Algorithms to elucidate causative gene mutations in rare disorders

PHEVOR & Phen-Gen

Journal Club

Marie-Angela Wulf

9.9.2014

Overview

Motivation

- is there a gene mutation leading to disease?
- if so, which one?
- is it inherited?
- is there a risk of replication in future siblings?
- possibility of pre-natal diagnostics?

[1]

[1] http://saalfeld.otz.de/web/lokal/leben/detail/-/specific/Familie-aus-Saalfeld-mit-behinderter-Tochter-braucht-Hilfe-291757532

Motivation

- is there a gene mutation leading to disease?
- if so, which one?
- is it inherited?
- is there a risk of replication in future siblings?
- possibility of pre-natal diagnostics?

2

- Problem: very rare diseases => very low number of cases
- only in approx. 50% of cases, the causal gene is known
- gap between sequencing and analysis capabilities

[2] http://saalfeld.otz.de/web/lokal/leben/detail/-/specific/Familie-aus-Saalfeld-mit-behinderter-Tochter-braucht-Hilfe-291757532

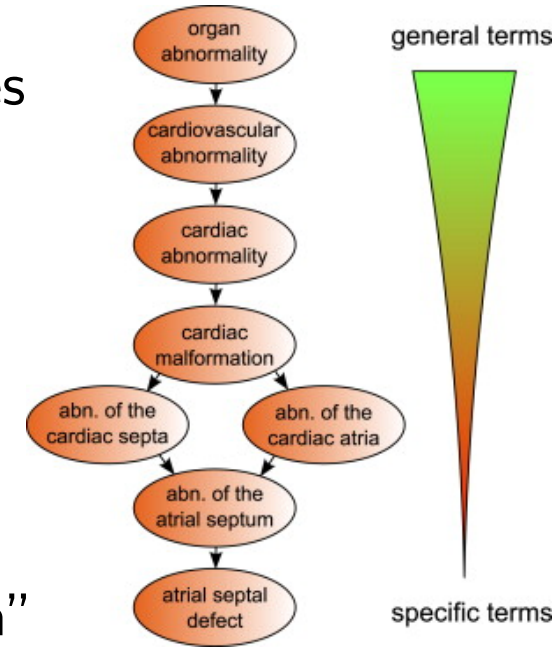
short repetition in genetics

Exome	coding region of the genome
start-loss	mutation in the ATG start codon that prevents the start of translation
stop-gain	mutation that induces a new stop-codon
stop-loss	loss of the normal stop codon, leading to larger transcript
splice site	where introns are cut out an exons are put together for the final mRNA
nonsynonymous	mutation leads to incorporation of a different amino acid
synonymous	mutation leads to incorporation of the same amino acid
indel	small insertion or deletion

Human Phenotype Ontology (HPO)

Ontology computational representation of a domain of knowledge based upon a controlled, standardized vocabulary for describing entities and the semantic relationships between them

Annotation is the process of assigning ontology terms (concepts) for the description of objects



- terms are related to parent terms by “is a” relationship
- different parent terms possible

Human Phenotype Ontology (HPO)

- introduced by Robinson et al in 2008
- standardized vocabulary of phenotypic abnormalities encountered in human disease
- Each term in the HPO describes a phenotypic abnormality
- ontology terms are used for annotation of diseases
- almost 50,000 annotations to 4813 diseases are provided

how to find the causative gene?

- sequencing
- comparison to reference data set
- but: also healthy individuals have about 100 genuine loss-of-function variants with about 20 genes completely inactivated
- $\Rightarrow$ need of a good post-sequencing processing step in order to find the causative mutation

algorithms that are currently available

- VAAST (2011)
- eXtasy (2013)
- FunSeq (2013)
- PHIVE (2014)
- PHEVOR (2014)
- Phen-Gen (2014)

algorithms that are currently available

- VAAST (2011)
- eXtasy (2013)
- FunSeq (2013)
- PHIVE (2014)
- *PHEVOR (2014)*
- *Phen-Gen (2014)*

principles of variant-prioritization tools at the example of VAAST

1. Sequencing
2. Comparison with reference genome or exome sequence => variants
3. ranking of the variants according to different criteria
    1. frequency of the variant in reference genomes
    2. predicted severity of amino acid changes
    3. location of the variant in the genome
    4. inheritance and penetrance patterns

This leads to causal gene top rank in 62% of cases for dominant disorders and 78% for recessive disorders.

PHEVOR

Phevor Combines Multiple Biomedical Ontologies for Accurate Identification of Disease-Causing Alleles in Single Individuals and Small Nuclear Families

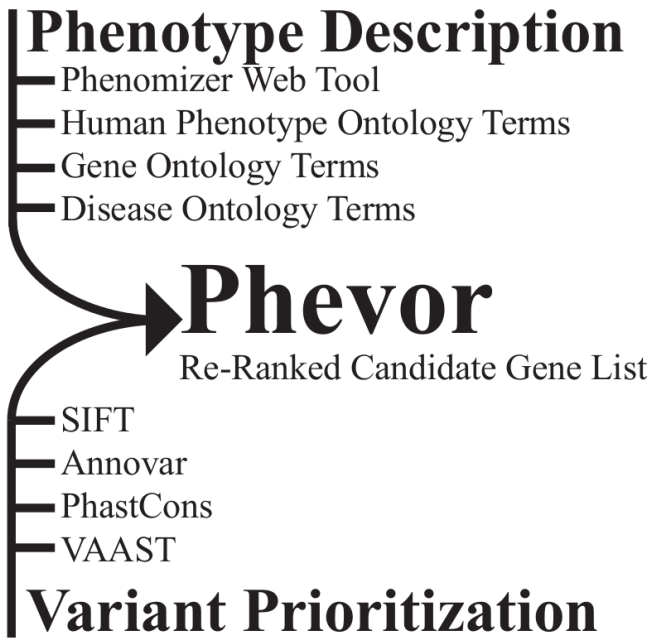
Marc V. Singleton,[1] Stephen L. Guthery,[5] Karl V. Voelkerding,[3,4] Karin Chen,[6] Brett Kennedy,[1] Rebecca L. Margraf,[4] Jacob Durtschi,[4] Karen Eilbeck,[2,7] Martin G. Reese,[8] Lynn B. Jorde,[1,2] Chad D. Huff,[9] and Mark Yandell[1,2,*]

[1]Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT 84112, USA; [2]Utah Science, Technology, and Research Center for Genetic Discovery, University of Utah, Salt Lake City, UT 84112, USA; [3]Department of Pathology, University of Utah, Salt Lake City, UT 84112, USA; [4]ARUP Institute for Clinical and Experimental Pathology, 500 Chipeta Way, Salt Lake City, UT 84108, USA; [5]Division of Pediatric Gastroenterology, Hepatology, and Nutrition, Department of Pediatrics, University of Utah, Salt Lake City, UT 84112, USA; [6]Division of Allergy, Immunology, and Rheumatology, Department of Pediatrics, University of Utah, Salt Lake City, UT 84112, USA; [7]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT 84112, USA; [8]Omicia Inc., 1625 Clay Street, Oakland, CA 94612, USA; [9]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, P.O. Box 301439, Houston, TX 77230, USA

The American Journal of Human Genetics *94*, 599–610, April 3, 2014

Introduction

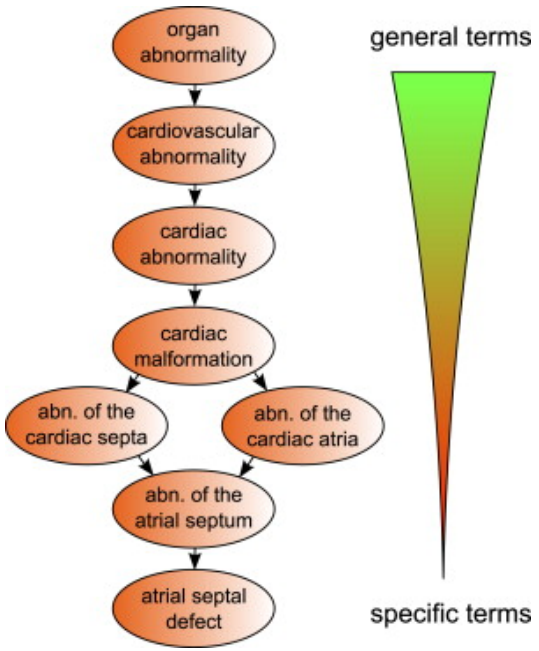
- PHEVOR = Phenotype Driven Variant Ontological Re-ranking tool
- Idea: combine the outputs of variant-prioritization tools with knowledge resident in biomedical ontologies [a]



[a] Robinson et al, 2008, Smith et al, 2012, Schriml et al, 2012, Ashburner et al, 2000

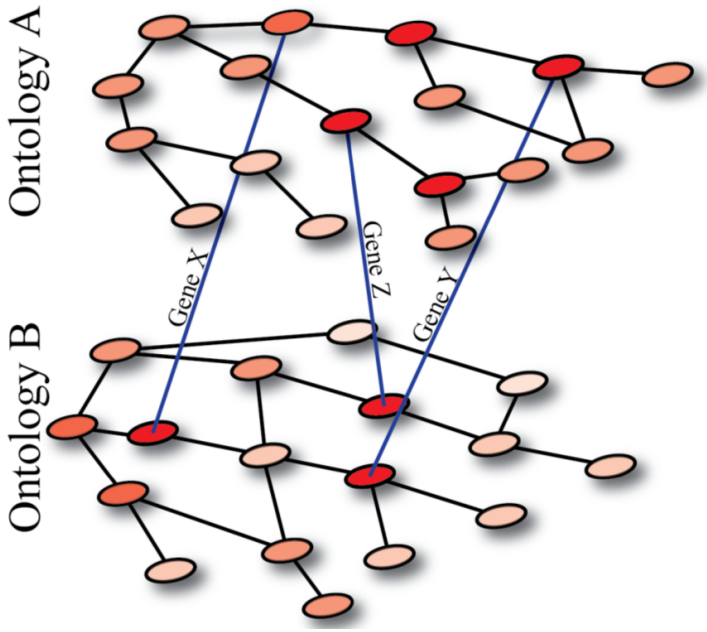
Workflow

1. user feeds the software with a list of phenotypic terms from an ontology (or a gene list)
2. Phevor saves the annotated genes in an internal list. If there are no annotated genes, Phevor iterates towards the root of the ontology and saves the annotated genes of the next nodes
3. the gene list is then fed into the other ontologies

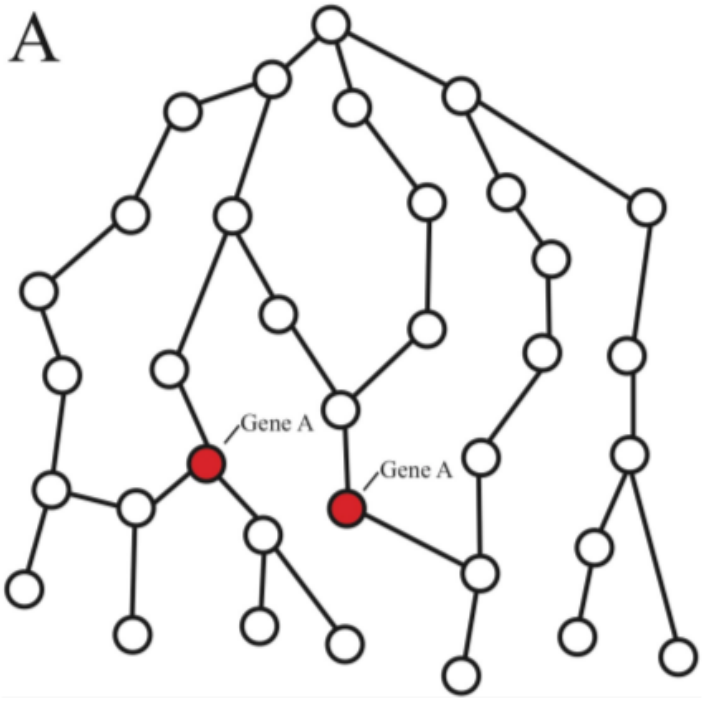


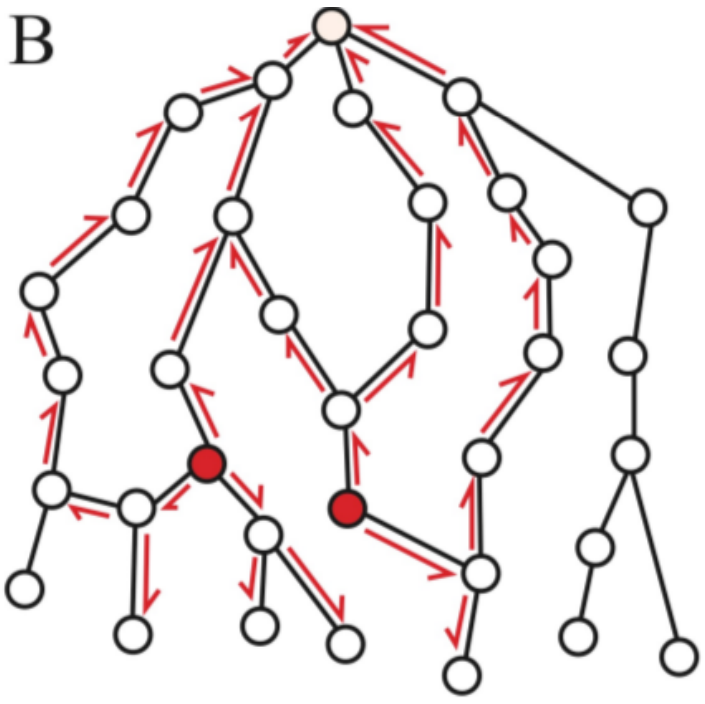
Workflow

1. user feeds the software with a list of phenotypic terms from an ontology (or a gene list)
2. Phevor saves the annotated genes in an internal list. If there are no annotated genes, Phevor iterates towards the root of the ontology and saves the annotated genes of the next nodes
3. the gene list is then fed into the other ontologies



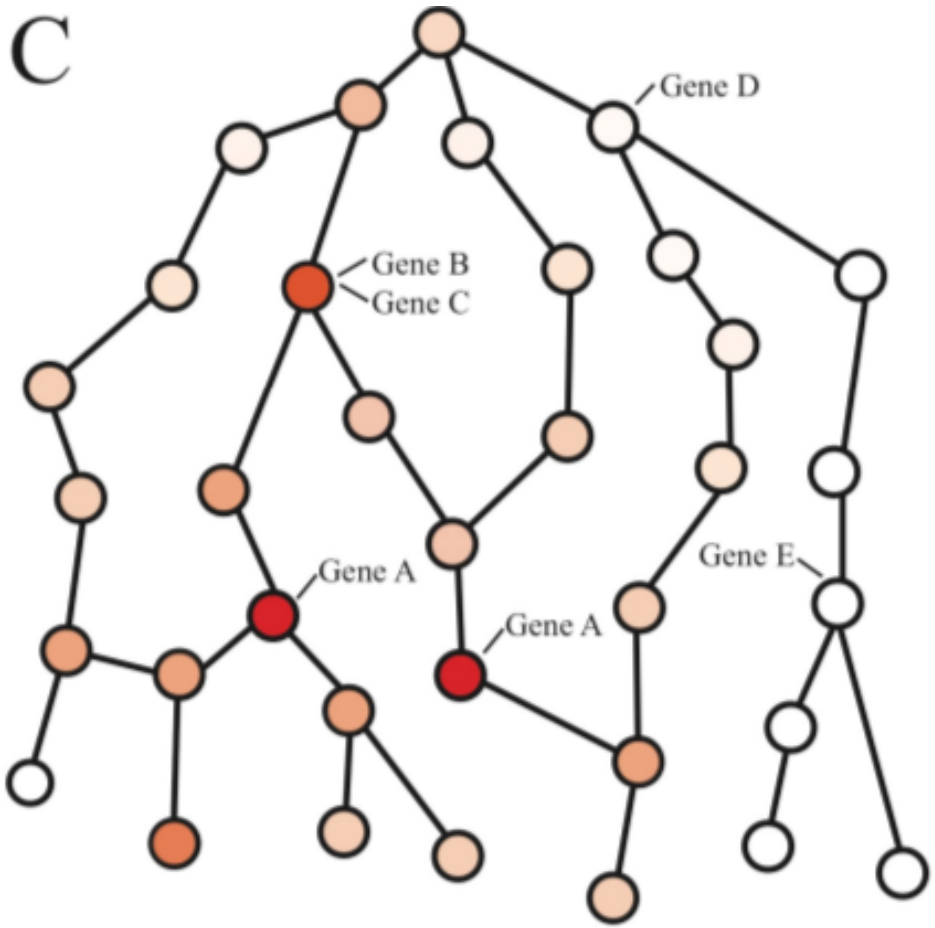
Ontological Propagation





each time an edge is crossed to a neighboring node, the current value of the previous node is divided by 2.

Ontological Propagation



Ontological Propagation

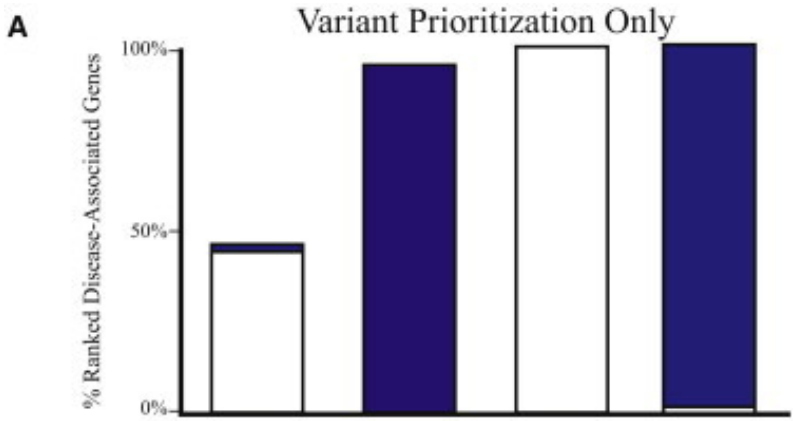
- Phevor renormalizes each node's value to between 0 and 1 by dividing it by the sum of all node scores in the ontology.
- Score(Gene) in a given ontology is the maximum score it achieved in any node
- final score is the sum of all the scores the gene achieved in the different ontologies
- normalized again to a value between 0 and 1

Combining Ontologies and Variant Data

- disease association score for each gene: $D_G = (1 - V_G) \times N_G$ where N_G is the percentile rank of the renormalized gene sum score as derived from the ontological combination and propagation procedures and V_G is the gene's percentile rank provided by the external variant-prioritization or search tool
- H_G, a second score summarizing the weight of evidence that the gene is not involved with the individual's illness: $H_G = V_G \times (1 - N_G)$
- Phevor score: $S_G = log \frac{D_G}{H_G}$

Results: gene known

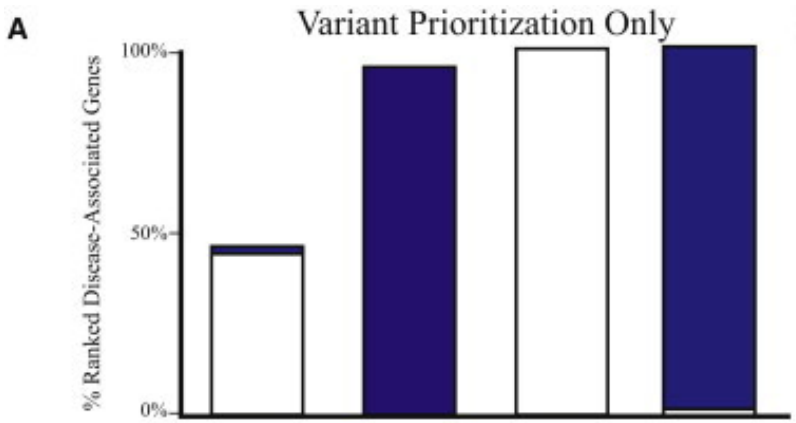
- two copies of a known disease-causing allele were inserted into an exome
- the process was repeated 100 times for 100 different genes with known disease associations



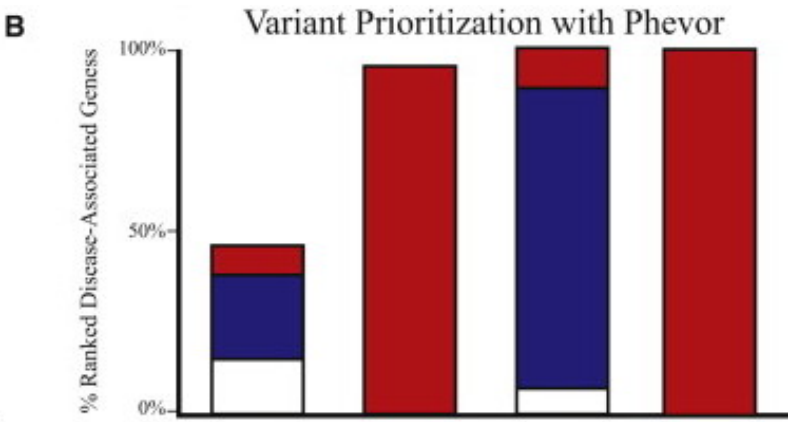
	SIFT	ANNOVAR	PhastCons	VAAST
Top 10 Rank	0 ± 0%	0 ± 0%	0 ± 0%	0 ± 0%
Top 100 Rank	2 ± 1%	95 ± 2%	0 ± 0%	99 ± 1%
Ranked	46%	95%	100%	100%
Average Rank	10,358 ± 978	1,141 ± 482	4,637 ± 140	83 ± 2

Results: gene known

- two copies of a known disease-causing allele were inserted into an exome
- the process was repeated 100 times for 100 different genes with known disease associations

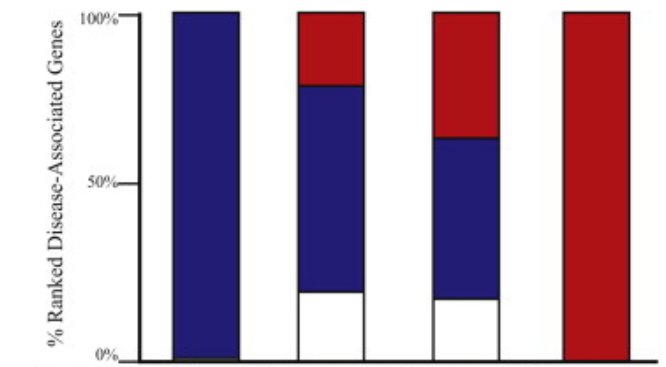


	SIFT	ANNOVAR	PhastCons	VAAST
Top 10 Rank	0 ± 0%	0 ± 0%	0 ± 0%	0 ± 0%
Top 100 Rank	2 ± 1%	95 ± 2%	0 ± 0%	99 ± 1%
Ranked	46%	95%	100%	100%
Average Rank	10,358 ± 978	1,141 ± 482	4,637 ± 140	83 ± 2



	SIFT	ANNOVAR	PhastCons	VAAST
Top 10 Rank	8 ± 3%	95 ± 4%	11 ± 3%	100 ± 0%
Top 100 Rank	31 ± 5%	95 ± 4%	93 ± 2%	100 ± 0%
Ranked	46%	95%	100%	100%
Average Rank	6,023 ± 588	552 ± 260	88 ± 34	1.8 ± 0.1

Results: gene not known



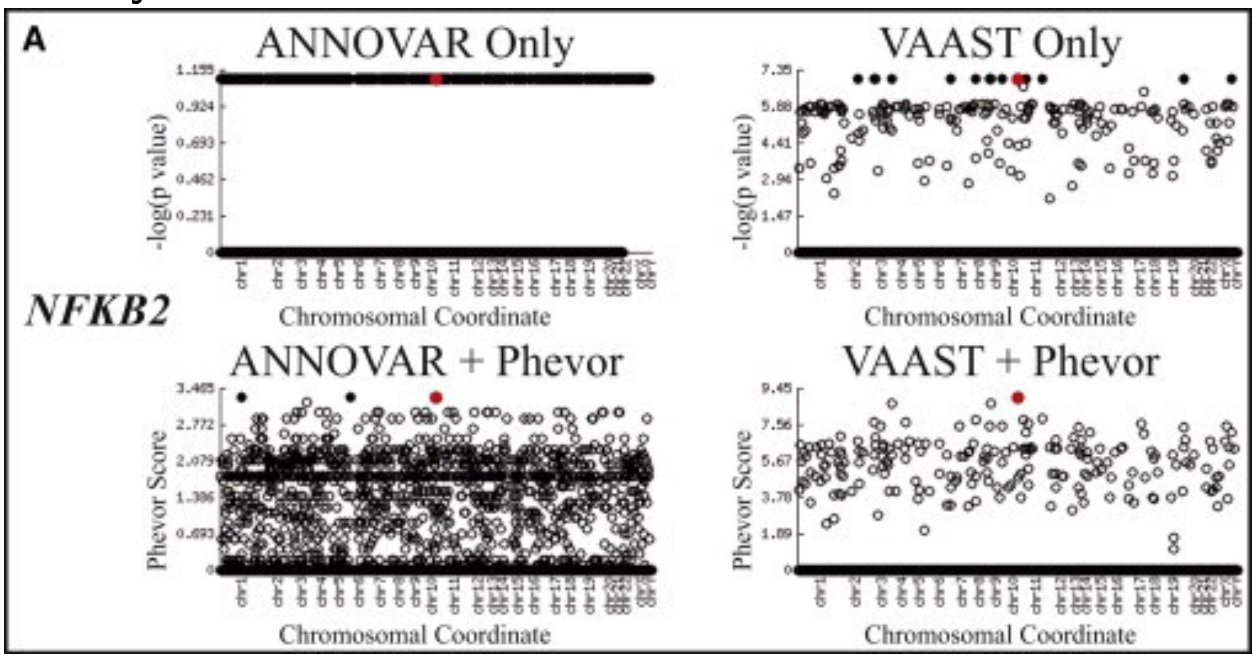
Ontologies Used	None	GO only	MPO, HPO, and DO	All
Top 10 Rank	0 ± 0%	21 ± 4%	36 ± 5%	100 ± 0%
Top 100 Rank	99 ± 1%	80 ± 4%	82 ± 4%	100 ± 0%
Average Rank	83 ± 2	226 ± 137	139 ± 73	1.8 ± 0.1

only VAAST genetic input

- none: no annotated genes
- GO only: annotated genes only in GO
- MPO, HPO & DO: annotated genes removed only in GO
- all: all annotated genes there

Results: Application in real patients 1

family affected by autosomal-dominant, early-onset hypogammaglobulinemia with variable autoimmune features and adrenal insufficiency

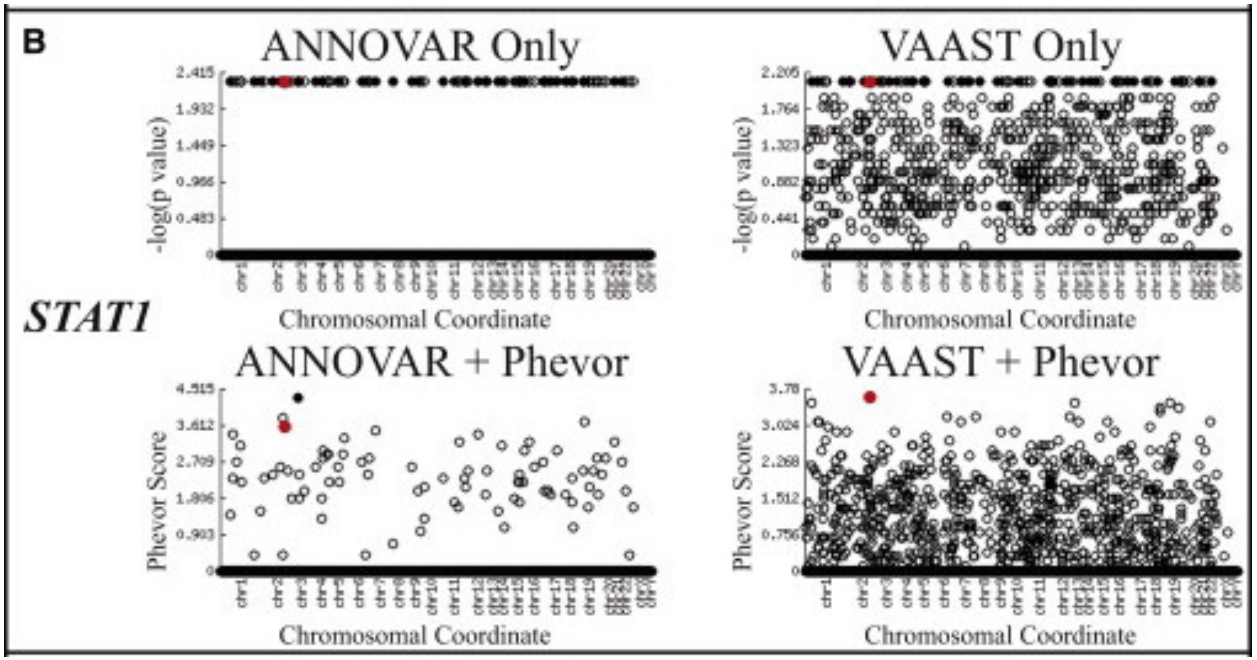


3

[3] Utah Genome Project

Results: Application in real patients 2

12-year-old male with severe diarrhea in the context of intestinal inflammation, total villous atrophy, and hypothyroidism
=> X-linked immunodysregulation, polyendocrinopathy, and enteropathy (IPEX syndrome)
BUT clinical sequencing of genes associated with IPEX revealed no pathologic variants



4

[4] Utah Genome Project

Summary

- PHEVOR ameliorates the performance of sequencing-based ranking algorithms
- can identify dominant mutations in a gene not previously associated with the disease phenotype in the ontologies (NFKB2) or a de novo dominant allele located in an known disease-associated gene (STAT1) and producing an atypical phenotype.
- if the gene is not annotated in the ontologies, the performance drops

Phen-Gen

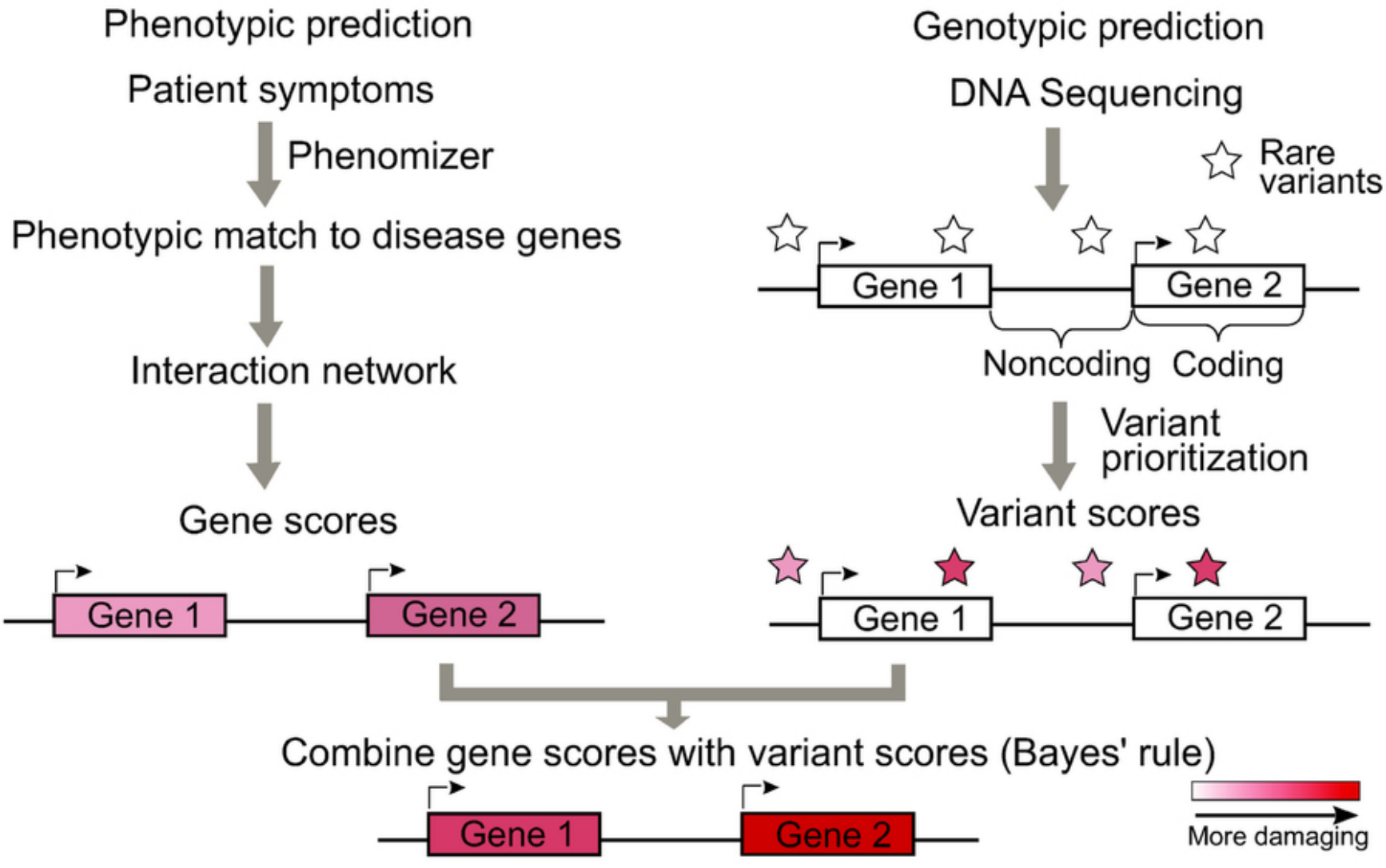
Phen-Gen: combining phenotype and genotype to analyze rare disorders

Asif Javed, Saloni Agrawal & Pauline C Ng

Computational and Systems Biology Group, Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore.

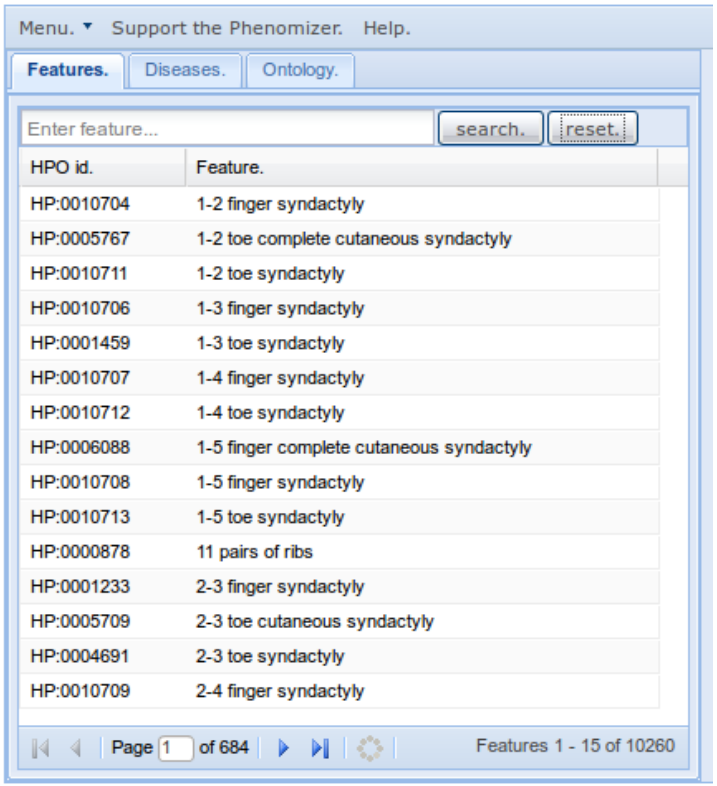
NATURE METHODS | VOL.11 NO.9 | SEPTEMBER 2014 | 935

overall workflow



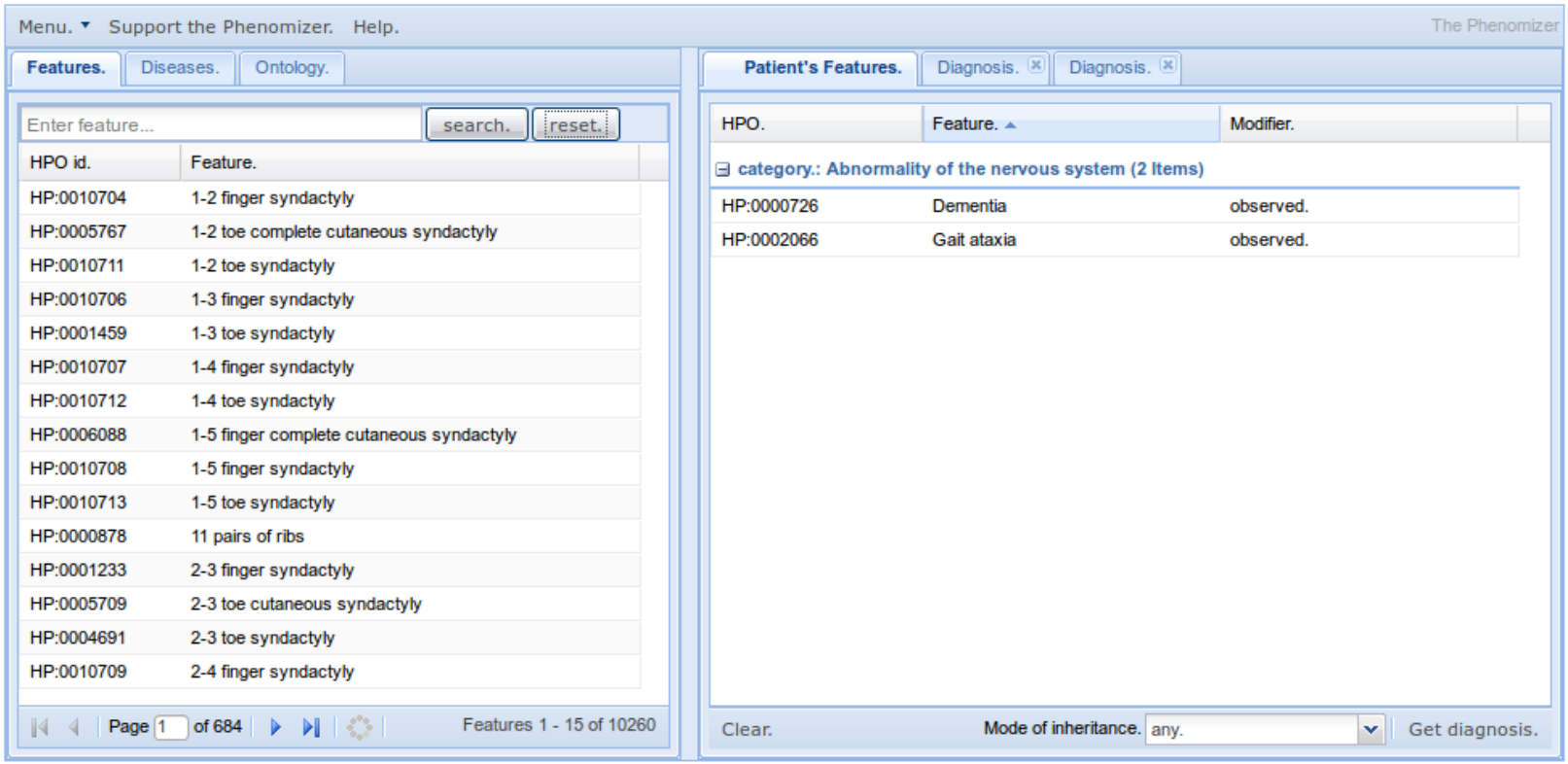
Phenomizer

- introduced by Kohler et al, 2009
- Aim: match the patient's symptomatology to the list of known disorders and estimate the significance of each disease match



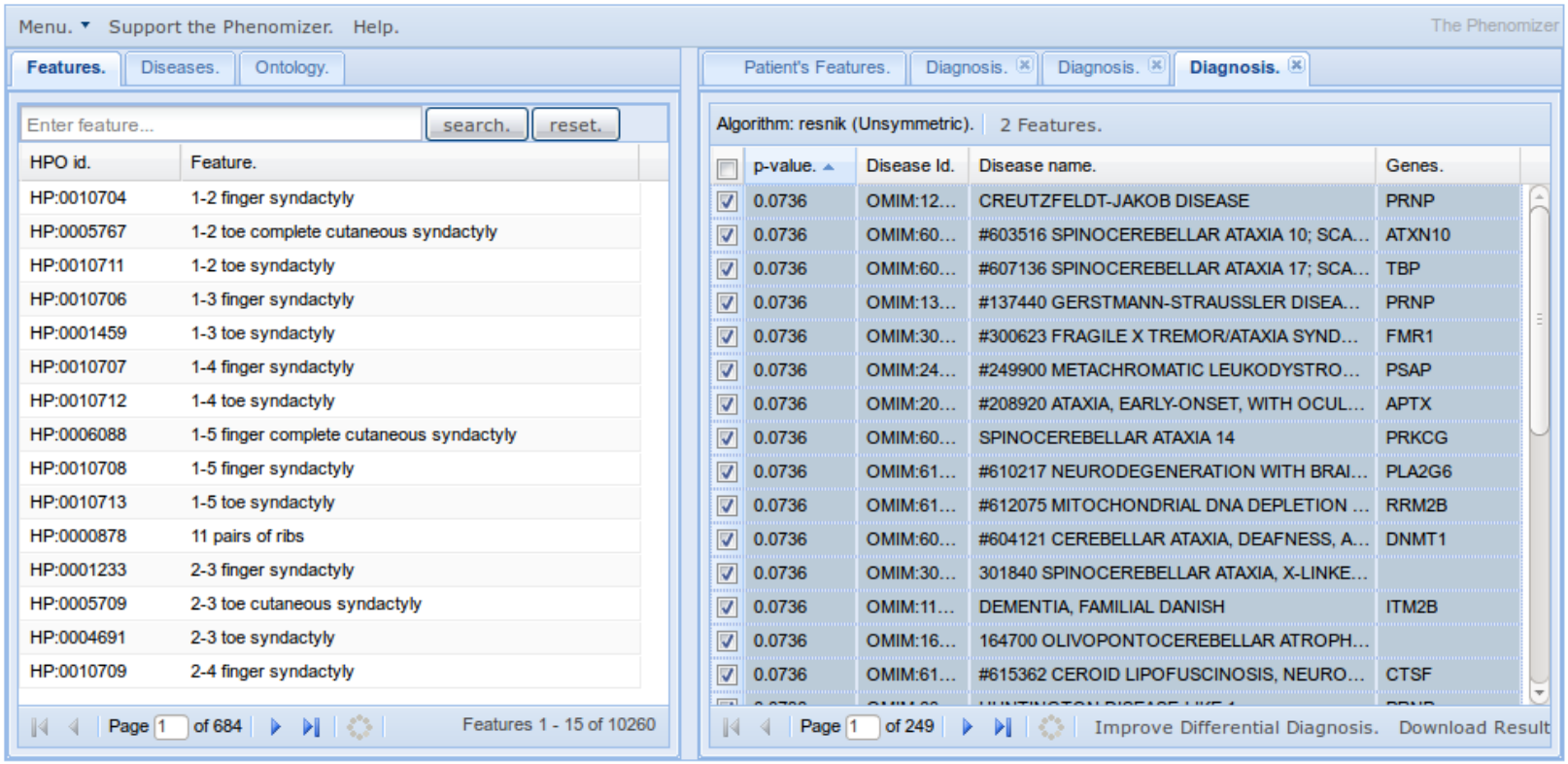
Phenomizer

- introduced by Kohler et al, 2009
- Aim: match the patient's symptomatology to the list of known disorders and estimate the significance of each disease match



Phenomizer

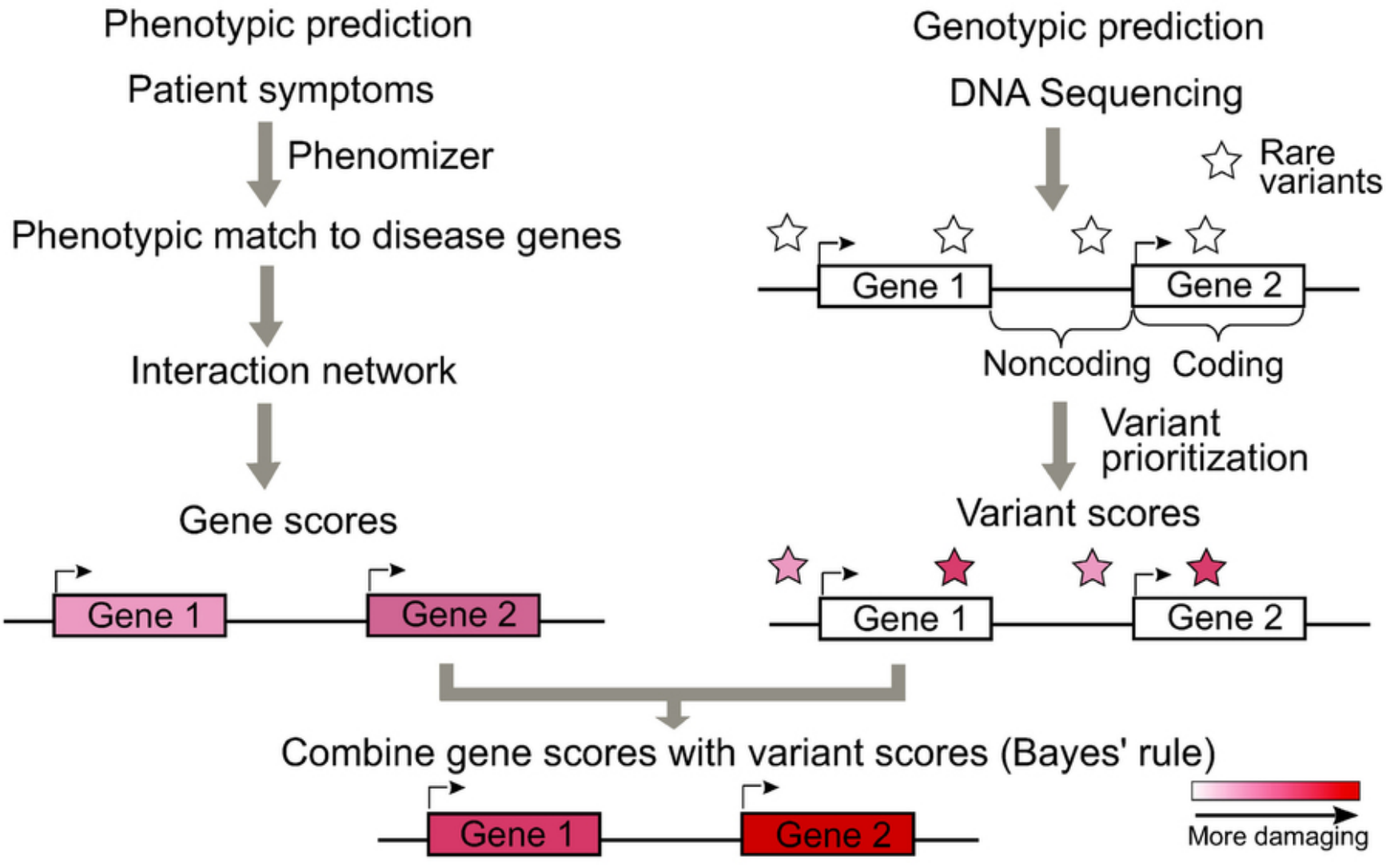
- introduced by Kohler et al, 2009
- Aim: match the patient's symptomatology to the list of known disorders and estimate the significance of each disease match



Phenomizer in Phen-Gen

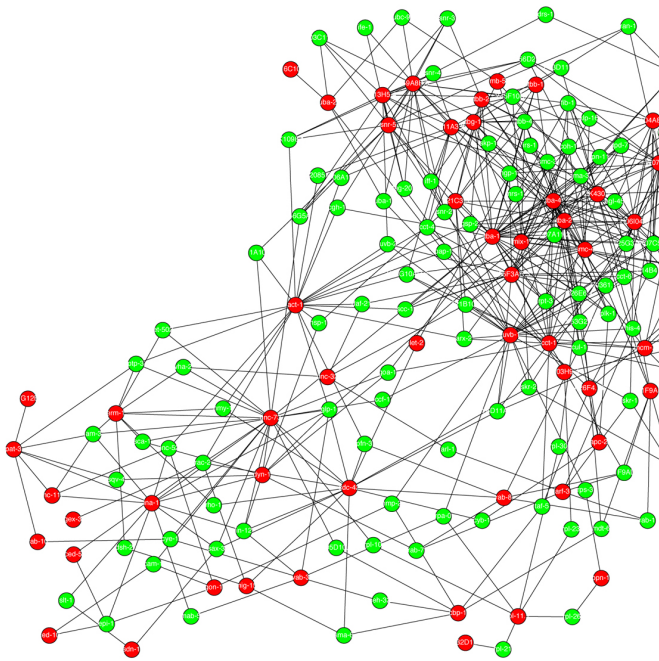
- translation of p-value to disease-probability
- Each disease probability is assigned to all genes implicated in the database for that disorder. If no genes are currently known, the probability is distributed uniformly across all genes.

overall workflow



Interaction Network

- Interaction: ability of two genes to affect the same underlying biology and thus lead to the same (or similar) disorders and symptoms
- Interaction data from REACTOME, KEGG, NCI-Nature, BioGRID, STRING, Gene Ontology domain annotations, COXPRESdb
- integrated into a gene-gene interaction network
- edge weights proportional to confidence in interaction
- 920,898 interactions in total

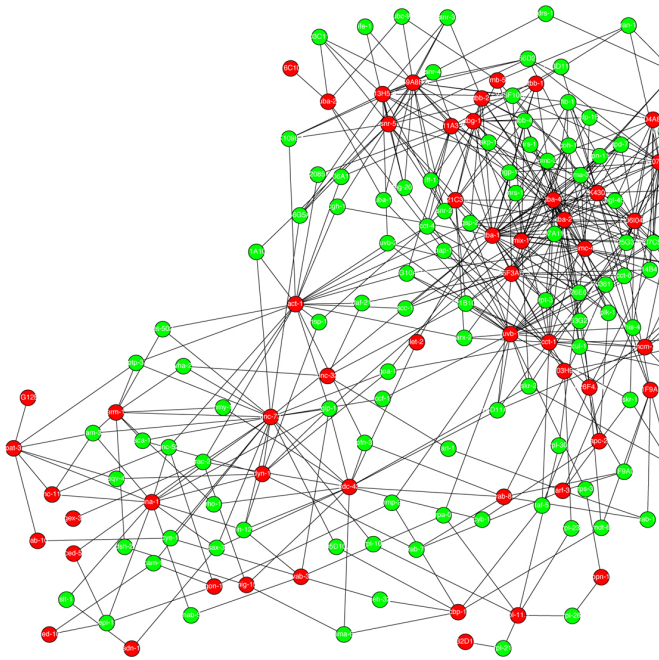


5

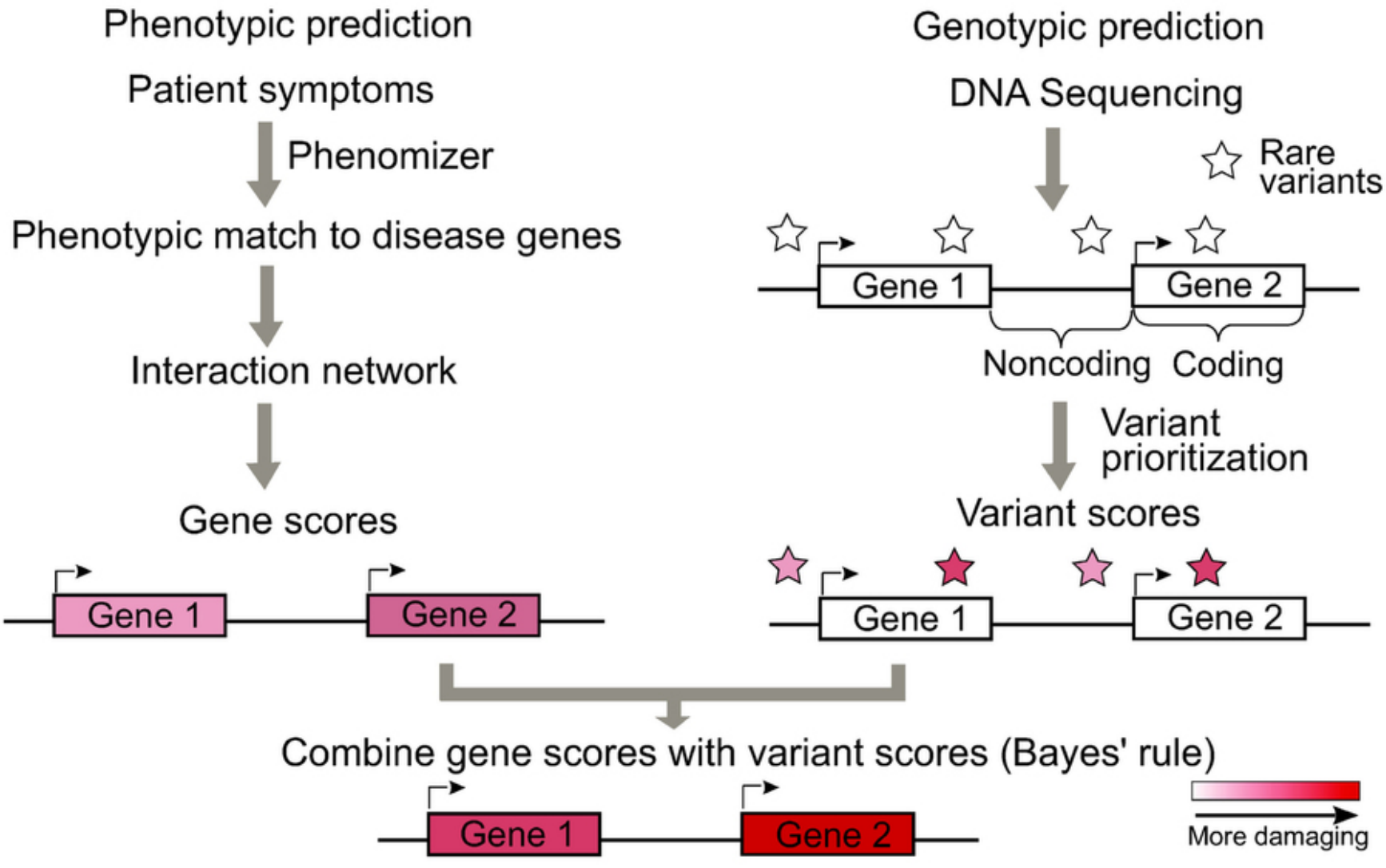
[5] Matthews et al, 2009, Kanehisa et al, 2012, Schaefer et al, 2009, Stark et al, 2006, Franceschini et al, 2013, Ashburner et al, 2000, Obayashi et al, 2013

Interaction Network

- random walk with restart on the gene-gene interaction network.
  - Each gene selected as starting point of the walk with probability P_{Gene} with restart probability of 90%
  - $=>$ 90% of the 'guilt', retained by the initial gene & 10% of it permeated across its known associates (with stronger evidence of interaction leading to higher probabilities)



overall workflow



Gene List

- four popular gene lists (Consensus CDS, RefSeq, Ensembl and UCSC known)
- Incomplete transcripts discarded
- only protein-coding genes considered
- Alternatively spliced transcripts reported in the same data source were assigned to the same gene
- Additionally, two genes were merged if any of their isoforms exhibited greater than 95% identity in the exon sequence in the same orientation
- The final gene list comprised 26,803 protein-coding genes
- key question: is the reported mutation harmful or not?

Variant Prediction - exome-centric approach

- each variant in the patient's genome (or exome) evaluated if
  - within the coding regions in a reported transcript of the gene list
  - within the splice-site definition of the intron-exon boundary
- coding variants subcategorized as start-loss, stop-gain, stop-loss, splice site, nonsynonymous, synonymous or indel.
- Each variant is assigned a probability of deleteriousness based on its highest estimated damaging impact.

Variant Prediction - exome-centric approach

Nonsynonymous variants For prediction of damaging impact, two commonly used algorithms were combined: SIFT and PolyPhen-2 $=>$ estimate the probability of a nonsynonymous mutation

Variant Prediction - exome-centric approach

Nonsynonymous variants For prediction of damaging impact, two commonly used algorithms were combined: SIFT and PolyPhen-2 $=>$ estimate the probability of a nonsynonymous mutation

Splice-site variants Mutations in the 8-bp locus surrounding the donor site or the 3-bp locus surrounding the acceptor site were considered splice disrupting

Variant Prediction

Start-loss, stop-gain & stop-loss variants are highly deleterious and assumed to be damaging

Variant Prediction

Start-loss, stop-gain & stop-loss variants are highly deleterious and assumed to be damaging

genic indels the prior probability of a genic indel being damaging was estimated to be 0.0787 based on analysis of a reference data set.
frameshift indels are assigned the probability of deleteriousness on the basis of the empirical distribution of its positive and neutral set combined
nonframeshift indel predictions are further refined by incorporating the importance of the affected locus (based on its tolerance to single-nucleotide mutations)

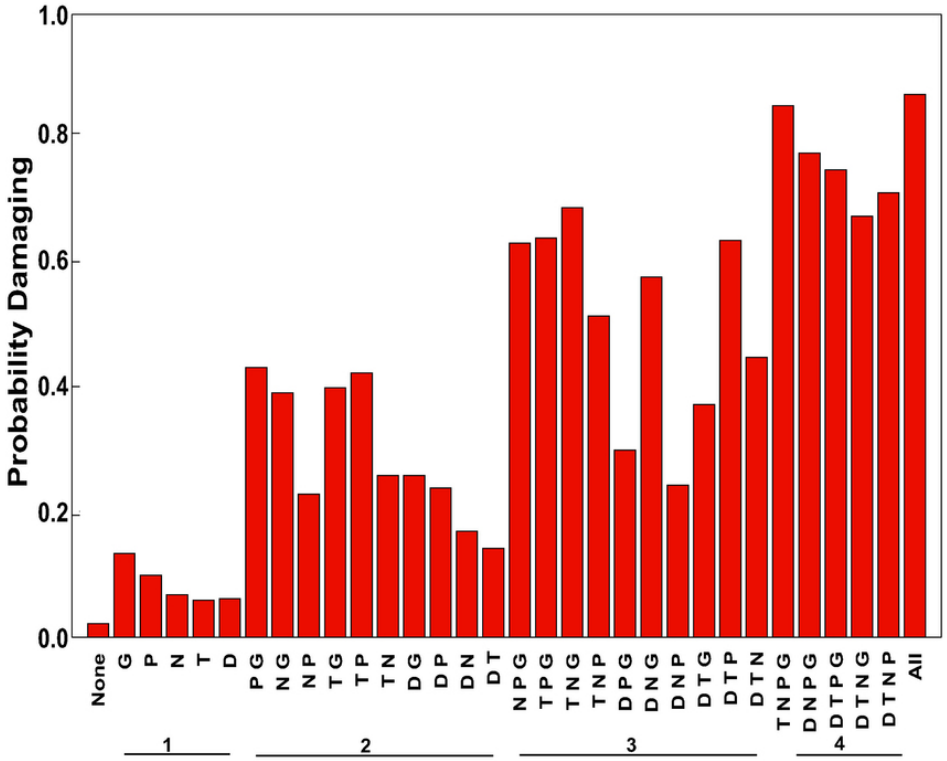
Variant Prediction - genome-centric approach

Genomic variants

aim: to estimate the putative functional role of all genomic variants, esp. those lying outside of the exome

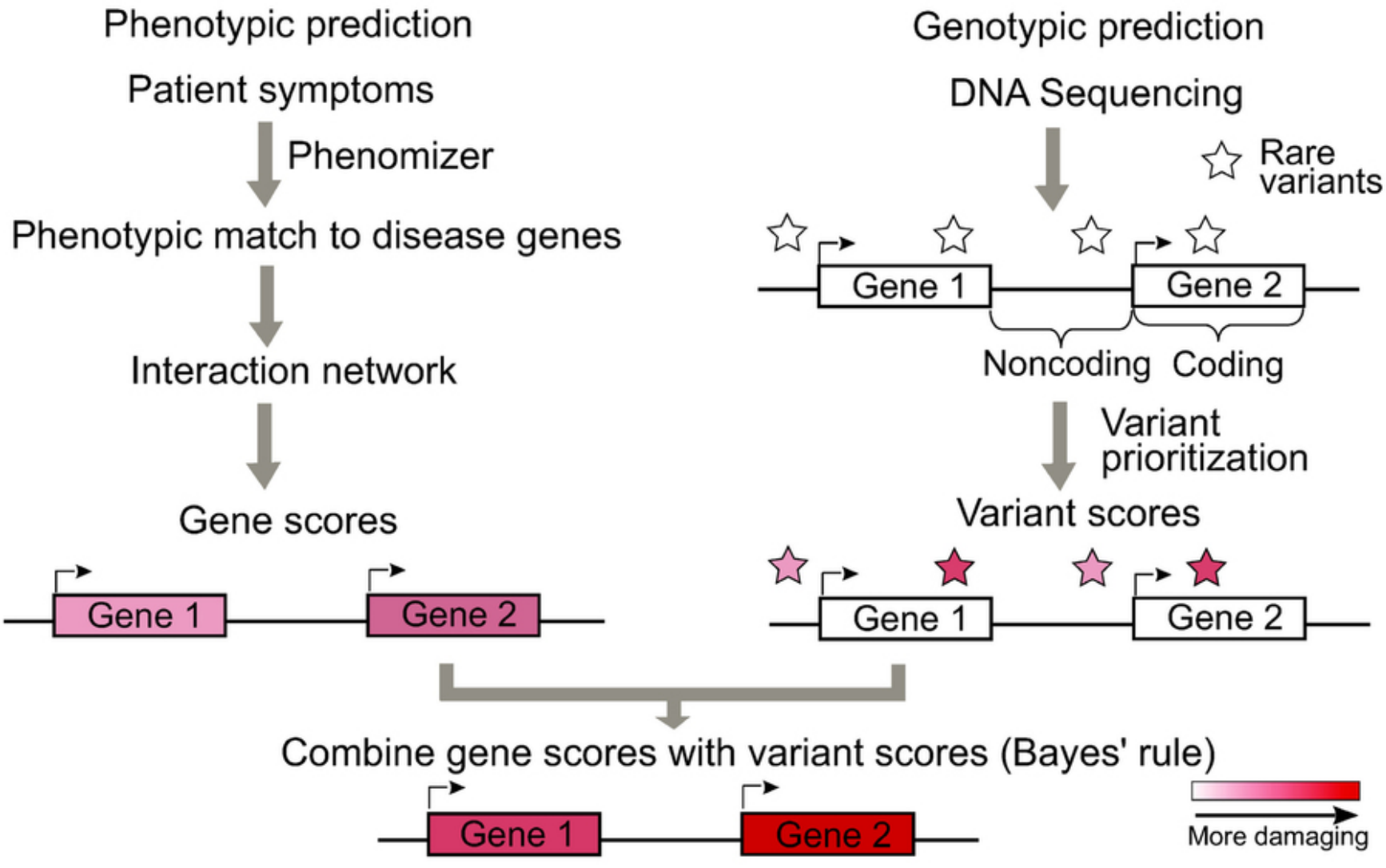
- conservation (using GERP++ and PhyloP)
- putative regulatory interactions (using ENCODE)
- proximity to genes (including coding sequence, UTRs and 70 bp at the start and end of each intron)
- => annotate the variants in order to estimate a locus's susceptibility for disease

Variant Prediction- genome-centric approach



G: GERP++, P: PhyloP, N: near-genic, T: transcription factor binding sites, and D: DNase hypersensitive sites

overall workflow

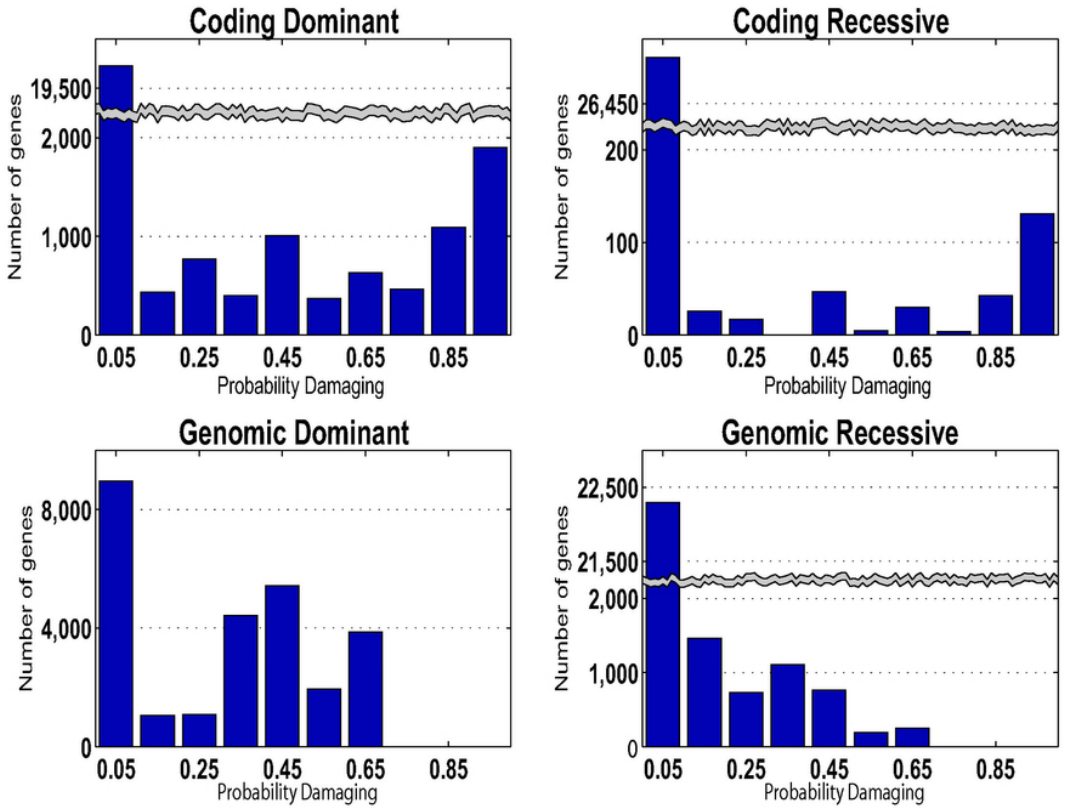


Pooling

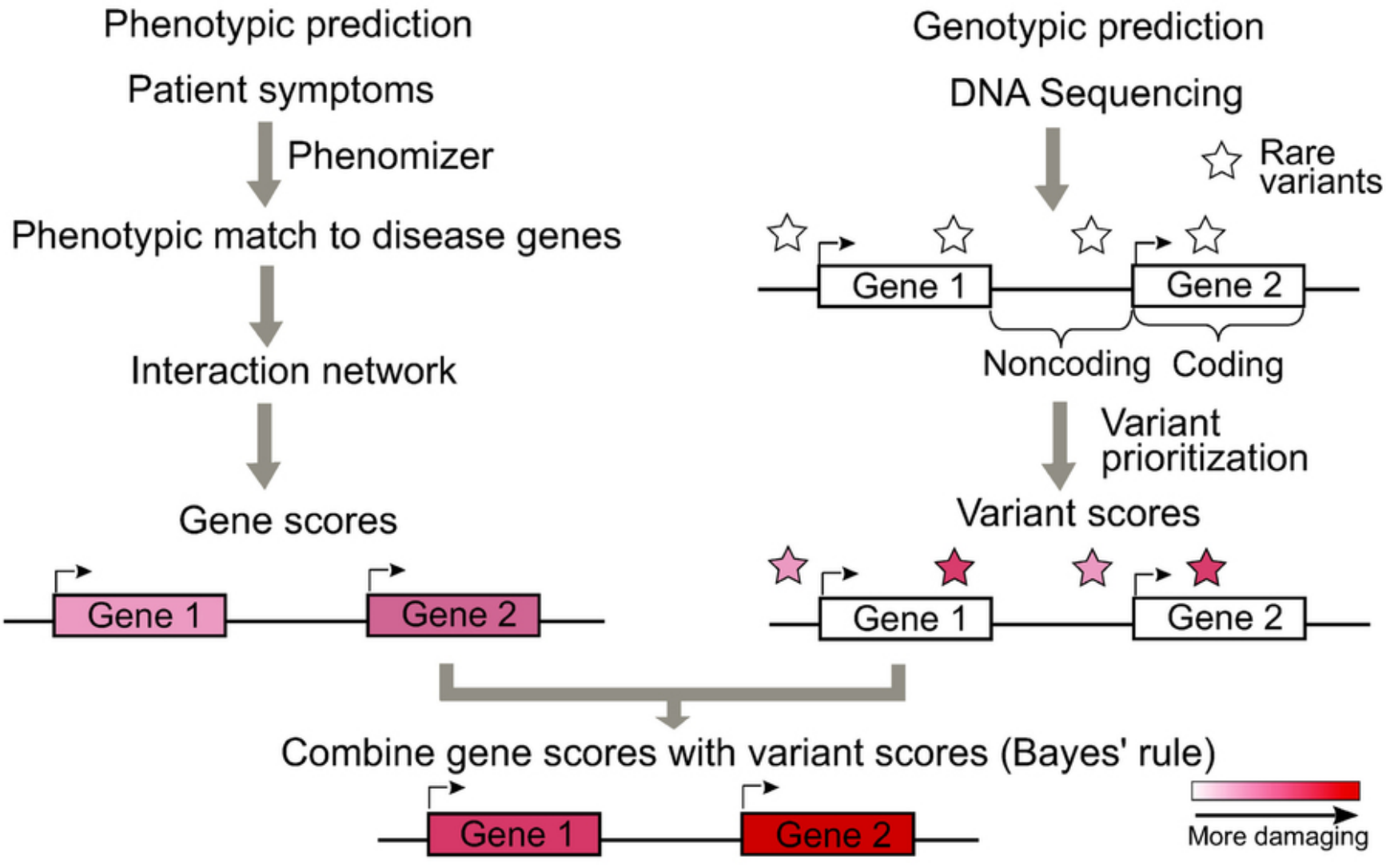
- The estimated loss of function at the genic level is predicted by pooling damaging variants within each gene considering the most damaging predicted variants for maternally and paternally inherited chromosomes.
- The inheritance pattern of the variants is compared against the disease inheritance pattern.
- Common variants are omitted from this pooling.
- only genes harboring variants that exceed the 99th percentile of the corresponding null distribution are considered for downstream analysis.

Pooling - null distribution

null distribution of predictor for each gene using the samples from 1000 healthy individuals (top 1 percentile of damaging variants)



overall workflow



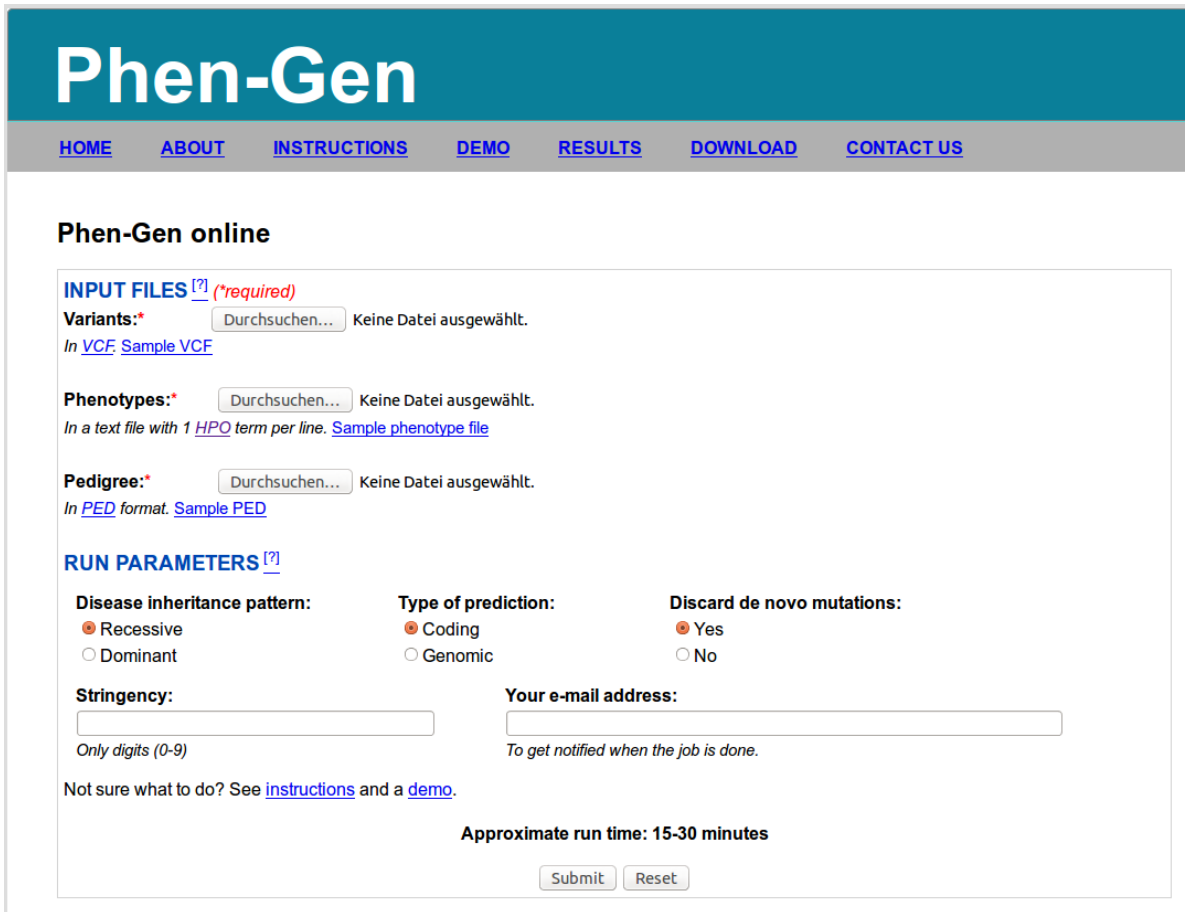
Combined Prediction

the probability that gene j is disease causal is given by

$$\frac{P_G(j) \times P_P(j)}{P_G(j) \times P_P(j) + (1 - P_G(j)) \times (1 - P_P(j))}$$

where $P_G(j)$ and $P_P(j)$ are the probabilities that gene j is disease causal on the basis of genotype data and phenotype data, respectively

Phen-Gen online



Results: Evaluation of exome-predictor with in-silico patients

- a patient's genetic data were created by adding a disease-causal mutation into a healthy individual's genome or exome
- phenotypic data was generated on the basis of reported disease symptoms

	Dominant			Recessive		
	Top gene (%)	Top 5 genes (%)	Top 10 genes (%)	Top gene (%)	Top 5 genes (%)	Top 10 genes (%)
Combined (missense, nonsense, splice site and indels)						
Phen-Gen	81	88	91	96	97	97
Genotype only	0	3	13	72	97	97
Phenotype only	31	69	74	24	51	68
Known	92	97	97	96	97	97
Unknown	43	61	73	92	96	96

Results: Evaluation of genomic predictor with in-silico patients

Performance if regulatory elements are affected

	Dominant			Recessive			Combined (dominant + recessive)		
	Top gene (%)	Top 5 genes (%)	Top 10 genes (%)	Top gene (%)	Top 5 genes (%)	Top 10 genes (%)	Top gene (%)	Top 5 genes (%)	Top 10 genes (%)
Regulatory									
Phen-Gen	58	64	65	40	61	72	49	62	69
Genotype only	1	2	2	34	51	62	19	29	34
Phenotype only	31	43	54	5	11	26	17	26	39

Results: Evaluation with real patients

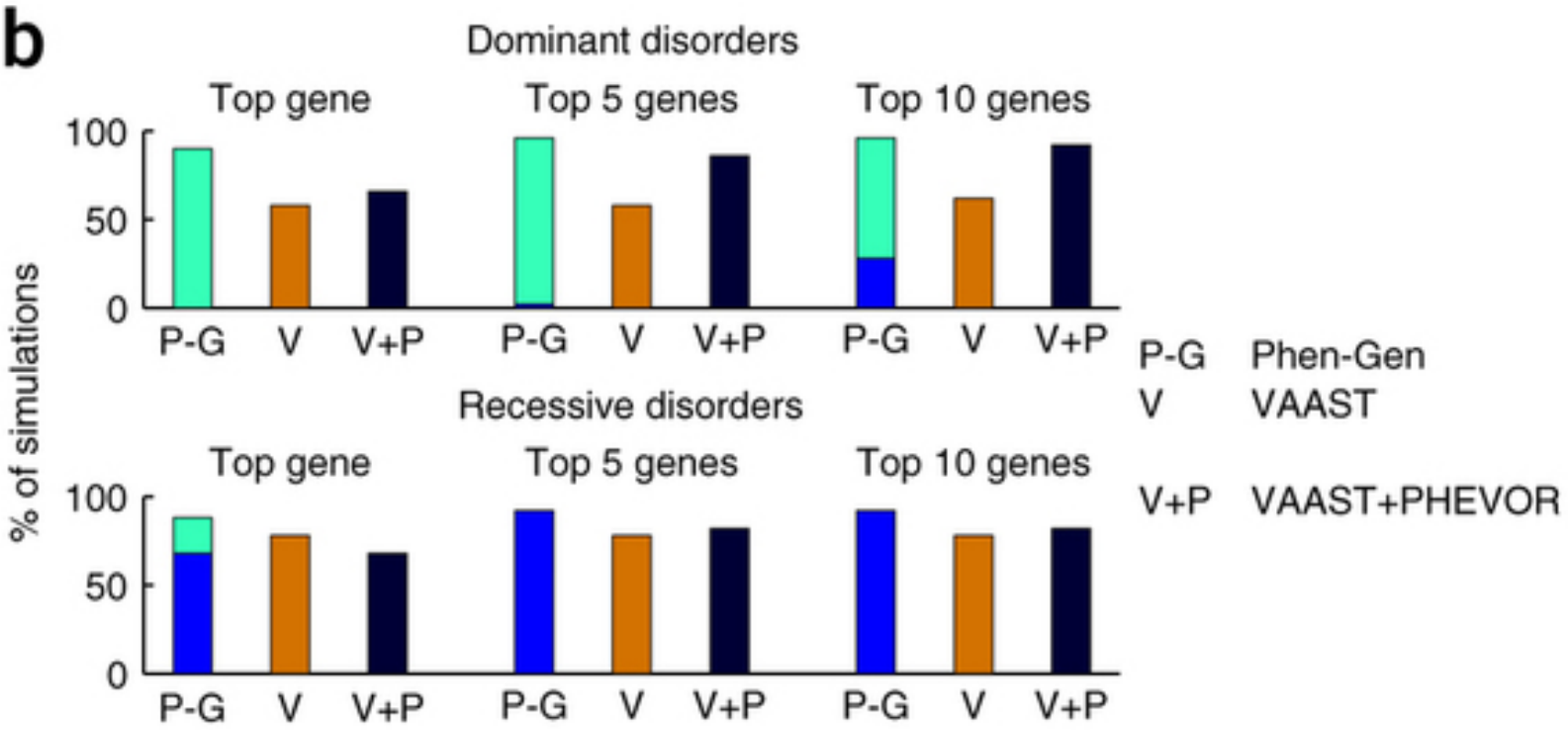
Phen-Gen applied to eleven families with variants implicated in recessive or X-mode inheritance

Trio ID	Gene	Classification	Phen-Gen's Rank	
			Inclusive of dbSNP MAF<1%	Filtration criteria from ref. 5
4	*FANCB*	Unknown	18	3
4	*PDHA1*	Known	5	2
4	*GUCY2F*	Unknown	4	1
16	*ENOX2*	Unknown	2	1
18	*ARHGEF9*	Known	8	1
25	*GPM6B*	Unknown	7	1
41	*ARHGEF9*	Known	1	1
42	*DDX26B*	Unknown	13	3
72	*PDZD11*	Unknown	3	3
93	*TRPC5*	Candidate	4	1
12	*SYCP2L*	Unknown	2	1
12	*VPS13B*	Known	14	4
12	*C8orf59*	Unknown	4	2
12	*PRUNE2*	Unknown	7	3
24	*PCNT*	Known	7	2
70	*IQGAP2*	Unknown	6	1

Filtration criteria: all mutations reported in the Single Nucleotide Polymorphism Database (dbSNP) or observed in the in-house data set were removed from further evaluation [6]

[6] de Ligt et al, 2012

Comparison of Phen-Gen with PHEVOR



Phen-Gen outperforms VAAST by 19-21% and PHEVOR by 20-26% when considering the first rank

Summary

- Phen-Gen top-ranks the causal gene in 49% (non-coding variants) to 97% (coding recessive variants)
- it outperforms other existing methods
- fast (15-30min)
- open source license
- clinical use needs to be confirmed
- a better performance with non-coding mutations would be nice

Discussion & Conclusion

Discussion

- need of very careful clinical diagnostics
- clinical value needs to be confirmed
- performance esp. in non-coding mutations might be ameliorated
- rely on further development of other existing tools

Discussion & Conclusion

Discussion

- need of very careful clinical diagnostics
- clinical value needs to be confirmed
- performance esp. in non-coding mutations might be ameliorated
- rely on further development of other existing tools

Conclusion

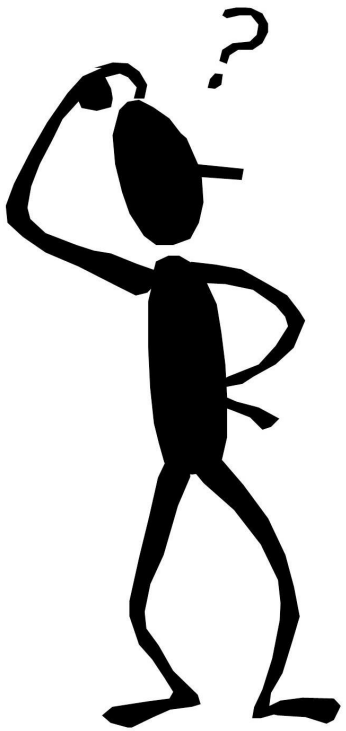
such algorithms might facilitate

- diagnosis of rare genetic diseases
- genetic counseling
- identification of candidate genes for further research
- drug development

List of Links

HPO:	http://www.human-phenotype-ontology.org/
PHEVOR:	http://weatherby.genetics.utah.edu/cgi-bin/Phevor/PhevorWeb.html
VAAST:	http://www.yandell-lab.org/software/vaast.html
Phenomizer:	http://compbio.charite.de/phenomizer/ or in App-/Play- Store
Phen-Gen:	http://phen-gen.org/

Questions?



Thank you!

Appendix: Bayes' Rule

- Bayes' Theorem: Rule for conditional probability i.e. the probability of A given B
  - calculation: $P(A \mid B) = (\frac{P(A) \times P(B|A)}{P(B)})$
- Bayes' Rule: relates the odds of event A_1 to the odds of event A_2 , before (prior to) and after (posterior to) conditioning on another event B
  - calculation: $\frac{P(A_1|B)}{P(A_2|B)} = (\frac{P(A_1) \times P(B|A_1)}{P(A_2) \times P(B|A_2)})$

Appendix: Phenomizer in detail

- uses HPO
- Calculation of the similarity score:
  - calculation of Information Content: $IC = -log(\frac{numberofdiseaseshavingtheterm}{numberofalldiseases})$
  - => the more specific the term, the higher the IC
  - similarity score: $sim = avg\,[\sum IC(MICA)]$ where MICA=most informative common ancestor
  - weighting function for additional features of the disease not covered by the entered terms
- Calculation of p-value:
  - based on the distribution of similarity scores that is obtained by randomly choosing combinations
  - If a given score is only rarely obtained by chance, then we consider it to be statistically significant.
  - Monte Carlo random sampling and corrected for multiple testing by the method of Benjamini and Hochberg