1. Introduction to the Gene Ontology*
2. SNPdb

*Many of the GO slides courtesy of Pascale Gaudet, dictyBase curator, Northwestern University, Chicago, IL
Too Much Published Literature

...to read in order to find out what your gene does....

- **PubMed**: db of over 15 million citations
  - **Basic search**: 
    - rad51 → 3929 articles
  - **Organism Limited search**: 
    - rad51 AND Human (organism) → 2488
  - **Disease Limited search**: 
    - rad51 AND cancer → 1909
Ontology

• Ontologies relate **facts** to **knowledge**

**facts**
- may be known/unknown/little known
- not attached to knowers
- unchanging

**knowledge**
- attached to knower
- may disappear
Gene Ontology

- Gene *annotation* system
- Controlled *vocabulary* that can be applied to all organisms
- Used to describe *gene products*
How do you answer: What is it?

Example:

What is a cell?
Cell
Cell
Cell
Cell

Image from http://microscopy.fsu.edu
The same name can be used to describe different concepts
What is this?

- Glucose synthesis
- Glucose biosynthesis
- Glucose formation
- Glucose anabolism
- Gluconeogenesis

All refer to the process of making glucose from simpler components
The same concept can be described with different names
How do we solve this?

• The same name can be used to describe different concepts

• The same concept can be described using different names

→ Comparison is difficult – in particular across species or across databases

→ Solving this problem will enable natural language processing (and more)
What is the Gene Ontology?

A part of the solution!

- A controlled vocabulary that can be applied to all organisms

- Used to describe gene products - proteins and RNA - in any organism

• An **ontology** provides structure to concepts.
Ontology

• In philosophy, the most fundamental branch of metaphysics. It studies **being** or **existence** as well as the basic categories thereof—trying to find out what **entities** and what **types of entities** exist.
  – Wikipedia

• Ontologies provide **controlled, consistent vocabularies** to describe concepts and relationships, thereby enabling **knowledge sharing**
  – Gruber 1993
Ontology

Includes:

1. Terms
2. Definitions (spectrum of *explicitity*)
3. Logical relationships

"Definition" spectrum

Vague  Explicit
Ontology graph structure

- Nodes = **concepts** in the ontology
- Edges = **relationships** between the concepts
Ontology edges

• Terms are linked by two relationships
  – is-a
  – part-of
Fusing databases

open biological ontologies

http://www.obofoundry.org/


The value of any kind of data is greatly enhanced when it exists in a form that allows it to be integrated with other data. One approach to integration is through the annotation of multiple bodies of data using common controlled vocabularies or 'ontologies'. Unfortunately, the very success of this approach has led to a proliferation of ontologies, which itself creates obstacles to integration. The Open Biomedical Ontologies (OBO) consortium is pursuing a strategy to overcome this problem. Existing OBO ontologies, including the Gene Ontology, are undergoing coordinated reform, and new ontologies are being created on the basis of an evolving set of shared principles governing ontology development. The result is an expanding family of ontologies designed to be interoperable and logically well formed and to incorporate accurate representations of biological reality. We describe this OBO Foundry initiative and provide guidelines for those who might wish to become involved.
core relationships

- OBO_REL:is_a
  - Name: is_a
  - Relation properties: [transitive] [reflexive] [anti-symmetric]
  - Definition: For continuants: C is_a C' if and only if: given any c that instantiates C at a time t, c instantiates C' at t. For processes: P is_a P' if and only if: that given any p that instantiates P, then p instantiates P'.

- OBO_REL:part_of
  - Name: part_of
  - Relation properties: [transitive] [reflexive] [anti-symmetric]
  - Definition: For continuants: C part_of C' if and only if: given any c that instantiates C at a time t, there is some c' such that c' instantiates C' at time t, and c *part_of* c' at t. For processes: P part_of P' if and only if: given any p that instantiates P at a time t, there is some p' such that p' instantiates P' at time t, and p *part_of* p' at t. (Here *part_of* is the instance-level part-relation.)
Simple hierarchies  (Trees)  

Relationships in OBO are directional. Cycles are not allowed.

Directed Acyclic Graphs

Single parent  

One or more parents
An ontology is a Directed Acyclic Graph (DAG)

- protein complex
- organelle
  - mitochondrion
    - fatty acid beta-oxidation multienzyme complex

- [other protein complexes]
- [other organelles]
True Path Rule

• The path from a child term all the way up to its top-level parent(s) must always be true
GO: the Gene Ontology

Composed of Three sub-ontologies

What does it do?

What processes is it involved in?

Where does it act?

gene product

Molecular Function

Biological Process

Cellular Component
GO: Cellular Component

• where a gene product acts
Mitochondrial membrane paths

- cellular_component: 0(+7268) genes
- cell: 1(+5434) genes
- intracellular: 46(+4974) genes
- organelle: 0(+3919) genes
- intracellular organelle: 0(+3919) genes
- membrane-bound organelle: 0(+3600) genes
- membrane: 83(+942) genes
- cytoplasm: 311(+2556) genes
- intracellular membrane-bound organelle: 0(+3600) genes
- organelle membrane: 0(+572) genes
- mitochondrion: 715(+382) genes
- mitochondrial membrane: 16(+194) genes
GO: Molecular Function

- A single reaction or activity, not a gene product
- A gene product may have several functions
- Sets of functions make up a biological process
Molecular Function

Phosphoenolpyruvate → Glucose 6-phosphate → Glucose

Pyruvate → Lactate

CO₂ → Acetyl coenzyme A → Oxaloacetate

Amino acid intermediates → KREBS CYCLE → CO₂

Citrate → Amino acid intermediates → Triacylglycerol metabolism

Glycerol
Carbonate dehydratase activity
Anatomy of a GO term

term: gluconeogenesis

identifier: GO:0006094

definition: The formation of glucose from noncarbohydrate precursors, such as pyruvate, amino acids and glycerol.
No GO Areas

- GO covers ‘normal’ functions and processes
  - No pathological processes
  - No experimental conditions
- NO evolutionary relationships
- NO gene products
- NOT a system of nomenclature
GO is a tree.

- **all : all (164672)**
  - **GO:0008150 : biological_process (116553)**
    - **GO:0007610 : behavior (3333)**
    - **GO:0000004 : biological process unknown (33114)**
    - **GO:0009987 : cellular process (69630)**
    - **GO:0007275 : development (11890)**
    - **GO:0040007 : growth (3071)**
    - **GO:0044419 : interaction between organisms (775)**
    - **GO:0007582 : physiological process (74243)**
    - **GO:0043473 : pigmentation (83)**
    - **GO:0050789 : regulation of biological process (12939)**
    - **GO:0000003 : reproduction (3821)**
    - **GO:0016032 : viral life cycle (278)**
  - **GO:0005575 : cellular_component (102004)**
    - **GO:0005623 : cell (72949)**
      - **GO:0008372 : cellular component unknown (25487)**
      - **GO:0031012 : extracellular matrix (566)**
      - **GO:0005576 : extracellular region (4320)**
      - **GO:0043226 : organelle (53431)**
      - **GO:0043234 : protein complex (10154)**
      - **GO:0019012 : virion (134)**
  - **GO:0003674 : molecular_function (114384)**
    - **GO:0016209 : antioxidant activity (457)**
    - **GO:0005488 : binding (29930)**
    - **GO:0003824 : catalytic activity (38141)**
    - **GO:0030188 : chaperone regulator activity (40)**
    - **GO:0030234 : enzyme regulator activity (1806)**
      - **GO:0005554 : molecular function unknown (34043)**
    - **GO:0003774 : motor activity (487)**
      - **GO:0045735 : nutrient reservoir activity (45)**
      - **GO:0031386 : protein tag (12)**
    - **GO:0004871 : signal transducer activity (8305)**
    - **GO:0005198 : structural molecule activity (3366)**
    - **GO:0030528 : transcription regulator activity (7662)**
Annotation of gene products with GO terms

Mitochondrial P450
GO terms for Mitochondrial P450

**Cellular component:**
mitochondrial inner membrane
GO:0005743

**Biological process:**
Electron transport
GO:0006118

**Molecular function:**
monooxygenase activity
GO:0004497

substrate + O_2 = CO_2 + H_2 O product
Two types of GO Annotations:

- Electronic Annotation
- Manual Annotation

All annotations must:

- be attributed to a source
- indicate what evidence was found to support the GO term-gene/protein association
Manual Annotations

• High–quality, specific gene/gene product associations made, using:
  • Peer-reviewed papers
  • Evidence codes to grade evidence

**BUT** – is very time consuming and requires trained biologists
Electronic Annotations

- Provides large-coverage
- High-quality

**BUT** – annotations tend to use high-level GO terms and provide little detail.
Electronic Annotations: Methods

1. Database entries
   • Manual mapping of GO terms to concepts external to GO (‘translation tables’)
   • Proteins then electronically annotated with the relevant GO term(s)

2. Automatic sequence similarity analyses to transfer annotations between highly similar gene products
Manual Annotations: Methods

1. Extract information from published literature

2. Curators performs manual sequence similarity analyses to transfer annotations between highly similar gene products (BLAST, protein domain analysis)
In this study, we report the isolation and molecular characterization of the *B. napus* PERK1 cDNA, that is predicted to encode a novel receptor-like kinase. We have shown that like other plant RLKs, the kinase domain of PERK1 has serine/threonine kinase activity, In addition, the location of a PERK1-GFP fusion protein to the plasma membrane supports the prediction that PERK1 is an integral membrane protein...these kinases have been implicated in early stages of wound response... for *B. napus* PERK1 protein (Q9ARH1)

Function: protein serine/threonine kinase activity  GO:0004674
Component: integral to plasma membrane  GO:0005887
Process: response to wounding  GO:0009611
Annotate to finest granularity

Annotating to GO:0030047 automatically annotates to all of its parents; thus a product is annotated to both protein modification AND cytoskeleton organization
Unknown vs. Unannotated

• “Unknown” is used when the curator has determined that there is no existing literature to support an annotation.
  – Biological process unknown GO:0000004
  – Molecular function unknown GO:0005554
  – Cellular component unknown GO:0008372

• Annotation “unknown” is NOT the same as having no annotation.
  – No annotation means that no one has looked yet. Unknown means they looked and did not find.
# GO Evidence Codes

<table>
<thead>
<tr>
<th>Code</th>
<th>Definition</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEA</td>
<td>Inferred from <strong>Electronic Annotation</strong></td>
<td></td>
</tr>
<tr>
<td>NAS</td>
<td><strong>Non-traceable Author Statement</strong></td>
<td></td>
</tr>
<tr>
<td>TAS</td>
<td><strong>Traceable Author Statement</strong></td>
<td></td>
</tr>
<tr>
<td>ND</td>
<td><strong>No Data</strong></td>
<td>Use with annotation to unknown</td>
</tr>
<tr>
<td>IDA</td>
<td>Inferred from <strong>Direct Assay</strong></td>
<td><strong>Manually annotated</strong></td>
</tr>
<tr>
<td>*IPI</td>
<td>Inferred from <strong>Physical Interaction</strong></td>
<td></td>
</tr>
<tr>
<td>*IGI</td>
<td>Inferred from <strong>Genetic Interaction</strong></td>
<td></td>
</tr>
<tr>
<td>IMP</td>
<td>Inferred from <strong>Mutant Phenotype</strong></td>
<td></td>
</tr>
<tr>
<td>IEP</td>
<td>Inferred from <strong>Expression Pattern</strong></td>
<td></td>
</tr>
<tr>
<td>*IC</td>
<td>Inferred from <strong>Curator</strong></td>
<td></td>
</tr>
<tr>
<td>*ISS</td>
<td>Inferred from <strong>Sequence Similarity</strong></td>
<td></td>
</tr>
</tbody>
</table>
### GO Evidence Codes

<table>
<thead>
<tr>
<th>Code</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>*IEA</td>
<td>Inferred from Electronic Annotation</td>
</tr>
<tr>
<td>IDA</td>
<td>Inferred from Direct Assay</td>
</tr>
<tr>
<td>IEP</td>
<td>Inferred from Expression Pattern</td>
</tr>
<tr>
<td>*IGI</td>
<td>Inferred from Genetic Interaction</td>
</tr>
<tr>
<td>IMP</td>
<td>Inferred from Mutant Phenotype</td>
</tr>
<tr>
<td>*IPI</td>
<td>Inferred from Physical Interaction</td>
</tr>
<tr>
<td>*ISS</td>
<td>Inferred from Sequence Similarity</td>
</tr>
<tr>
<td>TAS</td>
<td>Traceable Author Statement</td>
</tr>
<tr>
<td>NAS</td>
<td>Non-traceable Author Statement</td>
</tr>
<tr>
<td>*IC</td>
<td>Inferred from Curator</td>
</tr>
<tr>
<td>RCA</td>
<td>Inferred from Reviewed Computational Analysis</td>
</tr>
<tr>
<td>ND</td>
<td>No Data</td>
</tr>
</tbody>
</table>

**IDA:**
- Enzyme assays
- *In vitro* reconstitution (transcription)
- Immunofluorescence
- Cell fractionation

**TAS:**
- In the literature source the original experiments referred to are traceable (referenced).
GO Evidence Codes: with/from

Additional information required for certain evidence codes

<table>
<thead>
<tr>
<th>Code</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>*IEA</td>
<td>Inferred from Electronic Annotation</td>
</tr>
<tr>
<td>IDA</td>
<td>Inferred from Direct Assay</td>
</tr>
<tr>
<td>IEP</td>
<td>Inferred from Expression Pattern</td>
</tr>
<tr>
<td>*IGI</td>
<td>Inferred from Genetic Interaction</td>
</tr>
<tr>
<td>IMP</td>
<td>Inferred from Mutant Phenotype</td>
</tr>
<tr>
<td>*IPI</td>
<td>Inferred from Physical Interaction</td>
</tr>
<tr>
<td>*ISS</td>
<td>Inferred from Sequence Similarity</td>
</tr>
<tr>
<td>TAS</td>
<td>Traceable Author Statement</td>
</tr>
<tr>
<td>NAS</td>
<td>Non-traceable Author Statement</td>
</tr>
<tr>
<td>*IC</td>
<td>Inferred from Curator</td>
</tr>
<tr>
<td>RCA</td>
<td>Inferred from Reviewed Computational Analysis</td>
</tr>
<tr>
<td>ND</td>
<td>No Data</td>
</tr>
</tbody>
</table>

IGI:
- a gene identifier for the "other" gene involved in the interaction

*IPI:
- a gene or protein identifier for the "other" protein involved in the interaction

IC:
- GO term from another annotation used as the basis of a curator inference

*With column required
Evidence code, confidence

TAS/IDA
IMP/IGI/IPI
ISS/IEP
NAS
IEA

believe it
question it
Modifying the interpretation of an annotation: the Qualifier column

1. **NOT**
   - a gene product is NOT associated with the GO term
   - to document conflicting claims in the literature.

2. **Contributes to**
   - distinguishes between individual subunit functions and whole complex functions
   - used with GO Function Ontology

3. **Colocalizes with**
   - transiently or peripherally associated with an organelle or complex
   - used with GO Component Ontology
How to access the Gene ontology and its annotations

1. Downloads

   • Ontologies
   • Annotations: Gene association files
   • Ontologies and Annotations

2. Web-based access

   • AmiGO
     (http://www.godatabase.org)
   • QuickGO
     (http://www.ebi.ac.uk/ego)
   • GOrilla
     (http://cbl-gorilla.cs.technion.ac.il/)
   • Blast2GO
     (www.blast2go.com)
     among others…
Proteomic Enrichment

GO RILLA

Two sets of Genes

Enriched GO Terms Describing Function, Component, or Process

Gene Ontology enrichment analysis and visualization tool
• **Proteome** -- the protein content of the cell. Depends on species, cell type, conditions.

• **Transcriptome** -- the mRNA content of the cell. Depends on species, cell type, conditions.
RESEARCH ARTICLE

A novel hypothesis-unbiased method for Gene Ontology enrichment based on transcriptome data

Mario Fruzangohar¹,²*, Esmaeil Ebrahimie³,⁴,⁵,⁶, David L. Adelson¹,⁷*

¹ School of Biological Sciences, The University of Adelaide, Adelaide, South Australia, Australia, ² School of Agriculture, Food and Wine, The University of Adelaide, Adelaide, South Australia, Australia, ³ Australian Centre for Antimicrobial Resistance Ecology, School of Animal and Veterinary Sciences, The University of Adelaide, Adelaide, South Australia, Australia, ⁴ School of Medicine, Faculty of Health Sciences, The University of Adelaide, Adelaide, Australia, ⁵ School of Information Technology and Mathematical Sciences, Division of Information Technology, Engineering and the Environment, University of South Australia, Adelaide, Australia, ⁶ School of Biological Sciences, Faculty of Science and Engineering, Flinders University, Adelaide, Australia, ⁷ Zhendong Australia – China Centre for Molecular Chinese Medicine, The University of Adelaide, Adelaide, South Australia, Australia

* mario.fruzangohar@adelaide.edu.au (MF); david.adelson@adelaide.edu.au (DLA)

Abstract

Gene Ontology (GO) classification of statistically significantly differentially expressed genes is commonly used to interpret transcriptomics data as a part of functional genomic analysis. In this approach, all significantly expressed genes contribute equally to the final GO classification. Here we developed a novel hypothesis-unbiased method for GO enrichment analysis based on transcriptome data.
Materials and methods
Incorporation of mRNA expression levels into GO enrichment

Given \( N \) genes \( (g_1, \ldots, g_n) \) in \( K \) samples, we estimate the enrichment score (ES) of a GO term \( t \) in sample \( s \) \( ES_{t,s} \), when expression levels are given as RPKM (Reads per Kilo base per Million Reads)/FPKM (Fragments per Kilo base per Million Reads):

\[
ES_{t,s} = \sum_{i=1}^{n} \log_2 [e(i,s) + 1] \times I(i,t) \tag{1.1}
\]

or as microarray log fold change:

\[
ES_{t,s} = \sum_{i=1}^{n} \log_2 (2^{e(i,s)} + 1) \tag{1.2}
\]

Where \( e(i,s) \) is the expression level of gene \( g_i \) in sample \( s \) and \( I(i,t) \) is:

\[
I(i,t) = \begin{cases} 1, & \text{if } g_i \text{ annotated by GO term } (t) \\ 0, & \text{otherwise} \end{cases}
\]

We then define an intermediate value for fold change (\( F \)) of GO term \( t \) from sample \( s \) to sample \( s+1 \) (\( F_{t,s} \)):

\[
\frac{ES_{t,s+1}}{ES_{t,s}} \tag{2}
\]

Finally, the average fold change of GO term \( t \) across all samples is defined as:

\[
F_t = \sqrt[k-1]{\prod_{s=1}^{k-1} F_{t,s}} \tag{3.1}
\]

or log transformed as:

\[
F_t = \frac{1}{k-1} \sum_{s=1}^{k-1} \log_2 F_{t,s} \tag{3.2}
\]
Case studies

Case study 1: Comparison of whole transcriptome based GO enrichment between minimally and highly pathogenic *Salmonella enteritidis*. We used RNA-Seq data for six *Salmonella enteritidis* [12] strains. For each gene in both groups of strains the RPKM counts were averaged.

Case study 2: Comparison of whole transcriptome based GO enrichment between AD and normal human brain. The results of whole transcriptome GO classification followed by outlier testing in AD and normal brains in Biological Process, Molecular Function and Cellular Component are presented in Fig 2. The most significant functions in AD were inflammation and fatty acid related functions including granulocyte colony-stimulating factor receptor binding, interleukin-1, alcohol dehydrogenase, neuromedin U receptor activity, and norepinephrine transmembrane transporter activity.
Brief isoflurane anaesthesia affects differential gene expression, gene ontology and gene networks in rat brain

Damon A. Lowes \textsuperscript{a}, Helen F. Galley \textsuperscript{a} \& \textsuperscript{b}, Alessandro P.S. Moura \textsuperscript{b}, Nigel R. Webster \textsuperscript{a}

Abstract

Much is still unknown about the mechanisms of effects of even brief anaesthesia on the brain and previous studies have simply compared differential expression profiles with and without anaesthesia. We hypothesised that network analysis, in addition to the traditional differential gene expression and ontology analysis, would enable identification of the effects of anaesthesia on interactions between genes. Rats (\(n = 10\) per group) were randomised to anaesthesia with isoflurane in oxygen or oxygen only for 15 min, and 6 h later brains were removed. Differential gene expression and gene ontology analysis of microarray data was performed. Standard clustering techniques and principal component analysis with Bayesian rules were used along with social network analysis methods, to quantitatively model and describe the gene networks. Anaesthesia had marked effects on genes in the brain.
GO analysis of transcriptomics of human cells treated with natural product

Analysis of Functional Annotation – Downregulated Genes

GO terms
- 04110_cell cycle
- GO:0007067_mitosis
- GO:0000278_mitotic cell cycle
- GO:0000279_M phase
- PCNA
- E2F1
- CCNA2
- CDK2
- CDC6
- CACGTG_V$MYC_Q6
- V$E2F_Q6
- 00970_aminoacyl-tRNA biosynthesis
- GO:0005840_ribosome
- GO:0043039_tRNA aminoacylation
- GO:0006418_trNA aminoacylation for protein translation
- 03050_proteasome

courtesy of Shabana Shabeer, Albert Einstein School of Medicine
Review

1. What is an ontology?
2. Why do we need ontologies?
3. What are the three gene ontologies?
4. Give an example of a is-a relationship in the Cellular Component ontology
5. Give an example of the part-of relationship in the Cellular Component ontology.
6. ....in the Molecular Function ontology.
7. ....in the Biological process ontology.
8. What two things must every annotation have?
9. What are evidence codes?
10. What is the least confident evidence code?
SNPs
What are polymorphisms?

- Genetic differences between individuals in a population.
- Changes related to alleles
  - **Single nucleotide polymorphisms** (one base substitution)
    - Noncoding
    - Coding
      - synonymous -- same amino acid, different codon
      - non-synonymous
        - missense -- change in amino acid
        - nonsense -- stop codon
  - Frame-shifts
    - One or more base Insertion/deletion

SNPs = single nucleotide polymorphisms
How To View all SNPs associated with a gene
Starting with...

a gene name

1. Search the Gene database with the gene name. If you know the gene symbol and species, enter them as follows: tpo[gene] AND human[orgn]
2. Click on the Gene ID of the desired gene.
3. In the list of Links on the right, click "SNP:GeneView". If the link is not present, no SNPs are currently linked to this gene.

a nucleotide or protein accession number (e.g. NM_001126)

1. Search the Nucleotide or Protein database with the accession number.
2. In the Links menu in the upper right, click on "GeneView in dbSNP". If the link is not present, click on the "Gene" link in the same menu and continue at step 3 above under "a gene name".

a nucleotide sequence

1. Go to the BLAST home page and click "nucleotide blast" (blastn) under Basic BLAST.
2. Paste the sequence in the query box.
3. Enter the name of the organism of interest in the "Organism" box. Click the BLAST button.
4. Click on the desired sequence from the results.
5. Continue at step 2 under "a nucleotide or protein accession number" above.

a protein sequence

1. Go to the BLAST home page and click "protein blast" (blastp) under Basic BLAST.
2. Paste the sequence in the query box.
3. Enter the name of the organism of interest in the "Organism" box. Click the BLAST button.
4. Click on the desired sequence from the results.
5. Continue at step 2 under "a nucleotide or protein accession number" above.
Search Gene database

Gene ID

Contains basic description, gene location in chromosome coordinates
Mouse-over a SNP. Read it.
This one is “missense” Asp → Glu mutation.
### GeneView in dbSNP

<table>
<thead>
<tr>
<th>ID</th>
<th>rs</th>
<th>NAME</th>
<th>Minor Allele</th>
<th>MA</th>
<th>Annotation</th>
<th>Reference</th>
<th>Genomic Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>1414302</td>
<td>rs184331680</td>
<td>0.000</td>
<td></td>
<td></td>
<td>missense</td>
<td>A</td>
<td>Ile [I] 3 1</td>
</tr>
<tr>
<td>1414333</td>
<td>rs753695638</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>Met [M] 3 1</td>
</tr>
<tr>
<td>1414348</td>
<td>rs28909988</td>
<td>0.013</td>
<td></td>
<td></td>
<td>missense</td>
<td>T</td>
<td>Thr [T] 1 3</td>
</tr>
<tr>
<td>1414349</td>
<td>rs572288439</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>Ala [A] 1 3</td>
</tr>
<tr>
<td>1414352</td>
<td>rs547926253</td>
<td>0.006</td>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>Ala [A] 2 3</td>
</tr>
<tr>
<td>1414367</td>
<td>rs113120652</td>
<td>0.004</td>
<td></td>
<td></td>
<td>missense</td>
<td>C</td>
<td>Ala [A] 2 3</td>
</tr>
<tr>
<td>1414375</td>
<td>rs758529964</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>Ala [A] 3 3</td>
</tr>
<tr>
<td>1414378</td>
<td>rs758646981</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>Ala [A] 2 3</td>
</tr>
<tr>
<td>1414380</td>
<td>rs780565025</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>Ala [A] 2 3</td>
</tr>
<tr>
<td>1414382</td>
<td>rs747300142</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>Ala [A] 3 3</td>
</tr>
<tr>
<td>1414383</td>
<td>rs771399217</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>Ala [A] 3 3</td>
</tr>
<tr>
<td>1414396</td>
<td>rs755397803</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>Ala [A] 3 3</td>
</tr>
<tr>
<td>1414397</td>
<td>rs777940439</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>Val [V] 2 3</td>
</tr>
<tr>
<td>1414398</td>
<td>rs748390624</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>Val [V] 2 3</td>
</tr>
</tbody>
</table>

---

Go back to GenBank entry. On the right, find GeneView in the menu.
### Same SNP, in GeneView

<table>
<thead>
<tr>
<th>contig</th>
<th>reference</th>
<th>rs</th>
<th>SNP</th>
<th>MAF</th>
<th>effect</th>
<th>codon change</th>
<th>protein change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1496109</td>
<td>2218</td>
<td>rs199561216</td>
<td>0.000</td>
<td>0.0002</td>
<td>Yes</td>
<td>synonymous</td>
<td>synonym T, Phe [F]</td>
</tr>
<tr>
<td>1496115</td>
<td>2224</td>
<td>rs377151900</td>
<td>0.000</td>
<td></td>
<td></td>
<td>synonymous</td>
<td>synonym T, Val [V]</td>
</tr>
<tr>
<td>1496117</td>
<td>2236</td>
<td>rs732608</td>
<td>0.493</td>
<td>0.4389</td>
<td>Yes</td>
<td>synonymous</td>
<td>synonym A, Pro [P]</td>
</tr>
<tr>
<td>1496121</td>
<td>2237</td>
<td>rs138891531</td>
<td>0.001</td>
<td></td>
<td></td>
<td>missense</td>
<td>missense A, Lys [K]</td>
</tr>
<tr>
<td>1496132</td>
<td>2242</td>
<td>rs775227003</td>
<td>0.000</td>
<td></td>
<td></td>
<td>missense</td>
<td>missense G, Glu [E]</td>
</tr>
<tr>
<td>1496134</td>
<td>2243</td>
<td>rs199431413</td>
<td>0.000</td>
<td>0.0002</td>
<td>Yes</td>
<td>missense</td>
<td>missense G, Val [V]</td>
</tr>
<tr>
<td>1496137</td>
<td>2246</td>
<td>rs201782331</td>
<td>0.000</td>
<td>0.0002</td>
<td>Yes</td>
<td>missense</td>
<td>missense A, Lys [K]</td>
</tr>
<tr>
<td>1496141</td>
<td>2253</td>
<td>rs761527805</td>
<td>0.000</td>
<td></td>
<td></td>
<td>missense</td>
<td>missense A, Tyr [Y]</td>
</tr>
<tr>
<td>1496152</td>
<td>2261</td>
<td>rs764757383</td>
<td>0.000</td>
<td></td>
<td></td>
<td>missense</td>
<td>missense A, Ile [I]</td>
</tr>
</tbody>
</table>

**Back to SNP summary page**

**Back to the chromosome browser**
Population descriptors:
YRI: Yoruba in Ibadan, Nigeria, JPT: Japanese in Tokyo, Japan, CHB: Han Chinese in Beijing, China, CEU: CEPH (Utah residents with ancestry from northern and western Europe)
Linkage Disequilibrium

Triangle plot shows LD values using $r^2$ or D'/LOD scores in one or more HapMap populations.

Phased haplotype track shows all 120 chromosomes with alleles colored yellow and blue.
Tagging SNPs, tSNPs

- SNPs that are highly correlated are redundant information

- **tSNPs** are selected as the minimal non-redundant set of SNPs in a population, such that the genotypes can be reconstructed from the tSNPs.

- tSNPs allow genotyping with fewer steps
  - PCR amplification experiments determine which base is present.

- **Block based tagging**

Block based tagging requires that haplotype "blocks" first be inferred. In the majority of cases when you are investigating association within a candidate gene you are likely to start off with a large number of potential SNPs to choose from, and using various measures of linkage disequilibrium and inferred haplotypes it is possible to define 'haplotype blocks' of markers that are in strong LD with each other, but not with those in other blocks. The exact definition of a haplotype block is open to interpretation, and there are a number of different methods for choosing your haplotype blocks (Gabriel *et al* 2002, )
Tagging SNPs

tSNPs (tag SNPs, Tagging SNPs) are SNPs that correlate strongly with other SNP, therefore they give information on the haplotype without having to sequence the whole genome.

Highly correlated SNP have high LD. SVD finds repeating, nearly identical rows in the haplotype matrix.
Uses of SNPs

• Personalized medicine
  – SNP-associated Sensitivity to
    – diseases
    – drugs
    – chemicals
    – pathogens
    – vaccines

• Livestock breeding

• Human migrations
Human migration out of Africa

Each individual has a set of SNPs and a family history, marking geographic location. Tree is built ignoring the geographic location, to model migration paths.

Two Uniparental lineages

**mtDNA**: mitochondrial DNA
**NRY**: nonrecombining region of the Y chromosome