Methods for identifying disease causal or susceptibility mutations from high-throughput sequencing data

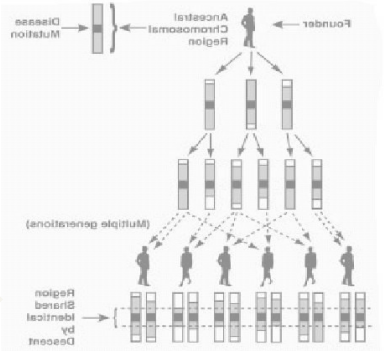
August 29-31, 2018

(Miaoxin Li; 李淼新)
Sun Yat-sen University & the University of Hong Kong

local genotypes

Probe Set ID	Genotypes
SNP_A-1813205	BB
SNP_A-1880143	AB
SNP_A-4215517	AA
SNP_A-1828242	AA
SNP_A-2029913	AA
SNP_A-1929900	BB
SNP_A-1818663	AB
SNP_A-2192352	AA
SNP_A-4218271	AA

Statistical Genetics + Bioinformatics

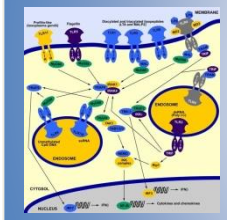


http://www.ncbi.nlm.nih.gov/bookshelf

statistical genetic mapping

Boost power!!!

Integrative computational frameworks



Biological resources

http://www.abcam.cn/index.html?pageconfig=resource&rid=12189&pid=10629

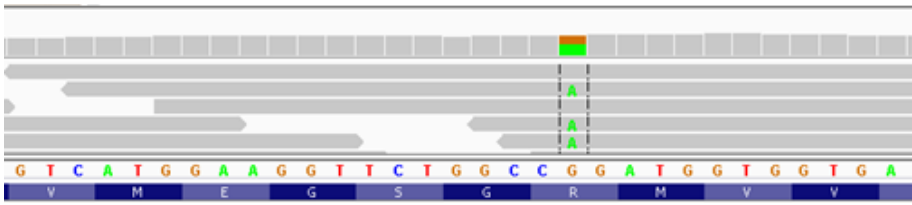
The general procedure of detecting disease-casual mutations using high-throughput sequencing data

Mapping

- Assemble short reads and map them onto reference genome

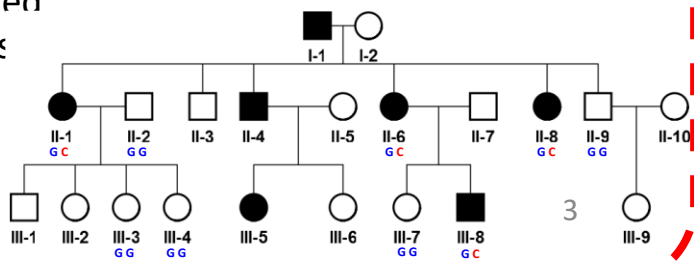
Calling

- Call sequence variants in samples



Identifying

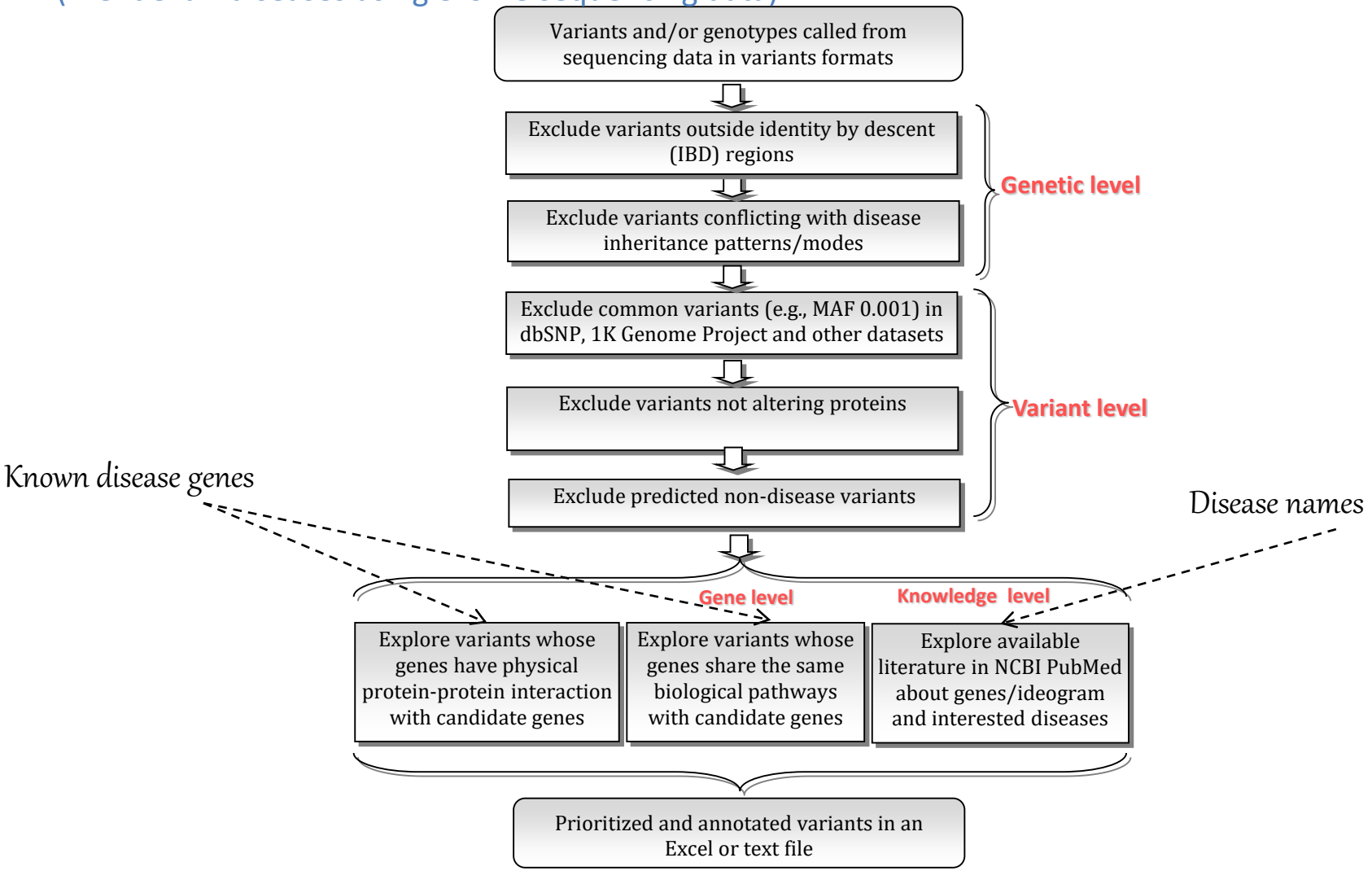
- Identify causal mutations in the called variants



downstream-analysis

A multilayer automated filtration and prioritization framework

(Mendelian-diseases using exome sequencing data)



Li MX, Gui HS, Kwan SH, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Research* (Nucleic Acids Res. 2012 Apr 1;40(7):e53)

Key functions of the three tools for filtration, prioritization and annotation

	KGGSeq	ANNOVAR (Version2)	VEP
Quality control	Systematic QC on genotype, variant and subject levels	Only by depth and sequencing quality	No
Use disease mode	Recessive, dominant compound heterozygous, de novo and runs of homozygosity.	Too simple, cannot directly take genotypes of controls; it only works for recessive and dominant modes	No
Variants reference databases	dbSNP, 1000 Genomes Project and ESP	Similar to wKGGSeq	Similar to wKGGSeq
Gene annotation	RefGene GENCODE UCSC knownGene	RefGene GENCODE UCSC knownGene ENSEMBL	ENSEMBL RefGene
Functional prediction	SLR SIFT Polyphen2_HDIV Polyphen2_HVAR LRT MutationTaster MutationAssessor FATHMM_score CADD_score GERP++_NR GERP++_RS PhyloP100way_vertebrate 29way_logOdds	Same as wKGGSeq	SIFT Polyphen2_HDIV Polyphen2_HVAR
Protein-protein interaction and pathway	Yes	No	No
Literature	PubMed	No	No
Management of input data	Yes	No	No
Disease-targeted prioritization	Yes	No	No

The summary filtration and prioritization results of three tools in three pedigrees

	KGGSeq	ANNOVAR	VEP
	Neonatal-onset Crohn disease		
Initial variants	1196282	68992[a]	68992[a]
Retained variants when the causal mutations were kept finally	3	53	4232
Variant hit by PPIs, pathways or PubMed search	-	-	-
Additional evidences to highlight the causal mutations	PPI+Pathway+PubMed	-	-
	Spinocerebellar ataxias		
Initial variants	1417935	82465[a]	82465[a]
Retained variants when the causal mutations were kept finally	17	29	6501
Variant hit by PPIs, pathways or PubMed search	3	-	-
Additional evidences to highlight the causal mutations	Pathway+PPI	-	-
	Familial spastic paraplegia		
Initial variants	1017018	63207[a]	63207[a]
Retained variants when the causal mutations were kept finally	3	7	5109
Variant hit by PPIs, pathways or PubMed search	2	-	-

Note: a: as wANNOVAR cannot effectively map variants on VCF data, KGGSeq was used to do the basic quality control on VCF data of affected samples.

Hum Mutat. 2015 May;36(5):496-503

When whole genome sequencing data come, …

1000 subjects

- Sequence variants: around 40 millions
- Genotypes + quality values in compressed format: 250GB

KGGSeq (v0.8-) needs 100+ GB RAM and 24+ hours to do downstream prioritization analysis!!!

I. How can the huge amount of sequencing data by analyzed with less RAM and faster speed?

Genotype bit-block encoding algorithm

a

##fileformat=VCFv4.0

subjects

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	A	B	C
chr1	4793	rs668235	A	G	596.57	PASS	AC=20	GT	1/1	0/0	./.
chr1	53560	.	T	C	573.51	PASS	AC=10;	GT	./.	0/0	0/1
chr1	12887054	rs338105	T	C,G	5572.99	PASS	AC=2,1	GT	1/2	0/0	0/0

	Reference homozygous	Heterozygous	Alternative homozygous	Missing
VCF genotype	A/A	A/G	G/G	./.
Bits	00	01	10	11

	Reference homozygous	Heterozygous	Heterozygous	Alternative homozygous	Missing
VCF genotype	A|A	A|G	G|A	G|G	.|.
Bits	000	001	010	011	100

Compare the size of genotypes in chromosome 1 of 1000 Genomes Project

	VCF format	Plink linkage format file set	Plink binary genotype format file set	BGT format	KGGSeq binary genotype format file set
Unphased	64.234GB	66.093GB	3.92GB	342MB	793MB
Phased	64.234GB	66.093GB	—[a]	268MB	990MB

~1%

Coding advantages

- Flexible for phased and unphased genotypes
- Flexible for variants with multiple alternative alleles
- Facilitate fast computing

Calculate genotypic correlation

Conventional methods

$$r_{i,j} = \frac{n\left(\sum x_i x_j\right) - \left(\sum x_i\right)\left(\sum x_j\right)}{\sqrt{\left[n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2\right] * \left[n\left(\sum x_j^2\right) - \left(\sum x_j\right)^2\right]}}$$

x_i and x_j are genotypes

Time used by a bit-block based algorithm and the conventional algorithm for computing Pearson correlation of genotypes

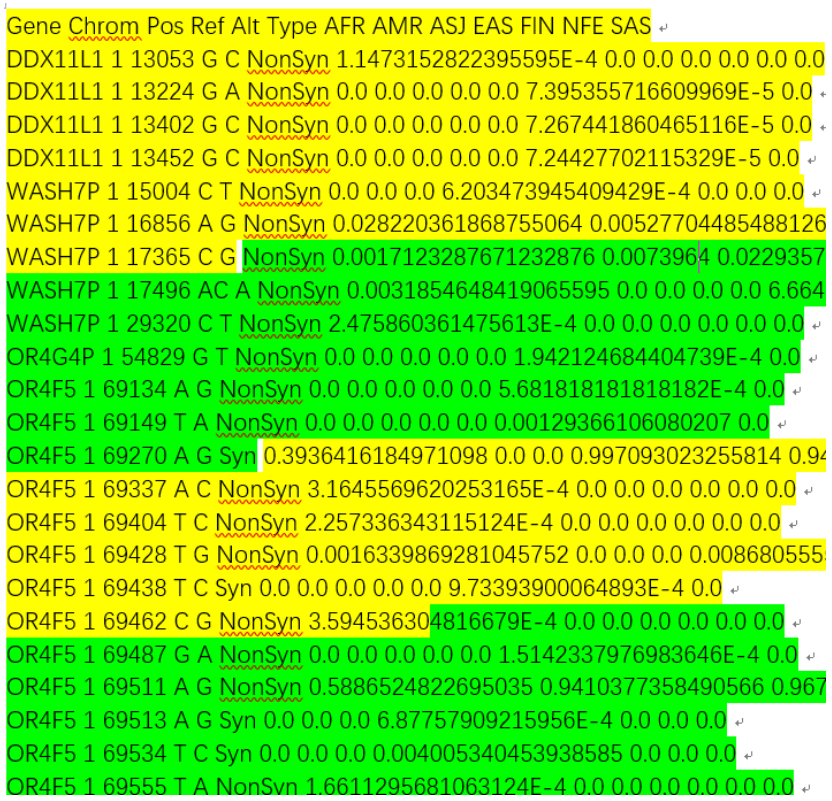
Number of variants	Bit-block	Conventional	Ratio
4000	0:0:2.59	50:01.36	1:1159
7000	0:0:7.59	2:31:56.48	1:1201
10000	0:0:14.99	5:08:54.14	1:1236
13000	0:0:23.60	8:43:17.78	1:1330
16000	0:0:35.18	13:18:33.96	1:1362
	Hour: Minute: Second	Hour: Minute: Second	

Nucleic Acids Res. 2017 Jan 23. pii: gkx019

Sectional accessing and optimized parsing text algorithm

- Sectional and random access of **compressed text lines (block-wised compressed format)**

1. Calculate initial positions in a compressed file to partition the file into approximately equal parts.
2. Search valid start and end reading positions around the initial positions and reserve the truncate lines
3. Read and parse the compressed data from the valid start reading positions by lines.
4. Merge all reserved data across the parts to splice the truncated lines between blocks after all parts have been read out.



Compare the speed to PLINK/SEQ and vcftools for parsing variants in VCF files

	Time	Maximal RAM used
kggseq (v1.0) 10 CPUs	35m	60MB*
kggseq (v1.0) 1 CPU	3h24m	36MB*
PLINK/SEQ (v0.10)^	24h25m	12MB
vcftools(v0.1.14)^	17h46m	6MB

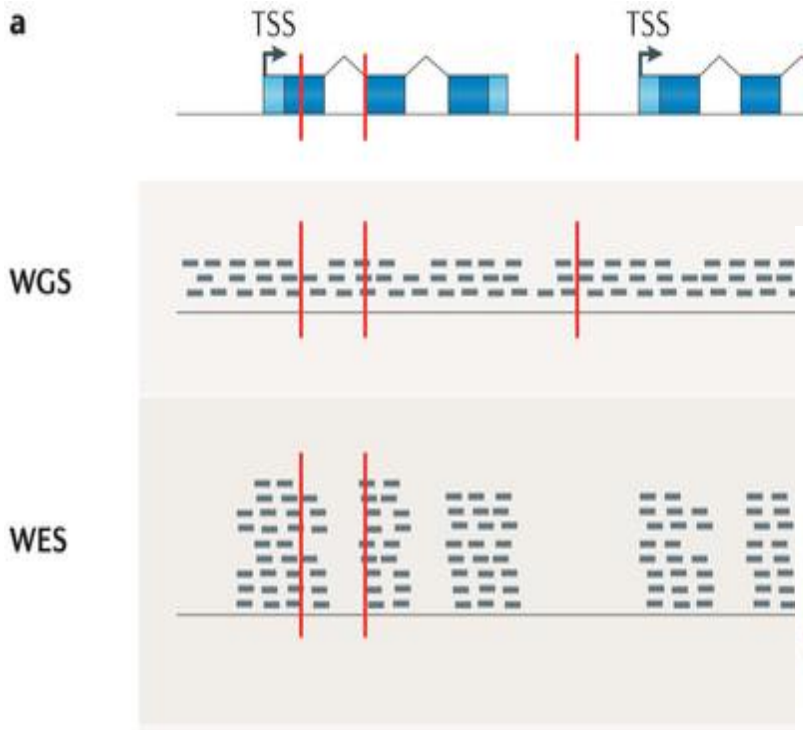
~40X

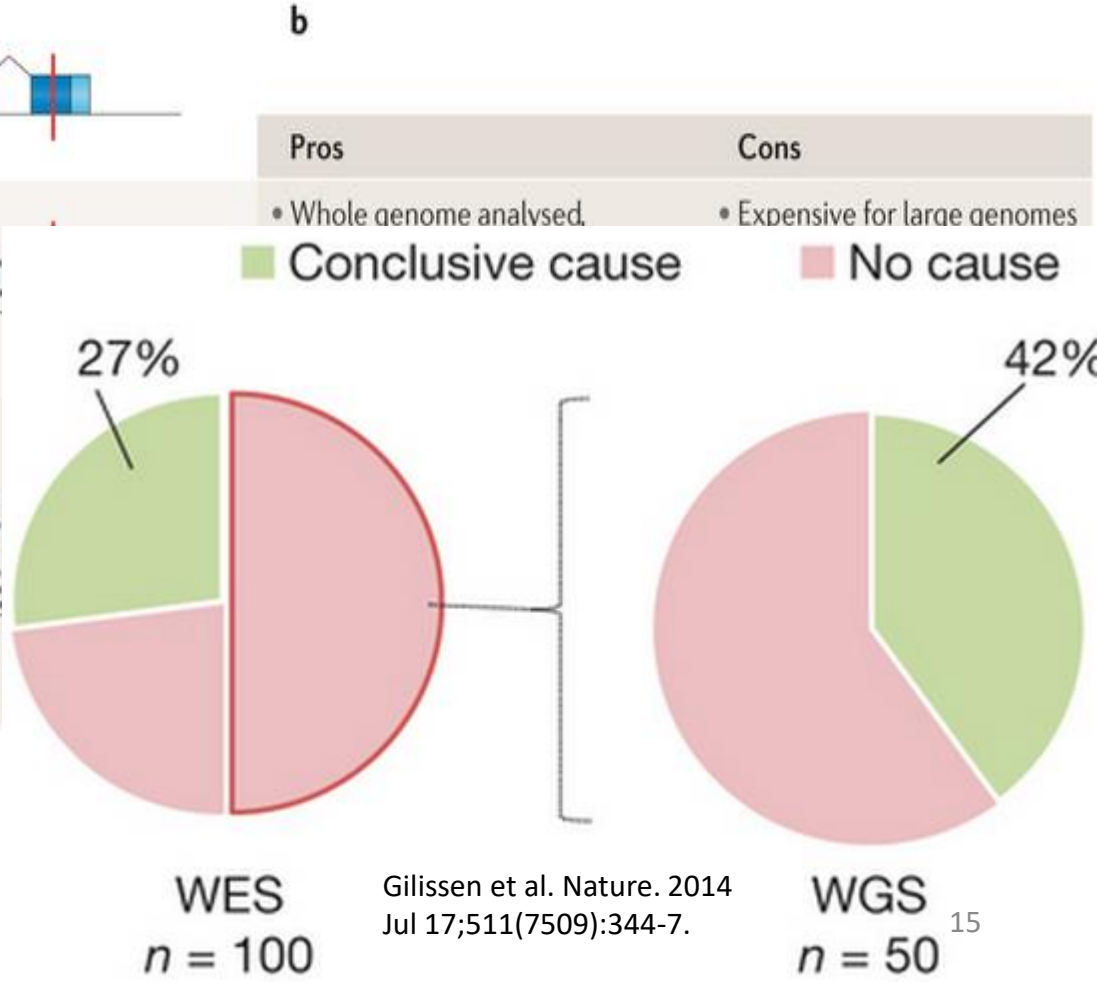
~8X

Testing dataset: 1KG whole genome sequencing data

NGS studies could fail …

Failed to find any candidate mutations

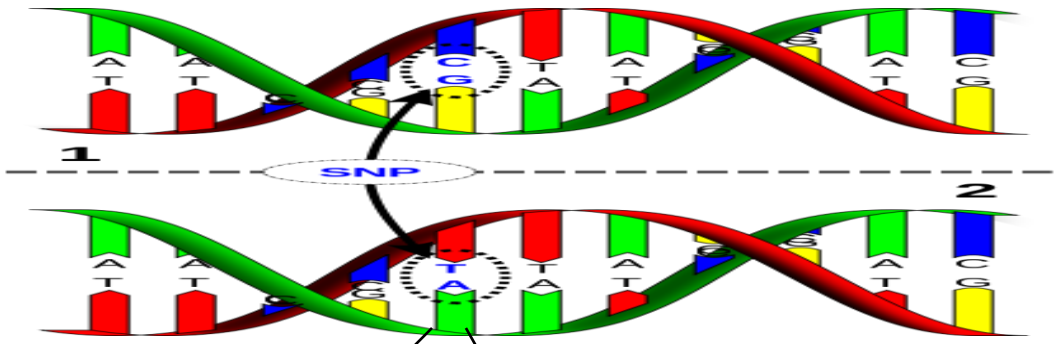


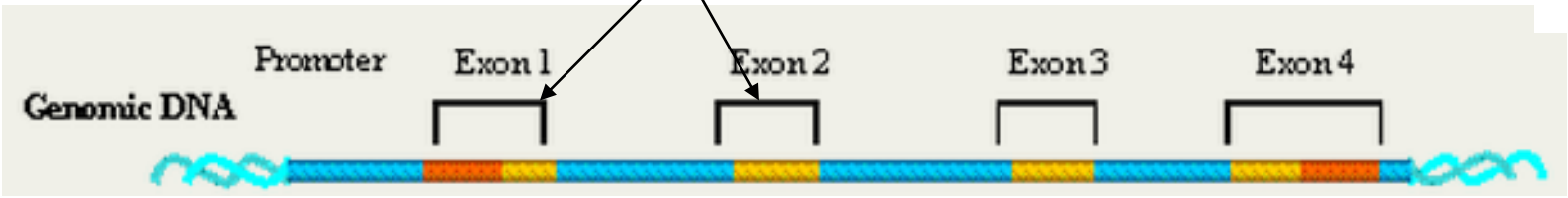


Gilissen et al. Nature. 2014 Jul 17;511(7509):344-7.

II. How can the data be analyzed more accurately?

Gene feature annotation-A seemingly simple question!





Feature	Explanation
Frameshift	Short insertion or deletion result in a completely different translation from the original.
Nonframeshift	Short insertion or deletion result in loss of amino acids in the translated proteins.
Startloss	Indels or nucleotide substitution result in the loss of start codon(ATG) (mutated into a non-start codon).
Stoploss	Indels or nucleotide substitution result in the loss of stop codons (TAG, TAA, TGA)
Stopgain	Indels or nucleotide substitution result in the new stop codons (TAG, TAA, TGA), which may truncate the protein.
Missense	Variants result in a codon coding for a different amino acid (missense)
Splicing	variant is within 2-bp of a splicing junction (use --splicing x to change this, the unit of x is base-pair)
Synonymous	Nucleotide substitution does not change amino acid.
Exonic	Due to loss of sequences, only map a variant into exonic region
UTR5	variant within a 5' untranslated region
UTR3	variant within a 3' untranslated region
Intronic	Variants within an intron
Upstream	variant overlaps 1-kb region upstream of transcription start site? (
Downstream	variant overlaps 1-kb region downtream of transcription enc
ncRNA	variant overlaps a transcript without coding annotation in the gene
Intergenic	variant is in intergenic region
Unknown	Variants failed to map

Various gene models:

RefSeq genes: NCBI Reference Sequence Database, 92,006 transcripts.

Ensembl genes: 196,501 transcripts

UCSC Known genes: 82,960 transcripts.

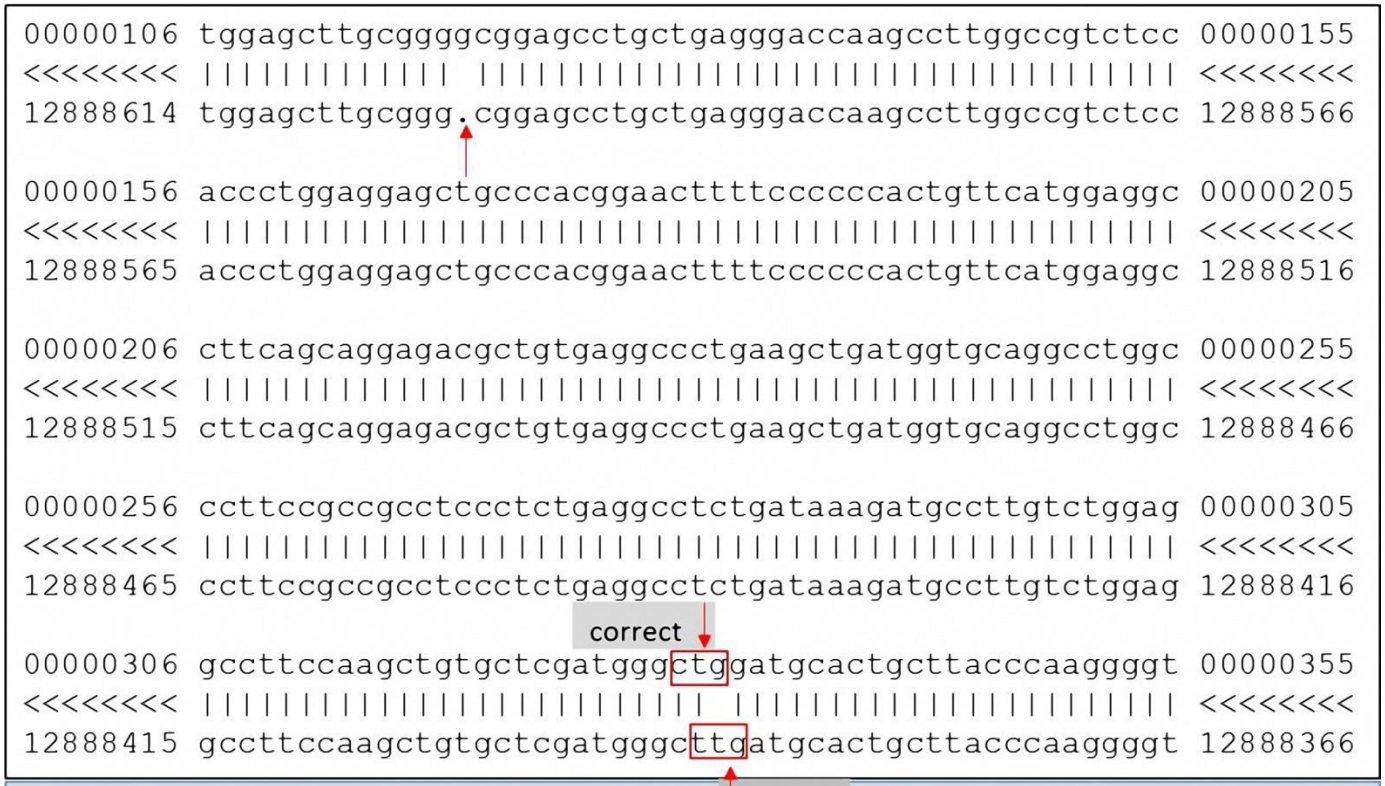
GEnocde: 196,520 transcripts

A sequence gap-filled gene feature annotation algorithm

RefSeq `cDNA NM_001146344`

`Reference genome hg19:chr1`

Chromosome	StartPositionHg19	ReferenceAlternativeAllele
1	69428	T/G
1	69534	T/C
1	865545	G/A
1	865664	C/T
1	871216	G/A
1	874762	C/T
1	874809	G/C
1	876592	C/T
1	878226	C/T
1	878254	C/T
1	878667	G/T
1	879180	C/T
1	879382	T/A



Variant: *rs1830486@chr1:12888389*

KGGSeq: PRAMEF11:NM_001146344:c.261T>G:p.L87L:synonymous

VEP: NM_001146344.1:synonymous_variant

ANNOVAR: exonic_unknown

SNPEff: NM_001146344::c.261T>G:p.L87W:missense

dbSNP: cds-synon: NM_001146344.2: 87 L⇒L

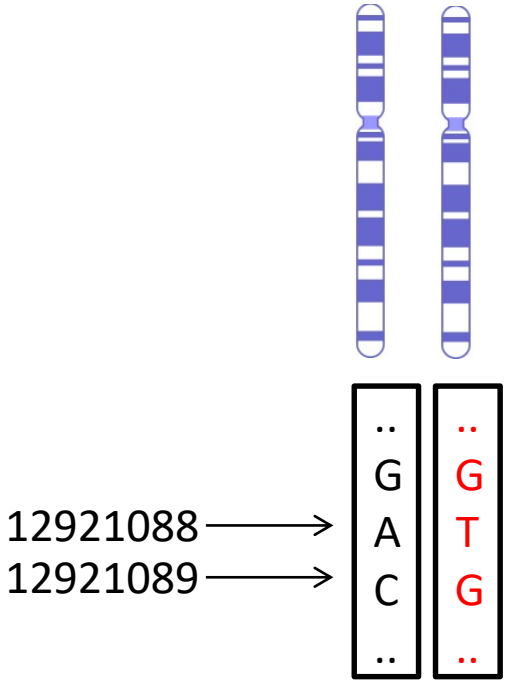
Number of exonic variants uniquely annotated by KGGSeq using gap-filled gene-feature annotation algorithm

	Unique without considering gap (dbSNP#[a], %[b])	overlapped	Unique with considering gap (dbSNP#[a], %[b])
startloss	0	492	0
Stoploss	0	223	2(1, 100%)
Stopgain	87(66, 1.52%)	3905	30(22, 86.36%)
Splicing	0	2287	0
Missense	829(651, 2%)	225145	549(443, 95%)
Synonymous	493(396, 0.25%)	166857	858(672, 93.75%)
Total	**1409**	**398910**	**1439**

Compare non-synonymous gene feature annotation of three popular tools

	RefGene								
	KGGSeq vs. ANNOVAR(2015Jun17)			KGGSeq vs. SNPEff(v4_1k)			KGGSeq vs. VEP(v81)		
	KGGSeq Unique (dbSNP#[a], %[b])	overlapped	ANNOVAR Unique (dbSNP#[a], %[b])	KGGSeq Unique (dbSNP#[a], %[b])	overlapped	SNPEff Unique (dbSNP#[a], %[b])	KGGSeq Unique (dbSNP#[a], %[b])	overlapped	VEP Unique (dbSNP#[a], %[b])
startloss[e]	—[c, d]	—[c]	—[c, d]	56	436	1	35	457	227
stoploss	7(5, 80%)	218	2(2, 100%)	21(15, 100%)	204	81(72, 2.78%)	15(10, 90%)	210	141(126, 62.7%)
stopgain	124(92, 71.74%)	3811	4(4, 50%)	171(139, 89.93%)	3764	128(118, 1.69%)	180(134, 83.58%)	3755	597(530, 83.58%)
splicing	96(90, 92.22%)	2191	3(2, 100%)	2286(2093, 98.09%)	1	1902(1764, 0.4%)	105(84, 73.81%)	2182	43322(40572, 3.18%)
missense	3965(87.08%)	222213	113(94, 55.32%)	6422(5868, 96.1%)	219272	1197(1027, 11.88%)	13021(10719, 94.88%)	212673	12978(11885, 94.26%)
total	**4192**	**228433**	**122**	**8956**	**223677**	**3309**	**13356**	**219277**	**18275**
	GENCODE(v19)								
startloss[e]	—[c, d]	—[c]	—[c, d]	73	580	94	0	653	44
stoploss	3(2, 100%)	337	7(5, 20%)	29(23, 52.17%)	311	129(115, 7.83%)	2(1, 100%)	338	40(25, 4%)
stopgain	30(23, 56.52%)	4222	156(52, 25%)	178(152, 84.21%)	4074	556(415, 13.01%)	10(7, 100%)	4242	163(58, 31.03%)
splicing	157(146, 73.29%)	2460	76(48, 31.25%)	2614(2360, 88.18%)	3	4283(3859, 0.34%)	18(14, 64.29%)	2599	49203(41541, 1.56%)
missense	680(509, 75.64%)	231245	5163(2121, 51.67%)	6773(6091, 94.07%)	225152	10948(7239, 22.82%)	6873(6314, 94.04%)	225052	4679(1582, 40.39%)
total	**870**	**238264**	**5402**	**9667**	**230120**	**16010**	**6903**	**232884**	**54129**

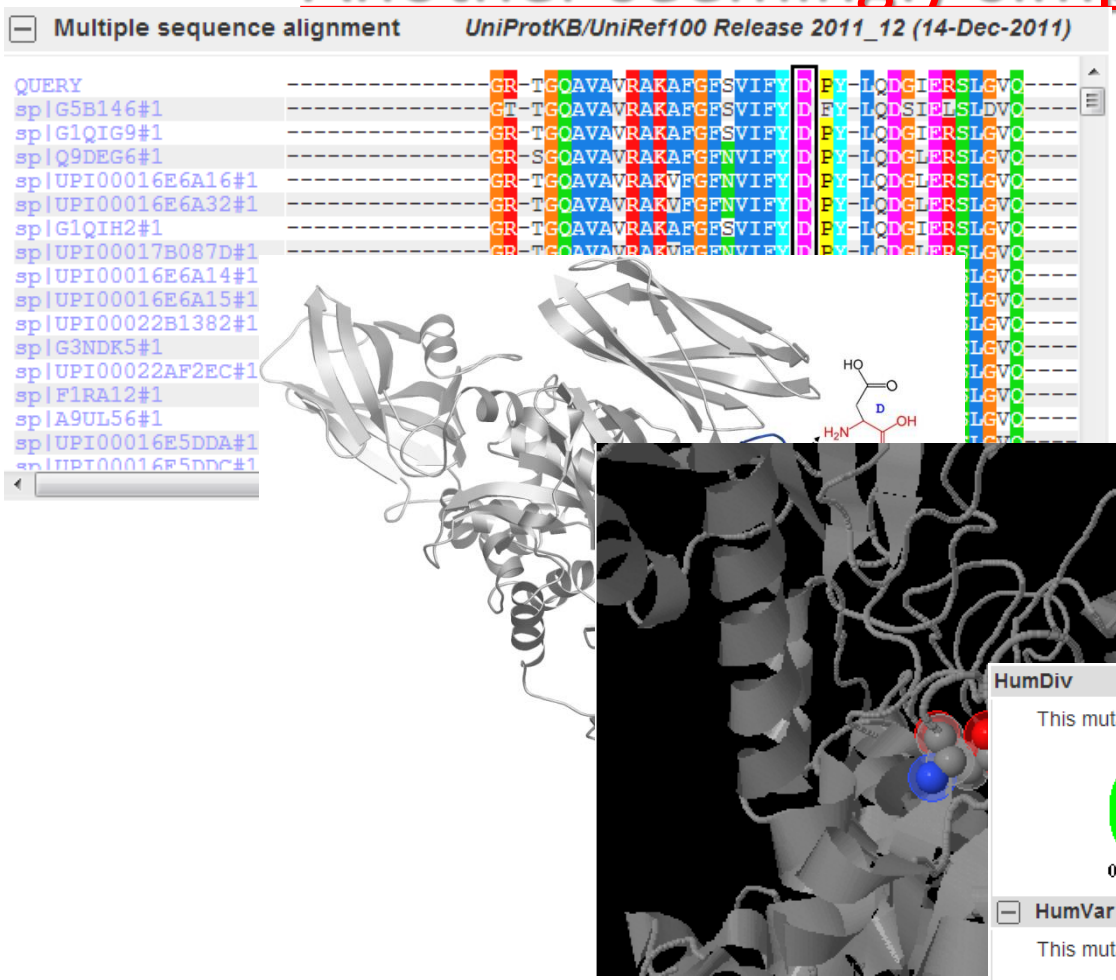
Multi-variants annotation

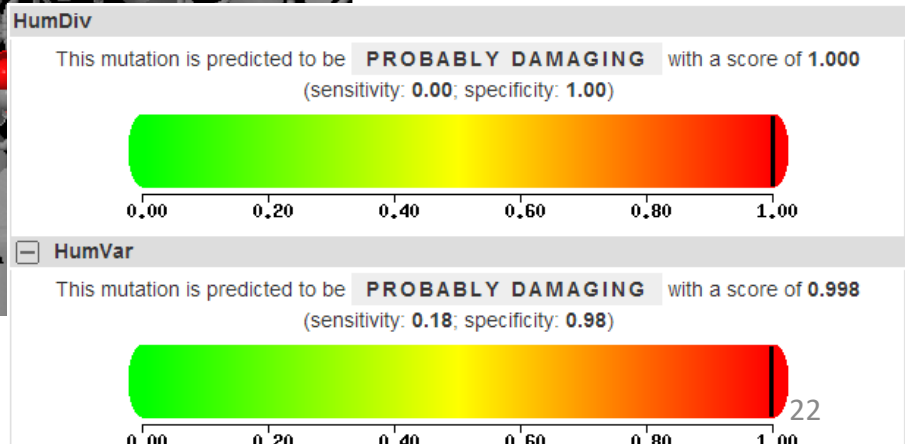


Chrom	Pos	Ref/Alt	Feature
1	12921087	A/T	PRAMEF2:NM_023014:c.878A>T:p.N293I:(4Exons):exon 4:missense
1	12921088	C/G	PRAMEF2:NM_023014:c.879C>G:p.N293K:(4Exons):exo n4:missense
1	12921087	AC/TG	PRAMEF2:NM_023014:c.878AC>TG:p.N293M:(4Exons): exon4:missense

Pathogenic prediction at protein coding mutations

Another seemingly simple question!





Pathogenic prediction at protein coding mutations

Chromosome	StartPosition	ReferenceAlternativeAllele	GeneSymbol	Polyphen2_HDIV_pred	Polyphen2_HVAR_pred	LRT_pred	MutationTaster_pred	MutationAssessor_pred
8	133251794	A/C	KCNQ3	D;D	D;D	D	D	medium
1	68669602	G/C	RPE65	B	B	N	D	low
17	7517842	A/G	TP53	P;P;P;P	P;P;D;P	N	D	medium
18	45817276	C/T	MYO5B	B;P	B;B	D	D	medium
10	127412194	C/T	C10orf137	D;D;D	D;D;P	D	D	low
3	170968107	C/T	ACTRT3	D	D	N	D	high

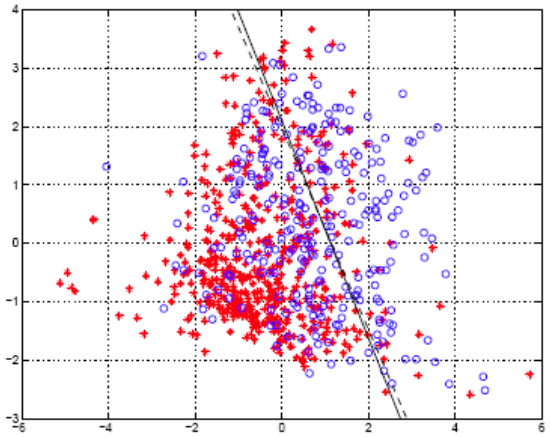
Population	Individual	Number of rare nsSNVs used[a]	% of rare variants predicted to be pathogenic	% of rare nsSNVs truly pathogenic	Total load of pathogenic derived alleles (95% CI)[b]
Caucasian	NA12156	384	20.3	5.5	21 (5, 36)
	NA12878	426	24.4	12.0	51 (33, 68)
Japanese	NA18956	356	19.1	3.6	13 (0, 26)
Chinese	NA18555	424	21.2	6.9	29 (12, 46)
African	NA18517	660	17.1	0.4	3 (0, 28)
	NA18507	623	20.9	6.4	40 (14, 64)
	NA19129	629	17.5	1.0	6 (0, 31)
	NA19240	688	18.3	2.3	16 (0, 42)

[a]The nsSNVs with missing scores at SIFT and/or MutationTaster were not used in the estimation.
[b]the 95% confidence interval was derived empirically from randomly repeating 10-fold cross-validation 200 times.
doi:10.1371/journal.pgen.1003143.t002

Combined prediction

$$combined = weight_1 * tool_1 + weight_2 * tool_2 + weight_3 * tool_3 + \cdots$$

Logistic regression model



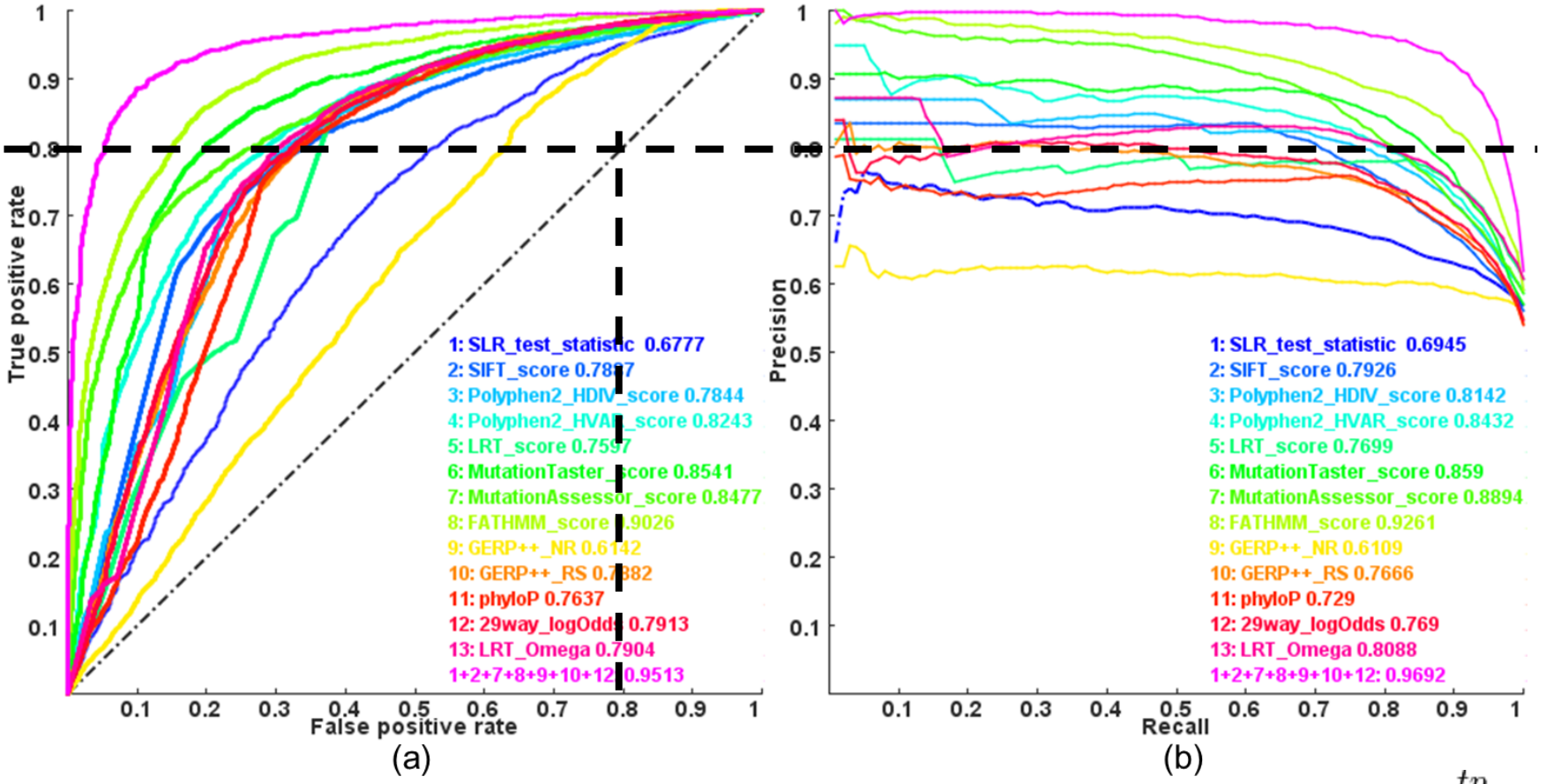
- Benchmark dataset

  5,340 disease-causal alleles vs. 4,752 **rare** non-disease-causal nsSNVs

$$\Pr(D = 1 | X = (S_1, S_2, \ldots S_n)) = ???$$

***Li et al*. PLOS Genetics; 2014; 9(1):e1003143**

Distinguishing pathogenic nsSNVs from other **rare** nsSNVs



receiver operating characteristic (ROC) curve

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp_{25}}{tp + fn}$$

***Li et al*. PLOS Genetics; 2014; 9(1):e1003143**

Pathogenic prediction by new tools

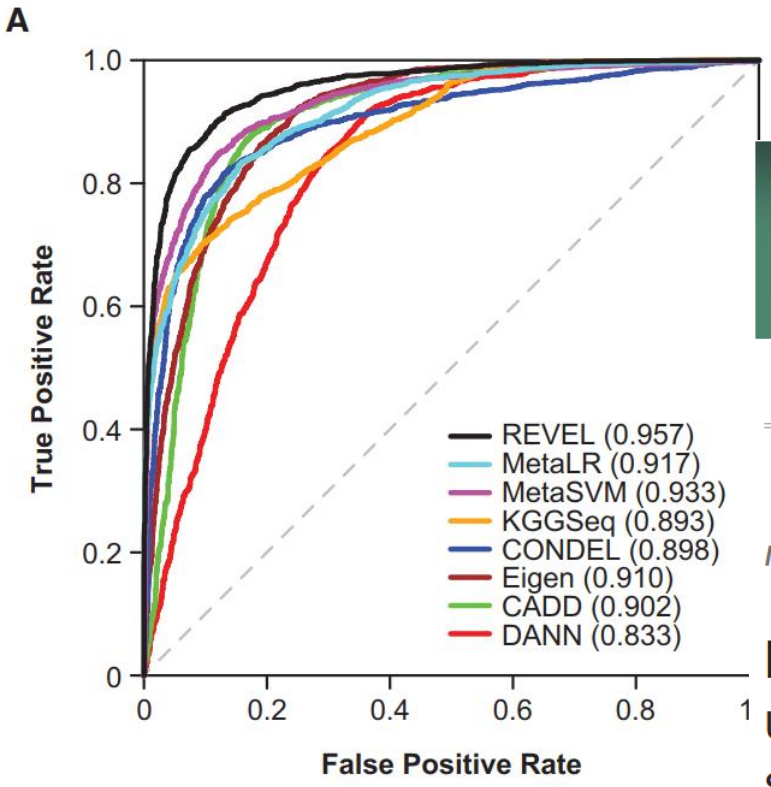


Figure 4. Performance of Ensemble Methods in an In
(A) ROC curves and the AUC for all variants.
(B) AUC for each ensemble method, stratified by neut

B

1.00

nature genetics

Home | Current issue | Comment | Research | Archive ▾ | Authors & referees ▾ | About the journal ▾

home ▸ advance online publication ▸ abstract

ARTICLE PREVIEW

view full access options ▸

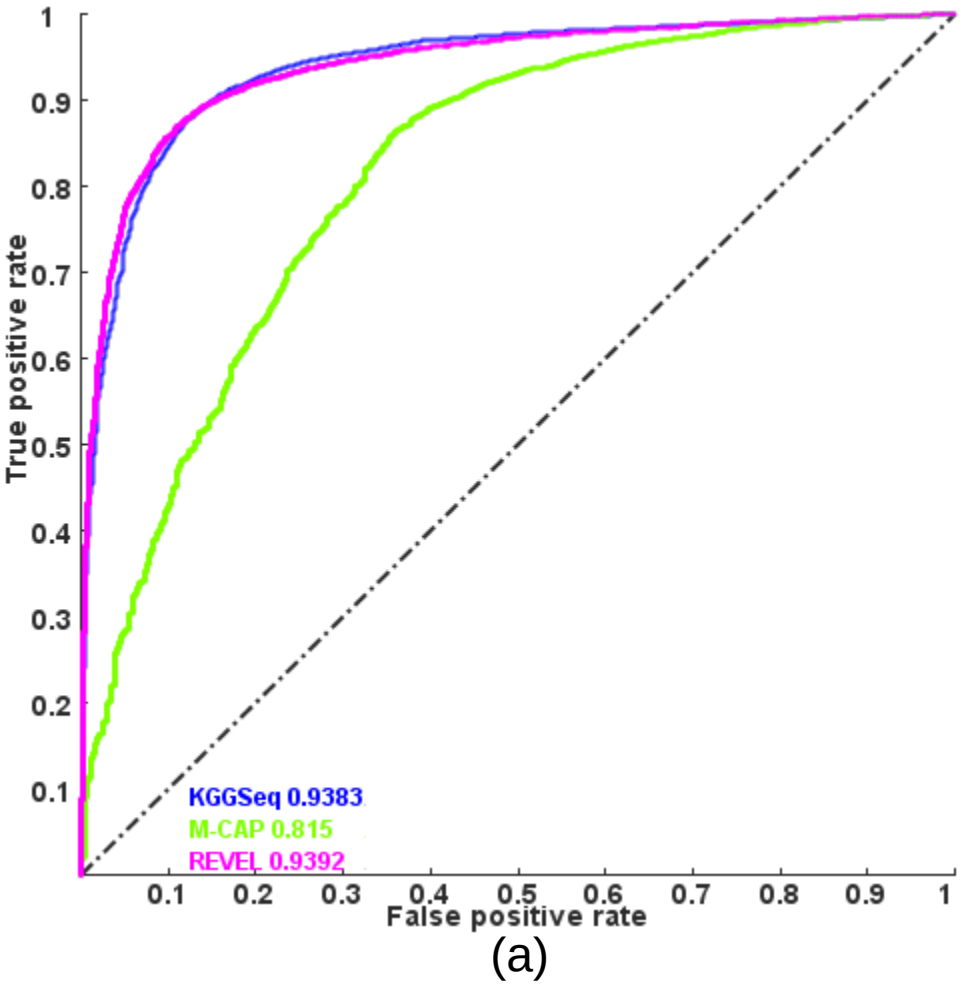
NATURE GENETICS | TECHNICAL REPORT

M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity

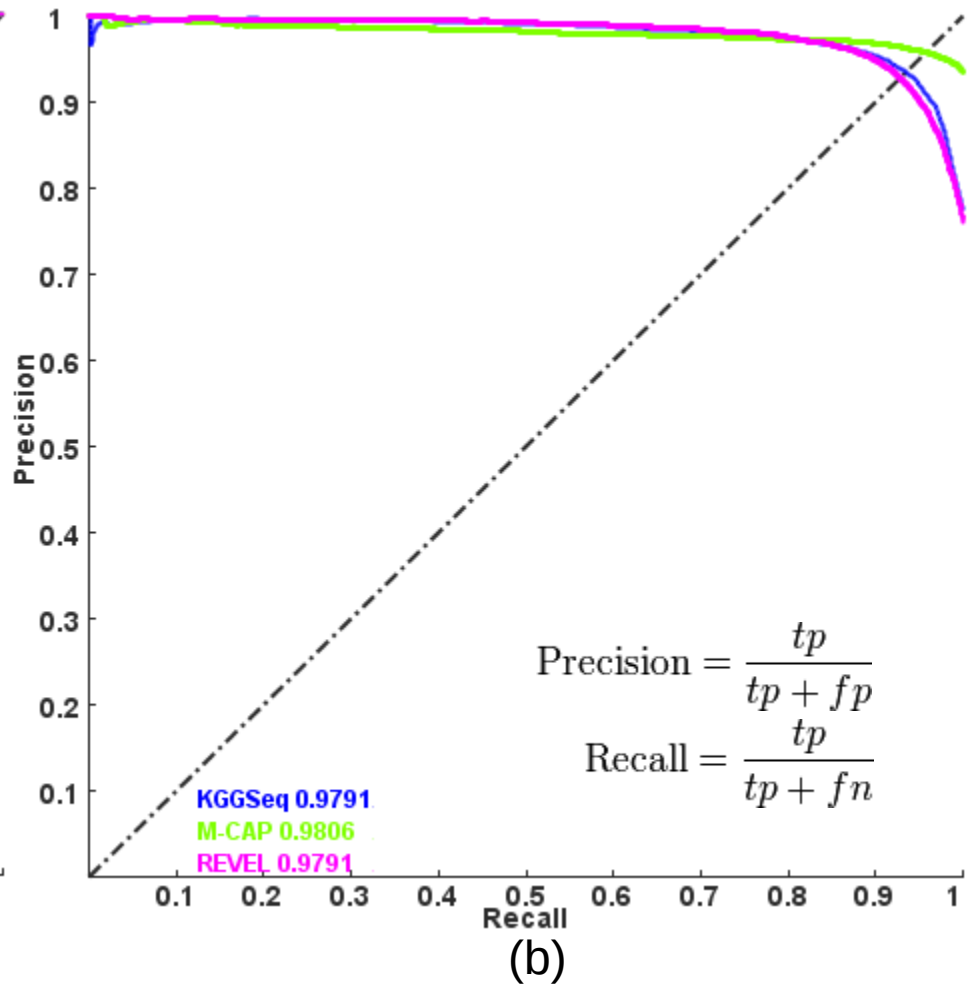
Karthik A Jagadeesh, Aaron M Wenger, Mark J Berger, Harendra Guturu, Peter D Stenson, David N Cooper, Jonathan A Bernstein & Gill Bejerano

Affiliations | Contributions | Corresponding author

Nature Genetics (2016) | doi:10.1038/ng.3703
Received 30 June 2016 | Accepted 26 September 2016 | Published online 24 October 2016



(a)



(b)

Evaluation with 1898 non-synonymous pathogenic variants and 2180 benign in ClinVar

KGGSeq: PLoS Genet. 2013;9(1):e1003143

M-CAP: Nat Genet. 2016 Dec;48(12):1581-1586.

REVEL: Am J Hum Genet. 2016;99(4):877-885

OXFORD

Human Molecular Genetics, 2015, Vol. 24, No. 8 2125–2137

doi: 10.1093/hmg/ddu733
Advance Access Publication Date: 30 December 2014
Original Article

ORIGINAL ARTICLE

Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies

Chengliang Dong[1,2,†], Peng Wei[4,6,†], Xueqiu Jian[5], Richard Gibbs[7], Eric Boerwinkle[4,5,7], Kai Wang[1,2,3,*] and Xiaoming Liu[4,5,*]

Abstract

Accurate deleteriousness prediction for nonsynonymous variants is crucial for distinguishing pathogenic mutations from background polymorphisms in whole exome sequencing (WES) studies. Although many deleteriousness prediction methods have been developed, their prediction results are sometimes inconsistent with each other and their relative merits are still unclear in practical applications. To address these issues, we comprehensively evaluated the predictive performance of 18 current deleteriousness-scoring methods, including 11 function prediction scores (PolyPhen-2, SIFT, MutationTaster, Mutation Assessor, FATHMM, LRT, PANTHER, PhD-SNP, SNAP, SNPs&GO and MutPred), 3 conservation scores (GERP++, SiPhy and PhyloP) and 4 ensemble scores (CADD, PON-P, KGGSeq and CONDEL). We found that FATHMM and KGGSeq had the highest discriminative power among independent scores and ensemble scores, respectively. Moreover, to ensure unbiased performance evaluation of these prediction scores, we manually collected three distinct testing datasets, on which no current prediction scores were tuned. In addition, we developed two new ensemble scores that integrate nine independent scores and allele frequency. Our scores achieved the highest discriminative power compared with all the deleteriousness prediction scores tested and showed low false-positive prediction rate for benign yet rare nonsynonymous variants, which demonstrated the value of combining information from multiple orthologous approaches. Finally, to facilitate variant prioritization in WES studies, we have pre-computed our ensemble scores for 87 347 044 possible variants in the whole-exome and made them publicly available through the ANNOVAR software and the dbNSFP database.

Pathogenic prediction at non-coding mutations

A difficult question!

Context-dependent epigenomic weighting improves identification of regulatory variants and disease-associated genes

Bench-mark datasets

- eQTLs fine mapping data from eleven studies on seven tissues/cell lines, and thirteen GTEx tissues

Key features

- 36 chromatin features to evaluate variant regulatory potential = (hit, score, centrality) * (DNase, H2AZ, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3, H4K20me1) for matched GTEx tissues/cell lines

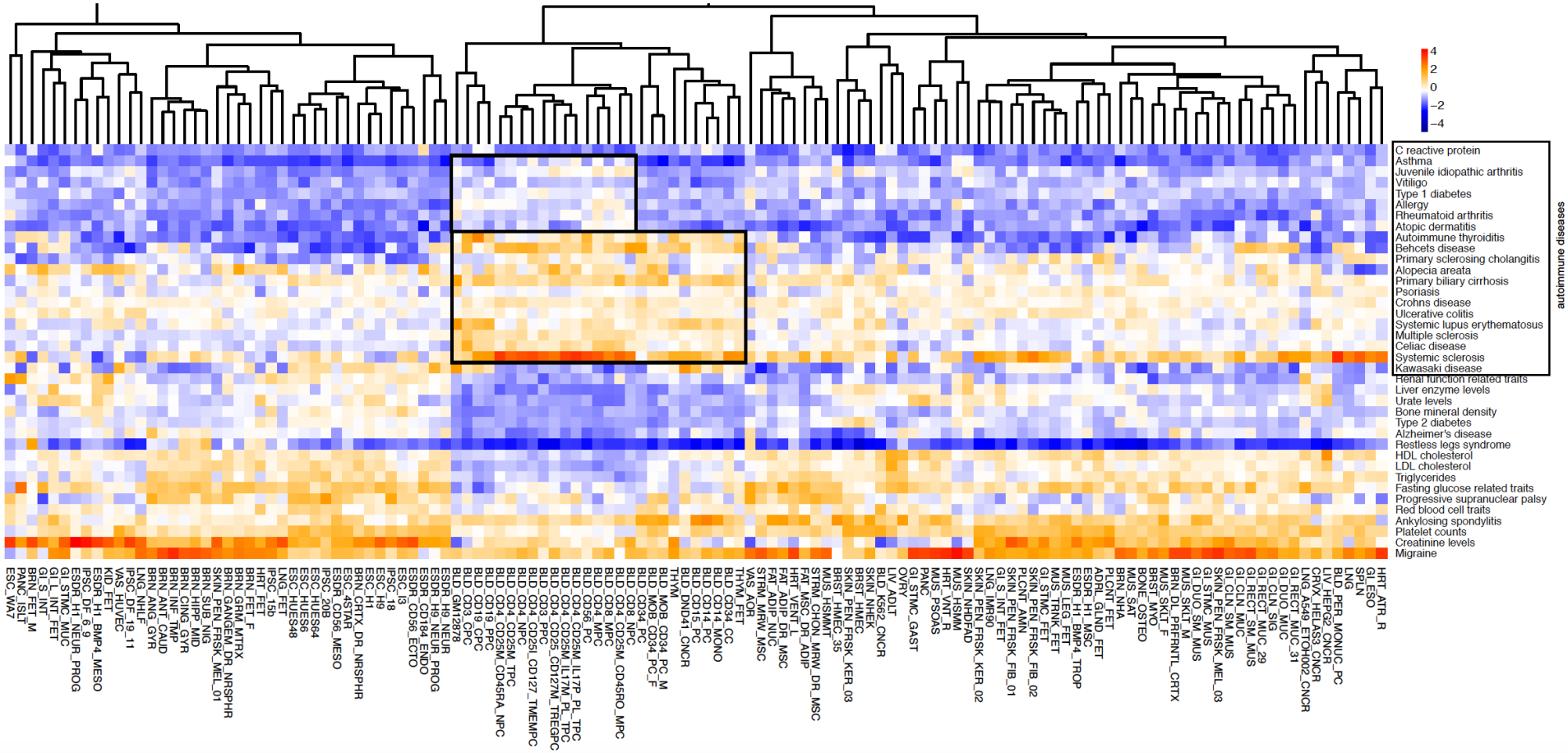
Models

- Logit model and model selection

$$P(causal|X) = \frac{1}{1 + e^{-(\alpha + \beta X)}}$$

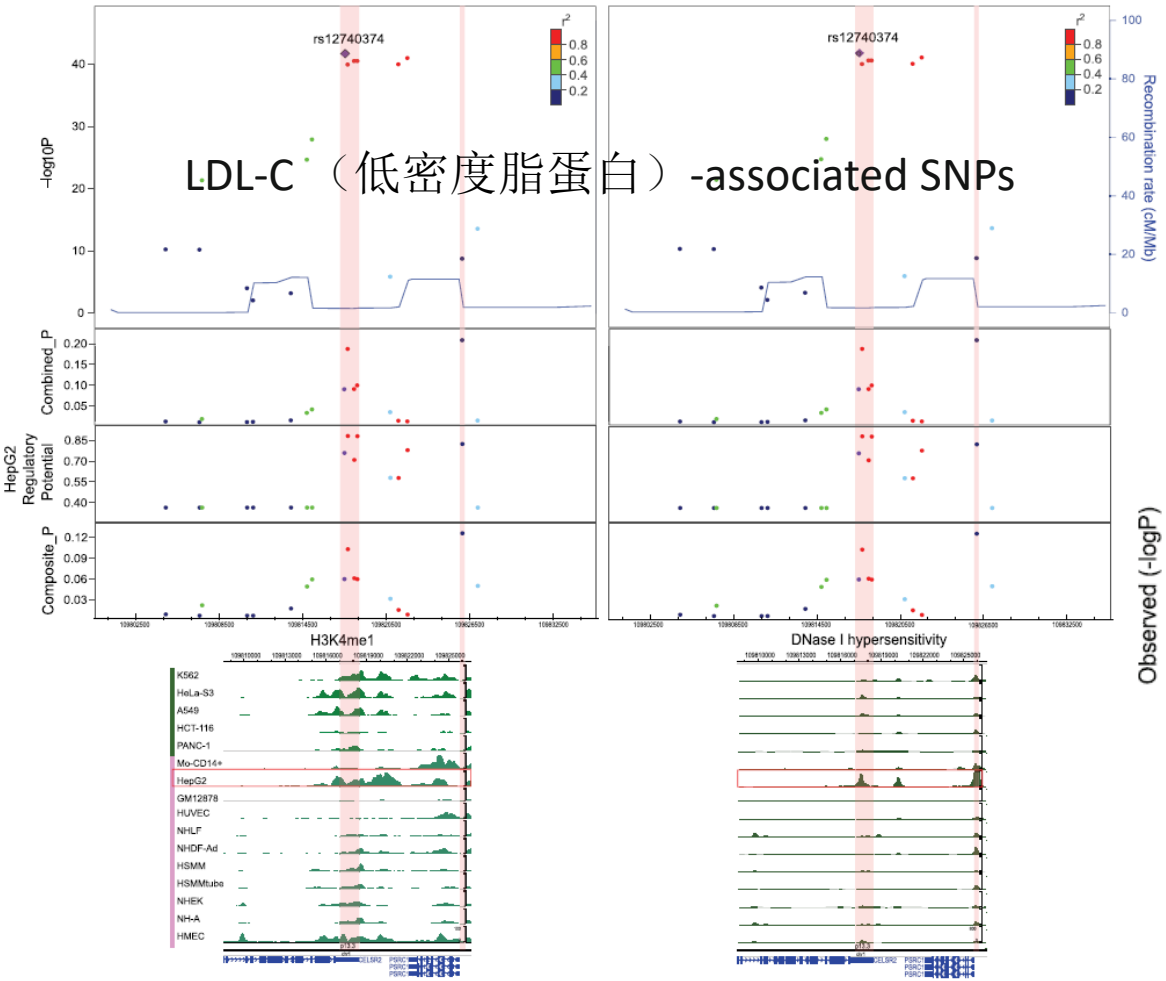
Li J*, Li M*, et al. *Genome Biol. 2017* Mar 16;18(1):52.

Context-dependent scoring of GWAS fine-mapped SNPs underlies phenotypic cell type specificity

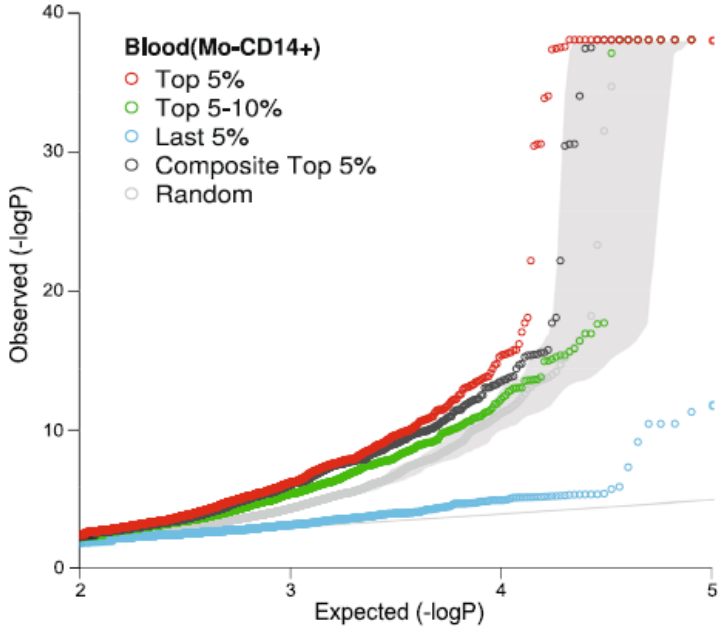


Li J*, Li M*, et al. *Genome Biol. 2017* Mar 16;18(1):52.

Improve the statistical power for identifying genes associated with complex diseases



Rheumatoid Arthritis （关节炎）-associated Genes



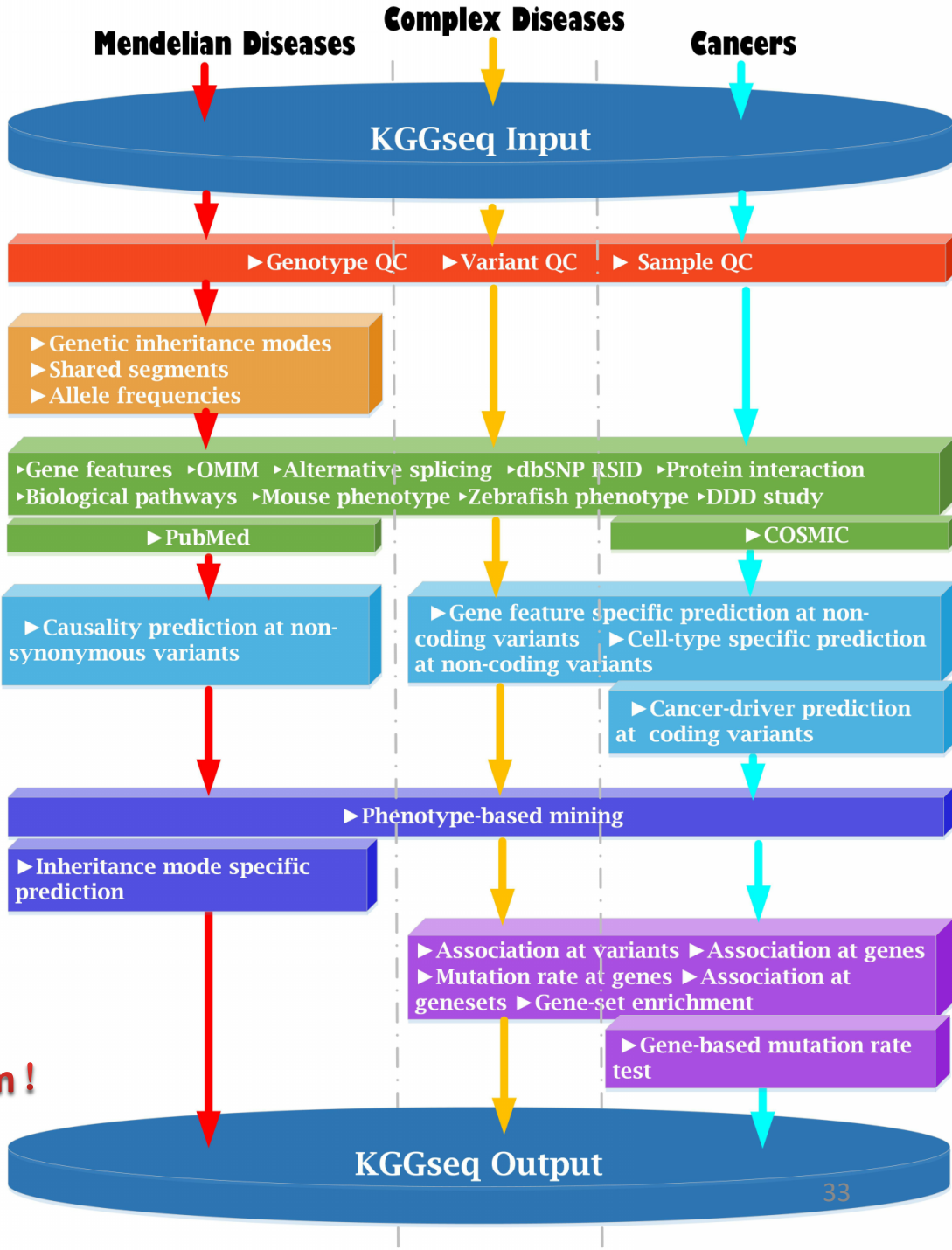
Li J*, Li M*, et al. *Genome Biol. 2017* Mar 16;18(1):52.

III. How can the high-throughput sequencing data be used for different types of genetic diseases?

KGGSeq V1.0+

- Quality control
  - Genotype QC
  - Variant QC
  - Sample QC
- Filtering
  - Genetic inheritance
  - Gene feature
  - Allele frequency
  - Type&region&gene
  - Supper duplicate
  - Gene's variant number
  - Phenotype
- Annotation
  - dbSNP RSID
  - Alternative splicing
  - Structure variation
  - Protein interaction
  - GeneSet
  - Mouse phenotype
  - Zebrafish Phenotype
  - DDD Study
  - PubMed literature
  - OMIM
  - COSMIC
- Prediction at variant
  - Mendelian disease
  - Cancer somatic driver
  - Complex disease
- Prediction at gene
  - Pathogenicity
  - Phenotype mining
  - Expression
- Statistical tests
  - Association at variant
  - Association at gene
    - SKAT
    - Rvtests
  - Mutation rate at gene
  - Association at geneset
    - SKAT
    - Rvtests
  - Enrichment at geneset
- Plotting
  - Allele frequency
  - Q-Q
- Miscellaneous
  - Linkage Disequilibrium
  - LD Pruning

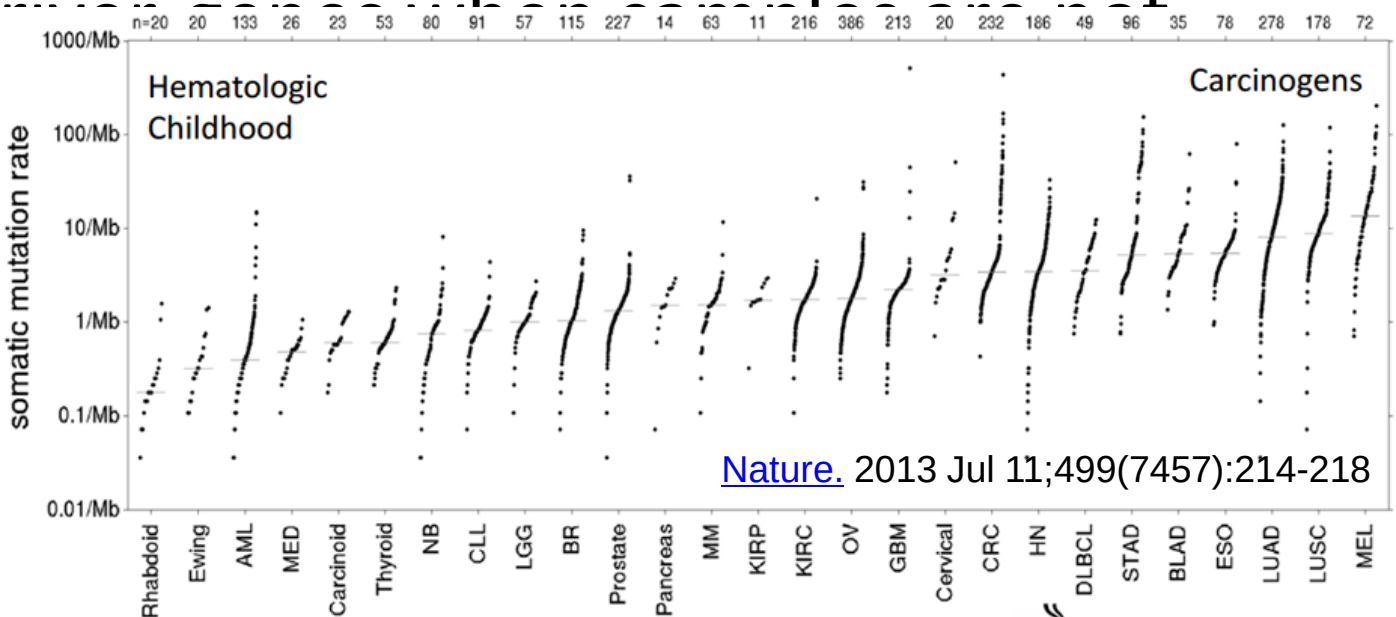
A comprehensive platform !



Identifying cancer-driver genes by somatic mutations

Challenges

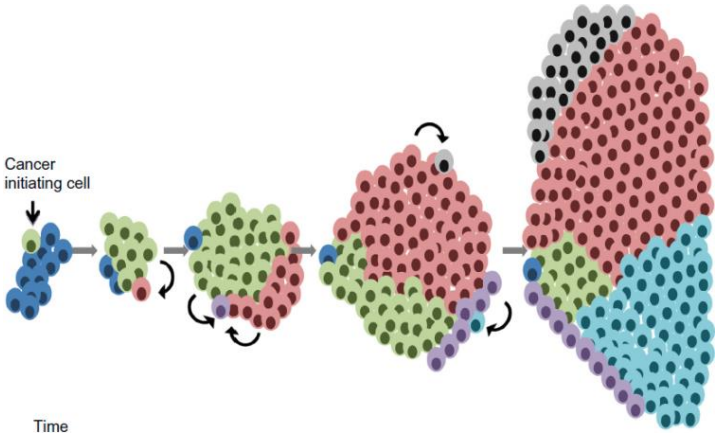
- The mutation rates vary across cancers and difficult to model somatic mutations;
- Existing methods are underpowered to identify d[illegible] huge.



A method for identifying cancer-driver genes by somatic mutations

Hypothesis :

- Genomic aberrations in somatic cells are major drivers of cancers;
- Driver-mutations conferring growth advantages of cancer cells and thus having higher mutations in cancer patients.



Authors Raza A, Iverson N, Ali A
Published 14 July 2014 Volume 2014:4
DOI https://doi.org/10.2147/AGG.S54184

A model of identifying rare mutations without using controls

Challenges:

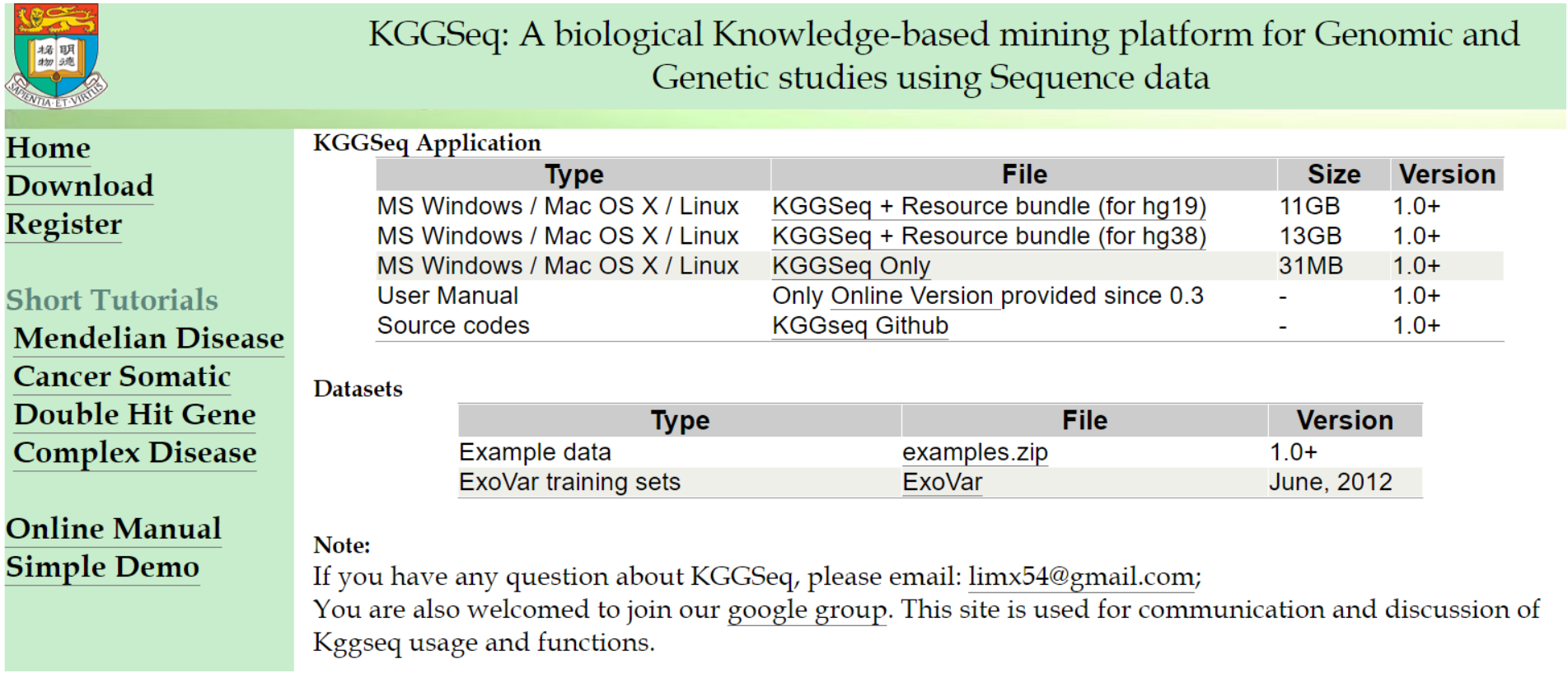
- Existing methods are underpowered to detect rare causal mutations unless the sample is huge.
- Spurious associations due to population structure

Hypothesis :

- Mutations of causal genes are excessive in human

AVAILABILITY

http://grass.cgs.hku.hk/limx/kggseq



KGGSeq: A biological Knowledge-based mining platform for Genomic and Genetic studies using Sequence data

Home
Download
Register

Short Tutorials
Mendelian Disease
Cancer Somatic
Double Hit Gene
Complex Disease

Online Manual
Simple Demo

KGGSeq Application

Type	File	Size	Version
MS Windows / Mac OS X / Linux	KGGSeq + Resource bundle (for hg19)	11GB	1.0+
MS Windows / Mac OS X / Linux	KGGSeq + Resource bundle (for hg38)	13GB	1.0+
MS Windows / Mac OS X / Linux	KGGSeq Only	31MB	1.0+
User Manual	Only Online Version provided since 0.3	-	1.0+
Source codes	KGGseq Github	-	1.0+

Datasets

Type	File	Version
Example data	examples.zip	1.0+
ExoVar training sets	ExoVar	June, 2012

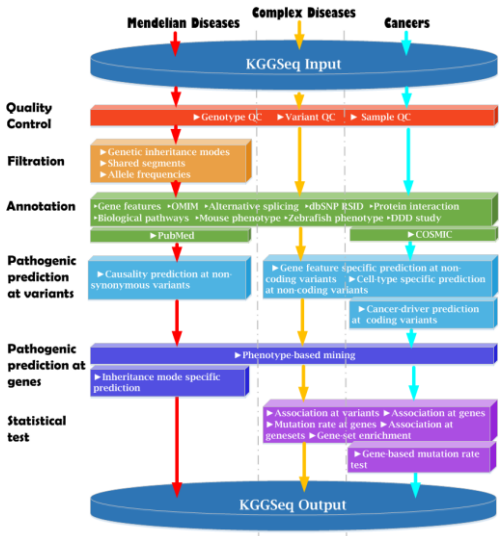
Note:
If you have any question about KGGSeq, please email: limx54@gmail.com;
You are also welcomed to join our google group. This site is used for communication and discussion of Kggseq usage and functions.

Methodological studies on KGGSeq(A biological Knowledge-based mining platform for Genomic and Genetic studies using Sequence data)

Methods developed by my research team for downstream analysis

- ✓ A powerful statistical model detecting rare-causal mutations of complex diseases using case-only samples (On-going)
- ✓ A powerful statistical model detecting cancer-driver genes using somatic mutations (In submission)
- ✓ Advanced algorithms for accurate and fast analyses of whole genome sequencing data of human diseases [Li M et al. *Nucleic Acids Res*. 2017 May 19;45(9):e75]
- ✓ Tissue-specific functional prediction of non-coding variants[,Li J*, Li M*, et al. *Genome Biol. 2017* Mar 16;18(1):52.]
- ✓ A multi-layer bioinformatics framework to prioritize Mendelian disease causal mutations with exome-sequencing data [Li M et al. Nucleic Acids Res. 2012; 40(7):e53]
- ✓ A method for accurate prediction of disease-causal mutations [Li M et al. *PLoS Genet*. 2013;9(1):e1003143]
- ✓ A inheritance-model based prediction model for disease causal genes [..., Li M. *Bioinformatics*. 2016;32(20):3065-3071]

KGGSeq (One of main-stream platforms in the world for d**ownstream-analysis of high-throughput sequencing data**)



Keep updating for 8+ years， downloaded for 3000+ times!

In Summary

- Effective to use the block-bit-based algorithm to reduce the computing space and time in NGS data analysis;
- The block-based compression algorithms to facilitate parallel computing;
- Subtly designed annotation and prioritization algorithms to remove noises;
- Advanced methods to models distributions of rare mutations for more powerful identification of disease associated genes

Acknowledgments

- KGGSeq team members:

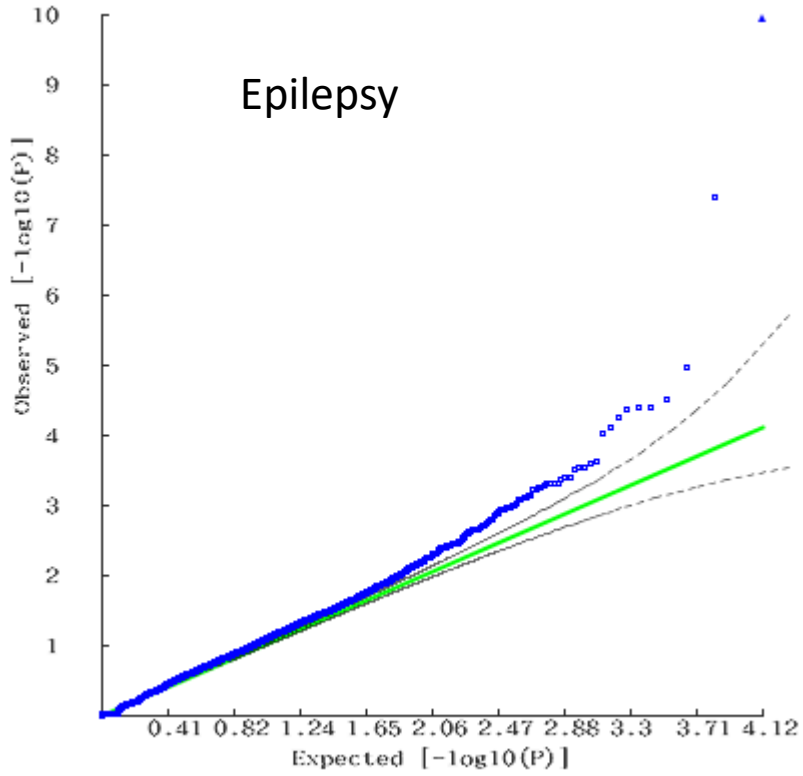
  Pak C Sham, Allen Hongsheng Gui, Suying Bao, Johnny Kwan, Li Jun, Li Yan, John…

- Clinical collaborators

- Active users of KGGSeq

100 Talents Project of Sun Yat-sen Univerity
HK GRF: 776412, 777511
HK HMRF: 02132236, 01121436
HKU SRT on Genomics
HKU Seed Funding: 201411159172; 201311159090; 201302159006

Application in real datasets



Unpublished!