Sino-German Workshop on
Multiscale Spatial Computational Systems Biology

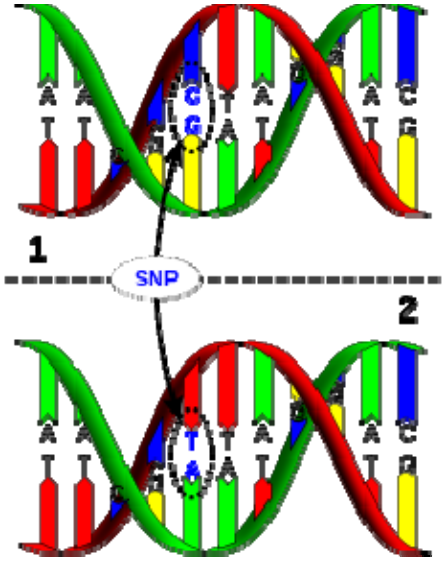
Identification of disease-causing single nucleotide variants in exome sequencing studies

Rui Jiang, PhD

Tsinghua University

Genotype-phenotype mapping

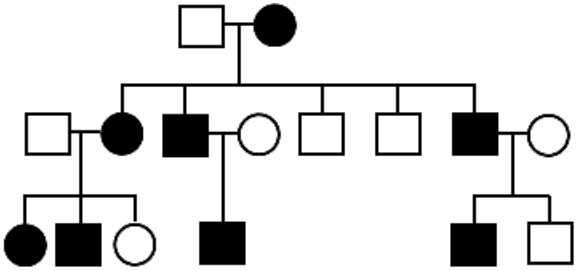
- Establishment of relationships between genotypes and phenotypes is a fundamental problem in genetics
- In the context of human inherited diseases, we care about **finding genetic variants underlying a specific disease**
  - Single nucleotide variants (nonsynonymous, synonymous, non-coding)
  - Insertions and deletions (indels)
  - Copy number variations (CNV)



http://upload.wikimedia.org/wikipedia/commons/thumb/2/2e/Dna-SNP.svg/220px-Dna-SNP.svg.png

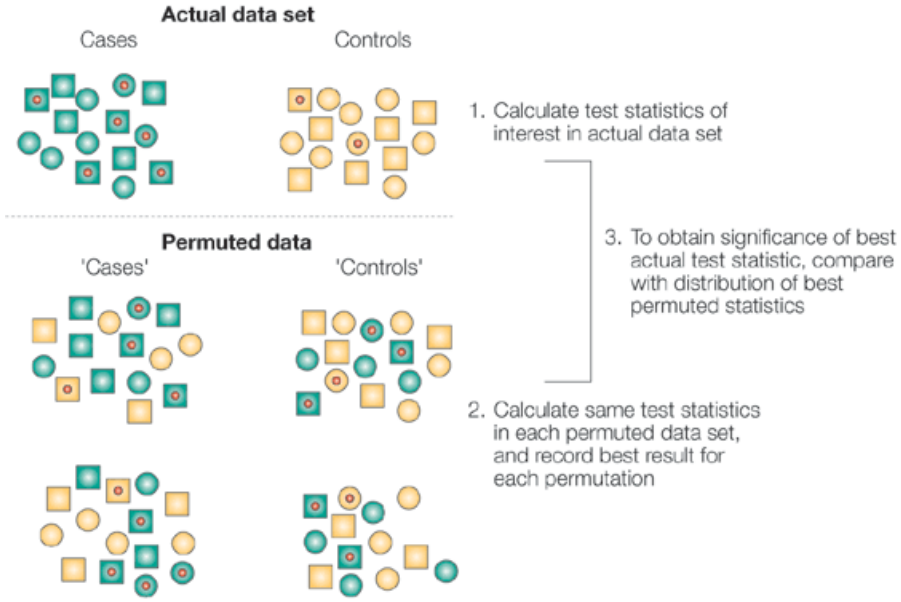
Traditional statistical approaches

- Family-based linkage analysis



http://www.sph.umich.edu/csg/abecasis/merlin/tour/parametric.gif

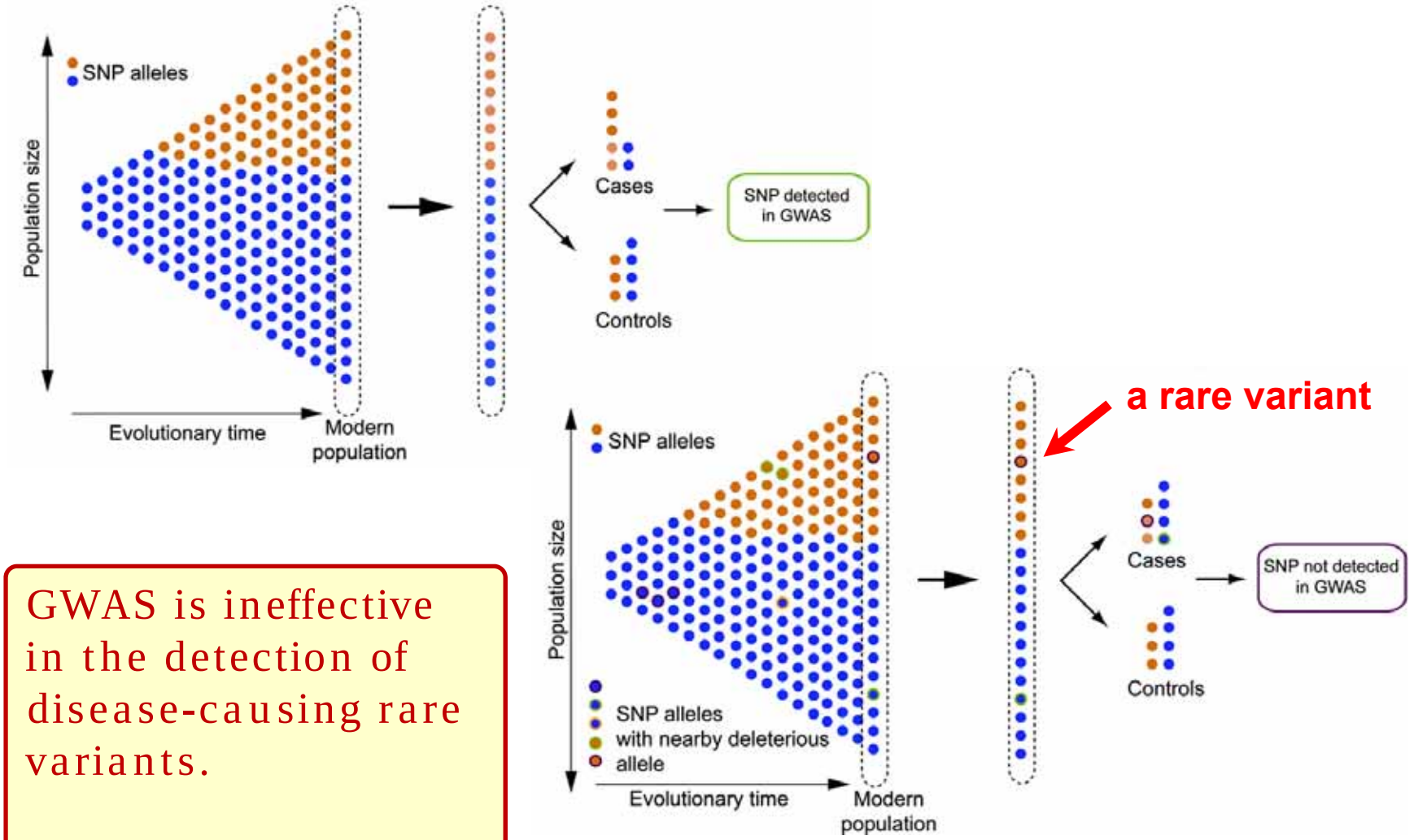
- Population-based association studies



http://www.nature.com/nrg/journal/v6/n2/images/nrg1521-i1.gif

Genome-wide association studies

- Common-disease common-variants assumption

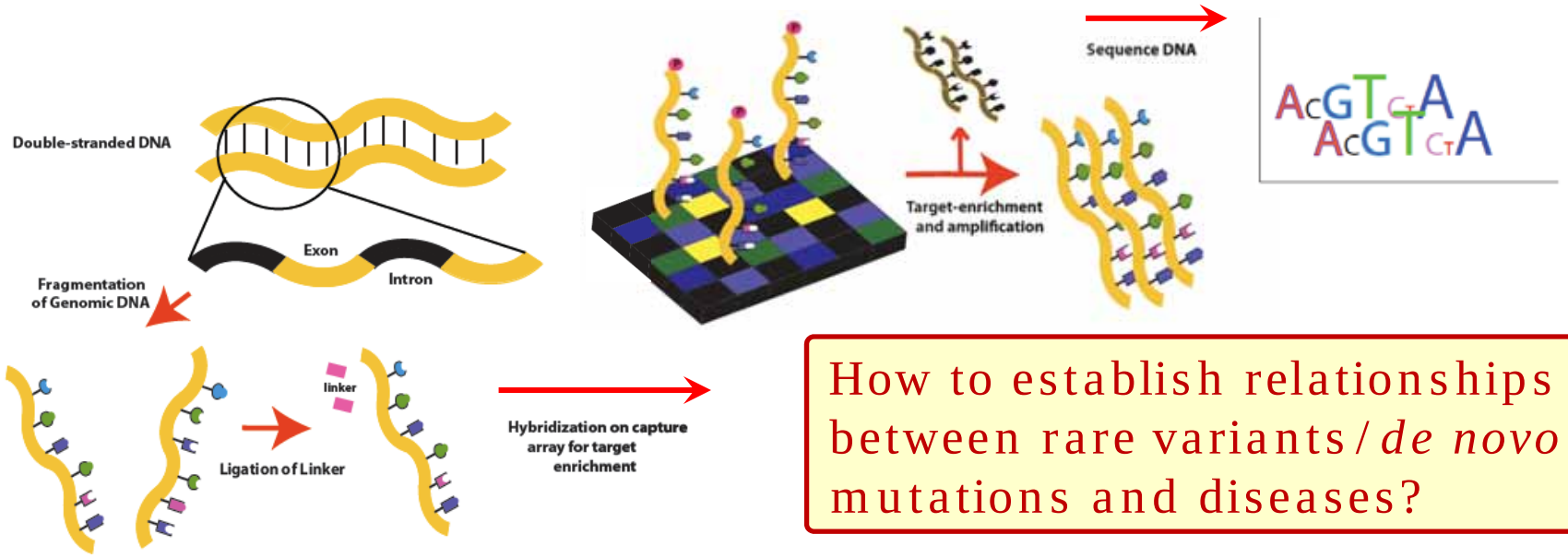


GWAS is ineffective in the detection of disease-causing rare variants.

http://www.neurology.org/content/72/6/558/F2.large.jpg

Exome sequencing

- Selectively sequence coding regions of a genome
- Effective in detecting nonsynonymous SNVs
  - Non-negligible fraction of rare SNVs
  - Presence of *de novo* mutations
- Succeeded in identifying causal variants for Mendelian disorder
- Has the potential to locate causative genes in complex diseases



How to establish relationships between rare variants/*de novo* mutations and diseases?

Prediction of deleterious variants

- Binary classification of functional implications of ns SNVs
  - Two categories: **Deleterious** / **Neutral**
  - Features: Sequence, structure, annotations, …
  - Machines: Naïve Bayes, logistic regression, SVM, random forest, …
- A lot of methods existing
  - SIFT
  - PolyPhen
  - GERP
  - PhyloP
  - MSRV
  - SinBaD
  - …
- Whole exome predictions
  - dbNSFP

Which disease is the alternation of the protein function responsible for?

Prioritization of candidate genes

- One-class novelty learning for disease genes for a query disease
  - One class: Disease-associated
  - Principle: Guilt-by-association (genes associated with the same disease are correlated in their functions)
  - Method: Calculate functional similarity of candidates to seed genes and then rank the candidates
- A lot of existing methods
  - Use **seed genes**: Endeavour, Random walk on a PPI network, …
  - Use **phenotype similarity**: CIPHER, AlignPI, MaxIF, Bridge, MaxDrive, …

Association between a gene and a disease does not mean every genetic variant in the gene is causative.

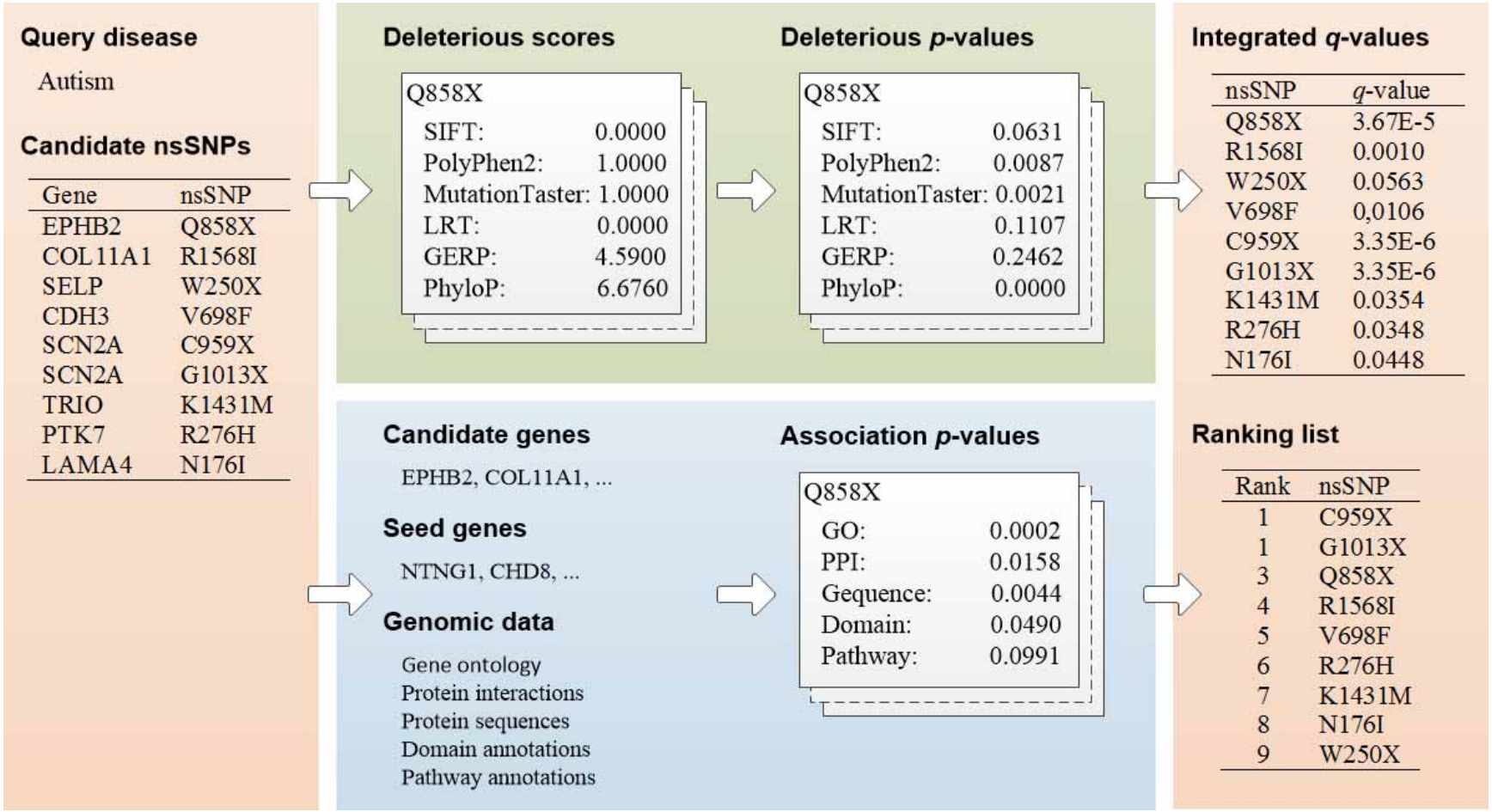
Limitations of existing approaches

- Genome-wide association studies
  - Ineffective in detecting disease-causing rare/de novo variants
- Prediction of deleterious variants
  - Unable to tell the specific disease that the alternation of the protein function is responsible for
- Prioritization of candidate genes
  - Unable to tell causative genetic variants in the disease-associated gene

We developed bioinformatics approaches, **SPRING and snvForest**, to **integrate functional damaging effects of SNVs and disease-gene association information to pinpoint disease-causing nonsynonymous SNVs in exome sequencing studies.**

SPRING

Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide polymorphisms



Wu et al. *PLoS Genetics*, 2014

Inputs

Query disease

Autism

Candidate nsSNPs

Gene	nsSNP
EPHB2	Q858X
COL11A1	R1568I
SELP	W250X
CDH3	V698F
SCN2A	C959X
SCN2A	G1013X
TRIO	K1431M
PTK7	R276H
LAMA4	N176I

Outputs

Query disease

Autism

Candidate nsSNPs

Gene	nsSNP
EPHB2	Q858X
COL11A1	R1568I
SELP	W250X
CDH3	V698F
SCN2A	C959X
SCN2A	G1013X
TRIO	K1431M
PTK7	R276H
LAMA4	N176I

Integrated *q*-values

nsSNP	*q*-value
Q858X	3.67E-5
R1568I	0.0010
W250X	0.0563
V698F	0,0106
C959X	3.35E-6
G1013X	3.35E-6
K1431M	0.0354
R276H	0.0348
N176I	0.0448

Ranking list

Rank	nsSNP
1	C959X
1	G1013X
3	Q858X
4	R1568I
5	V698F
6	R276H
7	K1431M
8	N176I
9	W250X

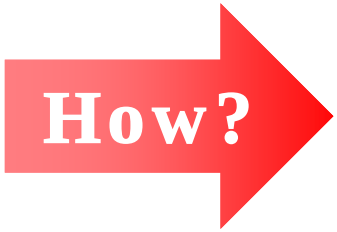
How to obtain outputs from inputs?

Query disease

Autism

Candidate nsSNPs

Gene	nsSNP
EPHB2	Q858X
COL11A1	R1568I
SELP	W250X
CDH3	V698F
SCN2A	C959X
SCN2A	G1013X
TRIO	K1431M
PTK7	R276H
LAMA4	N176I



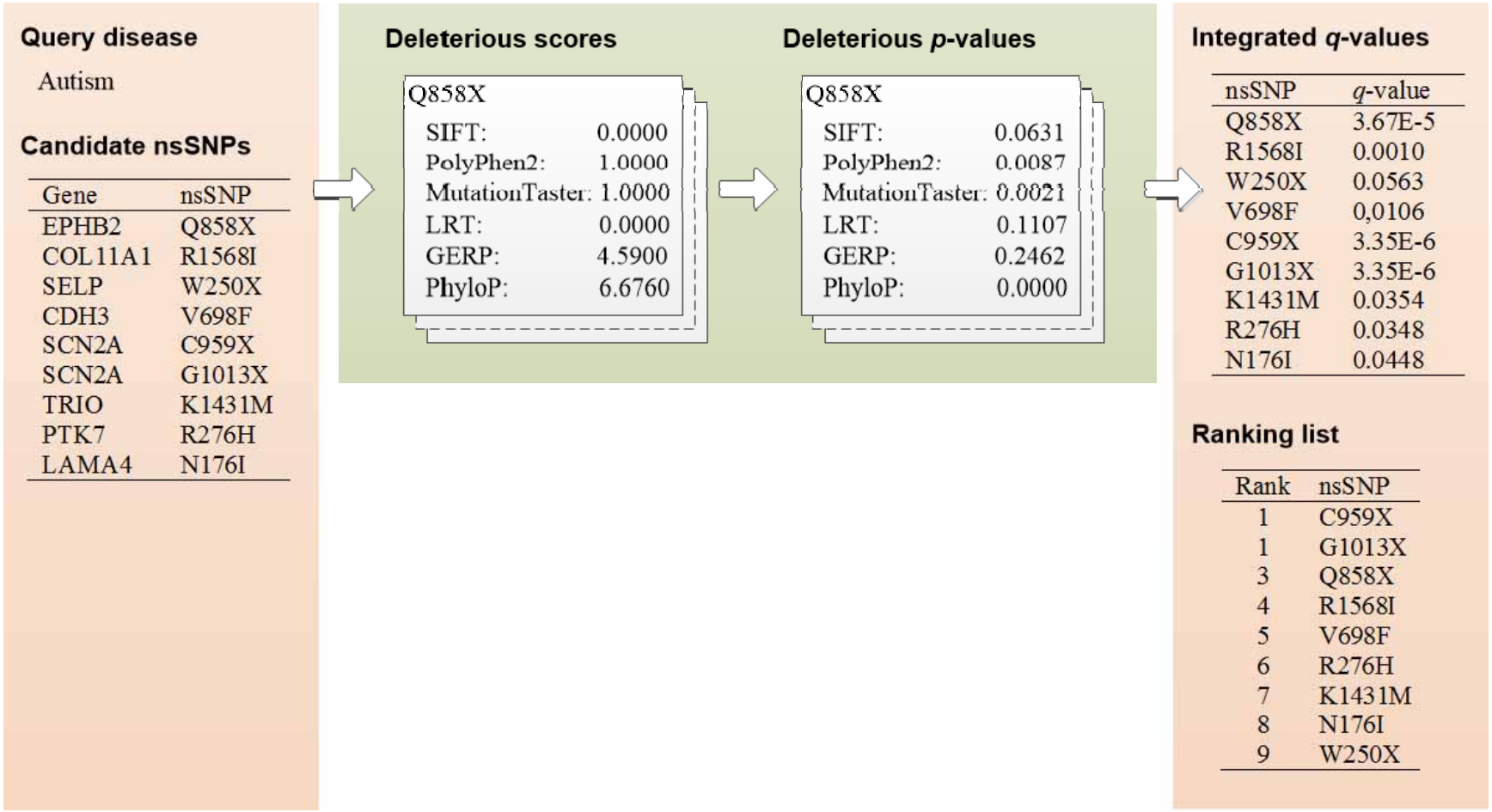
Integrated *q*-values

nsSNP	*q*-value
Q858X	3.67E-5
R1568I	0.0010
W250X	0.0563
V698F	0,0106
C959X	3.35E-6
G1013X	3.35E-6
K1431M	0.0354
R276H	0.0348
N176I	0.0448

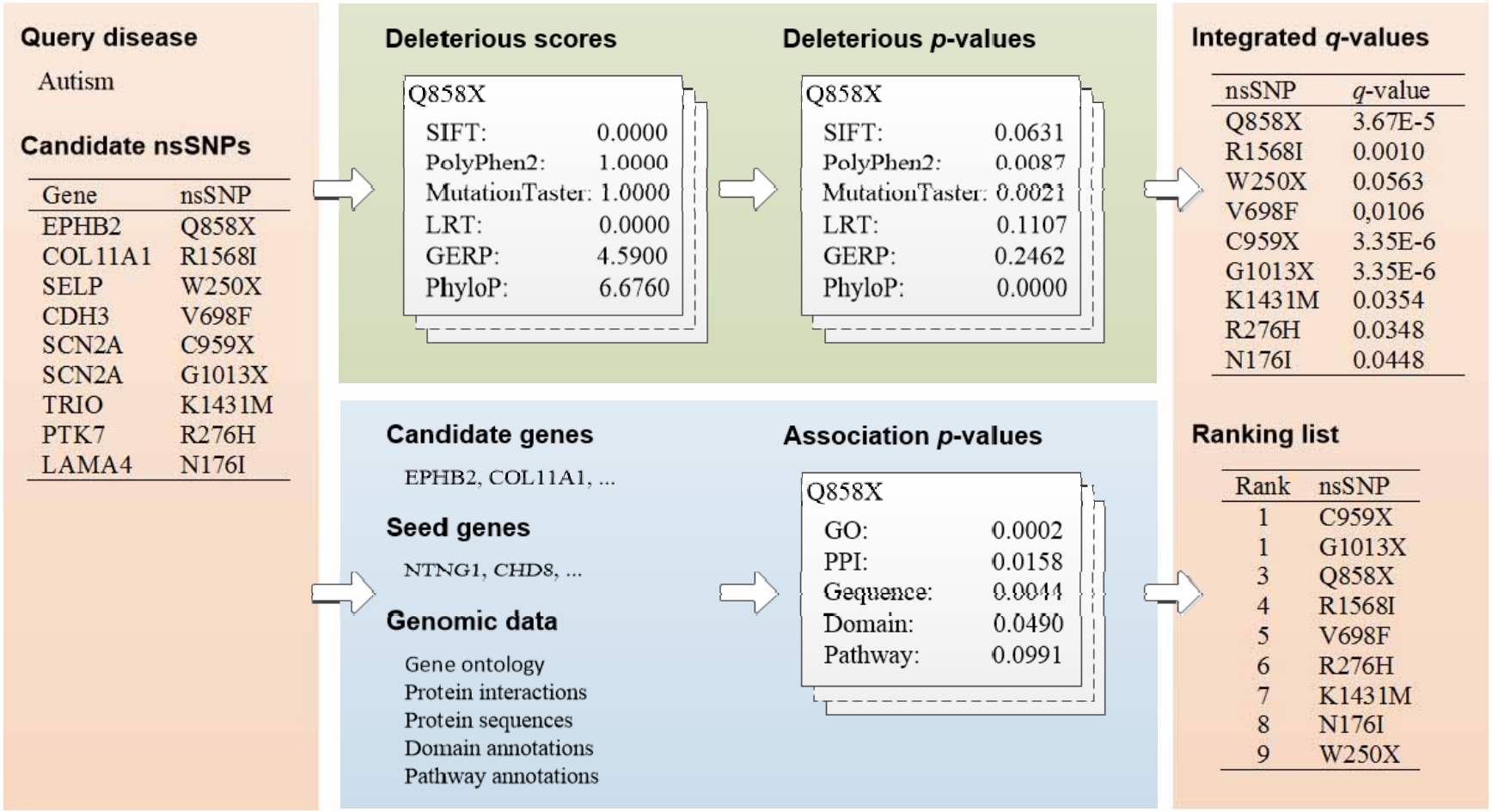
Ranking list

Rank	nsSNP
1	C959X
1	G1013X
3	Q858X
4	R1568I
5	V698F
6	R276H
7	K1431M
8	N176I
9	W250X

Deleterious scores (pre-calculated)



Association scores (derived)



Integration of multiple *p*-values

- Fisher's combined probability test

$$T = -2\sum_{i=1}^{k} \log p_i$$

- Independent case (chi-squared null)

$$T \sim \chi^2_{2k}$$

- Dependent case (assume scaled chi-squared null)

$$T \sim r\chi^2_{2v}$$

- Estimate the two parameters by the method of moments

$$\hat{r} = 1 + \frac{1}{2k}\sum_{i<j} \mathrm{Cov}(V_i, V_j) \text{ and } \hat{v} = 2k / r$$

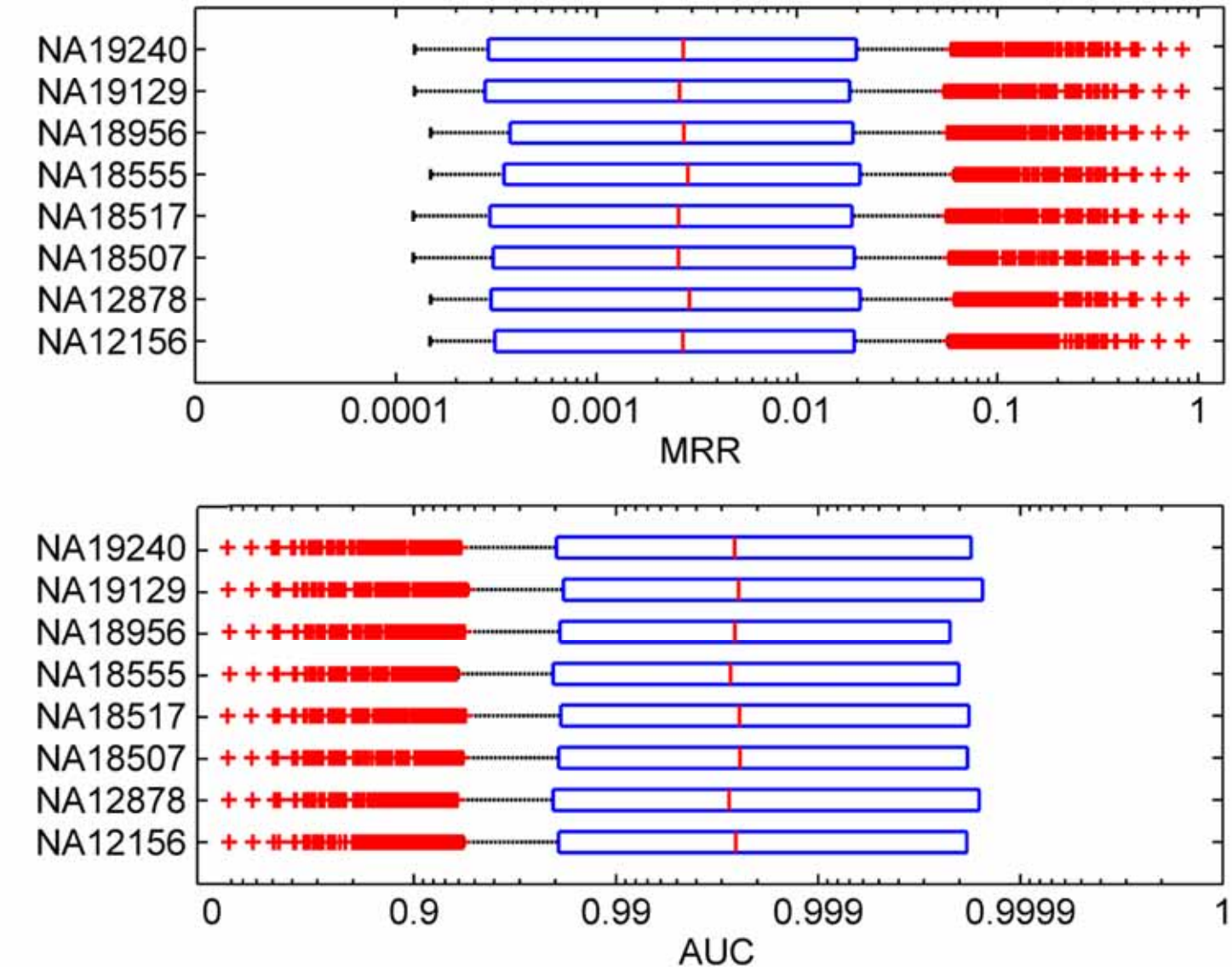
Other issues

- Missing data
  - Ignore the missing data sources in the calculation
  - Total number of *p*-values to be combined will then decrease accordingly
- Multiple testing correction
  - Control the positive false discovery rate (pFDR) by *q*-values

Data sources

- Diseases
  - 1,436 diseases (1378 Mendelian, 58 complex) from OMIM
- SNVs
  - 12,610 disease-causing and 23,403 neutral nsSNPs from Swiss-Prot
  - 6 types of deleterious scores extracted from dbNSFP
- Genomic data
  - PPI: similarities between 9,966 proteins according to STRING
  - GO: similarities between 14,283 genes according to GO
  - Sequence: similarities between 20,281 proteins according to UniProt
  - Domain: similarity between 12,713 proteins according to Pfam
  - Pathway: similarity between 5,951 genes according to KEGG

Effective in exome sequencing studies



Autism (MIM: 209850)

- A neurological and developmental disorder
- Usually appears in childhood, especially the first three years of life
- A child with autism appears to live in their own world, showing little interest in others and a lack of social awareness, and having problems in communication
- A complex genetic disease with a strong genetic basis
- Several genes and variants are involved in the development of autism

http://fitkidsok.org/stories-matter-autism-spectrum-disorders/

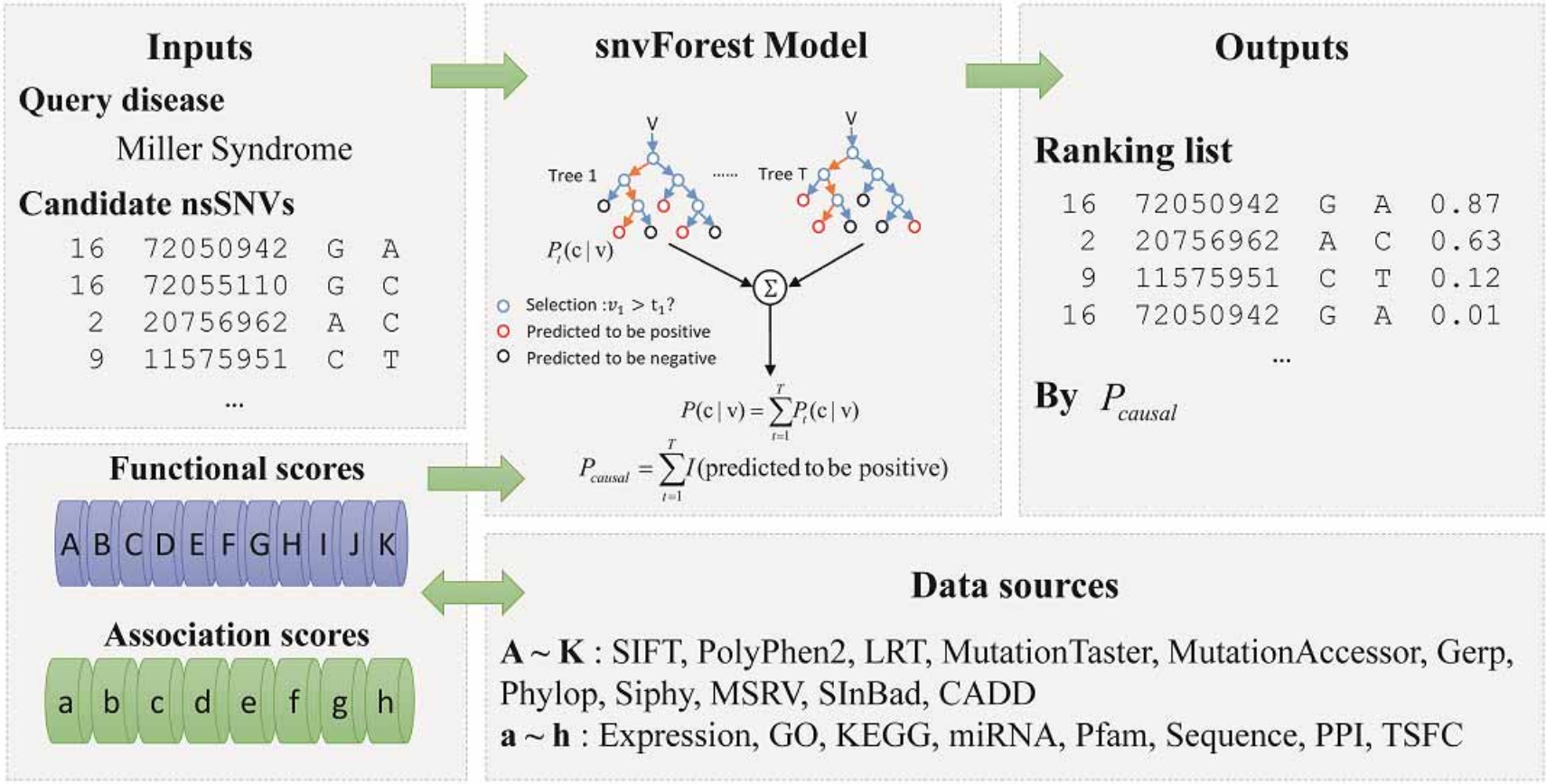
PMID 22495306

192 candidate nonsynonymous *de novo* mutations

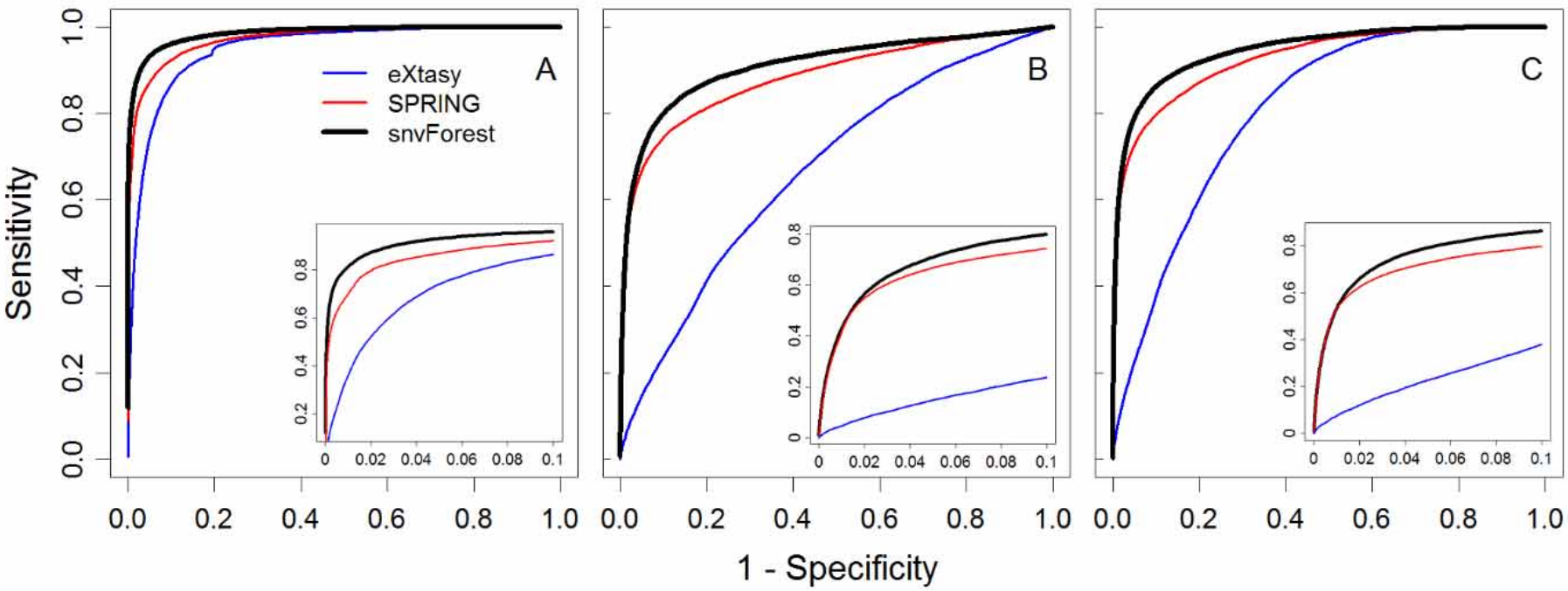
Chr	Pos	Ref	Alt	Gene	Type	Rank	pvalue	qvalue
2	**166210819**	**G**	**T**	**SCN2A**	**Nonsense**	1	6.8691E-13	1.4700E-10
2	**166201379**	**C**	**A**	**SCN2A**	**Nonsense**	2	3.1799E-12	3.4025E-10
1	23236941	C	T	EPHB2	Nonsense	3	1.0212E-10	7.2849E-09
3	62356982	G	A	FEZF2	Missense	4	4.7601E-10	2.5467E-08
1	231829616	C	T	DISC1	Missense	5	4.0285E-08	1.7242E-06
2	233396326	G	A	CHRND	Missense	6	1.7059E-05	6.0843E-04
X	41524649	C	G	CASK	Missense	7	2.6776E-05	8.1858E-04
5	92929487	G	A	NR2F1	Missense	8	6.6080E-05	1.7676E-03
3	7620458	G	A	GRM7	Missense	9	8.4673E-05	2.0133E-03
3	11067472	C	T	SLC6A1	Missense	10	1.0090E-04	2.0546E-03
5	1394870	G	A	SLC6A3	Missense	10	1.0561E-04	2.0546E-03
6	43099768	G	A	PTK7	Missense	12	1.4090E-04	2.5128E-03
6	112513029	T	A	LAMA4	Missense	13	1.7677E-04	2.9098E-03
5	14394220	A	T	TRIO	Missense	14	2.6395E-04	4.0346E-03
10	61833827	A	G	ANK3	Missense	15	3.1738E-04	4.5279E-03
14	95574729	G	A	DICER1	Missense	16	4.1978E-04	5.6145E-03
12	332363	G	A	SLC6A13	Missense	17	4.6704E-04	5.8793E-03
17	29684348	C	T	NF1	Missense	18	5.2068E-04	5.9417E-03
22	36684831	C	T	MYH9	Missense	18	5.2753E-04	5.9417E-03
10	102740667	T	C	SEMA4G	Missense	20	7.4687E-04	7.9915E-03

snvForest

- Fisher's method — unsupervised data integration
  - *p*-value of a data source is approximately the ranking score
- How to make use of label information (known causal / neutral)?

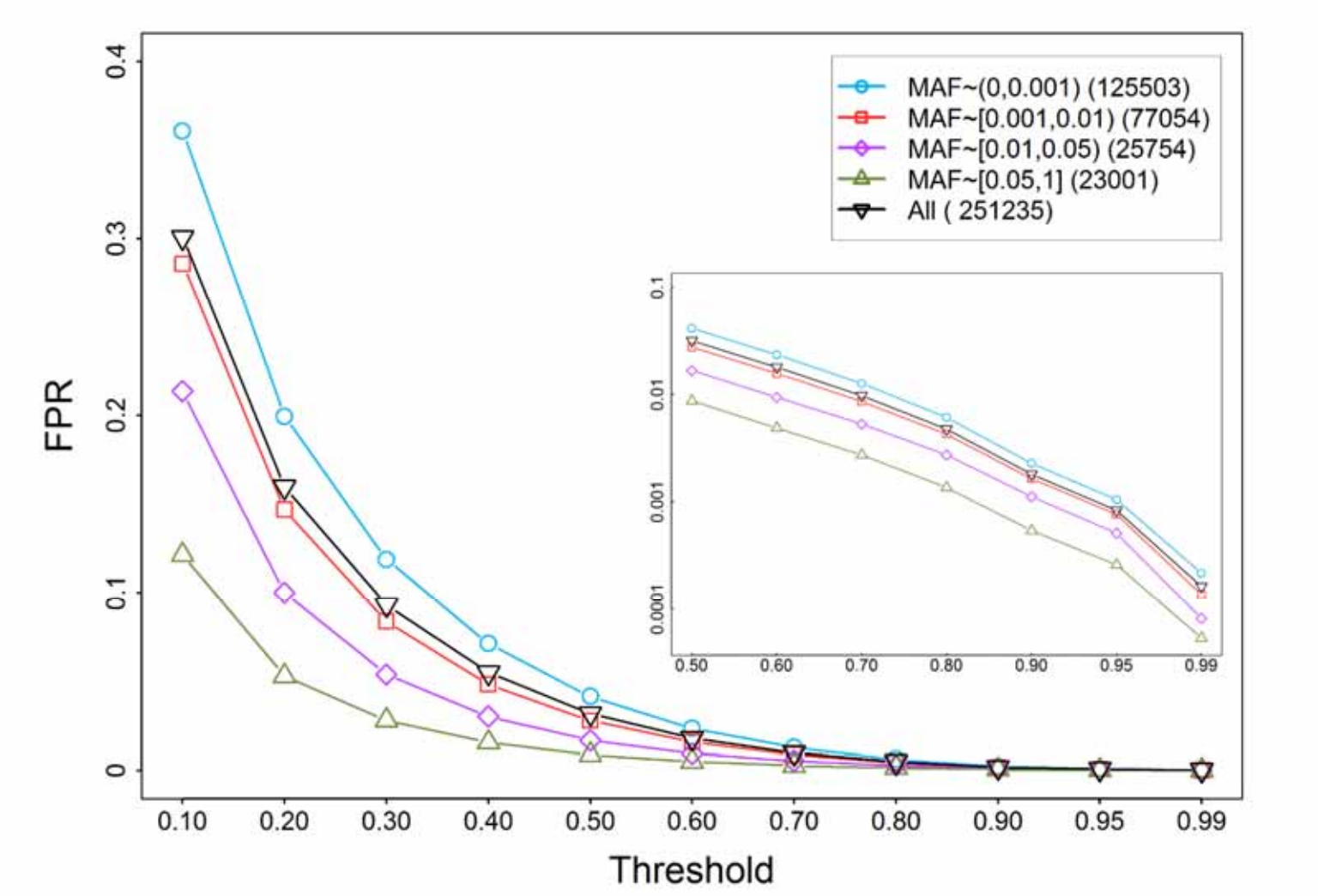


High performance



eXtasy: Sifrim, A. et al. *Nature Methods*, 2013
Spring: Wu et al. *PLoS Genetics*, 2014
snvForest: Wu et al. *Scientific Reports*, 2015

Well-controlled false positive rate



Real exome sequencing data

PMID	Candidate Mutations	Funcional Mutations	Rank		*p*-value	
			Top 10	Top 20	Top 10	Top 20
23934111	192	30	8	13	5.3×10^{-6}	9.4×10^{-8}
23033978	77	16	5	10	2.8×10^{-2}	5.5×10^{-4}
23020937	126	17	5	7	4.1×10^{-3}	6.4×10^{-3}

23934111: epileptic encephalopathies

23033978: Intellectual disability

23020937: Intellectual disability

dbWGFP

A database and web server of human whole-genome single nucleotide variants and their functional predictions

Introduction

The recent advancement of the next generation sequencing technology has enabled the fast and low-cost detection of all genetic variants spreading across entire human genomes, making the application of whole-genome sequencing a tendency in studies of disease-causing genetic variants. Nevertheless, there still lacks a repository that collects predictions of functionally damaging effects of human genetic variants, though it has been well recognized that such predictions plays a central role in the analysis of whole-genome sequencing data.

To fill in this gap, we developed **dbWGFP** (*a database of human whole-genome single nucleotide variants and their functional predictions*) that contains functional predictions and annotations of more than 8.5 billion possible human whole-genome single nucleotide variants. Specifically, this database integrates 32 functional predictions calculated by 13 popular computational methods, 15 conservation features derived from 4 conservation calculation approaches, and 44 valuable annotations obtained from the ENCODE project, accompanied with a highly efficient search program.

dbWGFP offers two web services for retrieving predictions and annotations for human whole-genome single nucleotide variants. In the step-by-step mode, you can upload a file containing variants and retrieve results online. In the batch mode, you can upload a file containing your email address and variants and retrieve results via your email.

dbWGFP offers two versions for downloading. The lite version includes prediction scores for human whole-genome single nucleotide variants. The full version includes both predictions and annotations. Both versions include a search program that can extract predictions and/or annotations in a highly efficient way. Different versions of dbWGFP are also archived for easy access.

If you have any comments, suggestions and questions, please do not hesitate to send us an email.

Please cite: Rui Jiang*, Jiaxin Wu, Lianshuo Li, Mengmeng Wu, Zhuo Liu, Wanwen Zeng, dbWGFP: a database of human whole-genome single nucleotide variants and their functional predictions, *In submission*, 2014.

dbWGFP

A database and web server of human whole-genome single nucleotide variants and their functional predictions

- Database
  - 8,576,251,873 (nearly **8.58 billion**) human SNVs
  - **48** prediction scores (lite and full versions)
  - **44** annotations (full version only)
- Web server
  - An ultra-fast search program
  - An online query service
  - A download service

33 functionally damaging effect scores

- Calculated by 13 prediction methods

Method	Source	Website
Grantham	CADD	—
SIFT	dbNSFP	http://sift.jcvi.org
PolyPhen-2	dbNSFP	http://genetics.bwh.harvard.edu/pph2
LRT	dbNSFP	http://www.genetics.wustl.edu/jflab/lrt_query.html
MutationTaster	dbNSFP	http://www.mutationtaster.org
Mutation Assessor	dbNSFP	http://mutationassessor.org
FATHMM	dbNSFP	http://fathmm.biocompute.org.uk
RadialSVM	dbNSFP	http://sites.google.com/site/jpopgen/dbNSFP
LR	dbNSFP	http://sites.google.com/site/jpopgen/dbNSFP
CADD	CADD	http://cadd.gs.washington.edu
GWAVA	GWAVA	https://www.sanger.ac.uk/sanger/StatGen_Gwava
MSRV	MSRV	http://bioinfo.au.tsinghua.edu.cn/msrv
SinBaD	SinBaD	http://tingchenlab.cmb.usc.edu/sinbad
phastCons	UCSC	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way
PhyloP	UCSC	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way
GERP++	GERP	http://mendel.stanford.edu/SidowLab/downloads/gerp
SiPhy	SiPhy	http://www.broadinstitute.org/genome_bio/siphy

15 conservation scores

- Calculated by 4 prediction methods

Method	Source	Website
Grantham	CADD	—
SIFT	dbNSFP	http://sift.jcvi.org
PolyPhen-2	dbNSFP	http://genetics.bwh.harvard.edu/pph2
LRT	dbNSFP	http://www.genetics.wustl.edu/jflab/lrt_query.html
MutationTaster	dbNSFP	http://www.mutationtaster.org
Mutation Assessor	dbNSFP	http://mutationassessor.org
FATHMM	dbNSFP	http://fathmm.biocompute.org.uk
RadialSVM	dbNSFP	http://sites.google.com/site/jpopgen/dbNSFP
LR	dbNSFP	http://sites.google.com/site/jpopgen/dbNSFP
CADD	CADD	http://cadd.gs.washington.edu
GWAVA	GWAVA	https://www.sanger.ac.uk/sanger/StatGen_Gwava
MSRV	MSRV	http://bioinfo.au.tsinghua.edu.cn/msrv
SinBaD	SinBaD	http://tingchenlab.cmb.usc.edu/sinbad
phastCons	UCSC	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way
PhyloP	UCSC	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way
GERP++	GERP	http://mendel.stanford.edu/SidowLab/downloads/gerp
SiPhy	SiPhy	http://www.broadinstitute.org/genome_bio/siphy

44 annotations

- dbSNP, CADD, ENCODE and 1000 Genomes Project
- Basic information
  - ID, ancestral base, annotation type, consequence type of the variants, ENSEMBL gene ID, ENSEMBL transcript ID, CCDS ID, gene name, protein accession number and ID in Uniprot database, reference codon, reference amino acid, substituted amino acid
- 1000 Genomes Project related annotations
  - validated status, project phase, common variant or not, and different types of allele frequency for different type of populations
- Advanced annotations
  - distance to the closest Transcribed Sequence Start (TSS), distance to the closest Transcribed Sequence End (TSE), amino acid position, codon position, base position from transcription start, relative position in transcript, base position from coding start, relative position in coding sequence, distance to splice site, closest splice site is ACCEPTOR or DONOR, total number of exons, and total number of introns

dbWGFP

A database and web server of human whole-genome single nucleotide variants and their functional predictions

Step 1 -- Upload Variants

Upload a txt or vcf file (uncompressed or compressed in gz, zip or rar format) by clicking Browse and then Submit (e.g., variants.txt, variants.vcf, variants.txt.gz, variants.vcf.gz, variants.zip, variants.rar, more examples).

Upload a file:

100MB maximum

Browse ... Submit

For testing purpose, enter a small number of variants in the text area below and then click Submit.

Enter variants:

```
# Replace text below with your variants.
# Lines starting with # are comments and will be ignored.
# Each line is an SNV, specified by DNA coordinate.
# Each line has two to four tab delimited columns.
# Two column format:    CHR POS
# Three column format: CHR POS REF
# Four column format:   CHR POS REF ALT
# where,
# CHR:  The chromosome (1-22, X, Y) in which the SNV occurs.
# POS:  The position where the SNV occurs.
# REF:  The reference nucleotide.
# ALT:  The alteration nucleotide.
```

1000 lines maximum

Submit

dbWGFP

A database and web server of human whole-genome single nucleotide variants and their functional predictions

Step 2 -- Check our website or Check your email

Your job has been submitted to the server. You can retrieve the result by visiting

http://bioinfo.au.tsinghua.edu.cn/dbwgfp/querytask.php?taskname=8467507133

Alternatively, enter your email address below, click Send email and check later.

Your name: Yourname

Your email: youremail@youruniversity.edu

Send email

Server messages:

```
[08/14/14 01:52:53] Search started.
[08/14/14 01:52:53] 192 SNVs loaded.
[08/14/14 01:52:55] 192 SNVs processed.
[14/08/14 01:52:55] Results generated.
```

95%

dbWGFP

A database and web server of human whole-genome single nucleotide variants and their functional predictions

Summary

Task: **8785606797** Submission time: **2014/08/14, 02:45:53** File format: **txt (4 columns)** File size: **3,507** bytes Task status: **Finished**

Results for downloading

Format	Predictions only		Predictions and annotations		Description
zip	dbWGFP.lite.zip	(21,011 bytes)	dbWGFP.full.zip	(39,794 bytes)	Suitable for Windows
rar	dbWGFP.lite.rar	(17,059 bytes)	dbWGFP.full.rar	(32,274 bytes)	Suitable for Windows
gz	dbWGFP.lite.txt.gz	(20,865 bytes)	dbWGFP.full.txt.gz	(39,648 bytes)	Suitable for MacOSX and Linux
bz2	dbWGFP.lite.txt.bz2	(18,121 bytes)	dbWGFP.full.txt.bz2	(34,568 bytes)	Suitable for MacOSX and Linux

Preview of the first 100 lines (show more in a separate window) (show all in a separate window)

Chrom	Pos	Ref	Alt	dbSNP_id	Anc	Type	Length	AnnoType	Consequence
2	166848563	C	G	.	C	SNV	0	CodingTranscript	NON_SYNONYMOU
2	166903480	G	A	rs121917929	G	SNV	0	CodingTranscript	NON_SYNONYMOU
2	166894356	C	T	.	C	SNV	0	CodingTranscript	NON_SYNONYMOU
2	166198975	G	A	.	G	SNV	0	CodingTranscript	NON_SYNONYMOU
2	166852575	G	T	.	G	SNV	0	CodingTranscript	NON_SYNONYMOU
12	52082568	G	A	.	G	SNV	0	CodingTranscript	NON_SYNONYMOU
12	52159534	T	A	.	T	SNV	0	CodingTranscript	NON_SYNONYMOU
9	130438189	G	A	.	G	SNV	0	CodingTranscript	NON_SYNONYMOU
X	18606157	G	A	.	G	SNV	0	CodingTranscript	NON_SYNONYMOU
9	130425622	C	T	.	C	SNV	0	CodingTranscript	NON_SYNONYMOU
9	130444768	G	A	rs121918317	G	SNV	0	CodingTranscript	NON_SYNONYMOU
9	130434370	C	T	.	C	SNV	0	CodingTranscript	NON_SYNONYMOU
X	18598064	C	T	.	C	SNV	0	CodingTranscript	NON_SYNONYMOU

Acknowledgements

- Prof. Yanda Li, Tsinghua University
- Jiaxin Wu, PhD student
- Mengmeng Wu, PhD student
- Wanwen Zeng, PhD student
- Lianshuo Li, Master student
- Zhuo Liu, Master student

- National Basic Research Program of China (2012CB316504)
- National High Technology Research and Development Program of China (2012AA020401)
- National Natural Science Foundation of China (61175002)
- Tsinghua National Laboratory for Information Science and Technology
- Tsinghua University

Thank you very much