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?

What do I need to know for selecting the "best" SNPs?
How do I pick the "best" SNPs?

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

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 #)

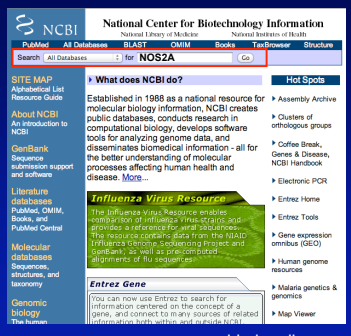
- Where is the SNP mapped? Exon, promoter, UTR, etc
  - ✓ picture of gene with mapped to the gene structure.
- How was it discovered? Method
- What assurances do you have that it is real? Validated how?
- What population – African, European, etc?
- What is the allele frequency of each SNP? Common (>10%),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

Finding SNPs: 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



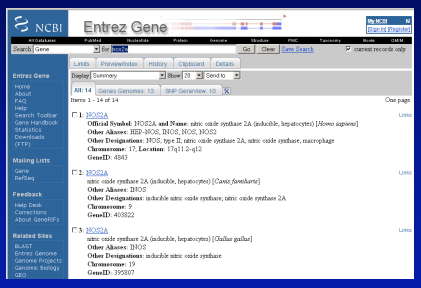
www.ncbi.nlm.nih.gov

Finding SNPs: Where do I start?

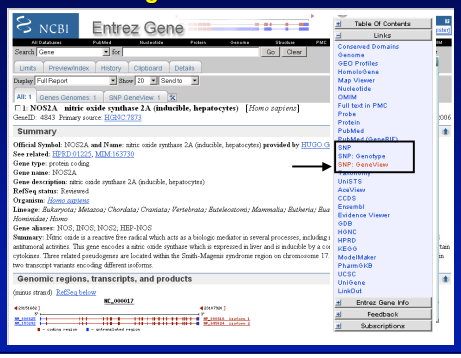
http://www.ncbi.nlm.nih.gov/gquery

Finding SNPs: Where do I start?

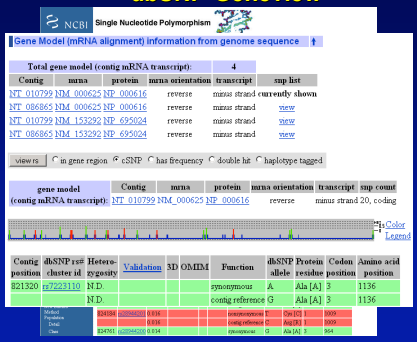
NCBI - Entrez Gene



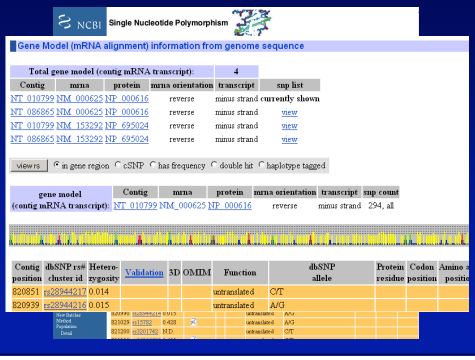
Finding SNPs: Entrez Gene



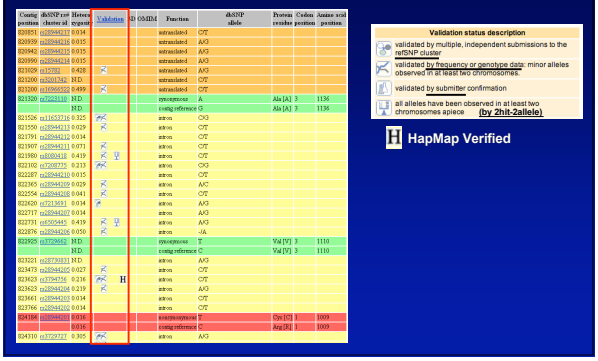
dbSNP Geneview

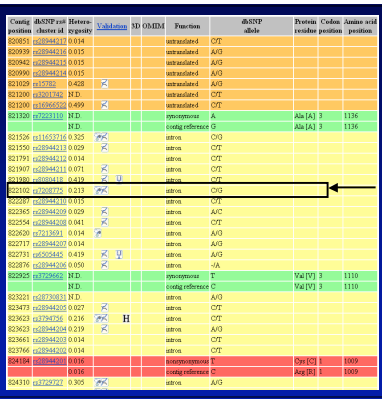


dbSNP Geneview

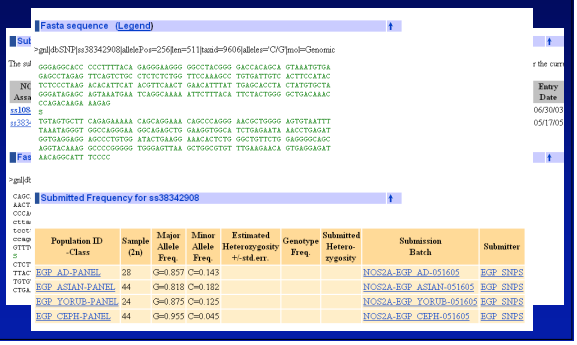


Finding SNPs: dbSNP validation

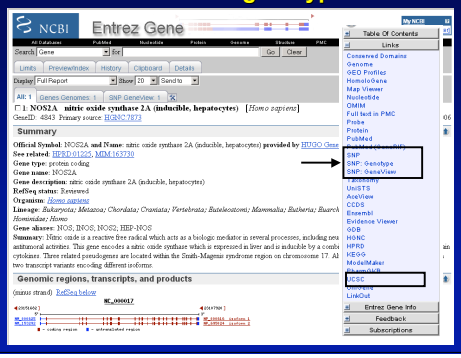




Finding SNPs: dbSNP database

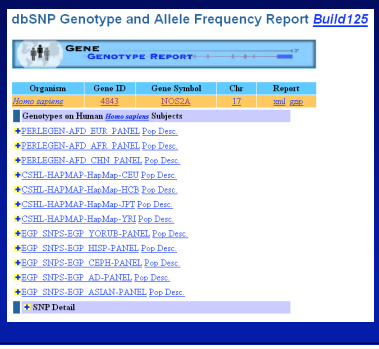


Entrez SNP - dbSNP genotype retrieval

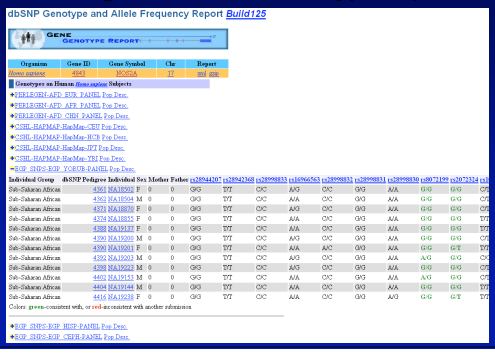


Finding SNPs - Gene Genotype Report

Finding SNPs - Gene Genotype Report



Finding SNPs - Gene Genotype Report



Minimal SNP information for genotyping/characterization

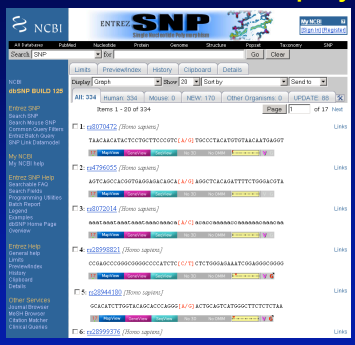
- What is the SNP? Flanking sequence and alleles.
  - ✓ FASTA format

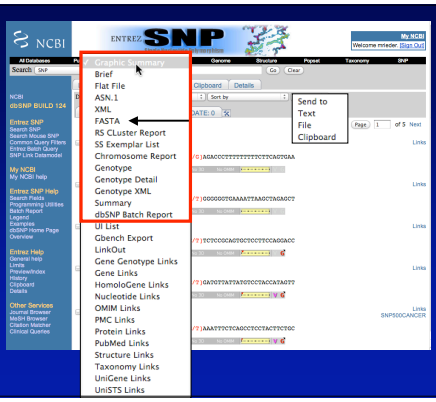
  >snp_name
  ACCGAGTAGCCAG
  [A/G]
  ACTGGGATAGAAC

  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!

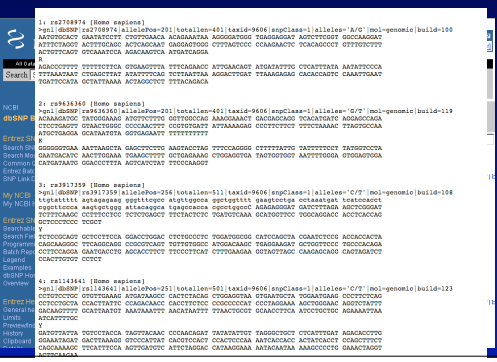
Entrez Gene Entry - Entrez SNP

Entrez SNP - direct dbSNP querying



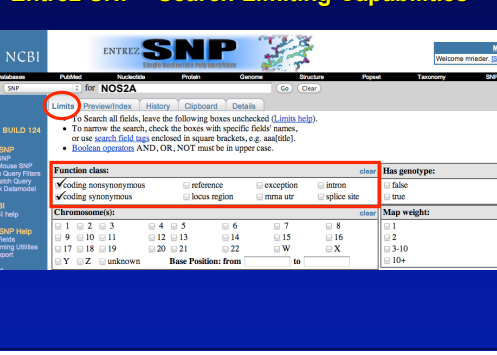


Entrez SNP - Parseable Multi-SNP reports

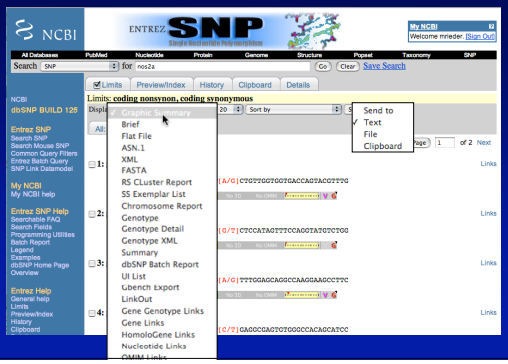


Entrez SNP - Parseable Multi-SNP reports

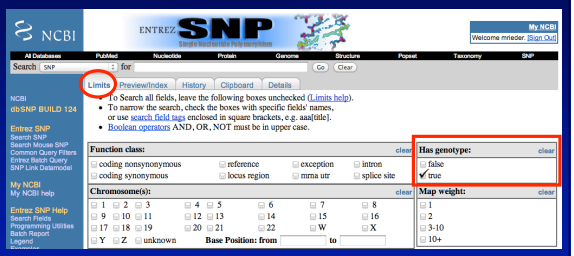
Entrez SNP - Search Limiting Capabilities



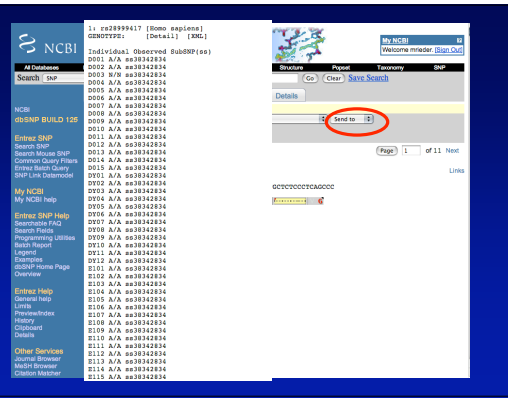
Entrez SNP - Search Limits



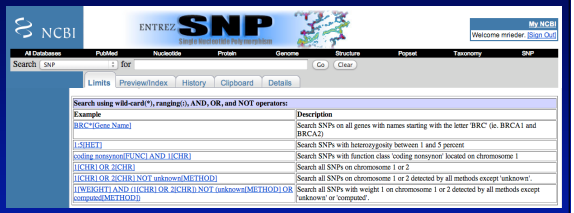
Entrez SNP - Search Limiting Capabilities



Entrez SNP - Search Limiting Capabilities



Entrez SNP - Query Term Capabilities



Entrez SNP - Search Terms Fields

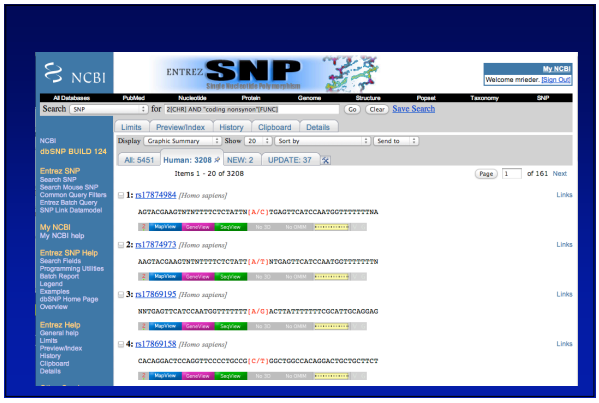
Search Field	Qualifier	Type	Description
Allele	[ALLELE],[VARIATION], [VARI]	IUPAC	[Observed allele(s)] Example: N[ALLELE]
Chromosome	[CHR]	Textnum	Mapped chromosome number Available values [1-22,W,Z, and Un (unknown)] Example: 2[CHR] or X[CHR]
Base Position	[CHRPOS],[BPOS]	Integer*	Mapped chromosome 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

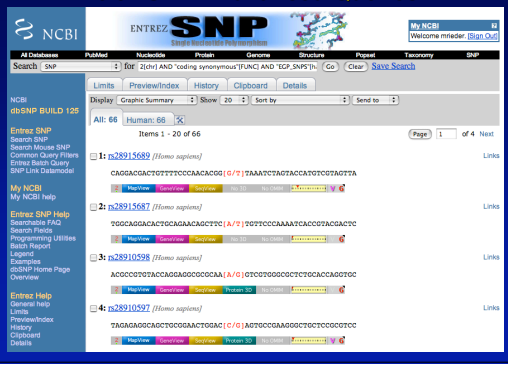
Submitter Handle	[HANDLE]	Text	Submitter handle Example: "LEE"[HANDLE]
Success Rate	[SUCCESS_RATE], [SRATE]	Integer*	Success rate(%) Example: 95[SRATE]
Organism	[ORGN],[TAX_ID]	Text	Organism name or Taxonomy ID number; use the prefix 'txid' before the Taxonomy ID number Example: "human"[ORGN] , "mouse"[ORGN], 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

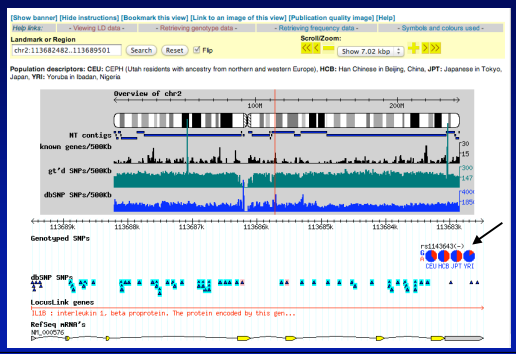
1. **dbSNP is useful for investigating detailed information on a small number SNPs - and its good for a picture of the gene**
2. **Entrez SNP is a direct, fast, database for querying SNP data.**
3. **Data from Entrez SNP can be retrieved in batches for many SNPs**
4. **Entrez SNP data can be "limited" 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

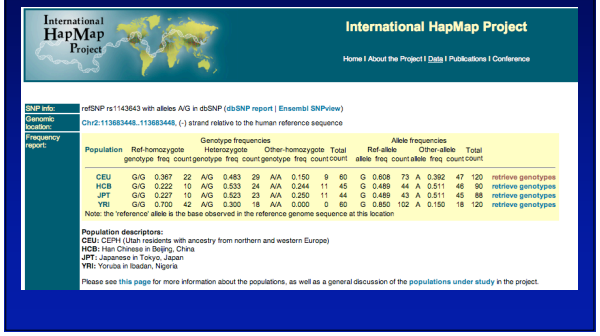
How do I find and download SNP data for analysis/genotyping?

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

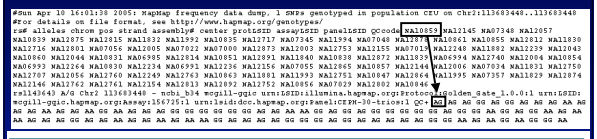
Finding SNPs: HapMap Browser



Finding SNPs: HapMap Browser



Finding SNPs: HapMap Genotypes

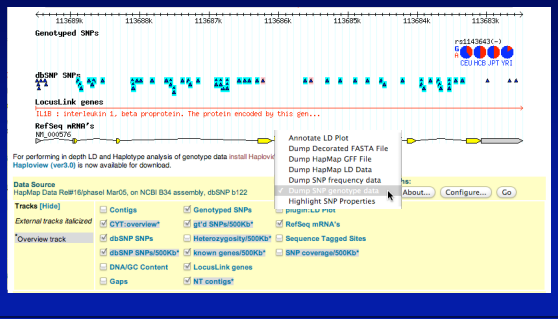


Bulk data downloads

The following directories contain project data that have been made publicly available. (See HapMap Data Access Policy for more information). More details about each dataset can be found in READMEs in the respective directories:

- **Genotypes:** Individual genotype data submitted to the DCC to date.
- **LD Data:** Linkage disequilibrium properties D', LOD , R^2 compiled from the genotype data to date
- **Allocated SNPs:** dbSNP reference SNP clusters that have been picked and prioritized for genotyping according to several criteria (see info on how SNPs were selected). The file 00README contains per-chromosome SNP counts and further details.
- **Frequencies:** Allele & genotype frequencies compiled from genotyping data submitted to the DCC to date. These have also been submitted to dbSNP and should be available in the next dbSNP build.
- **SNP assays:** Details about assays submitted to the DCC to date. PCR primers, extension probes etc., specific to each genotyping platform.
- **Protocols:** Information on assay design, genotyping and other protocols used in the project.
- **Samples/individuals:** Information on the samples used in the project and the individuals from which they were drawn. (See About the project: Which Populations are Being Sampled).
- **XML docs:** Documentation on the XML format used in the project.

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

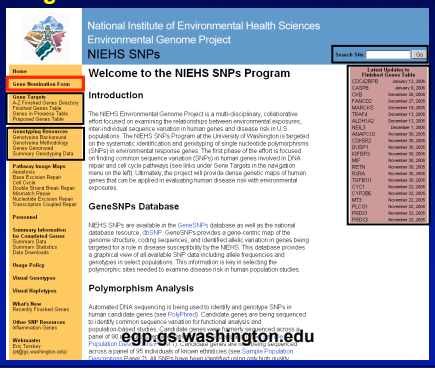
1. HapMap data sets are useful because individual genotype data can be used to determine optimal genotyping strategies (tagSNPs) or perform population genetic analyses (linkage disequilbrium)
2. Data are specific produced by those projects (not all dbSNP)
   - ✓ HapMap data is available in dbSNP
3. HapMap 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

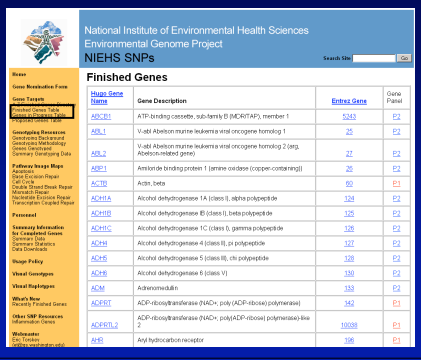
How do I find and download SNP data for analysis/genotyping?

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

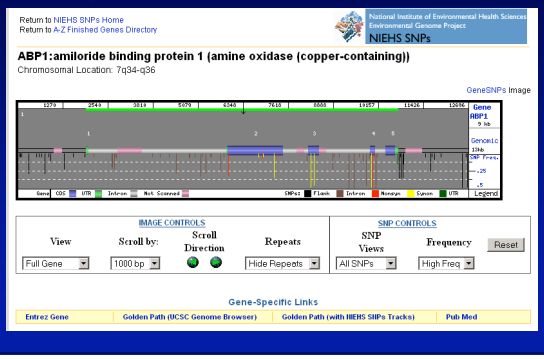
Finding SNPs: NIEHS SNPs Candidate Genes



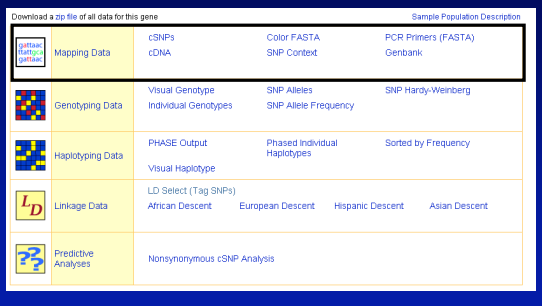
Finding SNPs: NIEHS SNPs Candidate Genes

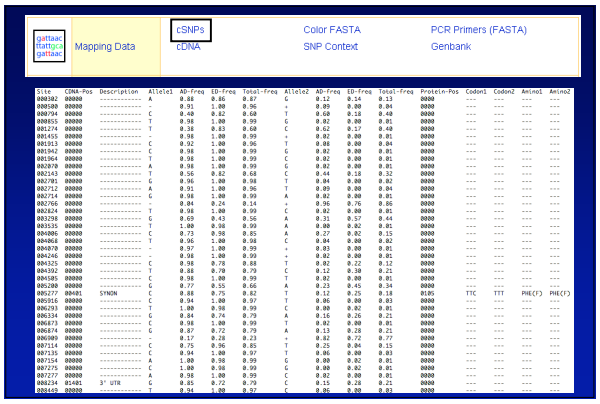


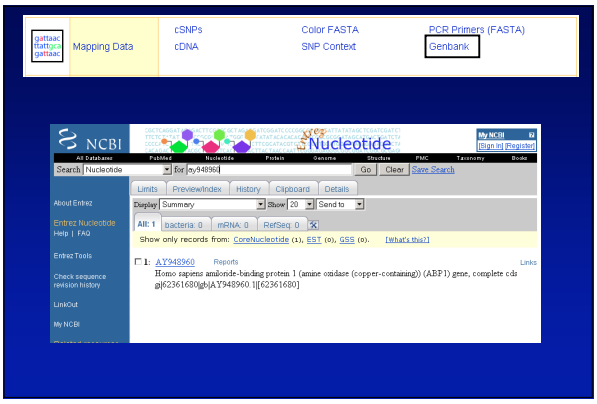
Finding SNPs: NIEHS SNPs Candidate Genes

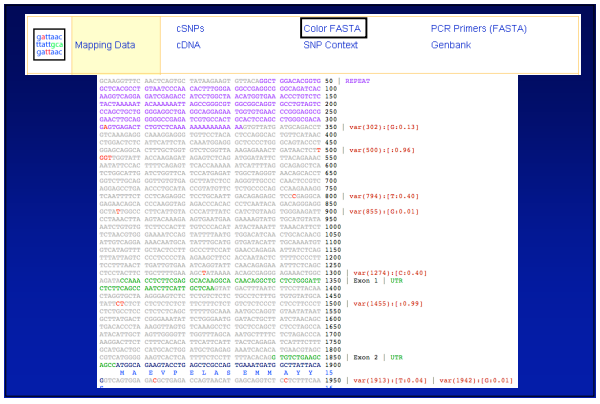


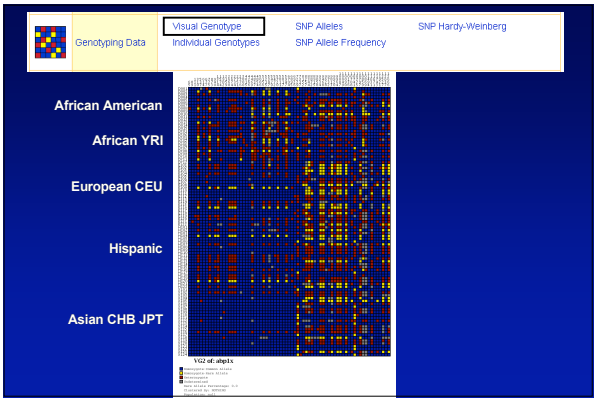
Finding SNPs: NIEHS SNPs Candidate Genes

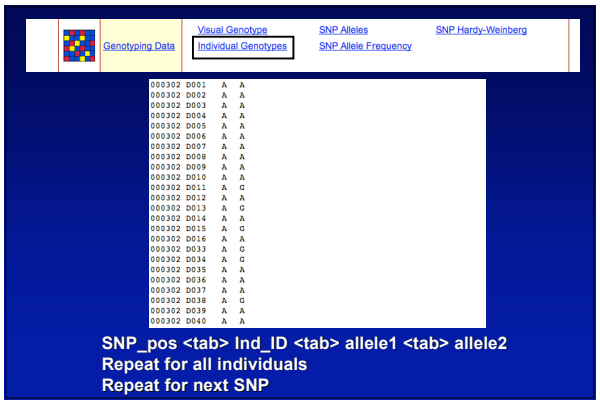


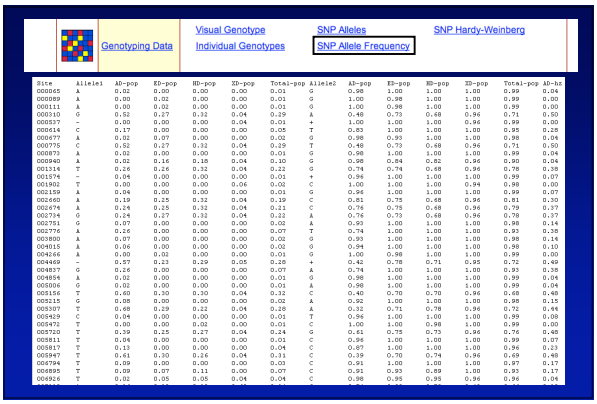

Mapping Data
cSNPs
cDNA
Color FASTA
SNP Context
PCR Primers (FASTA)
Genbank

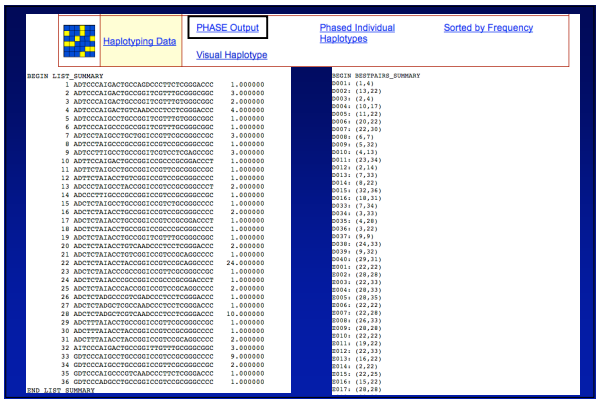

Mapping Data
cSNPs
cDNA
Color FASTA
SNP Context
PCR Primers (FASTA)
Genbank
NCBI
Nucleotide

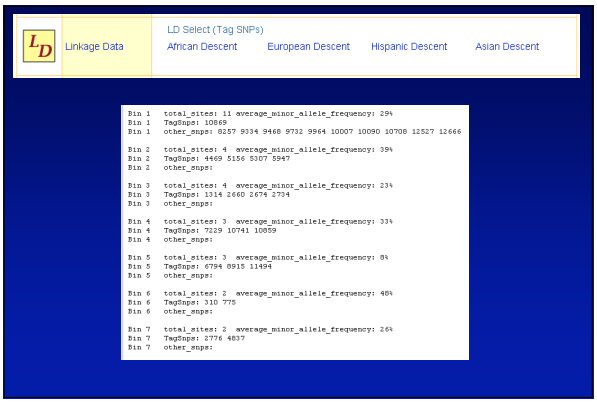

Mapping Data
cSNPs
cDNA
Color FASTA
SNP Context
PCR Primers (FASTA)
Genbank

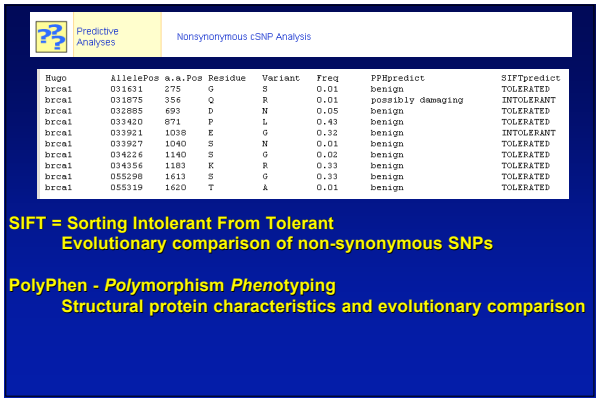

Genotyping Data
Visual Genotype
Individual Genotypes
SNP Alleles
SNP Allele Frequency
SNP Hardy-Weinberg
African American
African YRI
European CEU
Hispanic
Asian CHB JPT


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



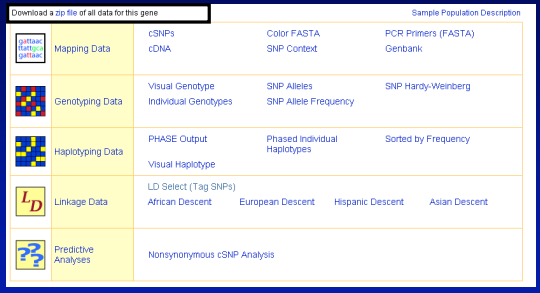




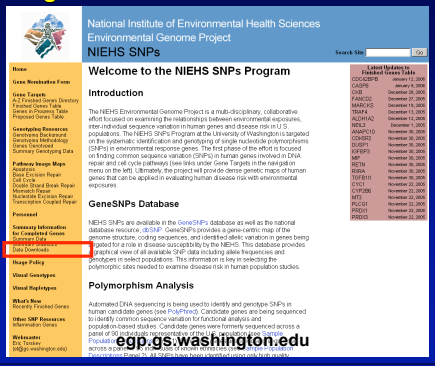
SIFT = Sorting Intolerant From Tolerant
Evolutionary comparison of non-synonymous SNPs

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

Finding SNPs: NIEHS SNPs Candidate Genes



Finding SNPs: NIEHS SNPs Candidate Genes



Finding SNPs: NIEHS SNPs Candidate Genes

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

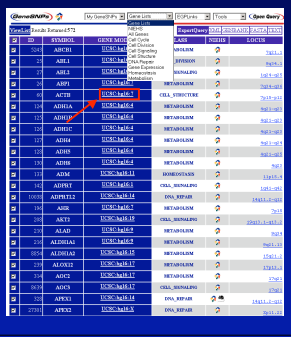
Graphic view of SNPs in context of gene elements
All NIEHS genes presented
- organized by pathway/function

SNPs from dbSNP
- organized by submitter handle

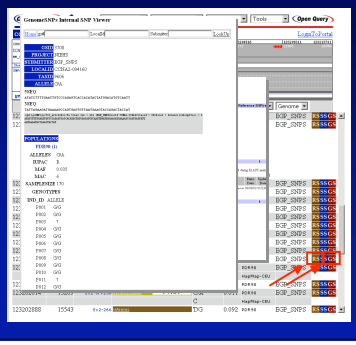
Sequence context of SNPs presented in Color Fasta format
Link-outs to EntrezSNP pages
Summary "Genome SNPs" internal SNP viewer for one-stop SNP shopping

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

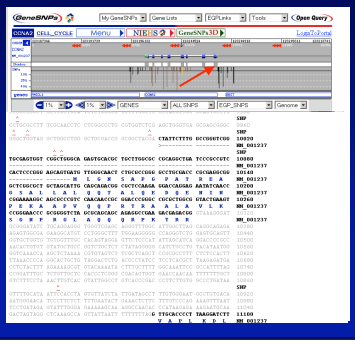
GeneSNPs: One stop shopping



GeneSNPs: One stop shopping



GeneSNPs: One stop shopping



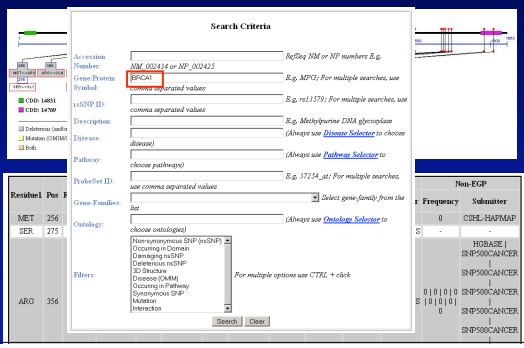
Polydoms

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

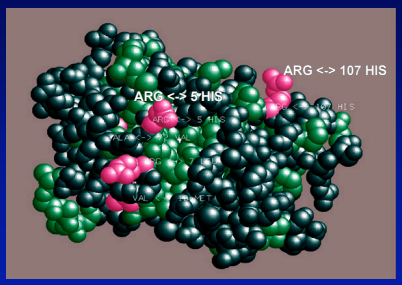
- SNPs are from dbSNP and GeneSNPs
- Domain structures from NCBI's Conserved Domain Database
- Functional predictions based on SIFT and PolyPhen
- 3 dimensional mapping of SNPs on protein structure using Chime viewer

http://polydoms.cchmc.org/polydoms/

Polydoms



Polydoms



Mapping of nsSNPS onto protein structure

TraFac: Transcription Factor Binding Site Comparison

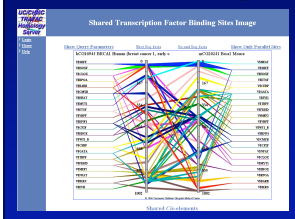
A tool for validating cis regulatory elements conserved between human and mouse

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

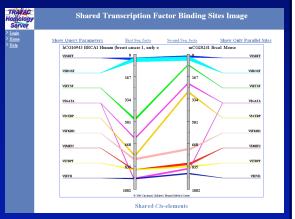
http://trafac.cchmc.org/trafac

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

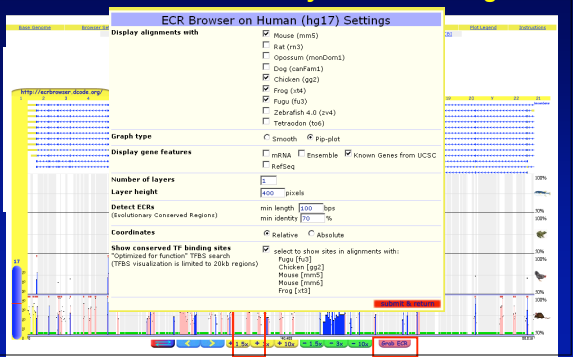
Gene annotations from UCSC Genome Browser

Easy retrieval of ECR sequences and alignments

Pre-computed transcription factor binding sites

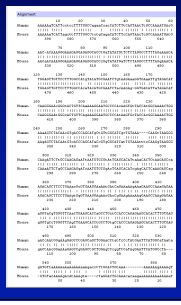
http://ecrbrowser.dcode.org

ECR Browser: Evolutionary Conserved Regions



ECR Browser: Evolutionary Conserved Regions

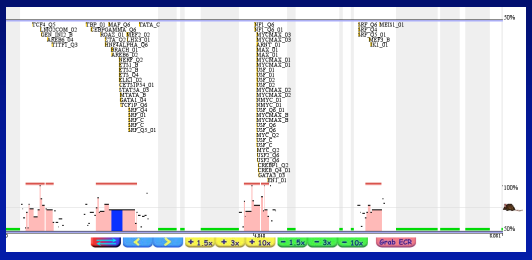
Human-mouse alignment



Fasta sequences

ECR Browser: Evolutionary Conserved Regions

Transcription Factor Binding Sites from Transfac

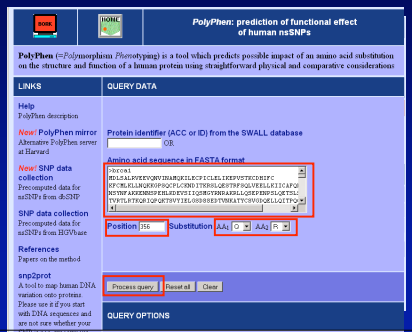


PolyPhen: Polymorphism Phenotyping-
prediction of functional effect of human nsSNPs

- 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 human nsSNPs

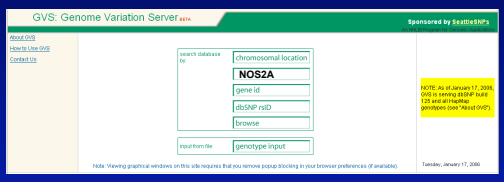


GVS: Genome Variation Server

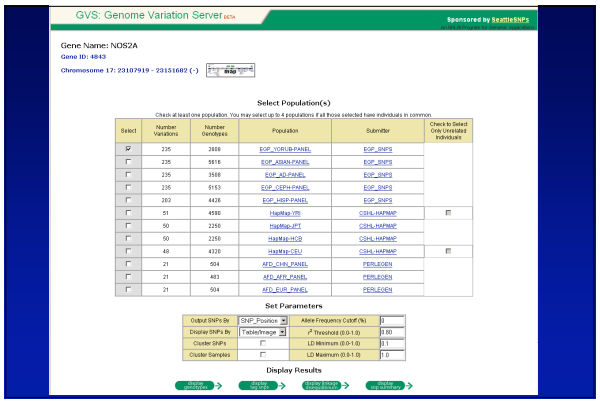
- Provides rapid analysis of 4.3 million SNPs from dbSNP and the HapMap
- Mapped to human genome build 35 (hg17)
- Displays genotype data in text and image formats
- Displays tagSNPs or clusters of informative SNPs in text and image formats
- Displays linkage disequilibrium (LD) in text and image formats

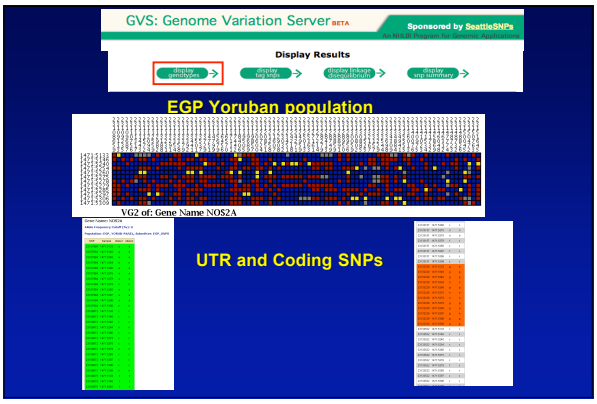
http://gvs.gs.washington.edu/GVS/

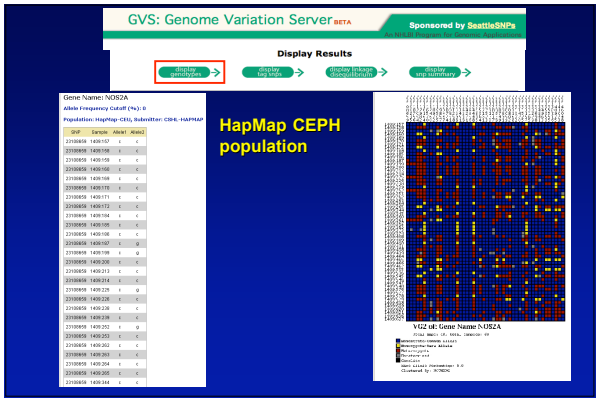
GVS: Genome Variation Server

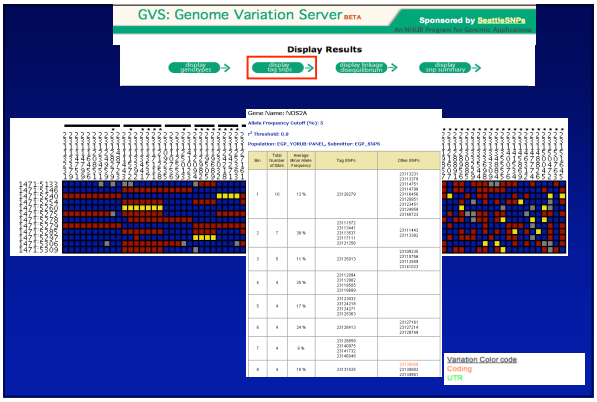


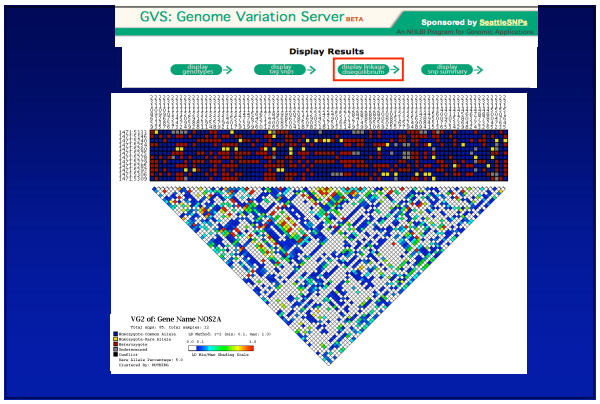
http://gvs.gs.washington.edu/GVS/

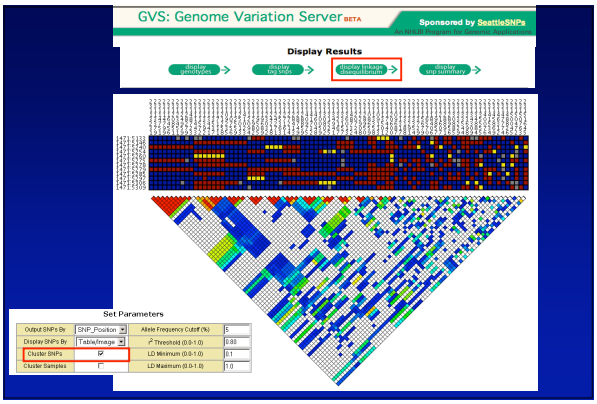

GVS: Genome Variation Server BETA
Sponsored by SeattleSNPs
Gene Name: NOS2A
Gene ID: 4843
Chromosome 17: 23107919 - 23151682 (-)
Select Population(s)
Set Parameters
Display Results

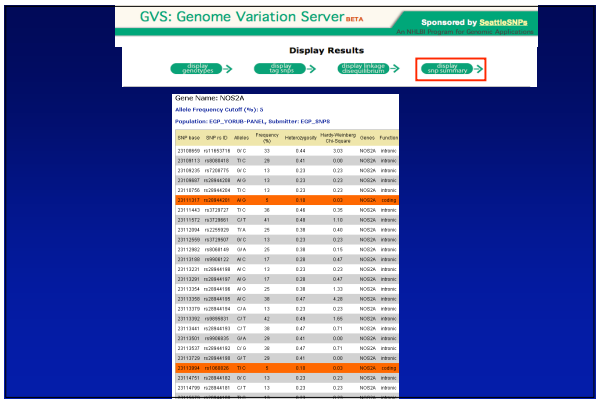

GVS: Genome Variation Server BETA
Sponsored by SeattleSNPs
Display Results
EGP Yoruban population
VG2 of: Gene Name NOS2A
UTR and Coding SNPs


GVS: Genome Variation Server BETA
Sponsored by SeattleSNPs
Display Results
HapMap CEPH 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