Example: 103[CBID]
Publication Date	[CREATEDATE],[CDAT],[PDAT], [PUBDATE]	Date*	SNP create/publication date Use the format YYYY/MM/DD; month and day are optional. Example: 2002/01[CDATE]
Function Class	[FXN_CLASS], [FUNC]	Text	Function Class: locus region, intron coding nonsynon, mrna utr coding synon, reference exception, splice site Example: "coding synon"[FUNC]
Gene Name	[GENE], [GENE_SYMBOL]	Textnum	Locus link symbol Example: LPL[GENE]
Protein Description	[PDSC]	Text	Protein description Example: "liver expressed"[PDSC]
Gene Description	[GDSC]	Text	Gene description Example: "kinase"[Gene Description]
GO Terms	[GO]	Text	Gene Ontology (GO) terms Example: "receptor"[GO term]
Genotype	[GENOTYPE], [GTYPE]	Boolean	Genotype Example: true[GTYPE]

Entrez SNP - Search Terms Fields

Submitter Handle	[HANDLE]	Text	Submitter handle Example: "LEE"[HANDLE]
Success Rate	[SUCCESS_RATE], [SRATE]	Integer*	Success rate(%) Example: 95[SRATE]
Organism	[ORGN],[TAX_ID]	Text	Organism name or Taxonomy ID number; use the prefix 'txid' before the Taxonomy ID number Example: "human"[ORGN] , "mouse"[ORGN], or "txid9606"[ORGN]
Update Build ID	[UPD_BUILD],[UBID]	Integer*	SNP update build ID Example: 102[UBID]
Modification Date	[UDAT], [UDAT], [MODDATE]	Date*	SNP modification/update date Use the format YYYY/MM/DD; month and day are optional. Example: 2002/01 [UDATE]
Validation	[VALIDATION]	Text	Validation status: by cluster, by submitter by frequency, no info Example: "by cluster"[VALIDATION]

More advanced queries:

2[CHR] AND "coding nonsynon"[FUNC]



Entrez SNP - Search Terms Fields

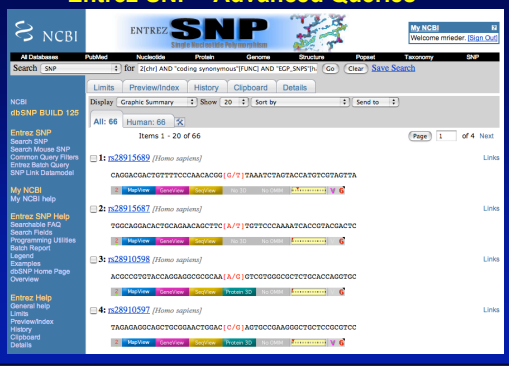
Submitter Handle	[HANDLE]	Text	Submitter handle Example: "LEE"[HANDLE]
Success Rate	[SUCCESS_RATE], [SRATE]	Integer*	Success rate(%) Example: 95[SRATE]
Organism	[ORGN],[TAX_ID]	Text	Organism name or Taxonomy ID number; use the prefix 'txid' before the Taxonomy ID number Example: "human"[ORGN] , "mouse"[ORGN], or "txid9606"[ORGN]
Update Build ID	[UPD_BUILD],[UBID]	Integer*	SNP update build ID Example: 102[UBID]
Modification Date	[UDATE], [UDAT], [MODDATE]	Date*	SNP modification/update date Use the format YYYY/MM/DD; month and day are optional. Example: 2002/01 [UDATE]
Validation	[VALIDATION]	Text	Validation status: by cluster by submitter by frequency no info Example: "by cluster"[VALIDATION]

More advanced queries:

2[CHR] AND "coding nonsynon"[FUNC] AND "EGP_SNPS"[HANDLE]

Note: Can also use wildcard (*) characters, AND, OR, and NOT operators

Entrez SNP - Advanced Queries



Minimal SNP information for genotyping/characterization

- What is the SNP? Flanking sequence and alleles.
  - ✓ FASTA format

  >snp_name
  ACCGAGTAGCCAG
  [A/G]
  ACTGGGATAGAAC

EntrezSNP - better!

- dbSNP reference SNP # (rs #)
- Where is the SNP mapped? Exon, promoter, UTR, etc
  - ✓ picture of gene with mapped to the gene structure.
- How was it discovered? Method
- What assurances do you have that it is real? Validated how?
- What population African, European, etc?
- What is the allele frequency of each SNP? Common (>10%), rare
- Are other SNPs associated redundant? Genotyping data!

Finding SNPs - Entrez SNP Summary

1. dbSNP is useful for investigating detailed information on a small number SNPs and its good for a picture of the gene
2. Entrez SNP is a direct, fast, database for querying SNP data.
3. Data from Entrez SNP can be retrieved in batches for many SNPs
4. Entrez SNP data can be limited to specific subsets of SNPs and formatted in plain text for easy parsing and manipulation
5. More detailed queries can be formed using specific field tags for retrieving SNP data

Finding SNPs: Databases and Extraction

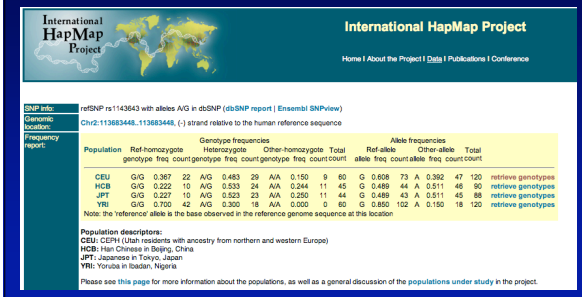
How do I find and download SNP data for analysis/genotyping?

1. Entrez Gene
   - dbSNP
   - Entrez SNP
2. HapMap Genome Browser
3. NIEHS Environmental Genome Project (EGP)
   Candidate gene website
4. NIEHS web applications and other tools
   GeneSNPS, PolyDoms, TraFac, PolyPhen, ECR Browser, GVS

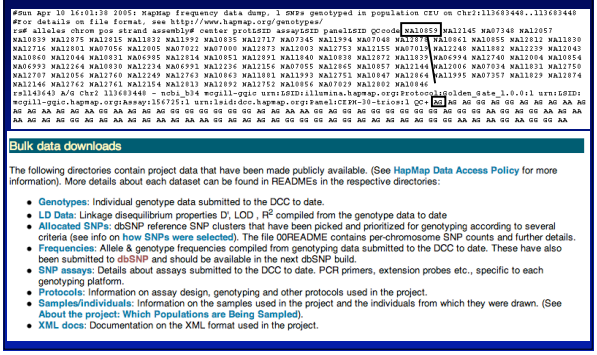
www.hapmap.org

Finding SNPs: HapMap Browser

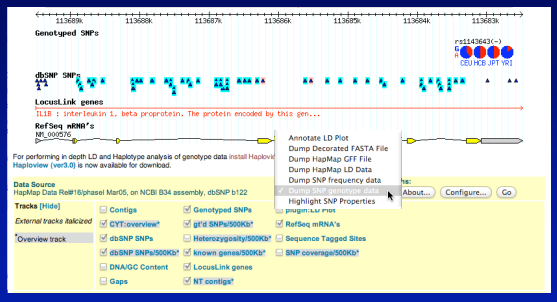
Finding SNPs: HapMap Browser



Finding SNPs: HapMap Genotypes



Finding SNPs: HapMap Browser



Minimal SNP information for genotyping/characterization

- What is the SNP? Flanking sequence and alleles.
  - ✓ FASTA format

    >snp_name
    ACCGAGTAGCCAG
    [A/G]
    ACTGGGATAGAAC
- dbSNP reference SNP # (rs #)
- Where is the SNP mapped? Exon, promoter, UTR, etc
  - ✓ picture of gene with mapped to the gene structure.
- How was it discovered? Method
- What assurances do you have that it is real? Validated how?
- What population African, European, etc?
- What is the allele frequency of each SNP? Common (>10%), rare
- Are other SNPs associated redundant? Genotyping data!

Finding SNPs: HapMap Browser

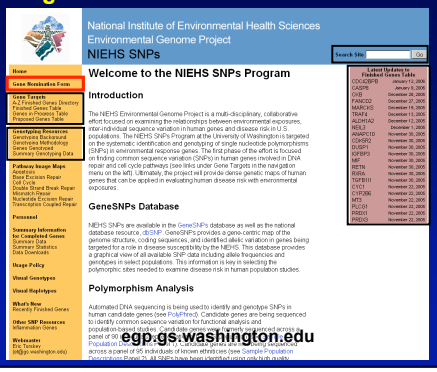
1. HapMap datasets are useful because individual genotype data can be used to determine optimal genotyping strategies (tagSNPs) or perform population genetic analyses (linkage disequilbrium)
2. Data are specific produced by those projects (not all dbSNP)
   - ✓ HapMap data is available in dbSNP
3. HapMap data (Phase II) can be accessed preleased prior to dbSNPs
4. Easier visualization of data and direct access to SNP data, individual genotypes, and LD analysis

Finding SNPs: Databases and Extraction

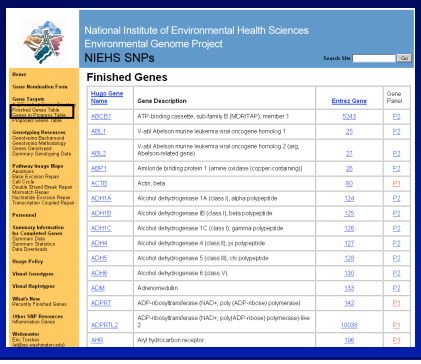
How do I find and download SNP data for analysis/genotyping?

1. Entrez Gene
   - dbSNP
   - Entrez SNP
2. HapMap Genome Browser
3. NIEHS Environmental Genome Project (EGP) Candidate gene website
4. NIEHS web applications and other tools GeneSNPS, PolyDoms, TraFac, PolyPhen, ECR Browser, GVS

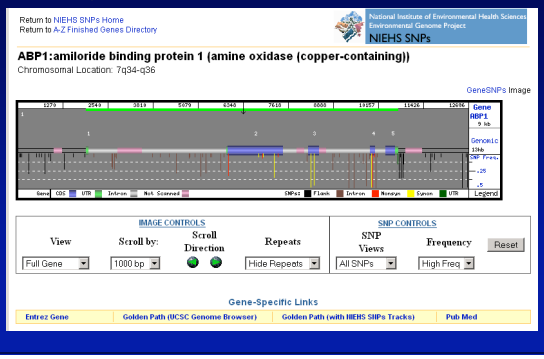
Finding SNPs: NIEHS SNPs Candidate Genes



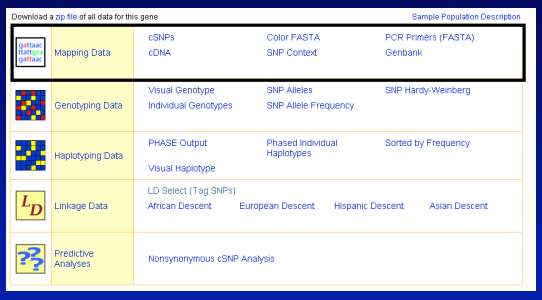
Finding SNPs: NIEHS SNPs Candidate Genes

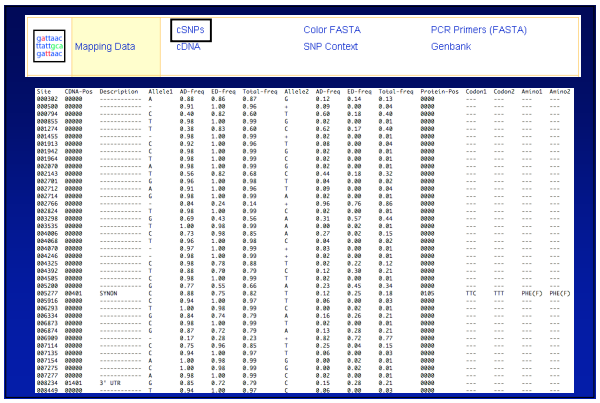


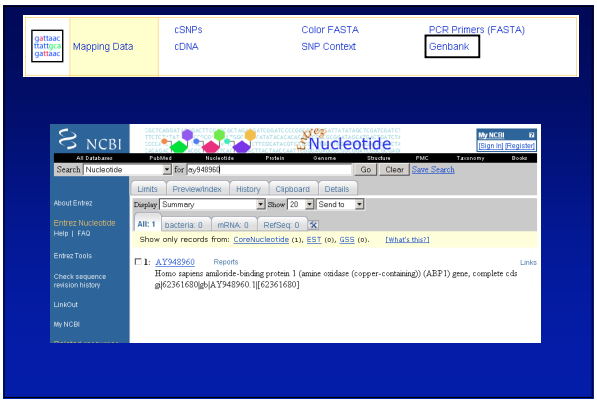
Finding SNPs: NIEHS SNPs Candidate Genes

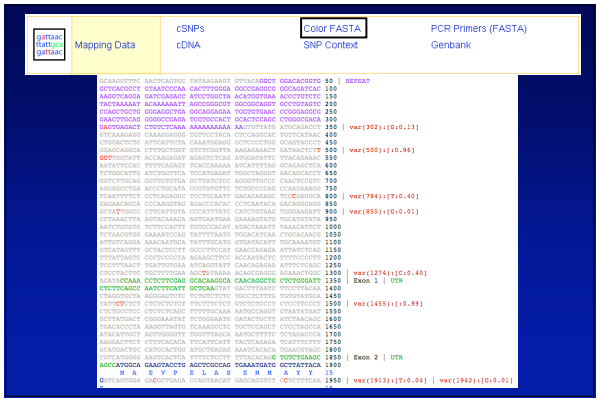


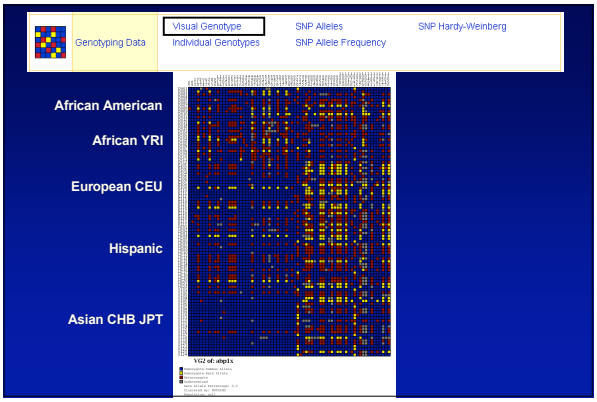
Finding SNPs: NIEHS SNPs Candidate Genes

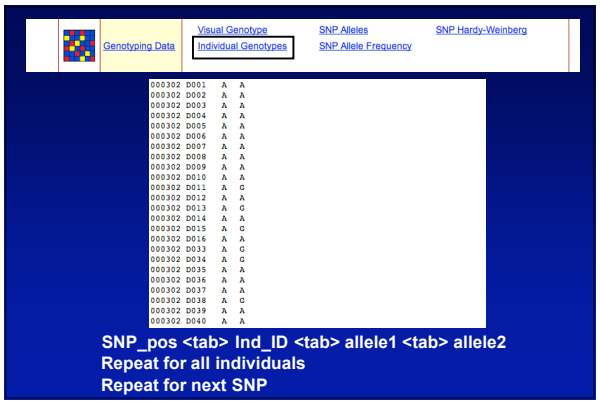


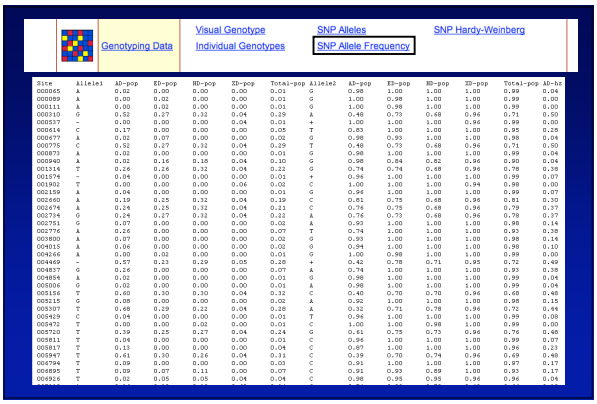

Mapping Data
cSNPs
cDNA
Color FASTA
SNP Context
PCR Primers (FASTA)
Genbank


Mapping Data
cSNPs
cDNA
Color FASTA
SNP Context
PCR Primers (FASTA)
Genbank
NCBI
Nucleotide
Search Nucleotide
Limits
Preview/Index
History
Clipboard
Details
Display Summary
Show 20
Send to
Show only records from: CoreNucleotide (1), EST (0), GSS (0).
1: AY948960
Homo sapiens amiloride-binding protein 1 (amine oxidase (copper-containing)) (ABP1) gene, complete cds


Mapping Data
cSNPs
cDNA
Color FASTA
SNP Context
PCR Primers (FASTA)
Genbank


Genotyping Data
Visual Genotype
Individual Genotypes
SNP Alleles
SNP Allele Frequency
SNP Hardy-Weinberg
African American
African YRI
European CEU
Hispanic
Asian CHB JPT


Genotyping Data
Visual Genotype
Individual Genotypes
SNP Alleles
SNP Allele Frequency
SNP Hardy-Weinberg
SNP_pos <tab> Ind_ID <tab> allele1 <tab> allele2
Repeat for all individuals
Repeat for next SNP


Genotyping Data
Visual Genotype
Individual Genotypes
SNP Alleles
SNP Allele Frequency
SNP Hardy-Weinberg

Haplotyping Data | PHASE Output | Phased Individual Haplotypes | Sorted by Frequency | Visual Haplotype

```
BEGIN LIST_SUMMARY
     1 ADTCCCAIGACTGCCAGDCCCTTCTCGGGACCC    1.000000
     2 ADTCCCAIGACTGCCGGITCCTTTGCGGGCCGC    3.000000
     3 ADTCCCAIGACTGCCGGITCCTTTGTCGGGCCGC    2.000000
     4 ADTCCCAIGACTGTCAADCCCTCCTCGGGACCC    4.000000
     5 ADTCCCAIGCCTGCCGGITCCTTTGCGGGCCGC    1.000000
     6 ADTCCCAIGCCCCGCCGGITCCTTTGCGGGCCGC    1.000000
     7 ADTCCTAIGCCCCGCCGGICCGTCCCGGGCCGC    3.000000
     8 ADTCCTAIGCCCCGCCGGICCGTCCCGGGCCGC    1.000000
     9 ADTCCTTIGCCTGCCGGITCCTCCTCGAGCCCGC    3.000000
    10 ADTTCCAIGACTGCCGGICCGCCCCGGGACCCT    1.000000
    11 ADTTCTAIGCCTGCCGGICCGTTCCGGGGCCGC    1.000000
    12 ADTTCTAIAACCTGCCGGICCGTCCCGGGGCCCC    1.000000
    13 ADCCCTAIGCCTACCGGICCGTCCCGCGGGCCCT    2.000000
    14 ADCCCTTIGCCCGCCGGICCGTCCCGCGGGCCGC    1.000000
    15 ADCTCTAIGCCTGCCGGICCGTCCCGCGGGCCGC    1.000000
    16 ADCTCTAIACCTGCCGGICCGTCCCGCGGGCCCC    2.000000
    17 ADCTCTAIACCTGCCGGICCGTCCCGGGACCCT    1.000000
    18 ADCTCTAIACCTGCCGGICCGCCCCGCGGGCCCT    1.000000
    19 ADCTCTAIACCTGCCGGITCCTTTGCCGGGCCGC    1.000000
    20 ADCTCTAIACCTGTCAADCCCTCCTCGGGACCCC    2.000000
    21 ADCTCTAIACCTGTCGGICCGTCCCGCAGGCCCC    1.000000
    22 ADCTCTAIACCTACCGGICCGTCCCGCAGGCCCC   24.000000
    23 ADCTCTAIACCCCGCCGGICCGTCCCGCGGGCCCT    1.000000
    24 ADCTCTAIACCCCGCCGGICCGCCCCGCGGACCCT    1.000000
    25 ADCTCTAIACCCCGCCGGICCGCCCCGCAGGCCCC    2.000000
    26 ADCTCTAIGCCCCTGCAADCCCTCCTCGGGACCC    1.000000
    27 ADCTCTAIGCCTGCCCAADCCCTCCTCGGGACCC    1.000000
    28 ADCTCTAIGCCTGCCCAADCCCTCCTCGGGACCC   10.000000
    29 ADCTTTAIACCTGCCGGICCGTTCCGCGGGCCGC    1.000000
    30 ADCTTTAIACCTACCGGICCGTCCCGCGGGCCCC    1.000000
    31 ADCTTTAIACCTACCGGICCGTCCCGCAGGCCCC    2.000000
    32 AITCCCAIGACTGCCGGITTCTTTGTCGGGCCGC    3.000000
    33 GDTCCCAIGCCTGCCGGICCGTCCCGCGGGCCCC    9.000000
    34 GDTCCCAIGCCTGCCGGICCGTTCCGCGGGCCGC    2.000000
    35 GDTCCCAIGCCCCTCAADCCCTCCTCGGGACCC    1.000000
    36 GDTCCCADGCCTGCCGGICCGTCCCGCGGGCCCC    1.000000
END LIST_SUMMARY
```

```
BEGIN BESTPAIRS_SUMMARY
0001: (1,4)
0002: (13,22)
0003: (2,4)
0004: (10,17)
0005: (11,22)
0006: (20,22)
0007: (22,36)
0008: (6,7)
0009: (5,32)
0010: (4,13)
0011: (23,34)
0012: (2,14)
0013: (7,33)
0014: (8,22)
0015: (32,36)
0016: (18,31)
0033: (7,34)
0034: (3,33)
0035: (4,28)
0036: (1,22)
0037: (9,9)
0038: (24,35)
0039: (9,32)
0040: (29,31)
0001: (22,22)
0002: (28,28)
0003: (22,33)
0004: (28,33)
0005: (28,35)
0006: (22,22)
0007: (22,28)
0008: (26,33)
0009: (28,28)
0010: (22,22)
0011: (19,22)
0012: (22,33)
0013: (18,22)
0014: (2,22)
0015: (22,25)
0016: (15,22)
0017: (28,28)
```

LD Linkage Data | LD Select (Tag SNPs) | African Descent | European Descent | Hispanic Descent | Asian Descent

```
Bin 1   total_sites: 11 average_minor_allele_frequency: 29%
Bin 1   TagSnps: 10869
Bin 1   other_snps: 8257 9334 9468 9732 9964 10007 10090 10708 12527 12666

Bin 2   total_sites: 4  average_minor_allele_frequency: 39%
Bin 2   TagSnps: 4469 5156 5307 5947
Bin 2   other_snps:

Bin 3   total_sites: 4  average_minor_allele_frequency: 23%
Bin 3   TagSnps: 1314 2660 2674 2734
Bin 3   other_snps:

Bin 4   total_sites: 3  average_minor_allele_frequency: 33%
Bin 4   TagSnps: 7229 10741 10859
Bin 4   other_snps:

Bin 5   total_sites: 3  average_minor_allele_frequency: 8%
Bin 5   TagSnps: 6794 8915 11494
Bin 5   other_snps:

Bin 6   total_sites: 2  average_minor_allele_frequency: 48%
Bin 6   TagSnps: 310 775
Bin 6   other_snps:

Bin 7   total_sites: 2  average_minor_allele_frequency: 26%
Bin 7   TagSnps: 2776 4837
Bin 7   other_snps:
```

Predictive Analyses | Nonsynonymous cSNP Analysis

Hugo	AllelePos	a.a.Pos	Residue	Variant	Freq	PPHpredict	SIFTpredict
brca1	031631	275	G	S	0.01	benign	TOLERATED
brca1	031875	356	Q	R	0.01	possibly damaging	INTOLERANT
brca1	032885	693	D	N	0.05	benign	TOLERATED
brca1	033420	871	P	L	0.43	benign	TOLERATED
brca1	033921	1038	E	G	0.32	benign	INTOLERANT
brca1	033927	1040	S	N	0.01	benign	TOLERATED
brca1	034226	1140	S	G	0.02	benign	TOLERATED
brca1	034356	1183	K	R	0.33	benign	TOLERATED
brca1	055298	1613	S	G	0.33	benign	TOLERATED
brca1	055319	1620	T	A	0.01	benign	TOLERATED

SIFT = Sorting Intolerant From Tolerant
Evolutionary comparison of non-synonymous SNPs

PolyPhen - *Poly*morphism *Phen*otyping
Structural protein characteristics and evolutionary comparison

Finding SNPs: NIEHS SNPs Candidate Genes

Download a zip file of all data for this gene | Sample Population Description

Mapping Data	cSNPs cDNA	Color FASTA SNP Context	PCR Primers (FASTA) Genbank	
Genotyping Data	Visual Genotype Individual Genotypes	SNP Alleles SNP Allele Frequency	SNP Hardy-Weinberg	
Haplotyping Data	PHASE Output Visual Haplotype	Phased Individual Haplotypes	Sorted by Frequency	
Linkage Data	LD Select (Tag SNPs) African Descent	European Descent	Hispanic Descent	Asian Descent
Predictive Analyses	Nonsynonymous cSNP Analysis			

Finding SNPs: NIEHS SNPs Candidate Genes

National Institute of Environmental Health Sciences
Environmental Genome Project
NIEHS SNPs

Search Site | Go

Home
Gene Nomination Form
Gene Targets
A-Z Finished Genes Directory
Finished Genes Table
Genes in Process Table
Proposed Genes Table
Genotyping Resources
Genotypes Background
Genotypes Methodology
Genes Deprived
Summary Genotyping Data
Pathway Image Maps
Apoptosis
Base Excision Repair
Cell Cycle
Double Strand Break Repair
Mismatch Repair
Nucleotide Excision Repair
Transcription Coupled Repair
Personnel
Summary Information for Completed Genes
Summary Data
Data Downloads
Image Policy
Visual Genotypes
Visual Haplotypes
What's New
Recently Finished Genes
Other SNP Resources
Inflammation Genes
Webmaster
Eric Tassier
(egp.gs.washington.edu)

Welcome to the NIEHS SNPs Program

Introduction

The NIEHS Environmental Genome Project is a multi-disciplinary, collaborative effort focused on examining the relationships between environmental exposures, inter-individual sequence variation in human genes and disease risk in U.S. populations. The NIEHS SNPs Program at the University of Washington is targeted on the systematic identification and genotyping of single nucleotide polymorphisms (SNPs) in environmental response genes. The first phase of the effort is focused on finding common sequence variation (SNPs) in human genes involved in DNA repair and cell cycle pathways (see links under Gene Targets in the navigation menu on the left). Ultimately, the project will provide dense genetic maps of human genes that can be applied in evaluating human disease risk with environmental exposures.

GeneSNPs Database

NIEHS SNPs are available in the GeneSNPs database as well as the national database resource, dbSNP. GeneSNPs provides a gene-centric map of the genome structure, coding sequences, and identified allelic variation in genes being proposed for a role in disease susceptibility by the NIEHS. This database provides graphical view of all available SNP data including allele frequencies and genotypes in select populations. This information is key in selecting the polymorphic sites needed to examine disease risk in human population studies.

Polymorphism Analysis

Automated DNA sequencing is being used to identify and genotype SNPs in human candidate genes (see PolyPhen). Candidate genes are being sequenced to identify common sequence variation for functional analysis and population-based studies. Candidate genes were formerly sequenced across a panel of 90 individuals representative of the U.S. population (see Sample Population Description). SNPs have been identified in all of the candidate genes across a panel of 24 individuals of known ethnicities (see Sample Population Description). Allele frequencies have been established in the majority

Latest Updates to Finished Genes Table

CDC42EP8	January 12, 2006
CASP8	January 9, 2006
CKB	December 28, 2005
FANCD2	December 27, 2005
MAPK3	December 19, 2005
TRAP1	December 13, 2005
ALDH1A2	December 12, 2005
NEIL3	December 1, 2005
ANAPC10	November 30, 2005
CDKSR2	November 30, 2005
FGF18	November 30, 2005
KYSPR	November 30, 2005
MIF	November 30, 2005
RETN	November 30, 2005
RIIRA	November 30, 2005
TOF(E11)	November 30, 2005
CYC1	November 22, 2005
CYP2B6	November 22, 2005
MT3	November 22, 2005
RUC01	November 22, 2005
PRD(1)	November 22, 2005
PRDC0	November 22, 2005

egp.gs.washington.edu

Finding SNPs: NIEHS SNPs Candidate Genes

Data Downloads

All NIEHS SNPs Variation Data

Bulk Download of All Variation Data Files

WARNING: This is a very large file and will take several minutes to download. The file is a compressed and "tarred" unix file containing the entire directory of text data files. These are the same text data files which appear in the data pages for each candidate gene in our Finished Genes Table. Please see our Usage Policy if this work is to be used in a publication.

Download of Variation Data (Single File)

Global Prettybase File

This is a tab delimited text file in our "prettybase" format that describes all SNP sites discovered by NIEHS SNPs. The format of this file is:

Line format:
<chromosome position chromosome-HUGO_NAME > <EGP Sample ID> <Allele1> <Allele2>

Example: 99504884-REV1L-2 DY07 G G

The 'chromosome position' is generated from mapping to the most recent genome assembly available from the UCSC Genome Assembly.

Download of EGP Variation Data by Chromosome

These are tab delimited text files in our "prettybase" format that describe all SNP sites discovered by NIEHS SNPs but separated into files based on chromosome. The format of this file is:

NIEHS SNPs SIFT/PolyPhen Data

Putative functional changes in a candidate gene's protein function were assessed by taking the nonsynonymous coding SNPs (cSNPs) for each gene and using both SIFT and PolyPhen. Generally, each nonsynonymous amino acid change is analyzed in the context of other evolutionarily similar proteins to determine the likelihood of the polymorphic nonsynonymous change and then statistically classified. These programs classify each coding SNP as tolerant or intolerant (SIFT), or as benign, possibly damaging, or probably damaging (PolyPhen).

Combined SIFT/PolyPhen Data for NIEHS SNPs Nonsynonymous SNPs

Combined SIFT/PolyPhen Data for NIEHS SNPs Nonsynonymous SNPs (Intolerant or Potentially Damaging)

SIFT Data for NIEHS SNPs Nonsynonymous SNPs (Potentially Intolerant)

PolyPhen Data for NIEHS SNPs Nonsynonymous Genes (Potentially Damaging)

Finding SNPs: Databases and Extraction

How do I find and download SNP data for analysis/genotyping?

1. Entrez Gene
   - dbSNP
   - Entrez SNP
2. HapMap Genome Browser
3. NIEHS Environmental Genome Project (EGP)
   Candidate gene website
4. NIEHS web applications and other tools
   GeneSNPS, PolyDoms, TraFac, PolyPhen, ECR Browser, GVS

GeneSNPs

Graphic view of SNPs in context of gene elements
All NIEHS genes presented
- organized by pathway/function

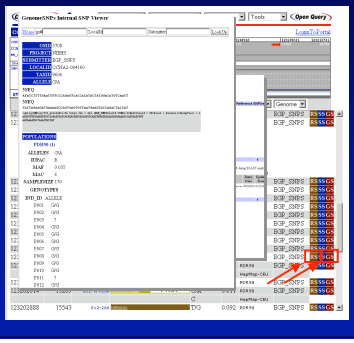
SNPs from dbSNP
- organized by submitter handle

Sequenc context of SNPs presented in Color Fasta format
Link-outs to EntrezSNP pages
Summary Genome SNPs internal SNP viewer for one-stop SNP shopping

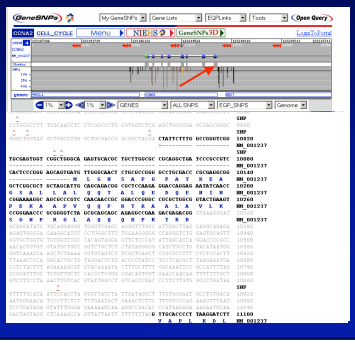
http://www.genome.utah.edu/genesnps/

GeneSNPs: One stop shopping

GeneSNPs: One stop shopping



GeneSNPs: One stop shopping



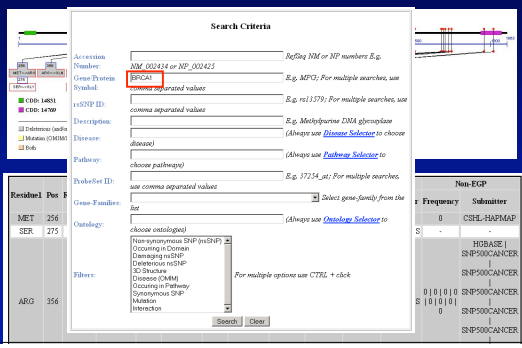
Polydoms

A web-based application that maps synonymous and non-synonymous SNPs onto known functional protein domains

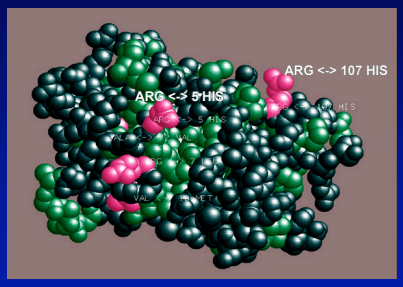
- SNPsare from dbSNP and GeneSNPs
- Domain structures from NCBI's Conserved Domain Database
- Functional predictions based on SIFT and PolyPhen
- 3 dimensional mapping of SNPs on protein structure using Chime viewer

http://polydoms.cchmc.org/polydoms

Polydoms



Polydoms



Mapping of nsSNPS onto protein structure

TraFac: Transcription Factor Binding Site Comparison

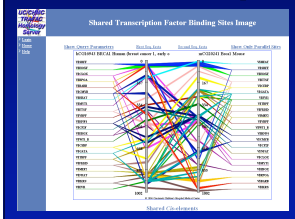
A tool for validating cis regulatory elements conserved between human and mouse

- Aligns human and mouse sequences using BLAST%
- Consensus transcription factor binding sequences from Transfac database

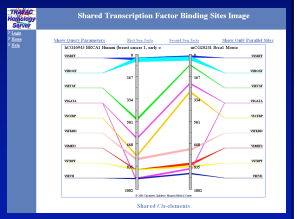
http://trafac.cchmc.org/trafac

TraFac: Transcription Factor Binding Site Comparison

All TFBS in common



TFBS in parallel



ECR Browser: Evolutionary Conserved Regions

Aligns sequences to Mouse, Rat, Dog, Opposum, Chicken, Fugu and Drosophila

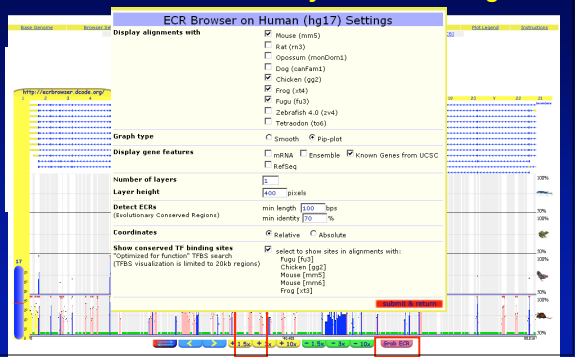
Gene annotations from UCSC Genome Browser

Easy retrieval of ECR sequences and alignments

Pre-computed transcription factor binding sites

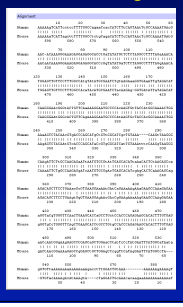
http://ecrbrowser.dcode.org

ECR Browser: Evolutionary Conserved Regions



ECR Browser: Evolutionary Conserved Regions

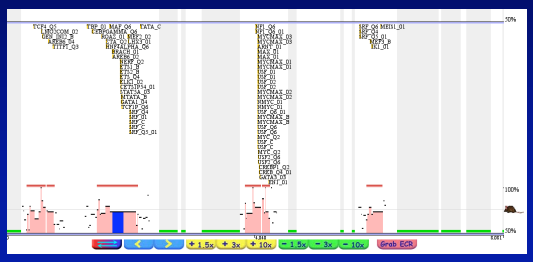
Human-mouse alignment



Fasta sequences

ECR Browser: Evolutionary Conserved Regions

Transcription Factor Binding Sites from Transfac

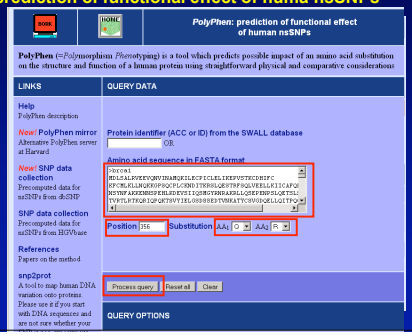


PolyPhen: Polymorphism Phenotyping-
prediction of functional effect of human nsSNPs

- Physical and comparative analyses used to make predictions
- Uses SwissProt annotations to identify known domains
- Calculates a substitution probability from BLAST alignments of homologous and orthologous sequences
- Ranks substitutions on scale of predicted functional effects from benign to probably damaging

http://tux.embl-heidelberg.de/ramensky

PolyPhen: Polymorphism Phenotyping
- prediction of functional effect of huma nsSNPs

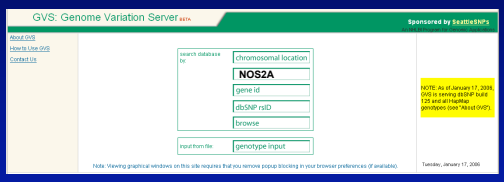


GVS: Genome Variation Server

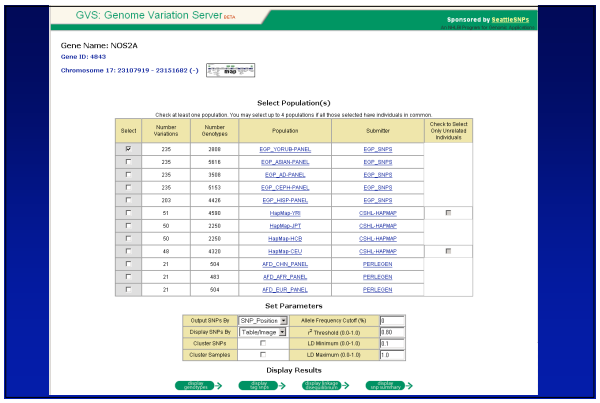
- Provides rapid analysis of 4.3 million SNPs from dbSNP and the HapMap
- Mapped to human genome build 35 (hg17) Displays genotype data in text and image formats Displays tagSNPs or clusters of informative SNPsin text and image formats Displays linkage disequilibrium (LD) in text and image formats

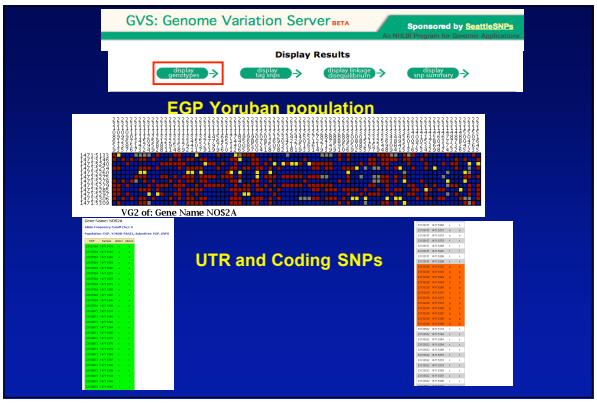
http://gvs.gs.washington.edu/GVS

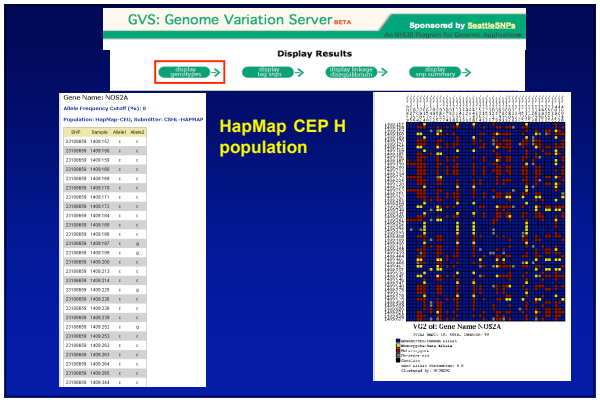
GVS: Genome Variation Server

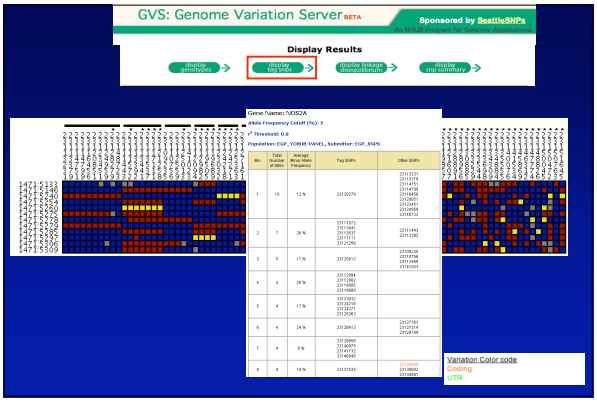


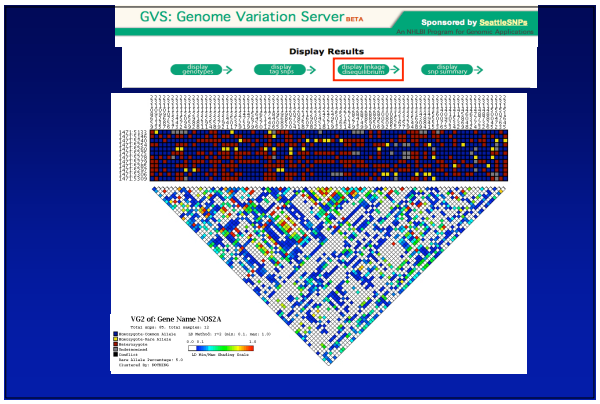
http://gvs.gs.washington.edu/GVS

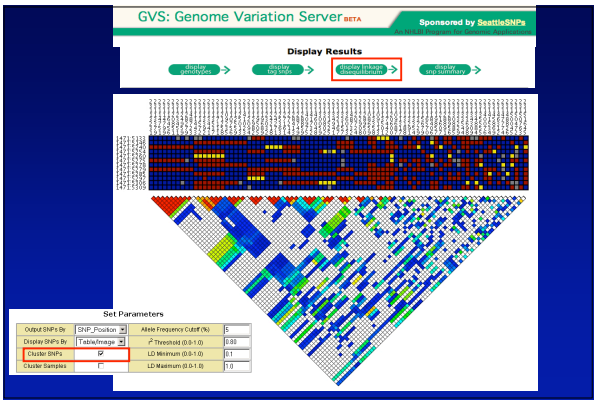

GVS: Genome Variation Server BETA
Sponsored by SeattleSNPs
Gene Name: NOS2A
Gene ID: 4843
Chromosome 17: 23107919 - 23151682 (-)
Select Population(s)
Check at least one population. You may select up to 4 populations if all those selected have individuals in common.
Set Parameters
Display Results

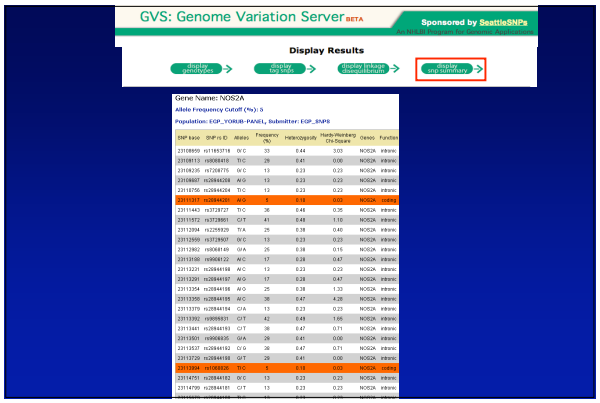

GVS: Genome Variation Server BETA
Sponsored by SeattleSNPs
Display Results
EGP Yoruban population
VG2 of: Gene Name NOS2A
UTR and Coding SNPs


GVS: Genome Variation Server BETA
Sponsored by SeattleSNPs
Display Results
HapMap CEP H population
VG2 of: Gene Name NOS2A


GVS: Genome Variation Server BETA
Sponsored by SeattleSNPs
Display Results
Variation Color code
Coding
UTR


GVS: Genome Variation Server BETA
Sponsored by SeattleSNPs
Display Results
VG2 of: Gene Name NOS2A


GVS: Genome Variation Server BETA
Sponsored by SeattleSNPs
Display Results
Set Parameters



Finding SNPs: Databases and Extraction

One stop shopping
- NIEHS SNPs and GeneSNPs

Prediction of functional variations
- Polydoms and PolyPhen

Identification of trancription factor binding sites in Evolutionary Conserved Regions
- TraFac and the ECR browser

Visualization and analysis of LD and TagSNPs
- GVS