Disease gene prioritization

Anil Jegga
Biomedical Informatics
http://anil.cchmc.org

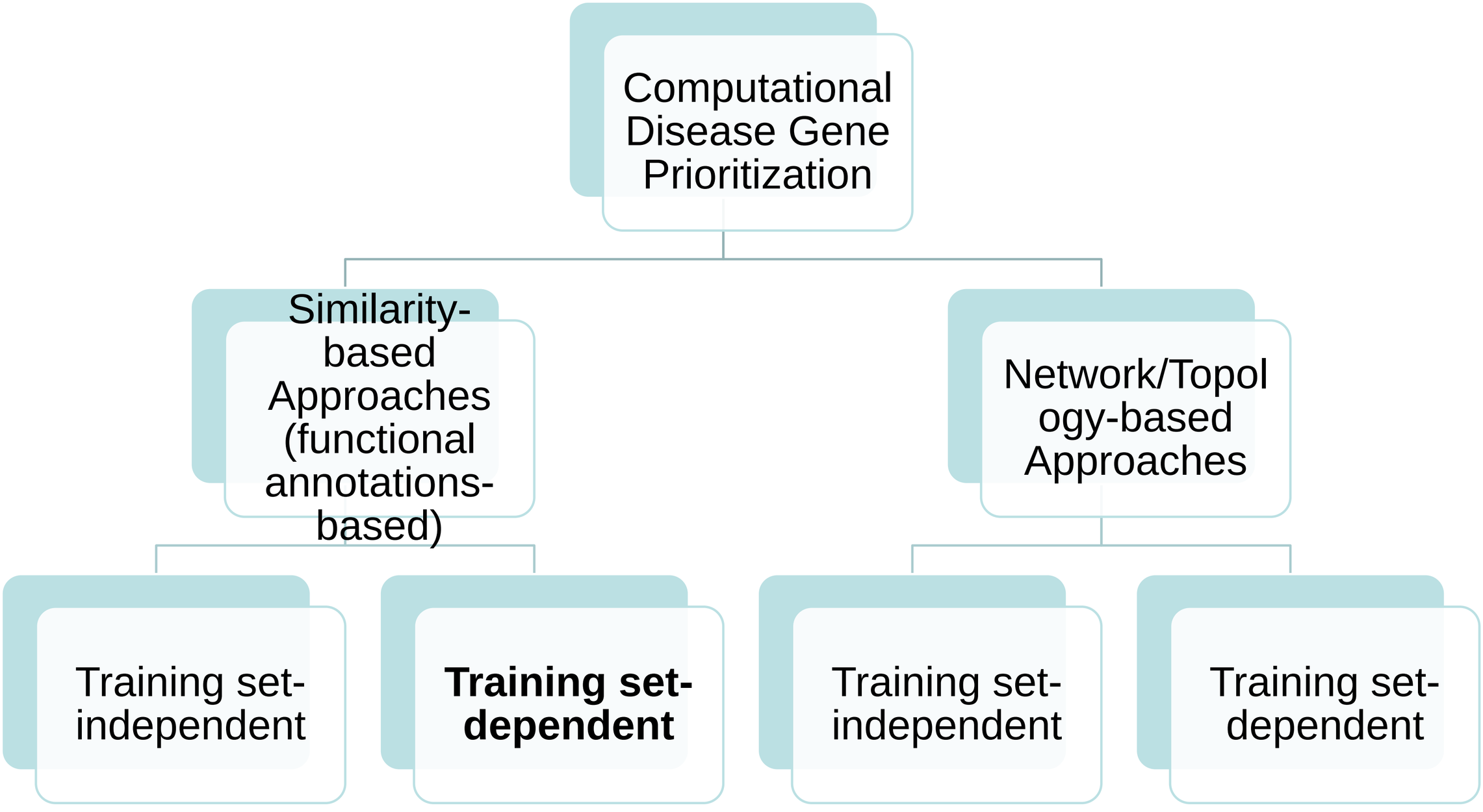
GLBIO-2014

What, Why, & How

Computational Disease Gene prioritization

- **What**: Computationally assigning likelihood of gene involvement in generating a disease phenotype
- **Why**: Narrows down the set of genes to be tested experimentally – saves time/resources.
- **How**: “Guilt by Association” - Gene “priority" in disease is assigned in a more “informed” way taking into account a set of relevant features or annotations (e.g., gene expression, function/processes, pathways, model organism phenotype, etc.) - Functional Similarity-based methods

Broad Classification



- Protein-Protein Interactions
- Protein Associations (Functional Linkage)

Functional annotation-based candidate disease gene prioritization

- *Guilt by association* - Reliable predictions about the disease involvement of a gene can be made if several of its partners (e.g., genes with correlated expression profiles or protein interactants or genes involved in same biological process or pathway) share a corresponding annotation.
- Incorporating the prior information or knowledge about a disease (e.g., known disease genes) is critical.
- Challenge: Gather, normalize, and integrate heterogeneous data from multiple sources (and keeping them current).

Functional annotation-based candidate disease gene prioritization – General workflow

- Step 1: List of candidate genes (**Test Set**) to prioritize - linkage regions, chromosomal aberrations, association study loci, differentially expressed gene lists or genes identified by sequencing variants, or the complete genome
- Step 2: **Seed** Genes or **Training Set**: Prior knowledge about the disease - known disease genes, or disease-relevant keywords, or biological processes or pathways.
- Step 3: **Prioritization** methods: Which one to select/use?
- Step 4: **Assessment** - Are the selected training/seed genes, keywords and tools suitable? Can reliable predictions be made using these?
- Step 5: Use multiple tools or multiple sets of seed gene or keywords - **Combine** the results to obtain a **consensus** result

What constitutes a “good” seed gene set?

- **Relevancy**: Review each gene - Domain experts especially for selecting keywords (e.g., disease-relevant phenotypes)

- **Size Matters**: Neither too small nor too large.
  - Too small - may be insufficiently informative
  - Too large - too heterogeneous pattern to be useful.
    - Break them down into multiple random sets
    - Filter them based on additional features (e.g., genes associated with a BP term + MP term)
  - Ideally 6 – 30

What constitutes a "good" seed gene set?

- **Robustness**: How robust are the ranking results using a particular seed set?

Cross-validation - Assess whether a set of seed genes provides a coherent pattern

- Create multiple sets of seed genes or keywords covering complementary phenotypic aspects of the disease and assess their performance separately.
- Negative control seed genes: Use genes for other unrelated diseases as training set.
  - Top-ranking candidates are same with negative control seed genes – suggests some systematic bias and prioritization results are probably unreliable

Other Quality Control Measures

Tool A ranks my "favorite" gene on/among top – Therefore tool A is the BEST!!!

- **Smaller Test Sets**: Perform prioritizations both on the actual set of candidates and on the whole genome *OR* on a larger set that includes the smaller set of candidates.
  - Are the top-ranking candidates from the small subset rank within the top 5–15% of the whole genome?
  - If not, the prioritization might not have been able to capture enough information to identify good candidates
- **Functional Coherence**: What are the enriched terms for the top ranked candidates? Do they match expectations for the biological process or phenotype of interest?

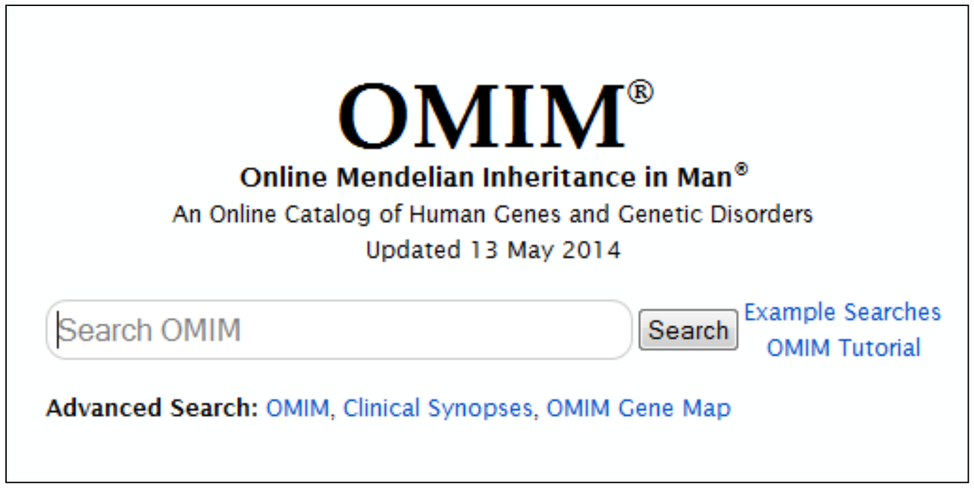
Moreau & Tranchevent, 2012

Cincinnati Children's

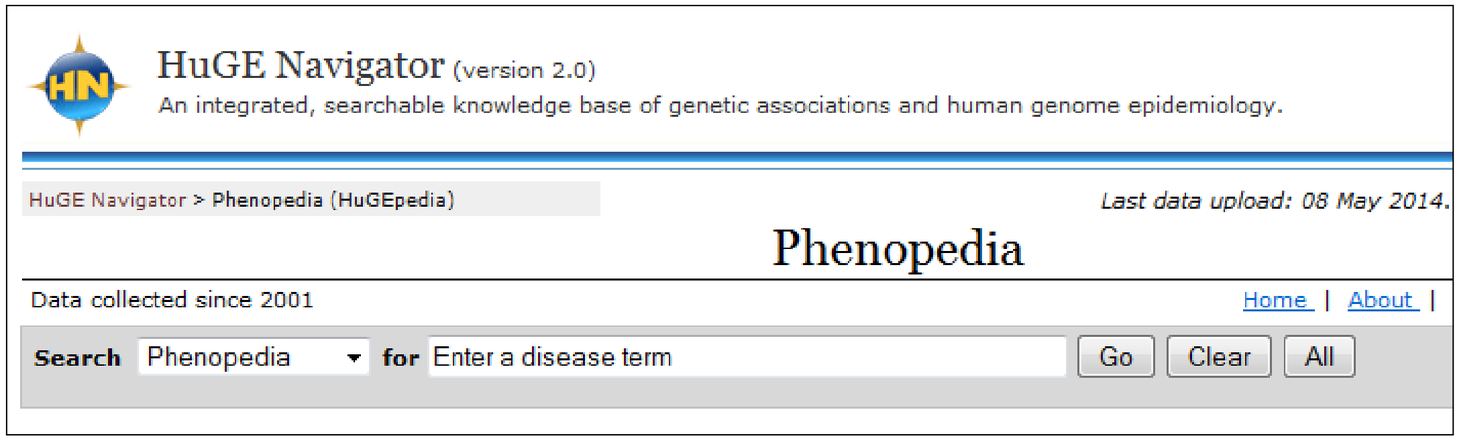
UNIVERSITY OF Cincinnati

Resources commonly used for compiling seed set

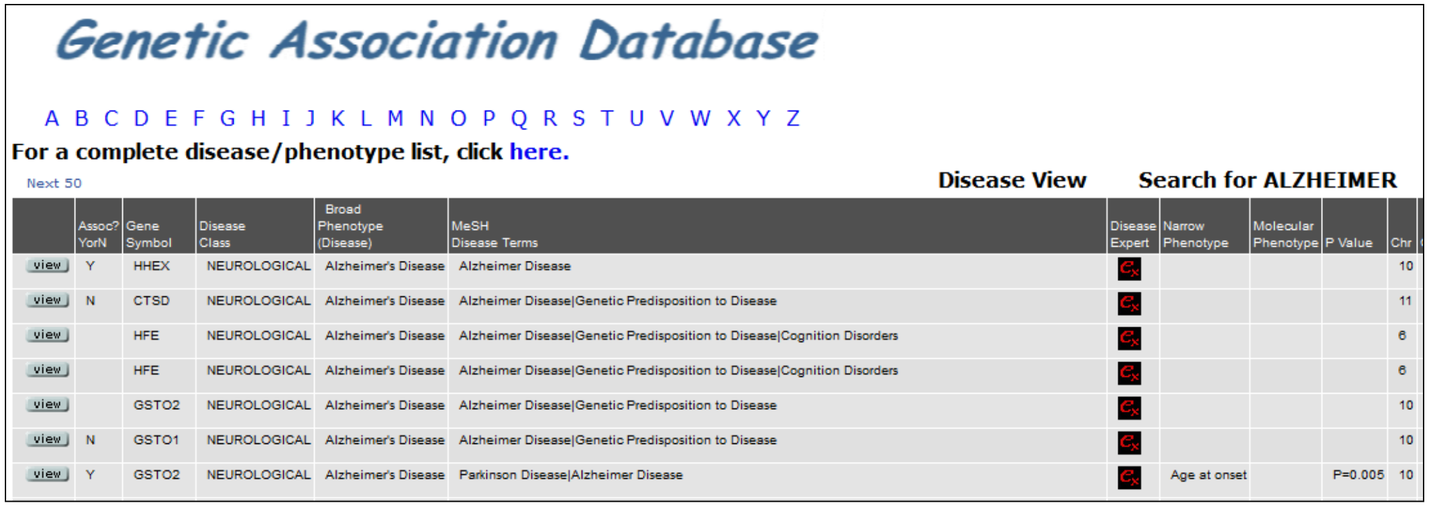
OMIM: http://omim.org



Phenopedia: http://hugenavigator.net

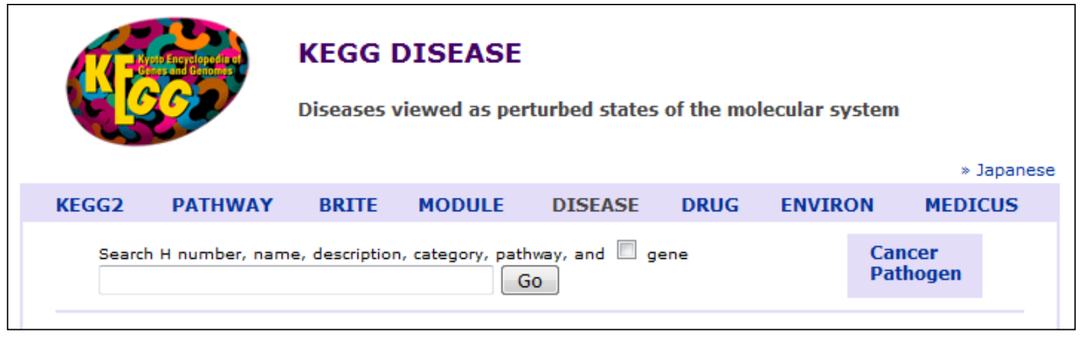


GAD: http://geneticassociationdb.nih.gov



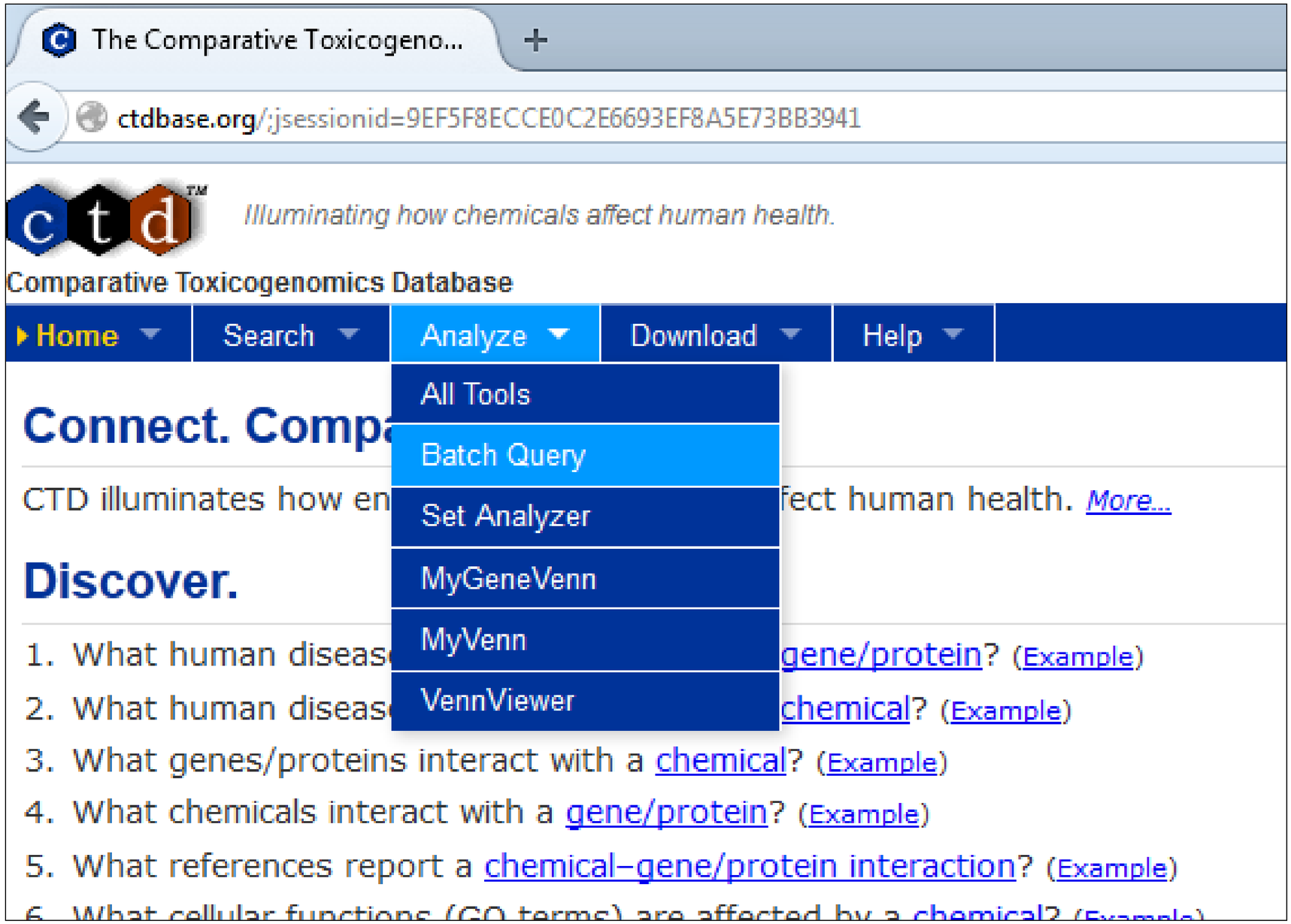
	Assoc? YorN	Gene Symbol	Disease Class	Broad Phenotype (Disease)	MeSH Disease Terms	Disease Expert	Narrow Phenotype	Molecular Phenotype	P Value	Chr
view	Y	HHEX	NEUROLOGICAL	Alzheimer's Disease	Alzheimer Disease					10
view	N	CTSD	NEUROLOGICAL	Alzheimer's Disease	Alzheimer Disease|Genetic Predisposition to Disease					11
view		HFE	NEUROLOGICAL	Alzheimer's Disease	Alzheimer Disease|Genetic Predisposition to Disease|Cognition Disorders					6
view		HFE	NEUROLOGICAL	Alzheimer's Disease	Alzheimer Disease|Genetic Predisposition to Disease|Cognition Disorders					6
view		GSTO2	NEUROLOGICAL	Alzheimer's Disease	Alzheimer Disease|Genetic Predisposition to Disease					10
view	N	GSTO1	NEUROLOGICAL	Alzheimer's Disease	Alzheimer Disease|Genetic Predisposition to Disease					10
view	Y	GSTO2	NEUROLOGICAL	Alzheimer's Disease	Parkinson Disease|Alzheimer Disease		Age at onset		P=0.005	10

KEGG Disease: http://www.genome.jp/kegg/disease/



Additional resources for compiling seed set

Comparative Toxicogenomics Database: http://ctdbase.org



Cincinnati Children's

UNIVERSITY OF Cincinnati

Comparative Toxicogenomics Database: http://ctdbase.org

Batch Query

1 Select your input type

- Chemicals (MeSH® names, synonyms, or IDs, or CAS RNs)
- Diseases (MeSH or OMIM names, synonyms, or IDs)
- Genes (NCBI symbols or IDs)
- Gene Ontology terms (GO names, synonyms, or IDs)
- Pathways (KEGG or REACTOME names or IDs)
- References (PubMed® IDs or DOIs)

2 Provide query terms (up to 4,000)

Return-, tab- or |-delimited

Or **upload** a tab-separated file: Browse... No file selected.

Identifiers column: 1

3 Choose data to download

Data

- Chemical–gene interactions
- Chemical associations
- Gene associations: Curated
- Disease associations: Inferred
- Pathway associations
- Gene Ontology associations

Format

- TSV (tab-separated values)
- CSV (comma-separated values)
- JSON
- XML

Download Clear

1 Select your input type

- Chemicals (MeSH® names, synonyms, or IDs, or CAS RNs)
- Diseases (MeSH or OMIM names, synonyms, or IDs)
- Genes (NCBI symbols or IDs)
- Gene Ontology terms (GO names, synonyms, or IDs)
- Pathways (KEGG or REACTOME names or IDs)
- References (PubMed® IDs or DOIs)

2 Provide query terms (up to 4,000)

Return-, tab- or |-delimited

Or **upload** a tab-separated file: Browse... No file selected.

Identifiers column: 1

3 Choose data to download

Data

- Chemical–gene interactions
- Chemical associations: All / Curated / Inferred
- Gene associations: All / Curated / Inferred
- Disease associations
- Pathway associations: Inferred
- Gene Ontology associations

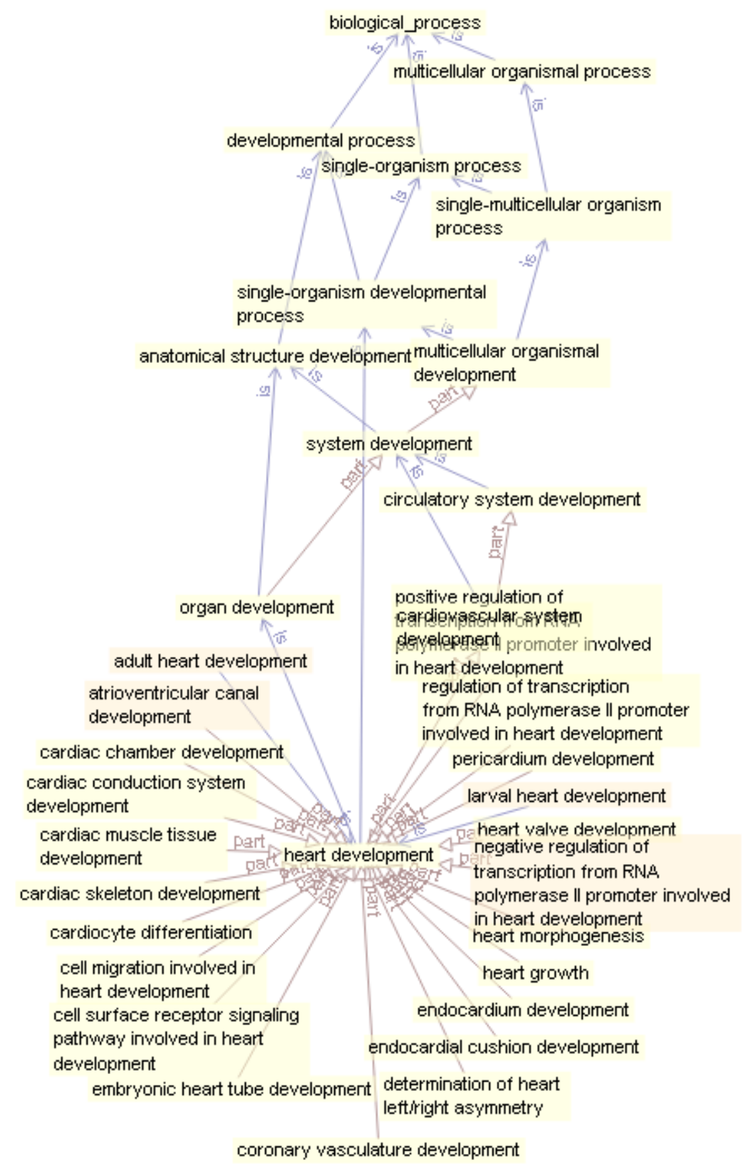
Format

- TSV (tab-separated values)
- CSV (comma-separated values)
- JSON
- XML

Download Clear

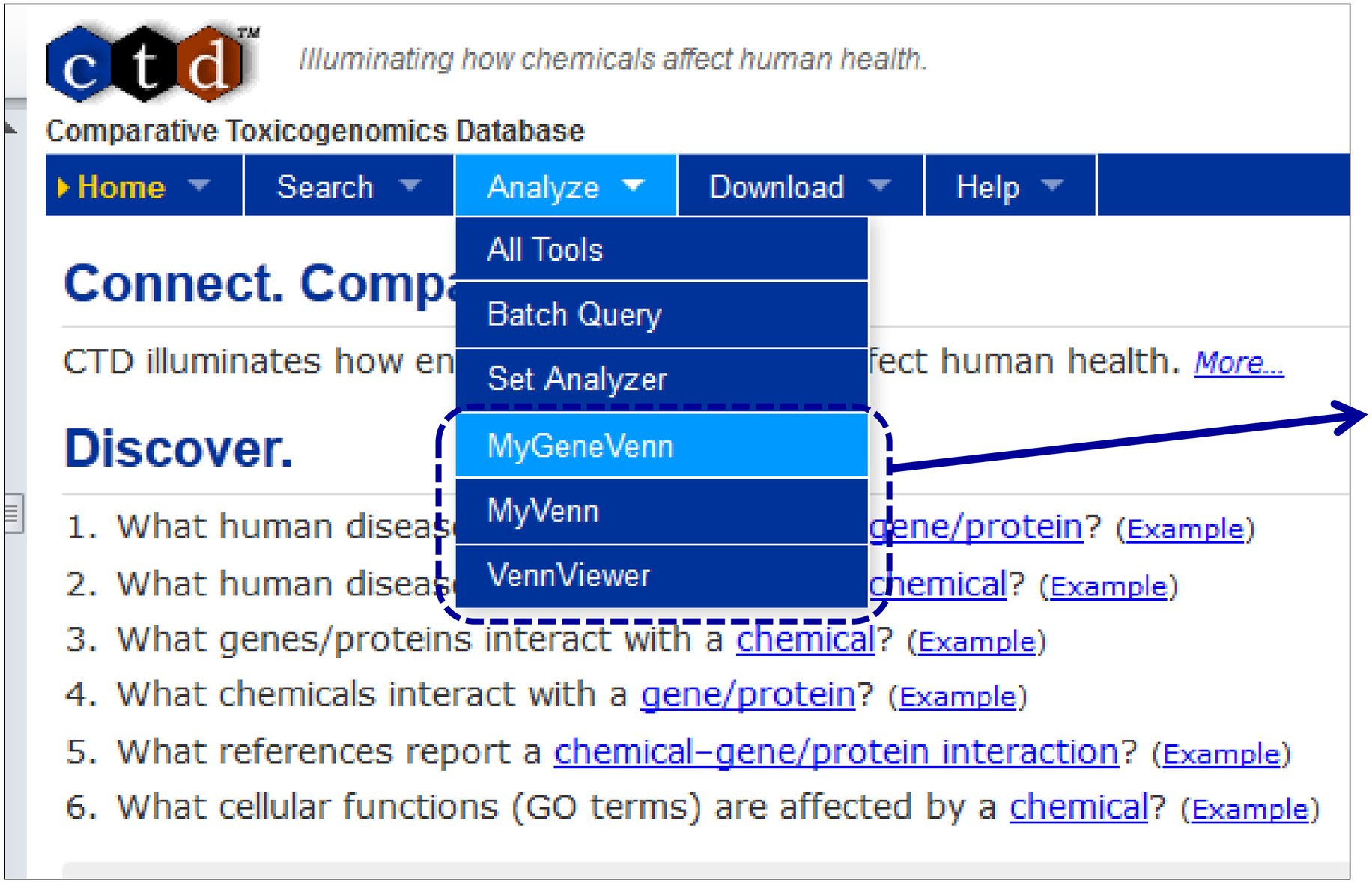
Comparative Toxicogenomics Database: http://ctdbase.org

# Input	GeneSymbol	GeneID	Ontology	GoTermName	GoTermID
go:0007507	TBX20	57057	Biological Process	aortic valve development	GO:0003176
go:0007507	EFNA1	1942	Biological Process	aortic valve morphogenesis	GO:0003180
go:0007507	GATA3	2625	Biological Process	aortic valve morphogenesis	GO:0003180
go:0007507	NOTCH1	4851	Biological Process	aortic valve morphogenesis	GO:0003180
go:0007507	TBX20	57057	Biological Process	aortic valve morphogenesis	GO:0003180
go:0007507	TWIST1	7291	Biological Process	aortic valve morphogenesis	GO:0003180
go:0007507	FHL2	2274	Biological Process	atrial cardiac muscle cell development	GO:0055014
go:0007507	MYH15	22989	Biological Process	atrial cardiac muscle cell development	GO:0055014
go:0007507	NKX2-5	1482	Biological Process	atrial cardiac muscle cell development	GO:0055014
go:0007507	NKX2-6	137814	Biological Process	atrial cardiac muscle cell development	GO:0055014
go:0007507	ISL1	3670	Biological Process	atrial cardiac muscle cell differentiation	GO:0055011
go:0007507	RBM24A	405801	Biological Process	atrial cardiac muscle tissue development	GO:0003228
go:0007507	RBM24B	562236	Biological Process	atrial cardiac muscle tissue development	GO:0003228
go:0007507	BMP10	27302	Biological Process	atrial cardiac muscle tissue morphogenesis	GO:0055009
go:0007507	MYH6	4624	Biological Process	atrial cardiac muscle tissue morphogenesis	GO:0055009
go:0007507	PITX2	5308	Biological Process	atrial cardiac muscle tissue morphogenesis	GO:0055009
go:0007507	PROX1	5629	Biological Process	atrial cardiac muscle tissue morphogenesis	GO:0055009
go:0007507	TNNT2	7139	Biological Process	atrial cardiac muscle tissue morphogenesis	GO:0055009
go:0007507	TNNT2A	58071	Biological Process	atrial cardiac muscle tissue morphogenesis	GO:0055009
go:0007507	WNT2	7472	Biological Process	atrial cardiac muscle tissue morphogenesis	GO:0055009
go:0007507	MYH6	4624	Biological Process	atrial cardiac myofibril assembly	GO:0055004
go:0007507	MYL7	58498	Biological Process	atrial cardiac myofibril assembly	GO:0055004
go:0007507	ANK2	287	Biological Process	atrial septum development	GO:0003283
go:0007507	DAND5	199699	Biological Process	atrial septum development	GO:0003283
go:0007507	GJA5	2702	Biological Process	atrial septum development	GO:0003283
go:0007507	NPHP3	27031	Biological Process	atrial septum development	GO:0003283
go:0007507	TBX5	6910	Biological Process	atrial septum development	GO:0003283
go:0007507	CFC1	55997	Biological Process	atrial septum morphogenesis	GO:0060413
go:0007507	CFC1B	653275	Biological Process	atrial septum morphogenesis	GO:0060413
go:0007507	CYR61	3491	Biological Process	atrial septum morphogenesis	GO:0060413
go:0007507	GATA4	2626	Biological Process	atrial septum morphogenesis	GO:0060413
go:0007507	GJA5	2702	Biological Process	atrial septum morphogenesis	GO:0060413
go:0007507	HEY2	23493	Biological Process	atrial septum morphogenesis	GO:0060413
go:0007507	ISL1	3670	Biological Process	atrial septum morphogenesis	GO:0060413
go:0007507	NKX2-5	1482	Biological Process	atrial septum morphogenesis	GO:0060413
go:0007507	NOTCH2	4853	Biological Process	atrial septum morphogenesis	GO:0060413
go:0007507	SMO	6608	Biological Process	atrial septum morphogenesis	GO:0060413
go:0007507	TBX20	57057	Biological Process	atrial septum morphogenesis	GO:0060413
go:0007507	TBX5	6910	Biological Process	atrial septum morphogenesis	GO:0060413
go:0007507	ZFPM1	161882	Biological Process	atrial septum morphogenesis	GO:0060413
go:0007507	ACVR1	90	Biological Process	atrial septum primum morphogenesis	GO:0003289
go:0007507	GATA4	2626	Biological Process	atrial septum primum morphogenesis	GO:0003289
go:0007507	SOX4	6659	Biological Process	atrial septum primum morphogenesis	GO:0003289
go:0007507	WHSC1	7468	Biological Process	atrial septum primum morphogenesis	GO:0003289
go:0007507	GATA4	2626	Biological Process	atrial septum secundum morphogenesis	GO:0003290
go:0007507	WHSC1	7468	Biological Process	atrial septum secundum morphogenesis	GO:0003290
go:0007507	GJA1	2697	Biological Process	atrial ventricular junction remodeling	GO:0003294
go:0007507	GJA5	2702	Biological Process	atrial ventricular junction remodeling	GO:0003294
go:0007507	PDLIM7	9260	Biological Process	atrial ventricular junction remodeling	GO:0003294
go:0007507	TBX5A	30071	Biological Process	atrial ventricular junction remodeling	GO:0003294
go:0007507	TBX3	6926	Biological Process	atrioventricular bundle cell differentiation	GO:0003167
go:0007507	ACTC1A	408256	Biological Process	atrioventricular canal development	GO:0036302
go:0007507	FOXN4	121643	Biological Process	atrioventricular canal development	GO:0036302
go:0007507	GATA4	2626	Biological Process	atrioventricular canal development	GO:0036302
go:0007507	HAS2	3037	Biological Process	atrioventricular canal development	GO:0036302
go:0007507	MMD	23531	Biological Process	atrioventricular canal development	GO:0036302
go:0007507	PTPN11	5781	Biological Process	atrioventricular canal development	GO:0036302



Ontological tree – Children nodes and their annotations also used

Comparative Toxicogenomics Database: http://ctdbase.org



Explore the Venn utilities – Handy for generating/comparing annotated gene lists (seed set selection)

Cincinnati Children's

UNIVERSITY OF Cincinnati

NCBI MedGen - http://www.ncbi.nlm.nih.gov/medgen

NCBI Resources How To Sign in to NCBI

MedGen MedGen Search

Limits Advanced Help

MedGen MedGen Search

Advanced Help

Limits

Search Field Tags

Field: All Fields

Chromosomes clear

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT Unknown

Show only clear

- Found in *GeneReviews*
- With genetic testing
- Associated with a gene

Record Category clear

- Disease Focus
- Clinical Features

Vocabulary clear

GTR HPO MeSH NCI OMIM

Clear Search

NCBI's portal to information related to Medical Genetics. Terms from the NIH Genetic Testing Registry (GTR), UMLS, HPO, ClinVar and other sources are aggregated into concepts and their gene annotations where available.

Cincinnati Children's

UNIVERSITY OF Cincinnati

NCBI MedGen - http://www.ncbi.nlm.nih.gov/medgen

Alzheimer's disease (AD)

MedGen UID: 1853 • Concept ID: C0002395 • Disease or Syndrome

Synonyms: AD; Alzheimer dementia (AD); ALZHEIMER DISEASE, FAMILIAL, 1, AUTOSOMAL RECESSIVE; Alzheimer scl
Alzheimer syndrome; Alzheimer"s disease; Alzheimer-type dementia (ATD); dementia; Dementia; Early-Onset F
Alzheimer Disease; familial Alzheimer disease (FAD); Presenile and senile dementia; Primary Senile Degenerati
Dementia

Modes of inheritance: Heterogeneous
Autosomal dominant inheritance

SNOMED CT: AD - Alzheimer's disease (26929004); Alzheimer disease (26929004); Alzheimers disease (26929004); Alzheime
disease (26929004)

Genes: PAXIP1; SORL1; PLAU; NOS3; MPO; HFE; ACE; BLMH; APP; APBB2; A2M

Cytogenetic locations: 10q22.2; 11q24.1; 12p13.31; 17q11.2; 17q22; 17q23.3; 21q21.3; 4p14-13; 6p22.2; 7q36.1; 7q36.2

OMIM: 104300

Genetic Testing Registry

- Deletion/duplication analysis (2)
- Mutation scanning of select exons (1)
- Research (1)
- Sequence analysis of the entire coding region (14)
- Targeted variant analysis (4
- See all (20)

- Disorder of nervous system
  - Disorder of the central nervous system
    - Disorder of brain
      - Delirium, Dementia, Amnestic, Cognitive Disorders
        - Cognitive disorder
          - Dementia
            - Alzheimer's disease
              - Alzheimer disease, type 6
              - Alzheimer disease, type 7
              - Alzheimer disease, type 8
              - Alzheimer disease, type 9
              - Alzheimer disease, type 10
              - Alzheimer disease, type 11
              - Alzheimer disease, type 12
              - Alzheimer disease, type 13
              - Alzheimer disease, type 14
              - Alzheimer disease, type 15
              - Alzheimer disease, type 16
            - Early-Onset Familial Alzheimer Disease
            - Late-Onset Familial Alzheimer Disease

Table of contents

- Disease characteristics
- Additional descriptions
- Clinical features
- Term Hierarchy
- Professional guidelines
- Recent clinical studies
- Recent systematic reviews

Molecular resources

- OMIM
- RefSeqGene
- View A2M variations in ClinVar
- View APBB2 variations in ClinVar
- View APP variations in ClinVar
- View BLMH variations in ClinVar
- View ACE variations in ClinVar
- View HFE variations in ClinVar
- View MPO variations in ClinVar
- View NOS3 variations in ClinVar
- View PLAU variations in ClinVar
- View SORL1 variations in ClinVar
- View PAXIP1 variations in ClinVar

NCBI BioSystems - http://www.ncbi.nlm.nih.gov/biosystems

NCBI Resources How To

BioSystems | BioSystems | Search

Limits Advanced

BioSystems | BioSystems | Search

Advanced

Limits

Taxonomic Scope

- Conserved BioSystems
- Organism Specific BioSystems

BioSystem Type

- Pathway
- Structural Complex
- Functional Set
- Signature Module

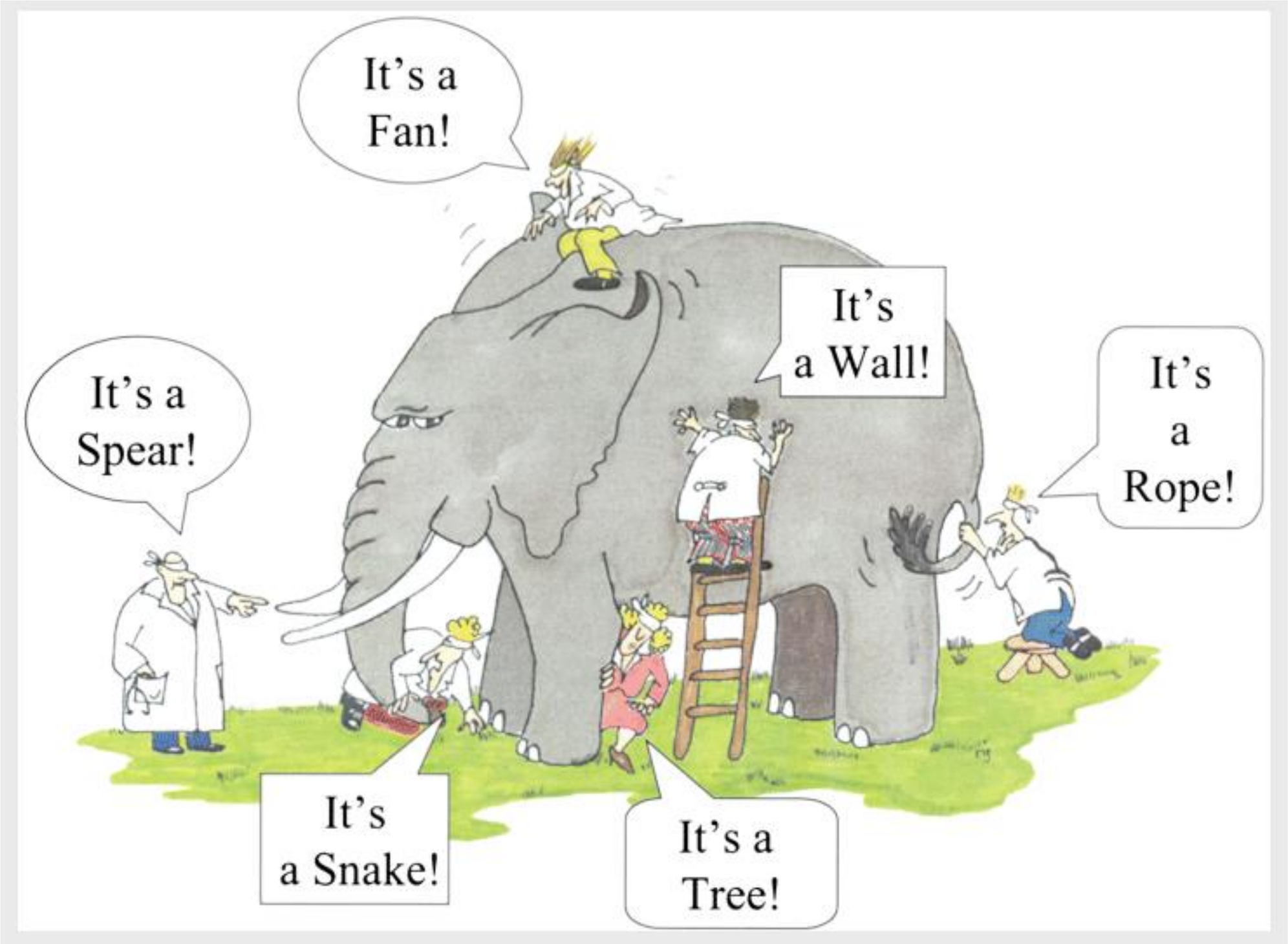
Search Field Tags

- KEGG
- BioCyc & MetaCyc
- Pathway Interaction Database (PID)
- Reactome
- WikiPathways
- Gene Ontology (GO)
- LIPID MAPS

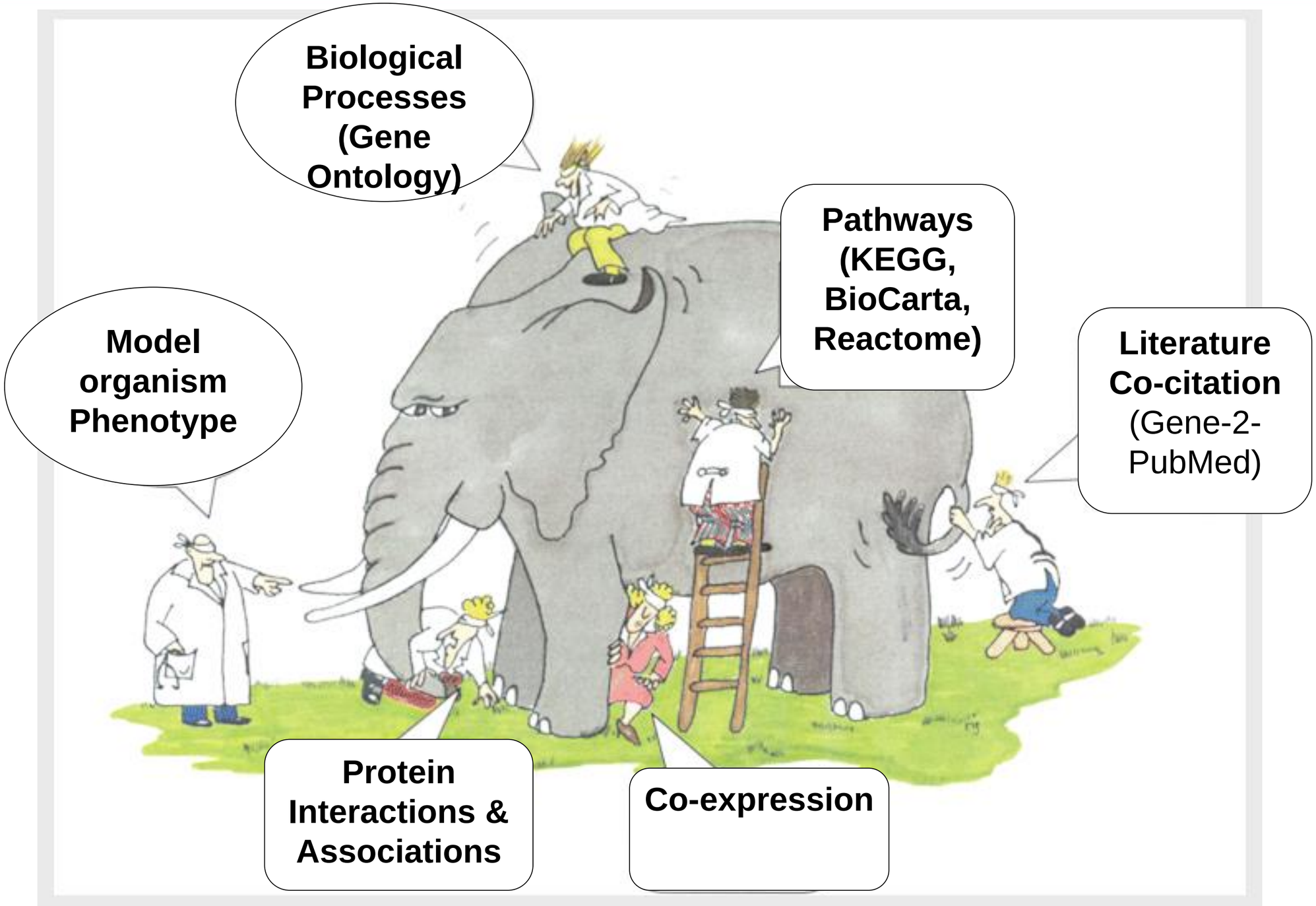
Reset Search

A group of genes that have a pathogenicity or other phenotype associated with them

Functional Similarity – What features to consider?

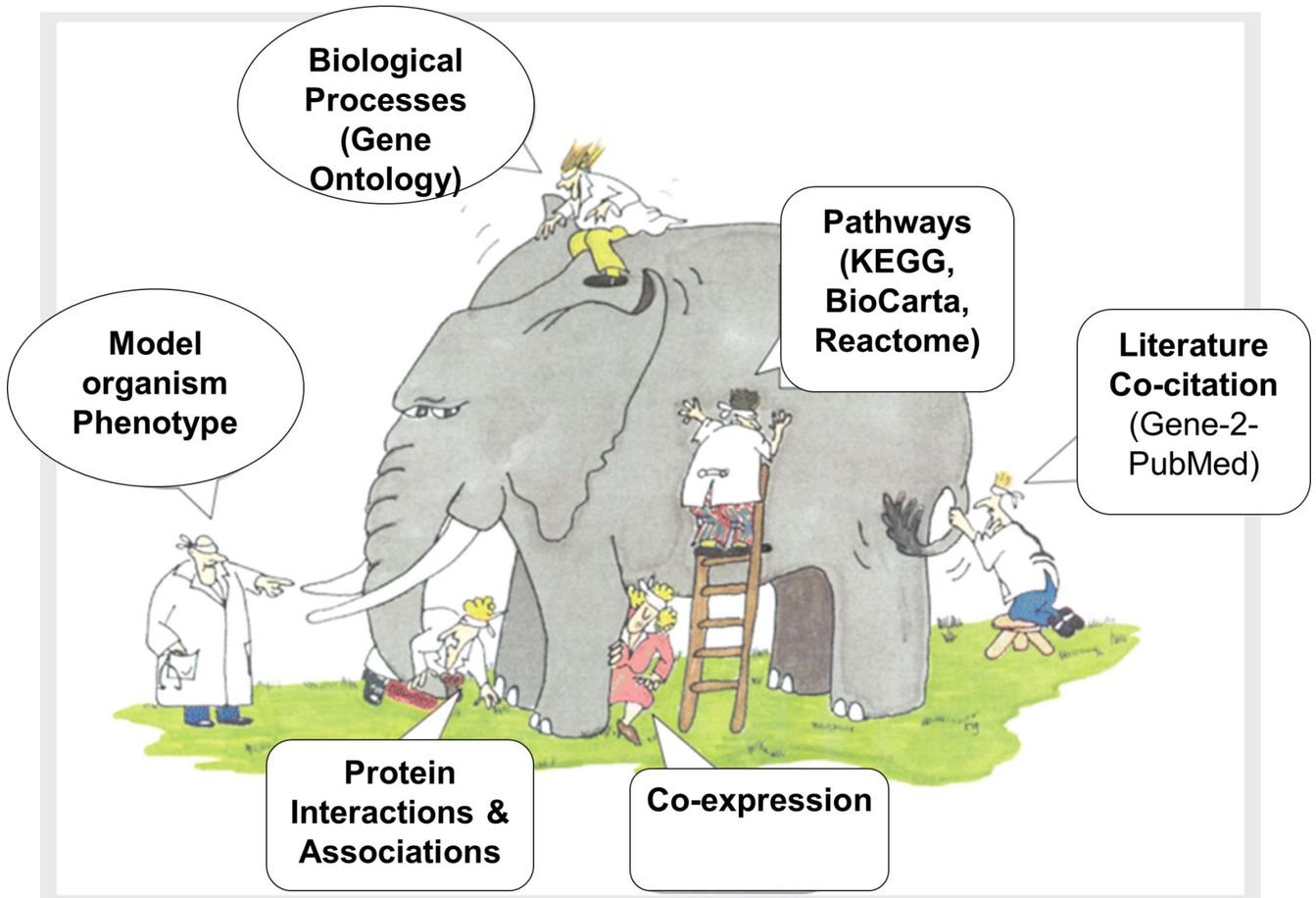


Cartoon: G. Renee Guzlas



- **No single source of data can be expected to capture all relevant relations**
- **Integrate multiple data sources: Better signal-to-noise ratio and improved prediction accuracy**

Cartoon - G. Renee Guzlas
Tranchevent & Moreau



Gene prioritization through **genomic data fusion**
S Aerts, D Lambrechts, S Maity, P Van Loo… - Nature …, 2006 - nature.com
Abstract The identification of genes involved in health and disease remains a challenge. We describe a bioinformatics approach, together with a freely accessible, interactive and flexible software termed Endeavour, to prioritize candidate genes underlying biological processes ...
Cited by 552 Related articles All 41 versions Web of Science: 341 Cite Save

ToppGene Suite for gene list enrichment analysis and candidate gene prioritization
J Chen, EE Bardes, BJ Aronow… - Nucleic acids research, 2009 - Oxford Univ Press
Abstract ToppGene Suite (http://toppgene. cchmc. org; this web site is free and open to all users and does not require a login to access) is a one-stop portal for (i) gene list functional enrichment,(ii) candidate gene prioritization using either functional annotations or network ...
Cited by 231 Related articles All 13 versions Web of Science: 148 Cite Save

- Guilt-by-association: Approaches differ by the strategy adopted in calculating similarity and by the data sources utilized
- With some exceptions (e.g., ENDEAVOUR, ToppGene), most of the existing approaches mainly focus on the combination of only a few data sources
- For methodological details & validation see:
  - Aerts et al., 2006 Nature Biotech.
  - Chen et al., 2007 BMC Bioinfo.

Cincinnati Children's

UNIVERSITY OF Cincinnati

ENDEAVOUR

Select All None

- Annotation - EnsemblEst
- Annotation - GeneOntology
- Annotation - Interpro
- Annotation - Kegg
- Annotation - Swissprot
- Blast
- CisRegModule
- Expression - SonEtAl
- Expression - SuEtAl
- Interaction - Bind
- Interaction - BioGrid
- Interaction - Hprd
- Interaction - InNetDb
- Interaction - Intact
- Interaction - Mint
- Interaction - String
- Motif
- Precalculated - Ouzounis
- Precalculated - Prospectr
- Text

ToppGene

- All
- GO: Molecular Function
- GO: Biological Process
- GO: Cellular Component
- Human Phenotype
- Mouse Phenotype
- Domain
  - Gene3D
  - InterPro
  - PROSITE
  - Pfam
  - ProDom
  - SMART
- Pathway
  - BioSystems: BIOCYC
  - BioSystems: KEGG
  - BioSystems: Pathway Interaction Database
  - BioSystems: REACTOME
  - BioSystems: WikiPathways
  - GenMAPP
  - MSigDB C2: BioCarta
  - MSigDB C2: SigmaAldrich
  - MSigDB C2: Signaling Gateway
  - MSigDB C2: Signaling Transduction KE
  - MSigDB C2: SuperArray
  - PantherDB
  - Pathway Ontology
  - SMPDB
- Pubmed
  - GeneRIF
  - Pubmed
- Interaction
- Cytoband
- Transcription Factor Binding Site
- Gene Family
- Coexpression
  - GeneSigDB
  - MSigDB C2: Aristoteles University of Thessaloniki
  - MSigDB C2: Broad Institute
  - MSigDB C2: Columbia University
  - MSigDB C2: Dana-Farber Cancer Institute
  - MSigDB C2: Giannina Gaslini Institute
  - MSigDB C2: Johns Hopkins University School of Medicine
  - MSigDB C2: Laboratoire CarMeN
  - MSigDB C2: Michigan State University
  - MSigDB C2: Stanford University
  - MSigDB C2: Steinbeis Transfer Center for Proteome Analysis
  - MSigDB C2: Telethon Institute for Child Health Research
  - MSigDB C2: University Pierre and Marie Curie
  - MSigDB C2: University Pierre and Marie Curie
  - MSigDB C2: University of Liverpool
  - MSigDB C2: University of Washington
  - MSigDB C6: Broad Institute
  - MSigDB C7: Nick Haining lab (DFCI)
- Coexpression Atlas
  - Body Map
  - FaceBase_RNAseq
  - FaceBase_ST1
  - Gudmap Mouse MOE430.2
  - Gudmap Mouse ST 1.0
  - Gudmap RNAseq
  - Immgen.org, GSE15907
  - PCBC
  - PCBC_AltAnalyze
- Computational
  - MSigDb: C4 - CGN: Cancer Gene Neighborhood
  - MSigDb: C4 - CM: Cancer Modules
- MicroRNA
  - MSigDB
  - MicroRNA.org
  - PITA
  - PicTar
  - TargetScan
  - miRTarbase
  - miRecords_TarBase
- Drug
  - CTD
  - CTD Marker
  - CTD Therapeutic
  - Drug Bank
  - Stitch
- Disease
  - CTD
  - Clinical Variations
  - GWAS
  - OMIM

ENDEAVOUR

http://homes.esat.kuleuven.be/~bioiuser/endeavour/index.php

Perform gene prioritization using our **Web server**

Perform gene prioritization using our **Java server**

Prioritize your candidates in 4 steps with the following wizard

1. Species | 2. Training genes | 3. Data sources used to build models | 4. Candidates

Species

- *Homo sapiens*
- *Rattus norvegicus*
- *Mus musculus*
- *Drosophila melanogaster*
- *Caenorhabditis elegans*

Supports multiple species

Ready

Prioritize your candidates in 4 steps with the following wizard

1. Species | 2. Training genes | 3. Data sources used to build models | 4. Candidates

Training genes (14 genes)

Gene (Reference ID)	Alias	Description
ENSG00000168646	AXIN2	Axin-2 (Axis inhibition protein 2) (Conductin) (Axin-like protein) (Axil). [Source:Uniprot/SWISSPROT;Acc:Q9Y2T1]
ENSG00000109472	CPE	Carboxypeptidase E precursor (EC 3.4.17.10) (CPE) (Carboxypeptidase H) (CPH) (Enkephalin convertase) (Prohormone-processing carboxypeptidase). [Source:Uniprot/SWISSPROT;Acc:P16870]
ENSG00000144891	AGTR1	Type-1 angiotensin II receptor (AT1) (AT1AR) (AT1BR). [Source:Uniprot/SWISSPROT;Acc:P30556]
ENSG00000168036	CTNNB1	Catenin beta-1 (Beta-catenin). [Source:Uniprot/SWISSPROT;Acc:P35222]
ENSG00000133895	MEN1	Menin. [Source:Uniprot/SWISSPROT;Acc:O00255]
ENSG00000159224	GIP	Gastric inhibitory polypeptide precursor (GIP) (Glucose-dependent insulinotropic polypeptide). [Source:Uniprot/SWISSPROT;Acc:P09681]

Add following genes

```
AXIN2
CPE
AGTR1
CTNNB1
MEN1
GIP
WISP2
LHCGR
```

Add

Little picky on the input types

Cincinnati Children's

UNIVERSITY OF Cincinnati

1. Species | 2. Training genes | 3. Data sources used to build models | 4. Candidates

Select All None

- [x] Annotation - EnsemblEst
- [x] Annotation - GeneOntology
- [x] Annotation - Interpro
- [x] Annotation - Kegg
- [x] Annotation - Swissprot
- [x] Blast
- [x] CisRegModule
- [x] Expression - SonEtAl
- [x] Expression - SuEtAl
- [x] Interaction - Bind
- [x] Interaction - BioGrid
- [x] Interaction - Hprd
- [x] Interaction - InNetDb
- [x] Interaction - Intact
- [x] Interaction - Mint
- [x] Interaction - String
- [x] Motif
- [x] Precalculated - Ouzounis
- [x] Precalculated - Prospectr
- [x] Text

1. Species | 2. Training genes | 3. Data sources used to build models | 4. Candidates

Candidate genes to prioritize (14 genes)

	Gene (Reference ID)	Alias	
	ENSG00000151617	EDNRA	Endothelin-1 receptor precursor (Endothelin A receptor) (
	ENSG00000183625	CCR3	C-C chemokine receptor type 3 (C-C CKR-3) (CC-CKR-3) (
	ENSG00000118523	CTGF	Connective tissue growth factor precursor (Hypertrophic c
	ENSG00000163638	ADAMTS9	ADAMTS-9 precursor (EC 3.4.24.-) (A disintegrin and met
	ENSG00000153208	MERTK	Proto-oncogene tyrosine-protein kinase MER precursor (E
	ENSG00000173598	NUDT4	Diphosphoinositol polyphosphate phosphohydrolase 2 (EC

Add following candidates

```
EDNRA
CCR3
CTGF
ADAMTS9
MERTK
NUDT4
IGF2
GPR125
F7
```

Add

Full genome

Cincinnati Children's

UNIVERSITY OF Cincinnati

Console (warnings and errors) clear

```
12:26:10 Updated training genes.
12:35:21 ⚠ no gene for CCNT
12:35:21 Updated candidate genes.
```

« Previous | Launch prioritization

- Little picky on the input types - e.g., gene symbols have to be HGNC approved.
- Supports chromosomal regions, e.g. chr:8p or chr:20p13 – will fetch all the genes in that region
- Doesn't support chromosomal coordinates

Results | Sprint plot | Table | Export | Discard results

Load previously saved results (exported as XML) Browse... No file selected. Load

Download the rankings table

What shared features make a test set gene rank at top <u>or</u> What features are shared between the training/seed set and test set gene are not explicit.

Legend

Note: The top 16 genes are attributed a color in order to easily find the rank of a given gene obtained for each model (selected data source).

ELN Candidates ranking in the top 16 are attributed a random background color.

ELN Candidates with global rank > 16 have a white background color.

ELN Candidates in red color achieved a maximum dissimilarity score (most dissimilar to the training genes).

~~ELN~~ Candidates displayed in a line through font were not scored for the data source (data missing).

Global prioritization	Annotation EnsemblEst	Annotation GeneOntology	Annotation Interpro	Annotation Kegg	Annotation Swissprot	Blast	CisRegModule	Expression SonEtAl	Expression SuEtAl	Interaction Bind	Interaction BioGrid	Interaction Hprd	Interaction InNetDb	Interaction Intact	Interaction Mint	Interaction String	Motif	Precalculated Ouzounis	Precalculated Prospectr	Text
EDNRA	EDNRA	EDNRA	CCR3	EDNRA	GPR125	CTGF	PVRL1	F7	ADAMTS9	F7	GPR125	MERTK	GPR125	GPR125	GPR125	LTBP1	F7	PVRL1	ADAMTS9	CCR3
MERTK	CALCB	CCR3	EDNRA	PVRL1	GPR116	CCR3	ADAMTS9	GPR116	PVRL1	ADAMTS9	EDNRA	EDNRA	CALCB	CALCB	CALCB	EDNRA	GPR116	MERTK	NUDT4	MERTK
PVRL1	CORIN	MERTK	CTGF	CCR3	CCR3	EDNRA	CORIN	CORIN	MERTK	PVRL1	F7	IGF2	EDNRA	EDNRA	EDNRA	CALCB	MERTK	F7	PVRL1	F7
GPR125	NUDT4	GPR125	GPR116	LTBP1	EDNRA	GPR125	LTBP1	NUDT4	F7	MERTK	ADAMTS9	GPR125	CTGF	CTGF	CTGF	CORIN	CCR3	EDNRA	MERTK	CTGF
CCR3	LTBP1	GPR116	GPR125	F7	MERTK	ADAMTS9	CTGF	EDNRA	CALCB	CALCB	PVRL1	F7	LTBP1	LTBP1	LTBP1	IGF2	CTGF	LTBP1	F7	IGF2
F7	CCR3	NUDT4	ADAMTS9	~~CTGF~~	PVRL1	LTBP1	NUDT4	LTBP1	CORIN	CORIN	MERTK	ADAMTS9	CORIN	CORIN	CORIN	ADAMTS9	ADAMTS9	CORIN	EDNRA	PVRL1
CTGF	F7	CTGF	MERTK	~~GPR116~~	F7	MERTK	EDNRA	PVRL1	GPR125	GPR125	CALCB	PVRL1	CCR3	CCR3	CCR3	CCR3	GPR125	IGF2	LTBP1	CORIN
ADAMTS9	GPR125	CORIN	NUDT4	~~GPR125~~	IGF2	GPR116	GPR125	GPR125	IGF2	IGF2	CORIN	CALCB	PVRL1	PVRL1	PVRL1	PVRL1	PVRL1	GPR125	CTGF	EDNRA
CORIN	IGF2	CALCB	CORIN	~~ADAMTS9~~	CTGF	F7	~~CCR3~~	MERTK	EDNRA	EDNRA	IGF2	CORIN	ADAMTS9	ADAMTS9	ADAMTS9	MERTK	NUDT4	CTGF	GPR125	ADAMTS9
IGF2	GPR116	IGF2	CALCB	~~MERTK~~	CORIN	NUDT4	~~MERTK~~	CTGF	CCR3	CCR3	CCR3	CCR3	IGF2	IGF2	IGF2	NUDT4	LTBP1	GPR116	IGF2	LTBP1
LTBP1	ADAMTS9	PVRL1	IGF2	~~NUDT4~~	LTBP1	PVRL1	~~GPR116~~	ADAMTS9	GPR116	GPR116	GPR116	GPR116	F7	F7	F7	GPR125	EDNRA	CALCB	CALCB	GPR125
GPR116	PVRL1	LTBP1	PVRL1	~~CORIN~~	ADAMTS9	IGF2	~~F7~~	CALCB	CTGF	CTGF	CTGF	CTGF	MERTK	MERTK	MERTK	GPR116	CORIN	CCR3	CORIN	NUDT4
CALCB	MERTK	ADAMTS9	LTBP1	~~CALCB~~	CALCB	CORIN	~~IGF2~~	IGF2	NUDT4	NUDT4	NUDT4	NUDT4	GPR116	GPR116	GPR116	CTGF	CALCB	ADAMTS9	CCR3	CALCB
NUDT4	~~CTGF~~	F7	F7	~~IGF2~~	NUDT4	CALCB	~~CALCB~~	~~CCR3~~	LTBP1	LTBP1	LTBP1	LTBP1	NUDT4	NUDT4	NUDT4	F7	~~IGF2~~	NUDT4	GPR116	GPR116

ToppGene – http://toppgene.cchmc.org

ToppGene Suite

A one-stop portal for gene list enrichment analysis and candidate gene prioritization based on functional annotations and protein interactions network

- **ToppFun**: Transcriptome, ontology, phenotype, proteome, and pharmacome annotations based gene list functional enrichment analysis

  Detect functional enrichment of your gene list based on Transcriptome, Proteome, Regulome (TFBS and miRNA), Ontologies (GO, Pathway), Phenotype (human disease and mouse phenotype), Pharmacome (Drug-Gene associations), literature co-citation, and other features.

- **ToppGene**: Candidate gene prioritization

  Prioritize or rank candidate genes based on functional similarity to training gene list.

- **ToppNet**: Relative importance of candidate genes in networks

  Prioritize or rank candidate genes based on topological features in protein-protein interactio

- **ToppGenet**: Prioritization of neighboring genes in protein-protein interaction network

  Identify and prioritize the neighboring genes of the seeds in protein-protein interaction netw

Select your gene identifier type, paste your training and test gene sets below or select example sets, then submit.

Example gene sets: HGNC Symbol Entrez ID
(click on "HGNC Symbol" or "Entrez ID" to use the example training and test set of genes)

Symbol Types	HGNC Symbol		HGNC Symbol
Training Gene Set:	AXIN2 CPE AGTR1 CTNNB1 MEN1 GIP WISP2 LHCGR AVPR1B MC2R AVPR1A ADRB1 GIPR AVPR2	Test gene set:	EDNRA CCR3 CTGF ADAMTS9 MERTK NUDT4 IGF2 GPR125 F7 LTBP1 GPR116 CORIN CALCB CCNT PVRL1

Clear Submit Query

Doesn't support other than human/mouse

Cincinnati Children's

UNIVERSITY OF Cincinnati

Training set (14 / 14)

Entered	Human Symbol	Gene ID
AXIN2	AXIN2	8313
CPE	CPE	1363
AGTR1	AGTR1	185
CTNNB1	CTNNB1	1499
MEN1	MEN1	4221
GIP	GIP	2695
WISP2	WISP2	8839
LHCGR	LHCGR	3973
AVPR1B	AVPR1B	553
MC2R	MC2R	4158
AVPR1A	AVPR1A	552
ADRB1	ADRB1	153
GIPR	GIPR	2696
AVPR2	AVPR2	554

Ignored

Entered	Status

Test set (14 / 15)

Entered	Human Symbol	Gene ID
EDNRA	EDNRA	1909
CCR3	CCR3	1232
CTGF	CTGF	1490
ADAMTS9	ADAMTS9	56999
MERTK	MERTK	10461
NUDT4	NUDT4	11163
IGF2	IGF2	3481
GPR125	GPR125	166647
F7	F7	2155
LTBP1	LTBP1	4052
GPR116	GPR116	221395
CORIN	CORIN	10699
CALCB	CALCB	797
PVRL1	PVRL1	5818

Ignored

Entered	Status
CCNT	Not Found

Find alternatives for missing symbols

Training parameters

Feature	Correction	p-Value cutoff	Gene Limits
All	FDR	0.05	1 ≤ n ≤ 1500
GO: Molecular Function	FDR	0.05	1 ≤ n ≤ 1500
GO: Biological Process	FDR	0.05	1 ≤ n ≤ 1500
GO: Cellular Component	FDR	0.05	1 ≤ n ≤ 1500
Human Phenotype	FDR	0.05	1 ≤ n ≤ 1500
Mouse Phenotype	FDR	0.05	1 ≤ n ≤ 1500
Domain	FDR	0.05	1 ≤ n ≤ 1500
Pathway	FDR	0.05	1 ≤ n ≤ 1500
Pubmed	FDR	0.05	1 ≤ n ≤ 1500
Interaction	FDR	0.05	1 ≤ n ≤ 1500
Cytoband	FDR	0.05	1 ≤ n ≤ 1500
Transcription Factor Binding Site	FDR	0.05	1 ≤ n ≤ 1500
Gene Family	FDR	0.05	1 ≤ n ≤ 1500
Coexpression	FDR	0.05	1 ≤ n ≤ 1500
Coexpression Atlas	FDR	0.05	1 ≤ n ≤ 1500
Computational	FDR	0.05	1 ≤ n ≤ 1500
MicroRNA	FDR	0.05	1 ≤ n ≤ 1500
Drug	FDR	0.05	1 ≤ n ≤ 1500
Disease	FDR	0.05	1 ≤ n ≤ 1500
CTD			
Clinical Variations			
GWAS			
OMIM			

Group results within a category by source.

Test parameter

Random sampling size: 1500 (6% of genome)

Min. feature count: 2

Home | Modify Query | Start

- Supports synonyms
- Presents suggestions/alternatives for unrecognized entries

Entered	Suggestions
CCNT	CCNT1 - cyclin T1 Human Synonym

Update | Check All

Cincinnati Children's

UNIVERSITY OF Cincinnati

Training Results [Expand All] [Download All] [Sparse Matrix] Display pValues and Scores as Scientific (4 significant digits) Table row limit 50

1: GO: Molecular Function [Display Chart] 132 annotations before applied cutoff / 18273 genes in category

	ID	Name	Source	pValue	FDR B&H	FDR B&Y	Bonferroni	Genes from Input	Genes in Annotation
1	GO:0017046	peptide hormone binding		1.118E-10	1.389E-8	7.589E-8	1.476E-8	5	43
2	GO:0008528	G-protein coupled peptide receptor activity		2.736E-10	1.389E-8	7.589E-8	3.612E-8	6	126
3	GO:0001653	peptide receptor activity		3.157E-10	1.389E-8	7.589E-8	4.167E-8	6	129
4	GO:0042562	hormone binding		1.979E-9	6.529E-8	3.568E-7	2.612E-7	5	75
5	GO:0005000	vasopressin receptor activity		1.251E-8	3.302E-7	1.804E-6	1.651E-6	3	7

Show 45 more annotations

2: GO: Biological Process [Display Chart] 1554 annotations before applied cutoff / 18138 genes in category

	ID	Name	Source	pValue	FDR B&H	FDR B&Y	Bonferroni	Genes from Input	Genes in Annotation
1	GO:0007186	G-protein coupled receptor signaling pathway		3.501E-10	5.441E-7	4.312E-6	5.441E-7	10	1057
2	GO:0030819	positive regulation of cAMP biosynthetic process		1.788E-9	6.796E-7	5.387E-6	2.779E-6	5	73
3	GO:0060341	regulation of cellular localization		2.187E-9	6.796E-7	5.387E-6	3.398E-6	9	875
4	GO:0030816	positive regulation of cAMP metabolic process		2.197E-9	6.796E-7	5.387E-6	3.413E-6	5	76
5	GO:0042592	homeostatic process		2.352E-9	6.796E-7	5.387E-6	3.655E-6	10	1284

Show 45 more annotations

3: GO: Cellular Component [Display Chart] 104 annotations before applied cutoff / 18818 genes in category

	ID	Name	Source	pValue	FDR B&H	FDR B&Y	Bonferroni	Genes from Input	Genes in Annotation
1	GO:0005768	endosome		3.932E-6	4.089E-4	2.137E-3	4.089E-4	6	650
2	GO:0005887	integral component of plasma membrane		2.019E-5	7.322E-4	3.827E-3	2.099E-3	7	1341
3	GO:0031226	intrinsic component of plasma membrane		2.589E-5	7.322E-4	3.827E-3	2.693E-3	7	1393
4	GO:0030877	beta-catenin destruction complex		2.816E-5	7.322E-4	3.827E-3	2.929E-3	2	11
5	GO:0031410	cytoplasmic vesicle		8.434E-4	1.754E-2	9.168E-2	8.771E-2	5	1093

Show 5 more annotations

4: Human Phenotype [Display Chart] 321 annotations before applied cutoff / 3811 genes in category

	ID	Name	Source	pValue	FDR B&H	FDR B&Y	Bonferroni	Genes from Input	Genes in Annotation
1	HP:0000850	Cushing syndrome		5.761E-5	1.849E-2	1.174E-1	1.849E-2	2	6
2	HP:0001578	Hypercortisolism		2.101E-4	2.248E-2	1.428E-1	6.745E-2	2	11
3	HP:0004316	Increased cortisol production		2.101E-4	2.248E-2	1.428E-1	6.745E-2	2	11
4	HP:0004421	Elevated systolic blood pressure		7.191E-4	3.521E-2	2.236E-1	2.308E-1	2	20
5	HP:0010514	Hyperpituitarism		9.545E-4	3.521E-2	2.236E-1	3.064E-1	2	23

Show 45 more annotations

5: Mouse Phenotype [Display Chart] 1062 annotations before applied cutoff / 7344 genes in category

	ID	Name	Source	pValue	FDR B&H	FDR B&Y	Bonferroni	Genes from Input	Genes in Annotation
1	MP:0005418	abnormal circulating hormone level		5.911E-8	6.278E-5	4.737E-4	6.278E-5	9	656
2	MP:0000639	abnormal adrenal gland morphology		1.454E-7	6.387E-5	4.819E-4	1.545E-4	5	86
3	MP:0003953	abnormal hormone level		1.804E-7	6.387E-5	4.819E-4	1.916E-4	9	745

Enrichment of Training/seed set:

- Helps in assessing the "quality" of the training set
- Can assist in selecting sub-sets of training set to perform prioritizations (e.g., large training set)

Cincinnati Children's

UNIVERSITY OF Cincinnati

Test Results [Hide Detail] [Download] [Show Network]

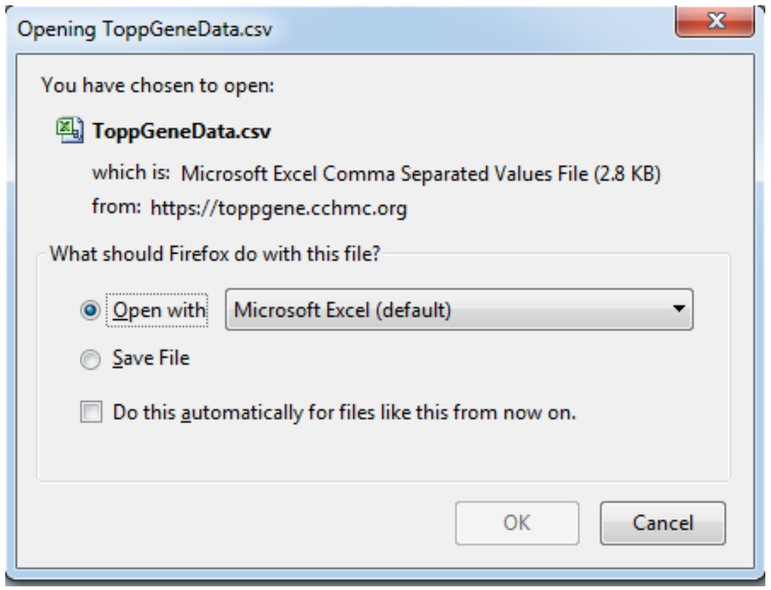
Rank (net)	Gene Symbol	Gene ID	GO: Molecular Function		GO: Biological Process		GO: Cellular Component		Human Phenotype		Mouse Phenotype		Domain		Pathway		Pubmed		Interaction		Cytoband		Transcription Factor Binding Site		Gene Family		Coexpression		Coexpression Atlas		Computational		MicroRNA		Drug		Disease		Average score	Overall P-value
			Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue		
1	EDNRA	1909	9.135E-1	5.232E-3	1.000E0	6.736E-2	8.078E-1	1.243E-2	1.323E-1	4.644E-2	3.591E-1	4.382E-2	1.142E-1	7.848E-3	9.932E-1	1.000E-6	1.000E0	1.112E-2	2.746E-1	4.055E-2	0.000E0	5.049E-1	0.000E0	4.997E-1			7.117E-1	1.000E-6	2.362E-1	5.886E-3	7.999E-1	1.000E-6	0.000E0	4.997E-1	1.000E0	9.810E-3	2.013E-1	6.540E-3	5.026E-1	8.680E-12
2	IGF2	3481	5.930E-1	4.709E-2	1.000E0	6.736E-2	9.065E-1	6.540E-4	1.000E0	6.540E-4	8.686E-1	1.897E-2	0.000E0	5.095E-1	0.000E0	5.154E-1	1.000E0	1.046E-2	7.393E-1	5.886E-3	0.000E0	5.049E-1	0.000E0	4.997E-1			7.597E-1	1.000E-6	3.295E-1	3.924E-3	0.000E0	5.003E-1	0.000E0	4.997E-1	1.000E0	9.810E-3	8.721E-1	1.308E-3	5.335E-1	3.530E-8
3	CCR3	1232	9.135E-1	5.232E-3	1.000E0	6.736E-2	8.029E-1	1.243E-2			2.507E-1	5.952E-2	1.142E-1	7.848E-3	8.511E-1	2.616E-3	9.931E-1	2.420E-2	0.000E0	5.258E-1	0.000E0	5.049E-1	0.000E0	4.997E-1	0.000E0	4.997E-1	5.732E-1	2.616E-3	0.000E0	5.173E-1	7.999E-1	1.000E-6	0.000E0	4.997E-1	8.654E-1	1.014E-1	0.000E0	5.069E-1	4.254E-1	3.308E-6
4	CTGF	1490	7.846E-1	1.635E-2	1.000E0	6.736E-2	4.245E-1	1.354E-1			3.729E-1	4.251E-2	9.998E-1	6.540E-4	0.000E0	5.154E-1	1.000E0	1.046E-2	0.000E0	5.258E-1	0.000E0	5.049E-1	0.000E0	4.997E-1			6.812E-1	6.540E-4	2.362E-1	5.886E-3	0.000E0	5.003E-1	0.000E0	4.997E-1	1.000E0	1.046E-2	4.526E-1	3.924E-3	4.345E-1	1.498E-5
5	GPR116	221395	6.602E-1	3.009E-2	9.964E-1	1.485E-1	6.169E-1	5.101E-2			3.507E-1	5.298E-2	9.704E-1	2.616E-3	0.000E0	5.154E-1	3.163E-1	1.236E-1			0.000E0	5.049E-1	0.000E0	4.997E-1	0.000E0	4.997E-1	0.000E0	5.154E-1	2.362E-1	5.886E-3			0.000E0	4.997E-1	4.710E-1	1.687E-1			3.630E-1	5.249E-3
6	CORIN	10699	2.504E-1	1.367E-1	1.000E0	6.736E-2	8.981E-1	1.308E-3	1.247E-3	5.494E-2	1.181E-1	7.129E-2	0.000E0	5.095E-1	0.000E0	5.154E-1	9.970E-1	2.093E-2			0.000E0	5.049E-1	0.000E0	4.997E-1			0.000E0	5.154E-1	3.295E-1	3.924E-3	0.000E0	5.003E-1	0.000E0	4.997E-1	7.081E-1	1.308E-1	0.000E0	5.069E-1	2.773E-1	5.666E-3
7	GPR125	166647	6.602E-1	3.009E-2	7.881E-1	2.485E-1	3.164E-1	1.609E-1			1.000E-2	7.848E-2	9.721E-1	1.962E-3			6.835E-1	7.587E-2	0.000E0	5.258E-1	0.000E0	5.049E-1					0.000E0	5.154E-1	0.000E0	5.173E-1			0.000E0	4.997E-1	0.000E0	6.167E-1			3.061E-1	2.118E-2
8	F7	2155	4.430E-1	7.260E-2	1.000E0	6.802E-2	6.354E-1	4.120E-2	2.679E-1	3.859E-2	2.507E-1	5.952E-2	0.000E0	5.095E-1	0.000E0	5.154E-1	1.000E0	1.112E-2	0.000E0	5.258E-1	0.000E0	5.049E-1					0.000E0	5.154E-1	0.000E0	5.173E-1	0.000E0	5.003E-1	0.000E0	4.997E-1	9.911E-1	4.774E-2	0.000E0	5.069E-1	2.870E-1	2.533E-2
9	CALCB	797	3.463E-1	9.614E-2	9.987E-1	1.328E-1	0.000E0	6.566E-1					0.000E0	5.095E-1	7.597E-1	5.232E-3	8.666E-1	5.363E-2			0.000E0	5.049E-1	0.000E0	4.997E-1			0.000E0	5.154E-1	0.000E0	5.173E-1	0.000E0	5.003E-1	0.000E0	4.997E-1	1.000E0	9.810E-3			3.105E-1	2.890E-2
10	CCNT1	904	8.203E-1	1.373E-2	1.000E0	7.194E-2	5.235E-1	6.671E-2					0.000E0	5.095E-1	0.000E0	5.154E-1	7.583E-1	6.998E-2	4.588E-1	2.158E-2	0.000E0	5.049E-1	0.000E0	4.997E-1			0.000E0	5.154E-1	0.000E0	5.173E-1	0.000E0	5.003E-1	0.000E0	4.997E-1	4.071E-1	1.838E-1			3.080E-1	3.093E-2
11	PVRL1	5818	8.505E-1	1.046E-2	1.000E0	7.979E-2	4.833E-1	9.745E-2	8.204E-2	4.905E-2	1.000E-2	7.848E-2	0.000E0	5.095E-1	0.000E0	5.154E-1	9.972E-1	2.027E-2	0.000E0	5.258E-1	0.000E0	5.049E-1	0.000E0	4.997E-1	0.000E0	4.997E-1	0.000E0	5.154E-1	0.000E0	5.173E-1	0.000E0	5.003E-1	0.000E0	4.997E-1	8.866E-1	9.810E-2	0.000E0	5.069E-1	2.426E-1	5.302E-2
12	N							-2	9.733E-2	2.976E-1			0.000E0	5.095E-1	0.000E0	5.154E-1	0.000E0	5.618E-1			0.000E0	5.049E-1	0.000E0	4.997E-1	0.000E0	4.997E-1	5.136E-1	2.616E-3	0.000E0	5.173E-1	0.000E0	5.003E-1	0.000E0	4.997E-1	6.224E-1	1.445E-1			1.917E-1	9.202E-2
13	M							-2	6.682E-1	3.205E-2	1.247E-3	5.494E-2	5.149E-1	3.728E-2	0.000E0	5.095E-1	0.000E0	5.154E-1	0.000E0	5.618E-1	0.000E0	5.258E-1	0.000E0	5.049E-1	0.000E0	4.997E-1			0.000E0	5.154E-1	0.000E0	5.173E-1	0.000E0	5.003E-1	0.000E0	4.997E-1	5.005E-1	1.629E-1	0.000E0	5.069E-1	2.024E-1	9.464E-2
14	L							-1	0.000E0	6.566E-1			2.507E-1	5.952E-2	0.000E0	5.095E-1	0.000E0	5.154E-1	7.945E-1	6.344E-2	0.000E0	5.258E-1	0.000E0	5.049E-1	0.000E0	4.997E-1			0.000E0	5.154E-1	0.000E0	5.173E-1	0.000E0	5.003E-1	0.000E0	4.997E-1	6.978E-1	1.321E-1			2.382E-1	1.203E-1
15	ADA							-1	3.867E-1	1.406E-1			1.000E-2	7.848E-2	0.000E0	5.095E-1	0.000E0	5.154E-1	5.696E-1	9.418E-2			0.000E0	5.049E-1	0.000E0	4.997E-1	0.000E0	5.154E-1	0.000E0	5.173E-1			0.000E0	4.997E-1	8.444E-2	2.341E-1	0.000E0	5.069E-1	1.750E-1	1.736E-1

Rank (net)	Gene Symbol	Gene ID
1	EDNRA	1909
2	IGF2	3481
3	CCR3	1232
4	CTGF	1490
5	GPR116	221395
6	CORIN	10699
7	GPR125	166647
8	F7	2155
9	CALCB	797
10	CCNT1	904
11	PVRL1	5818
12	NUDT4	11163
13	MERTK	10461
14	LTBP1	4052
15	ADAMTS9	56999

Why is a test set gene ranked at top <u>or</u> What features are shared between the training/seed set and ranked test set gene are presented both as a network and tabular format.

- Select the ranked genes
- Resulting training set, shared annotations, and the ranked gene(s) can be downloaded as an XGMML or GEXF file (Cytoscape/Gephi import)

Download the rankings table



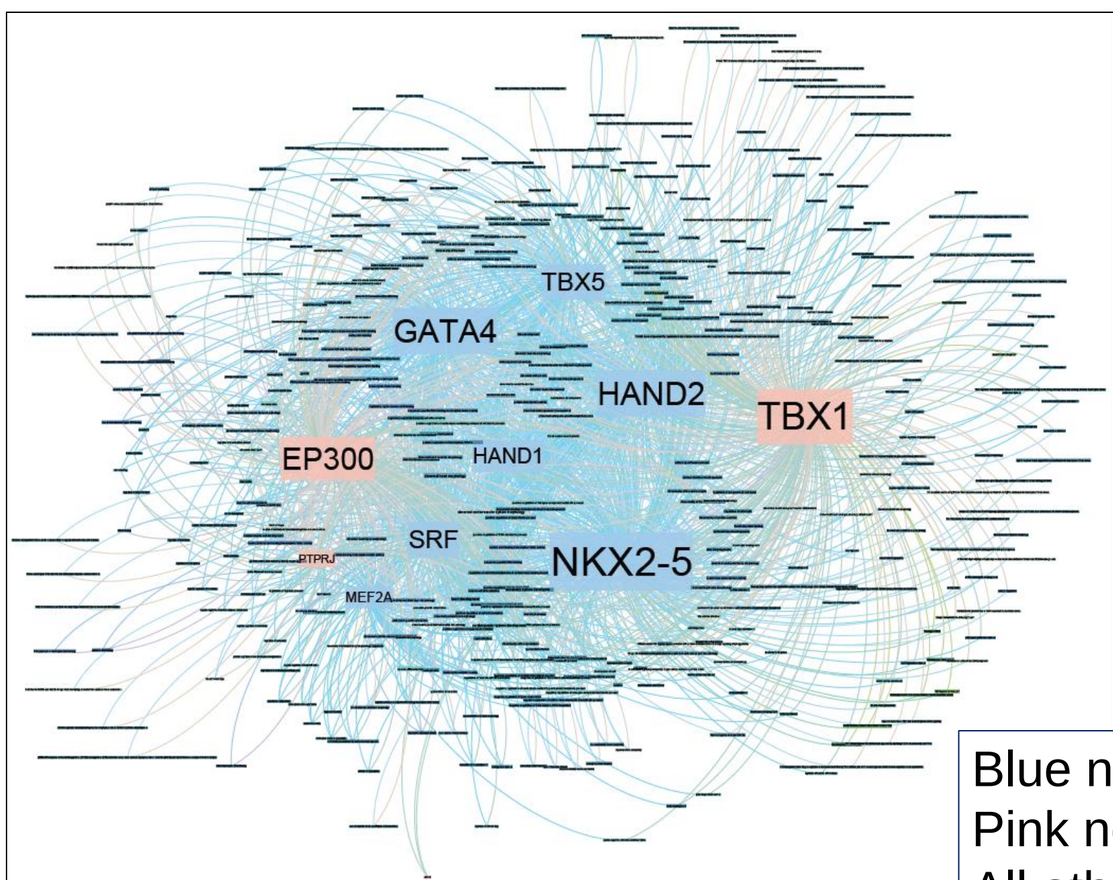
Cincinnati Children's

UNIVERSITY OF Cincinnati

Supporting Details

Selected genes shown in **bold**

Feature	ID	Name	Genes
GO: Molecular Function	GO:0008528	G-protein coupled peptide receptor activity	AGTR1 AVPR1A AVPR1B AVPR2 **EDNRA** LHCGR MC2R
GO: Molecular Function	GO:0001653	peptide receptor activity	AGTR1 AVPR1A AVPR1B AVPR2 **EDNRA** LHCGR MC2R
GO: Molecular Function	GO:0004930	G-protein coupled receptor activity	ADRB1 AGTR1 AVPR1A AVPR1B AVPR2 **EDNRA** GIPR LHCGR MC2R
GO: Molecular Function	GO:0004888	transmembrane signaling receptor activity	ADRB1 AGTR1 AVPR1A AVPR1B AVPR2 **EDNRA** GIPR LHCGR MC2R
GO: Molecular Function	GO:0038023	signaling receptor activity	ADRB1 AGTR1 AVPR1A AVPR1B AVPR2 **EDNRA** GIPR LHCGR MC2R
GO: Biological Process	GO:0043084	penile erection	AVPR1A **EDNRA**
GO: Biological Process	GO:0051482	positive regulation of cytosolic calcium ion concentration involved in phospholipase C-activating G-protein coupled signaling pathway	AGTR1 **EDNRA**
GO: Biological Process	GO:0051281	positive regulation of release of sequestered calcium ion into cytosol	**EDNRA** LHCGR
GO: Biological Process	GO:0007620	copulation	AVPR1A **EDNRA**
GO: Biological Process	GO:0007190	activation of adenylate cyclase activity	ADRB1 AVPR2 **EDNRA** GIPR LHCGR
GO: Biological Process	GO:0045762	positive regulation of adenylate cyclase activity	ADRB1 AVPR2 **EDNRA** GIPR LHCGR
GO: Biological Process	GO:0031281	positive regulation of cyclase activity	ADRB1 AVPR2 **EDNRA** GIPR LHCGR
GO: Biological Process	GO:0090183	regulation of kidney development	AGTR1 CTNNB1 **EDNRA**
GO: Biological Process	GO:0051349	positive regulation of lyase activity	ADRB1 AVPR2 **EDNRA** GIPR LHCGR
GO: Biological Process	GO:0007202	activation of phospholipase C activity	AVPR1A AVPR1B **EDNRA**
GO: Biological Process	GO:0010863	positive regulation of phospholipase C activity	AVPR1A AVPR1B **EDNRA**
GO: Biological Process	GO:0045761	regulation of adenylate cyclase activity	ADRB1 AVPR2 **EDNRA** GIPR LHCGR
GO: Biological Process	GO:1900274	regulation of phospholipase C activity	AVPR1A AVPR1B **EDNRA**
GO: Biological Process	GO:0007200	phospholipase C-activating G-protein coupled receptor signaling pathway	AGTR1 **EDNRA** LHCGR



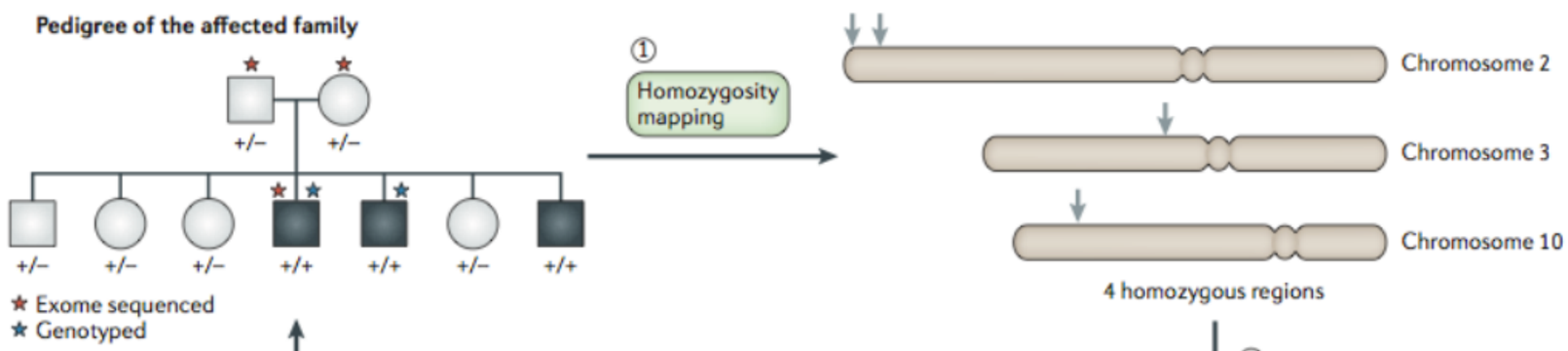
Shared annotations between training set & ranked test set gene

Gene in bold is ranked test set gene; rest are training/seed genes

Blue nodes: Seed genes
Pink nodes: ranked candidates
All other nodes – shared annotations

Combining gene level information with genomic variant information – Few case studies

Exome sequencing and gene prioritization



	Rank				
	1	2	3	4	5
Suspects	*KIF1A*	*D2HGDH*	*HDLBP*	*ING5*	*ATG4B*
ToppGene	*KIF1A*	*D2HGDH*	*ATG4B*	*LOC728846*	*SNED1*
Endeavour	*KIF1A*	*PASK*	*D2HGDH*	*HDLBP*	*ATG4B*
Combination	*KIF1A*	*D2HGDH*	*ATG4B*	*HDLBP*	*PASK*

Nature Reviews | Genetics

CSH PRESS GENOME RESEARCH

Exome sequencing and disease-network analysis of a single family implicate a mutation in *KIF1A* in hereditary spastic paraparesis

Yaniv Erlich, Simon Edvardson, Emily Hodges, et al.

Genome Res. 2011 21: 658-664 originally published online April 12, 2011

Moreau & Tranchevent, 2012

Some more examples of published studies that used Endeavour and ToppGene for candidate gene prioritization

Familial Cancer

10.1007/s10689-013-9642-y

Original Article

Whole exome sequencing identifies mutation of EDNRA involved in ACTH-independent macronodular adrenal hyperplasia

Jie Zhu[1], Liang Cui[1], Wei Wang[1], Xing-Yi Hang[2], A-Xiang Xu[1], Su-Xia Yang[1], Jing-Tao Dou[1], Yi-Ming Mu[1], Xu Zhang[1] and Jiang-Ping Gao[1]

The rank of candidate genes in different diseases prediction algorithms

Rank	SUSPECTS	ToppGene	Endeavour
1	EDNRA	EDNRA	EDNRA
2	CCR3	CCR3	GRM1
3	CTGF	PTAFR	GPR125
4	ADAMTS9	GRID2	ACVR1B
5	MERTK	CASR	CCR3
6	NUDT4	GPR18	LGR5
7	IGF2	CALCB	CUL1
8	GPR125	ADCY8	F7
9	F7	GRM1	GRHPR
10	LTBP1	WFS1	PTK2

AJMG AMERICAN JOURNAL OF medical genetics

Research Article

Identification of novel candidate gene loci and increased sex chromosome aneuploidy among infants with conotruncal heart defects

Kazutoyo Osoegawa[1,*], David M. Iovannisci[1], Bin Lin[1], Christina Parodi[1], Kathleen Schultz[1], Gary M. Shaw[2] and Edward J. Lammer[1]

Article first published online: 11 OCT 2013
DOI: 10.1002/ajmg.a.36291

Issue
American Journal of Medical Genetics Part A
Volume 164, Issue 2, pages 397–406, February 2014

Exome Sequencing and Systems Biology Converge to Identify Novel Mutations in the L-Type Calcium Channel, *CACNA1C*, Linked to Autosomal Dominant Long QT Syndrome

Nicole J. Boczek, Jabe M. Best, David J. Tester, John R. Giudicessi, Sumit Middha, Jared M. Evans, Timothy J. Kamp and Michael J. Ackerman

Circ Cardiovasc Genet. published online May 15, 2013;

OPEN ACCESS Freely available online

Exome-Sequencing Confirms *DNAJC5* Mutations as Cause of Adult Neuronal Ceroid-Lipofuscinosis

Bruno A. Benitez[1], David Alvarado[2], Yefei Cai[1], Kevin Mayo[1], Sumitra Chakraverty[1], Joanne Norton[1], John C. Morris[3], Mark S. Sands[2,4], Alison Goate[1,3,4,5], Carlos Cruchaga[1,4]*

Table 3. Gene Ranking Using Bioinformatic/Systems Biology Algorithms

Gene	Rank: Endeavour[30]	SUSPECTS[31]	Topp Gene[32]	Total Rank
CACNA1C	1	1	1	1
SLC26A1	3	2	2	2
MAP1A	2	6	3	3
DMRTB1	8	3	4	4
DSCAML1	6	4	6	5
GLTSCR1	5	8	5	6
PPP1R12C	4	7	8	7
MARCKS	7	5	7	8

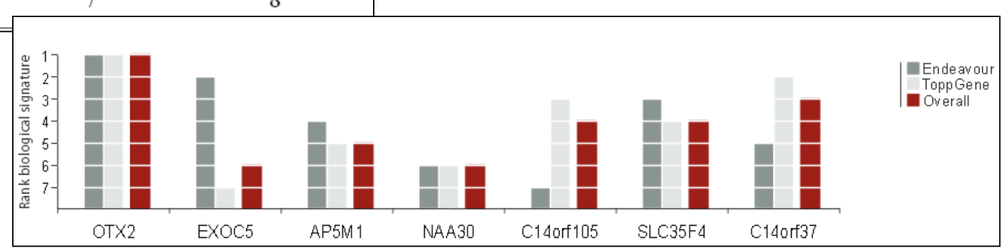
Human Genetics
March 2013, Volume 132, Issue 3, pp 285-292

Variants in *GATA4* are a rare cause of familial and sporadic congenital diaphragmatic hernia

Lan Yu, Julia Wynn, Yee Him Cheung, Yufeng Shen, George B. Mychaliska, Timothy M. Crombleholme, Kenneth S. Azarow, Foong Yen Lim, Dai H. Chung, Douglas Potoka, Brad W. Warner, Brian Bucher, Charles Stolar, Gudrun Aspelund, Marc S. Arkovitz, Wendy K. Chung

Table S1. Summary of comparison of results from ToppGene and Endeavour gene prioritization analysis

Gene Symbol	Toppgene Results: Rank Net	Average Score	Overall pValue	Endeavour results: Global prioritization score	Rank	Combined score
DNAJC5	1	0.7031	7.07E-05	2	0.117	1.5
CDC16	3	0.4483	0.0133049	3	0.167	3
PDCD6IP	5	0.3351	0.0246281	1	0.071	3
NPY1R	2	0.6797	2.30E-04	5	0.249	3.5
LIPJ	4	0.3063	0.015721	4	0.215	4
ZNF717	8	0.1556	0.1493499	7	0.29	7.5
LCMT1	10	0.1008	0.1746657	8	0.312	9
MUC3A	12	0.1445	0.2637284	6	0.256	9
ENPP3	7	0.2334	0.1058833	13	0.531	10
ABCC1	11	0.2423	0.2043685	9	0.337	10
ARHGAP10	9	0.1517	0.1615053	12	0.479	10.5
TNRC6A	6	0.3164	0.0609277	16	0.705	11
LOC100287879	14	-1	0.3513417	11	0.41	12.5
CCDC74B	15	-1	0.3513417	10	0.355	12.5
TTBK2	13	0.1311	0.2709082	15	0.689	14
LOC389906	16	-1	0.3513417	14	0.69	15



OTX2 Duplication Is Implicated in Hemifacial Microsomia

Dina Zielinski[1], Barak Markus[1], Mona Sheikh[1], Melissa Gymrek[1,2,3,4], Clement Chu[5], Marta Zaks[6], Balaji Srinivasan[5], Jodi D. Hoffman[7], Dror Aizenbud[6,9], Yaniv Erlich[1,*,9]

1 Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, United States of America, 2 Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, Massachusetts, United States of America, 3 Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America, 4 Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts, United States of America, 5 Counsyl, South San Francisco, California, United States of America, 6 Rambam Health Care Campus, Haifa, Israel, 7 Division of Genetics, Tufts Medical Center, Boston, Massachusetts, United States of America

Limitations & Points to Remember

- Bias towards the training set: Disease genes yet to be discovered will be consistent with what is already known about a disease and/or its genetic basis – assumption not always true.
- Bias towards selecting better annotated genes: “true” candidate can be missed if it lacks “sufficient” annotations.
- Accuracy depends on the quality (and coverage) of underlying original sources from which the annotations are retrieved.
- Appropriate or “true representative” training set selection: Using larger training sets (>100 genes) decreases the sensitivity and specificity of the prioritization compared to smaller training sets (6 to 30 genes).
- Coding-gene-centric: Complex traits result more often from noncoding regulatory variants than from coding sequence variants

Cincinnati Children's

UNIVERSITY OF Cincinnati

Gene Prioritization Portal

http://homes.esat.kuleuven.be/~bioiuser/gpp/index.php

Gene Prioritization Portal

Home | Tools | Stats | Search | Contact

Gene prioritization

One of the major challenge in human genetics is to find the genetic cause underlying disorders in order to unravel the molecular basis of these diseases and eventually elaborate medical treatments. High-throughput technologies such as linkage analysis and association studies are usually able to associate a chromosomal region with a genetic condition. One can also use expression arrays and obtain a list of transcripts differentially expressed in a disease sample with respect to a reference sample. A common characteristic of these methods is usually the large number of genes returned: for instance hundreds of differentially expressed genes are often reported. The working hypothesis is often that only one or a few candidates are really of primary interest. Identifying the most promising candidates among such large lists of genes is a challenging and time consuming task.The bioinformatics community has therefore introduced the concept of gene prioritization that uses computational methods and genomic data to identify the most promising candidates.

The gene prioritization portal

In the last decade, many different strategies have been developed, some of which have been implemented into web applications and eventually experimentally validated. This website represents an regularly updated electronic review in that field. The aim is to help biologists who need to select the most promising genes from large gene lists by describing what are the current available options. We describe more than 20 strategies (see the tool section) and give more general information to support user decision (see the statistics section).

References

Tranchevent LC., Bonachela Capdevila F., Nitsch D., De Moor B., De Causmaecker P., Moreau Y. A guide to web tools to prioritize candidate genes. *Brief. Bioinform.* (2010) link to paper

Bornigen D., Tranchevent LC., Bonachela-Capdevila F., Devriendt K., De Moor B., De Causmaecker P., Moreau Y. An unbiased evaluation of gene prioritization tools. *Bioinformatics* (2012) link to paper

Summary:

46 tools
12 data sources
113 publications
39 validations

A random tool:

SNPs3D
More details

Gene Prioritization Portal

Home | Tools | Stats | Search | Contact

List of gene prioritization tools

aGeneApart
Read the detailed page or browse the corresponding website

BioGraph
Read the detailed page or browse the corresponding website

Biomine
Read the detailed page or browse the corresponding website

Bitola
Read the detailed page or browse the corresponding website

Candid
Read the detailed page or browse the corresponding w

CGI
Read the detailed sheet

DIR
Read the detailed page or browse the corresponding w

DomainRBF
Read the detailed page or browse the corresponding w

Endeavour
Read the detailed page or browse the corresponding w

G2D
Read the detailed page or browse the corresponding w

GeneDistiller
Read the detailed page or browse the corresponding w

Summary:

46 tools
12 data sources
113 publications
39 validations

Cincinnati Children's

UNIVERSITY OF Cincinnati

Gene Prioritization Portal

Gene Prioritization Portal

Home | Tools | Stats | Search | Contact

General statistics

This section contains different statistics about the gene prioritization portal. These statistics are automatically retrieved from our frequently updated database and thus they reflect the current state of the art in gene prioritization. Currently, the portal contains :

- 46 gene prioritization tools
- 12 data sources
- 113 gene prioritization related publications
- 39 gene prioritization validations

Data sources statistics

This section contains details about the different datasources and their use by the gene prioritization tools. The following table corresponds to Table 2 of the paper Tranchevent *et al.* (2010). Further details about the data sources can be found below the table.

Tools	Functional annotations	Expression	Regulatory information	Text (co-citation)	Text (functional)	Interactions	Pathways	Sequence	Phenotype	Conservation/homology	Disease probabilities	Chemicals	
aGeneApart					X								1
BioGraph	X				X	X	X					X	5
Biomine	X	X			X	X	X		X	X		X	8
Bitola					X								1
Caesar	X	X			X	X	X			X	X		7
Candid		X			X	X		X		X	X		6
CGI		X				X							2
DGP								X		X			2
DIR		X				X	X						3
DomainRBF								X					1
Endeavour	X	X	X		X	X	X	X			X		8
eResponseNet		X		X		X	X						4
G2D	X			X		X		X					4
GeneDistiller	X	X		X		X	X		X				6
GeneFriends		X											1
GeneProspector					X								1
GeneRank	X	X											2
GeneRanker	X			X	X	X			X				5
GeneSeeker		X			X				X	X			4
GeneWanderer	X	X		X		X	X						5

Inputs / outputs statistics

This section contains details about the different inputs needed by the gene prioritization tools and about the different outputs they produce. The following two tables corresponds to Table 3 of the paper Tranchevent *et al.* (2010). Further details about the inputs / outputs can be found below the tables. The first table describes the inputs of the gene prioritization softwares.

	Training data			Candidate genes		
Tools	Known genes	Keywords	Expression dataset	Region	DEG	Genome
aGeneApart		X				X
BioGraph		X				X
Biomine	X	X			X	
Bitola		X		X		X
Caesar		X		X	X	X
Candid		X		X	X	X
CGI			X			X
DGP						X
DIR	X				X	
DomainRBF		X				X
Endeavour	X	X		X	X	X
eResponseNet	X				X	
G2D	X	X		X		
GeneDistiller	X	X		X		
GeneFriends	X					X
GeneProspector		X				X
GeneRank			X			X
GeneRanker	X	X				X
GeneSeeker		X		X		X
GeneWanderer	X			X	X	
Génie		X			X	X
GenTrepid		X		X	X	

Cincinnati Children's

UNIVERSITY OF Cincinnati

Gene Prioritization Portal

Detailed view of a gene prioritization tool

Previous tool Back to list Next tool

Name
Pinta

URL
http://www.esat.kuleuven.be/pinta/

Description
PINTA identifies the most promising candidates within a region when only sparse information about the phenotype is available by replacing this knowledge by experimental data on differential gene expression between affected and healthy individuals. Considering the problem from the perspective of a gene/protein network, we assess the relevance of a candidate gene by considering the level of differential expression in its network neighborhood under the assumption that strong candidates tend to be surrounded by differentially expressed neighbors.

Software modes

Web based software	The software contains a web based interface and therefore can be used from a web browser.

Datasources

Functional annotations	Keywords associated to the genes and describing their functions or the ones of the corresponding proteins. Often an organized vocabulary (ontology) is used. The most used annotation source is Gene Ontology.
Expression	Data covering the expression level (usually at the transcript level) of many genes (usually genome-wide) over a wide range of tissues/conditions.
Text (co-citation)	Links between genes obtained by applying text mining techniques in order to find co-citations occurring in the scientific literature.
Interactions	Collection of gene-gene (or protein-protein) interactions coming from biological experiments. Two examples of such data sources are DIP and BioGrid.

Inputs / outputs

Training data: expression dataset
Candidate genes: region
Candidate genes: DEG
Candidate genes: genome

Prioritized list of candidates
Test Statistics

Publications

Nitsch *et al.*, **PINTA: a web server for network-based gene prioritization from expression data**, *Nucleic Acids Res. (2011)*
Show abstract, link to paper

Benchmarks

Nitsch *et al.*, **Candidate Gene Prioritization by Network Analysis of Differential Expression using Machine Learning Approaches**, *BMC Bioinformatics (2010)*
Show abstract, link to paper

Tools used: Pinta.

Critical Assessment

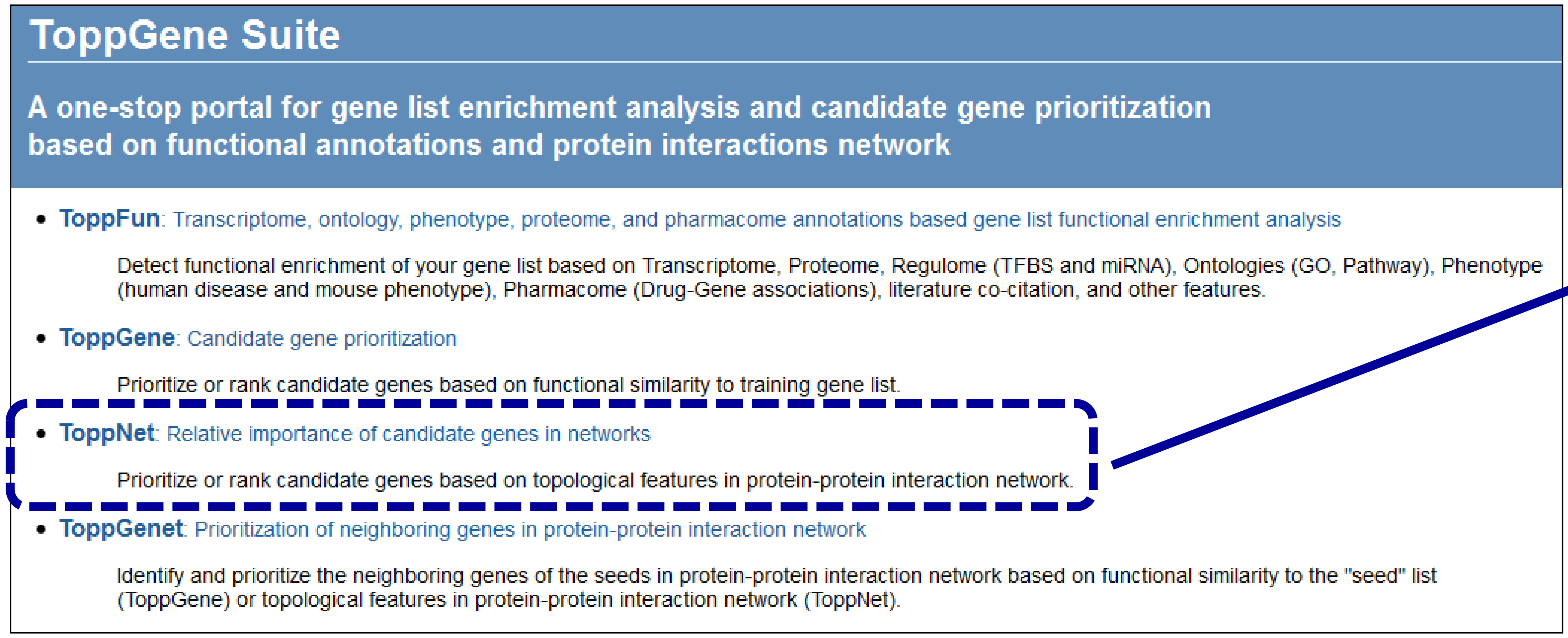
Results of a critical assessment performed on 8 gene prioritization softwares (reference to come). Regular benchmarks are usually run on know data; oppositely this benchmark mimics the discovery of a novel disease gene by performing predictive prioritizations. More precisely, 42 novel disease gene associations were used to benchmark the tools as soon as they were published and therefore before their inclusions in the databases used by these tools.

This plot presents the results of this benchmark for Pinta. Several performance indices are computed: median over the 42 rank ratios, AUC (Area Under the ROC Curve), and True Positive Rates at 10% and 30%.

Disease Gene Prioritization - Network-based strategies

- Candidate genes are ranked based on their topological relevance (e.g., distance) to known disease genes (Training/seed genes) in a network.
  - Protein-protein interactions network (BioGrid, BIND, HPRD, etc.)
  - Protein association network (STRING)
- Random-walk (or PageRank) approaches outperform clustering and neighborhood approaches.

ToppNet (http://toppgene.cchmc.org)



ToppGene Suite

A one-stop portal for gene list enrichment analysis and candidate gene prioritization based on functional annotations and protein interactions network

- **ToppFun**: Transcriptome, ontology, phenotype, proteome, and pharmacome annotations based gene list functional enrichment analysis

  Detect functional enrichment of your gene list based on Transcriptome, Proteome, Regulome (TFBS and miRNA), Ontologies (GO, Pathway), Phenotype (human disease and mouse phenotype), Pharmacome (Drug-Gene associations), literature co-citation, and other features.

- **ToppGene**: Candidate gene prioritization

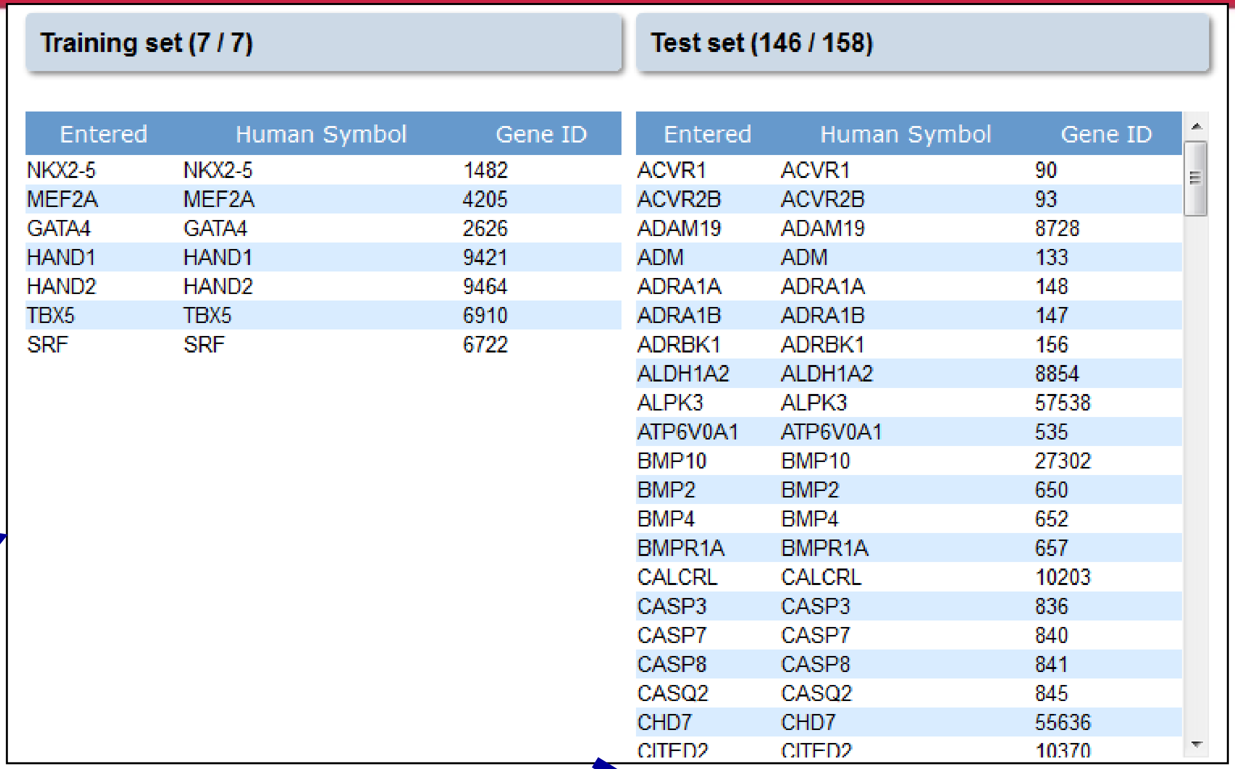
  Prioritize or rank candidate genes based on functional similarity to training gene list.

- **ToppNet**: Relative importance of candidate genes in networks

  Prioritize or rank candidate genes based on topological features in protein-protein interaction network.

- **ToppGenet**: Prioritization of neighboring genes in protein-protein interaction network

  Identify and prioritize the neighboring genes of the seeds in protein-protein interaction network based on functional similarity to the "seed" list (ToppGene) or topological features in protein-protein interaction network (ToppNet).

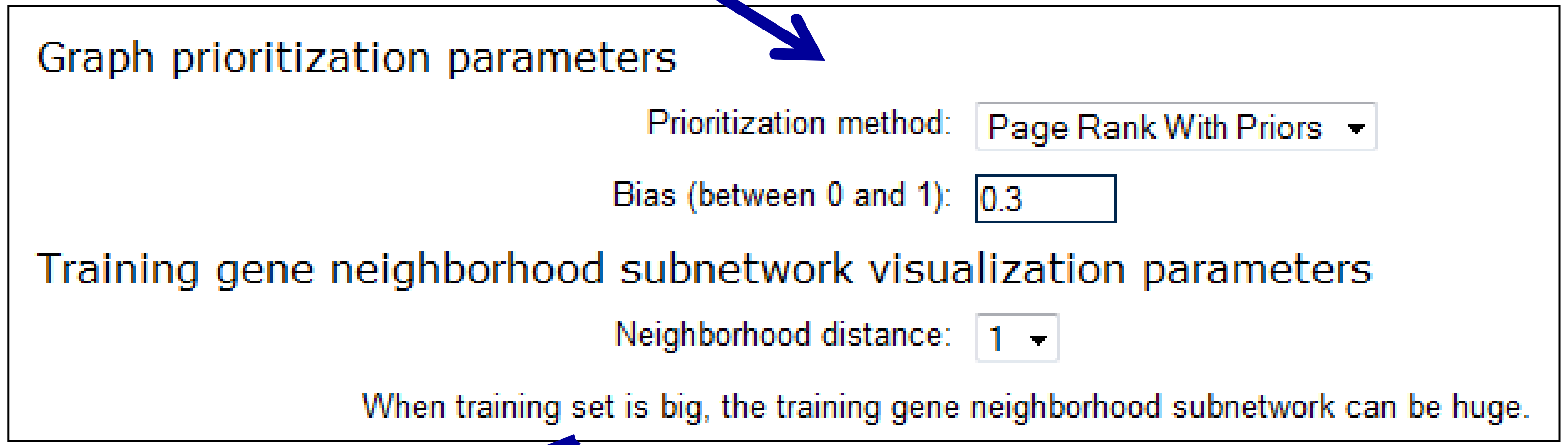


Training set (7 / 7)

Entered	Human Symbol	Gene ID
NKX2-5	NKX2-5	1482
MEF2A	MEF2A	4205
GATA4	GATA4	2626
HAND1	HAND1	9421
HAND2	HAND2	9464
TBX5	TBX5	6910
SRF	SRF	6722

Test set (146 / 158)

Entered	Human Symbol	Gene ID
ACVR1	ACVR1	90
ACVR2B	ACVR2B	93
ADAM19	ADAM19	8728
ADM	ADM	133
ADRA1A	ADRA1A	148
ADRA1B	ADRA1B	147
ADRBK1	ADRBK1	156
ALDH1A2	ALDH1A2	8854
ALPK3	ALPK3	57538
ATP6V0A1	ATP6V0A1	535
BMP10	BMP10	27302
BMP2	BMP2	650
BMP4	BMP4	652
BMPR1A	BMPR1A	657
CALCRL	CALCRL	10203
CASP3	CASP3	836
CASP7	CASP7	840
CASP8	CASP8	841
CASQ2	CASQ2	845
CHD7	CHD7	55636
CITED2	CITED2	10370



Graph prioritization parameters

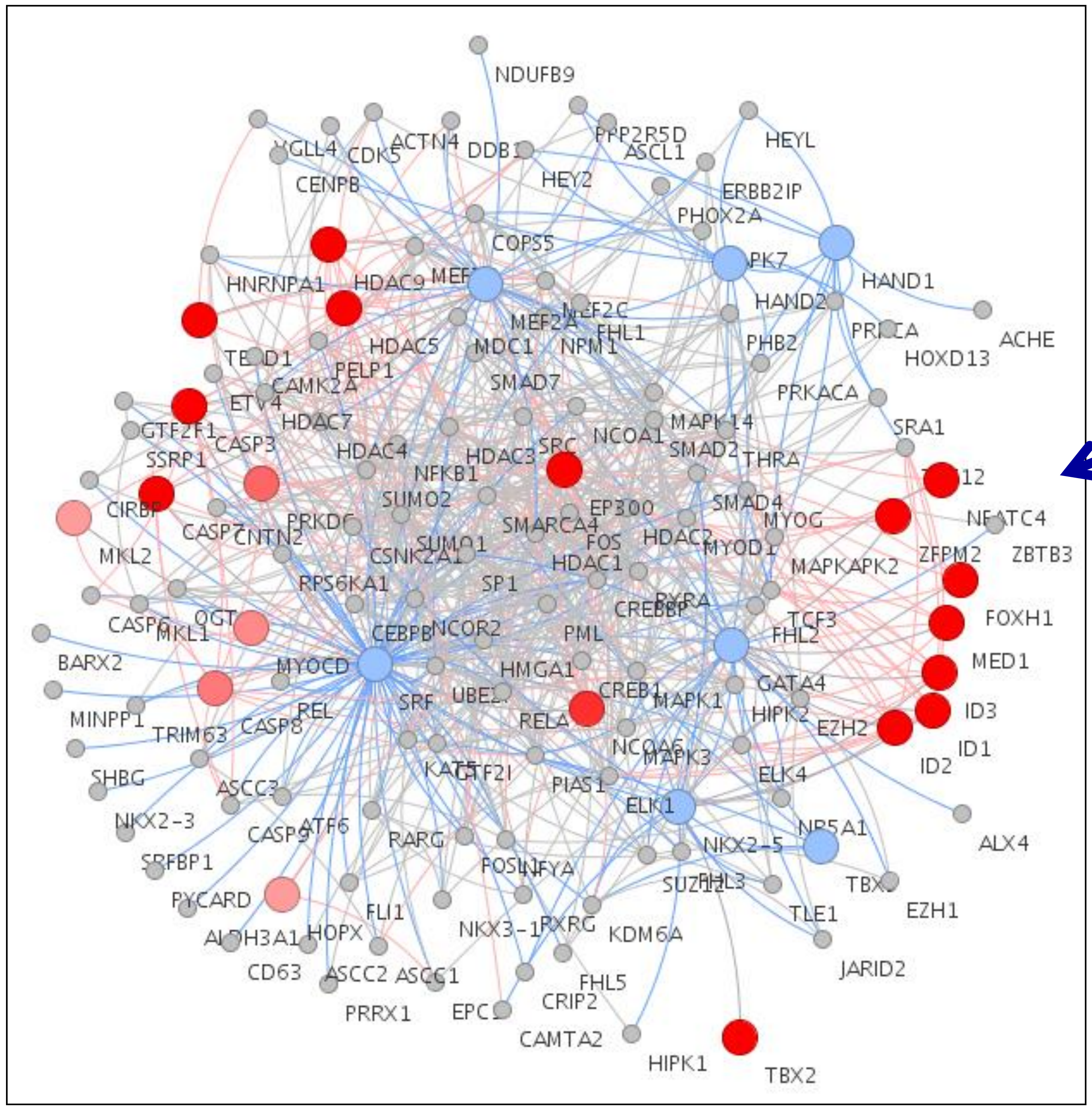
Prioritization method: Page Rank With Priors

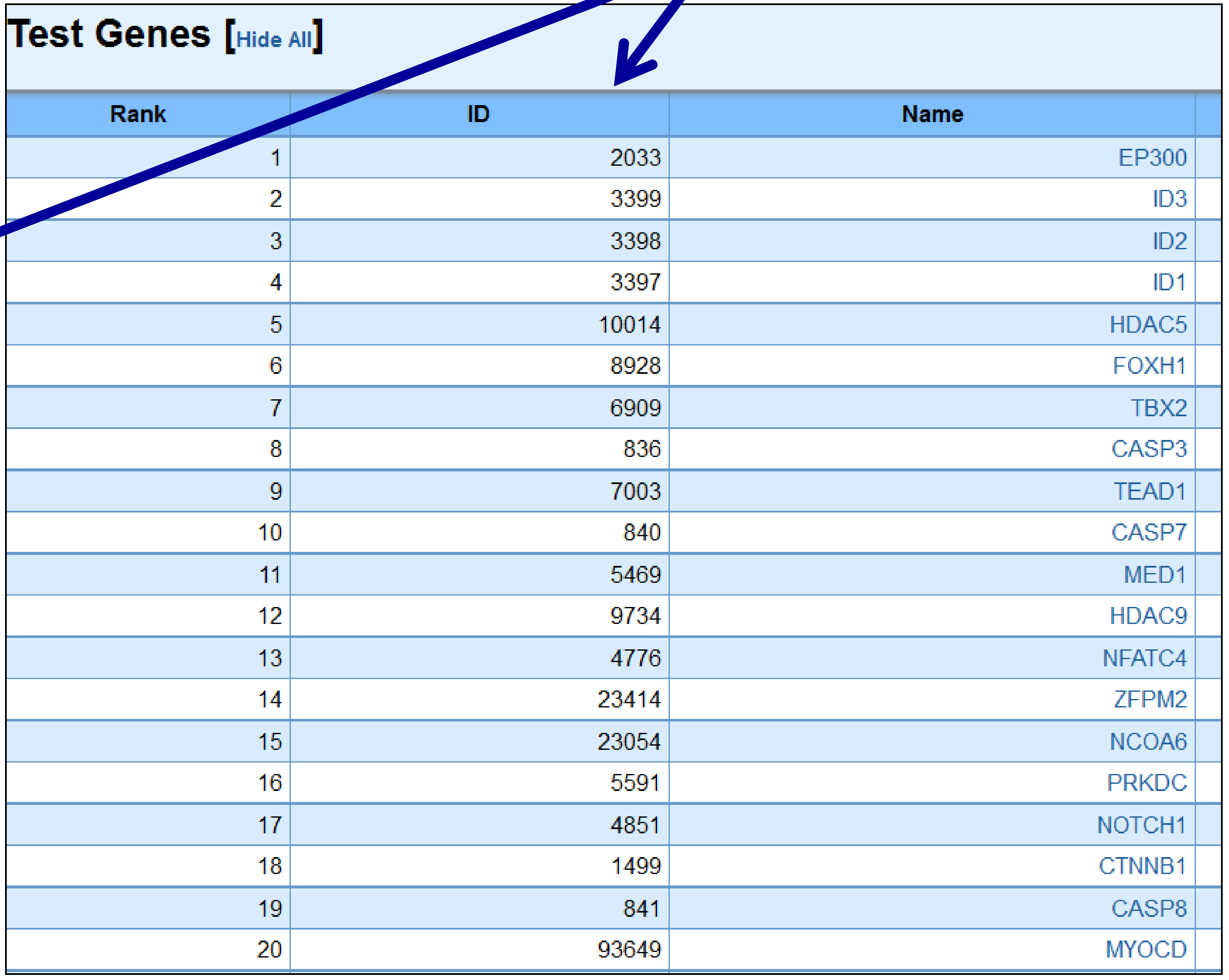
Bias (between 0 and 1): 0.3

Training gene neighborhood subnetwork visualization parameters

Neighborhood distance: 1

When training set is big, the training gene neighborhood subnetwork can be huge.





Test Genes [Hide All]

Rank	ID	Name
1	2033	EP300
2	3399	ID3
3	3398	ID2
4	3397	ID1
5	10014	HDAC5
6	8928	FOXH1
7	6909	TBX2
8	836	CASP3
9	7003	TEAD1
10	840	CASP7
11	5469	MED1
12	9734	HDAC9
13	4776	NFATC4
14	23414	ZFPM2
15	23054	NCOA6
16	5591	PRKDC
17	4851	NOTCH1
18	1499	CTNNB1
19	841	CASP8
20	93649	MYOCD

Chen et al., 2009

New tools – Variant prioritization

http://homes.esat.kuleuven.be/~bioiuser/eXtasy/

eXtasy Variant Prioritization

KU LEUVEN

eXtasy: Variant Prioritization by Genomic Data Fusion

What is eXtasy?
eXtasy is a pipeline for ranking nonsynonymous single nucleotide variants given a specific phenotype. It takes into account the putative deleteriousness of the variant, haploinsufficiency predictions of the underlying gene and the similarity of the given gene to known genes in the given phenotype.

Who develops eXtasy?
eXtasy was developed in the Bioinformatics group at the Department of Electrical Engineering of the University of Leuven (part of the iMinds Future Health Department). It was implemented by Alejandro Sifrim and Dusan Popovic under the supervision of Prof. Jan Aerts, Prof. Bart de Moor and Prof. Yves Moreau.

Run eXtasy online:

HPO term(s): (Comma-separated)

VCF file: Browse... No file selected.

Email: (Optional)

Output file name: (Optional)

Submit

http://compbio.charite.de/ExomeWalker

exomewalker

Prioritization of whole-exome data by random-walk analysis of protein-protein interactions

EXOME WALKER is a computational method to prioritise a set of candidates in exome sequencing projects that aim to identify novel Mendelian disease genes. Our approach involves filtering a Variant Call Format (VCF) file according to a number of user-definable criteria, for instance, off-target variants (those that are not located within or close to protein-coding exons) are removed.

Genes are prioritised according to a variant score (predicted pathogenicity, rarity, pattern of variants compatible with the assumed mode of inheritance) and to their vicinity to other genes that belong to the same disease-gene family within the protein protein interaction (PPI) network, using the Random-Walk method as described in Köhler et al. (2008) to determine similarity within the PPI network on the basis of the global characteristics of the network.

1. Upload VCF Data

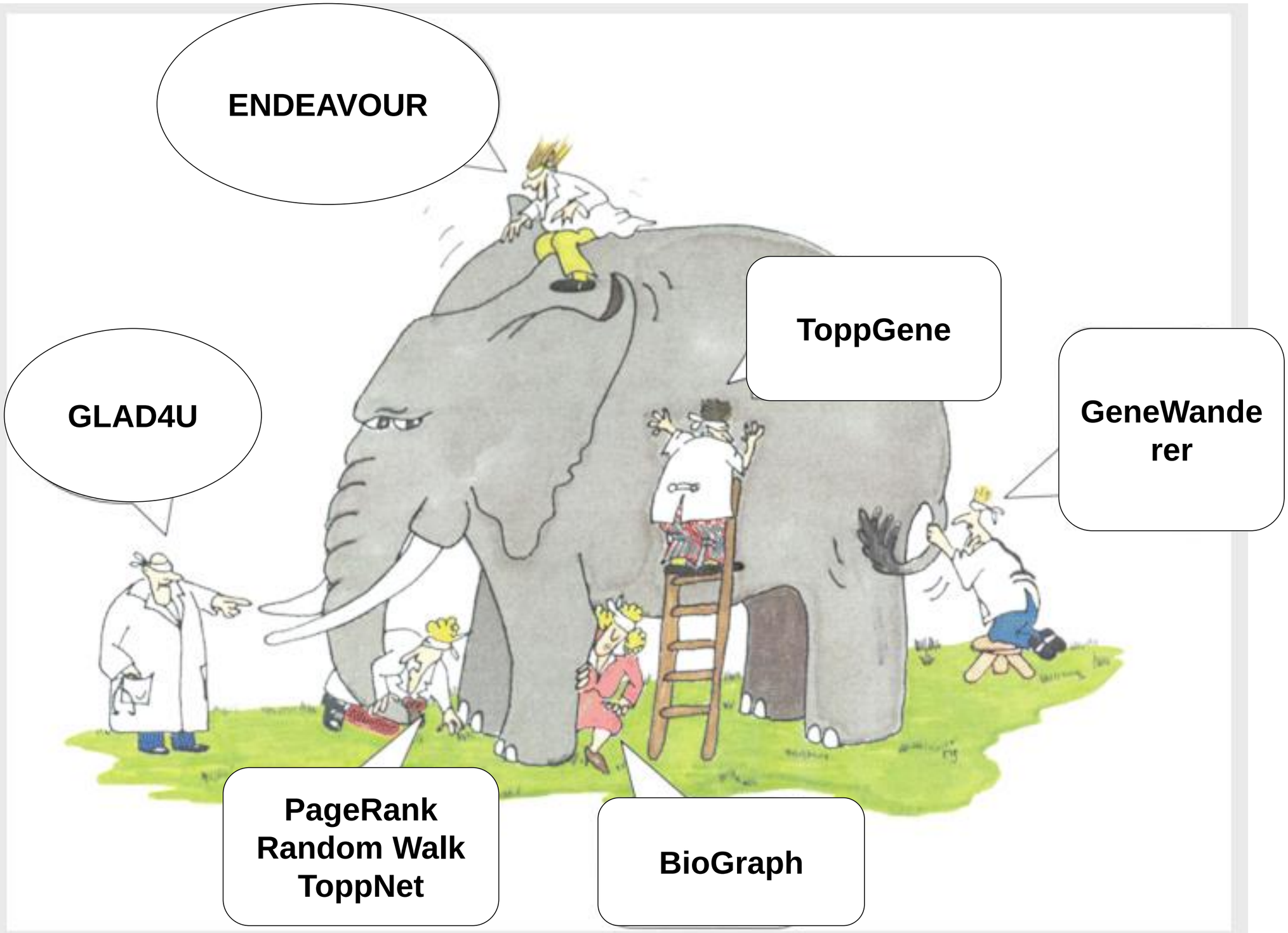
VCF files for single or multiple samples are supported. If you upload a multiple-sample VCF file, you will also be required to enter pedigree data in PED file format. The file suffix must be "VCF" or "vcf".

Choose VCF file: Browse... No file selected.

Send File

References & Further Reading

- Kann MG. 2010. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. Brief Bioinform. 11(1):96-110.
- Piro, R. M. & Di Cunto, F. 2012. Computational approaches to disease-gene prediction: rationale, classification and successes. FEBS J. 279: 678–696.
- ❖ Moreau Y, Tranchevent LC. 2012. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet 13: 523–536.
- ❖ Bromberg Y. 2013. Chapter 15: disease gene prioritization. PLoS Comput Biol. 9(4):e1002902.
- Navlakha S, Kingsford C. 2010. The power of protein interaction networks for associating genes with diseases. Bioinformatics 26(8):1057-63.
- Gonzalez MW, Kann MG. 2012. Chapter 4: Protein interactions and disease. PLoS Comput Biol. 8(12):e1002819
- Gilissen C, Hoischen A, Brunner HG, Veltman JA. 2012. Disease gene identification strategies for exome sequencing. Eur J Hum Genet. 20(5):490-7.



Cartoon - G. Renee Guzlas