Table S1. GeMSTONE settings analogous to analysis guidelines for investigating causality of sequence variants in human disease

The table presents how GeMSTONE's centralized workflow aligns to guidelines and consensus recommendations for interpreting variant causality in human [as listed in Table 1 and BOX 2 of MacArthur *et al.* paper (19)].

		Guideline Highlights (Table 1 of MacArthur *et al.*)	**Guideline Explanation (BOX 2 or Text of MacArthur *et al.*)**	**Relevant GeMSTONE Settings**	**Description of Settings (* Indicates Unique Feature of GeMSTONE)**
Variant level	**Genetic**	**Association**: the variant is significantly enriched in cases compared to controls.	Determine and report the formal statistical evidence for segregation or association of each variant.	GENE ANNOTATION (4) + Genetic Association Test + Single variant association	Fisher's exact test to perform allelic, dominant and recessive single variant tests by PLINK/SEQ.
		Segregation: the variant is co-inherited with disease status within affected families and additional co-segregating pathogenic variants are unlikely or have been excluded.	Replication of newly implicated disease genes in independent families or population cohorts is critical supporting evidence, and in most cases essential for a novel gene to be regarded as convincingly implicated in disease.	PEDIGREE + Inheritance Model + Recurrence FILE INPUT + File Upload + Pedigree File Upload (Specifics on .ped format customization)	Co-segregation analysis for six inheritance models with customization on stringency (e.g. inclusion of variants shared with siblings who may have not developed the disease phenotype) via user-customized pedigree file. Recurrence constraint to eliminate false positives and for novel gene discovery*.
		Population frequency: the variant is found at a low frequency, consistent with the proposed inheritance model and disease prevalence, in large population cohorts with similar ancestry to patients.	Determine and report the frequency of each variant in large control populations matched as closely as possible to patients in terms of ancestry.	*To annotate*: ALLELE FREQUENCY DATABASES *To filter*: FILE INPUT + File Upload + Pedigree File Upload VCF FILTERS + Allele Frequency Upperbound	Annotation of all variants with allele frequencies from 4 different databases of large control populations; and/or filter by allele frequency in relation to any sub-population within them in an sample-specific manner to match as closely as possible to samples' ethnicities*.
	Computational Genomic	**Conservation**: the site of the variant displays evolutionary conservation consistent with deleterious effects of sequence changes at that location.	Predict variant deleteriousness with comparative genomic approaches, but avoid considering any single method as definitive or multiple methods as independent lines of evidence for implication.	VARIANT FUNCTION + Conservation Scores	23 different *in silico* prediction algorithms and data libraries with customizable thresholds of each individual predictor. A 'global deleteriousness filter' to set a threshold on the number of selected predictors needed in order for a variant to pass the filter, allowing adjustment of stringency and balancing and/or investigating inconsistency among different predictions*.
		Predicted effect on function: variant is found at the location within the protein predicted to cause functional disruption (for example, enzyme active site, protein-binding region).		VARIANT FUNCTION + Functional Predictions GENE ANNOTATION (1) + Protein Domain Annotation and/or Filter	

Gene level	**Genetic**	**Gene burden**: the affected gene shows statistical excess of rare (or *de novo*) probably damaging variants segregating in cases compared to control cohorts or null models.	In all cases in which it is possible, apply statistical methods to compare the distribution of variants in patients with large matched control cohorts or well-calibrated null models.	GENE ANNOTATION (4) + Genetic Association Test + BURDEN + calpha + vt + skat	Gene-based association tests by PLINK/SEQ.
	Functional Prediction	**Protein interactions**: the gene product interacts with proteins previously implicated (genetically or biochemically) in the disease of interest. **Biochemical function**: the gene product performs a biochemical function shared with other known genes in the disease of interest, or consistent with the phenotype. **Expression**: the gene is expressed in tissues relevant to the disease of interest and/or is altered in expression in patients who have the disease.	Experimental data can be used to demonstrate that the normal function of the gene is consistent with the known biology of the disease process. In presumed monogenic-disease cases, evaluate genes previously implicated in similar phenotypes before exploring potential new genes.	GENE ANNOTATION (1) + Protein-protein Interactions + Genotype-phenotype Databases + DISEASE GENE FILE + Gene Ontology (GO) GENE ANNOTATION (2) + Pathway Databases GENE ANNOTATION (3) + GTEx (The Genotype-Tissue Expression Project) + HPA (The Human Protein Atlas)	Four protein-protein interaction databases; Gene Ontology and three pathway databases parsed from MSigDB; three genotype-phenotype (disease) databases (as well as a user-defined gene list of interest). All databases are stored via MySQL for quick query. Information can be combined across libraries—for instance, known disease gene annotation on candidates can to be supplemented with their interaction partners as reported in other databases*.
			Genetic evidence implicating a variant must be assessed within the context of the considerable background of rare genetic variants in humans. Genes differ markedly in their tolerance to variation and rare variants predicted to be damaging in disease-associated genes are often observed even in population controls.	GENE ANNOTATION (4) + GDI (Gene damage index) + RVIS (Residual Variation Intolerance Score)	Two methods to quantitatively assess gene tolerance to variation in general population, predicting whether a gene is likely to harbor disease-causing mutations.

Table S2. Information of external resources used in GeMSTONE

The table provides information of all external tools and databases used in GeMSTONE, categorized by their applications. Methods or datasets complied by a software/data library are in *italic* and listed under the corresponding master resource. Both latest versions that GeMSTONE uses by default and older versions available for users' selection are listed below at the time GeMSTONE published. Please refer to GeMSTONE manual page for the latest updates (http://gemstone.yulab.org/manual.html). Sizes of programs and datasets are listed based on the memory space taken on our web server.

External Resources	**GeMSTONE Application**	**Version(s)**		**Size**
		Latest in use	**Older available**	
VT	Variant Normalization	v0.5		4.3MB
VCFtools	Variant Quality Filter, Control Filter	0.1.14		13MB
SnpEff	Variant Consequence Annotation	4.3k		773MB
BCFtools	Co-segregation Analysis	1.4		4.0MB
GEMINI	Co-segregation Analysis	0.19.1		7.6MB
ExAC	Variant Annotation	0.3.1		5.8GB
ESP	Variant Annotation	v.0.0.30		195MB
1000 Genomes	Variant Annotation	Phase 3		15GB
TAGC	Variant Annotation	EGAD00001000781		6.0GB
Rosetta ddG	Variant Annotation	Rosetta 3		16GB
Pfam	Variant Annotation	31.0		4.5MB
dbNSFP	Variant/Gene Annotation	v3.4	v3.3	76GB
SIFT	Variant Annotation	ensembl 66		
PROVEAN	Variant Annotation	v1.1		
PolyPhen-2	Variant Annotation	v2.2.2		
LRT	Variant Annotation	2009-11		
Mutation Taster	Variant Annotation	2015		
Mutation Assessor	Variant Annotation	3		
FATHMM	Variant Annotation	v2.3		
FATHMMMKL	Variant Annotation	2015		

VEST3	Variant Annotation	v3.0		
CADD	Variant Annotation	v1.3		
DANN	Variant Annotation	2015		
MetaSVM	Variant Annotation	2015		
MetaLR	Variant Annotation	2015		
fitCons	Variant Annotation	v1.01		
GERP++	Variant Annotation	2010		
phyloP	Variant Annotation	2015		
phastCons	Variant Annotation	2015		
SiPhy	Variant Annotation	0.5		
IntAct	Gene Annotation	2016-03		
BioGRID	Gene Annotation	3.4.134		
ConsensusPathDB	Gene Annotation	31		
MSigDB	Gene Annotation	v6.0		4.8MB
Gene Ontology	Gene Annotation	2016-05		
KEGG	Gene Annotation	82.0		
BioCarta	Gene Annotation	2016-05		
Reactome	Gene Annotation	v59		
HGMD	Gene Annotation	2015.3		458MB
OMIM	Gene Annotation	2017-04		1.3MB
ClinVar	Gene Annotation	2017-04	2017-03, 2017-02	701KB
MGI	Gene Annotation	6.08		2.7MB
GDI	Gene Annotation	Itan_et_al_2015		4.5MB
RVIS	Gene Annotation	Petrovsk_et_al_2013		1.3MB
PLINK/SEQ	Gene Annotation	0.10		16MB
BURDEN	Gene Annotation	0.10		

vt	Gene Annotation	2010		
Calpha	Gene Annotation	2011		
SKAT	Gene Annotation	2011		
GTEx	Gene Annotation	V6p		481MB
HPA	Gene Annotation	16.1		4.3MB
HINT	Gene Annotation	Version 4		5.7MB

Supplementary Table S3. Detailed reasons for less powerful powerful functionalities of other tools comparing to GeMSTONE

The table details the reasons why other tools were considered less powerful in specific functionalities when compared to GeMSTONE.

Functionality		Detailed reasons for less powerful functionalities of other tools	GeMSTONE
Raw Data Input and Prioritization	Multi-Sample VCF File Input	Restricted to a single family with multiple samples or a single sample per VCF each run, or limited VCF file size (<100M).	Supporting multiple families AND multiple sporadic samples in one VCF file input.
	Customizable Genes-of-Interest	User-defined disease-gene family disables selecting from the disease database; only supporting Entrez gene IDs.	Providing a separate entry for user-defined gene list, independent of other three provided disease databases; allowing annotating user's notes relating to the genes (if any); Ensembl, Gene name, HGNC symbol and Entrez gene ID are all valid identifiers.
	SNV	Restricted to certain types of variants; not allow selecting specific mutation types.	Variant Consequence filter allows selecting from a comprehensive list of mutation types.
	INDEL		
Knowledge-based Annotation and Prioritization	Ethnicity-Specific Allele Frequency	Not allowing AF filtering on certain sub-population, or not to a per-sample resolution.	Allowing AF filtering for each sample in relation to his/her ethnicity.
	Variant Functional Prediction	Fewer *in silico* prediction algorithms provided.	19 variant function/conservation predictors with customizable thresholds of each individual predictor.
	Variant Conservation		
	Disease-Gene Annotation	Fewer disease databases provided.	Three high-profile human disease databases (HGMD, ClinVar, OMIM) and a database of mouse disease model data (MGI).
	Protein Interaction	Fewer protein-protein interaction (PPI) databases provided.	10 PPI databases (HINT, IntAct, BioGRID, Consensus PathDB; HINT integrates interactions from BioGRID, MINT, iRefWeb, DIP, IntAct, HPRD, MIPS and the PDB).
	Gene Association Tests	Only providing a basic statistical test comparing allele frequencies between cases and controls.	Three published gene association tests (vt, calpha, SKAT) with different algorithms, as well as a basic association test (BURDEN) comparing allele frequencies between cases and controls.
	Interactive Variant Visualization	Not providing filtered versus unfiltered statistics comparison, or/and not "interactive", or/and fewer features.	Interactively showing variant statistics for both raw data (unfiltered) and prioritized data (filtered).
Inheritance Models	Inheritance Models	Limitation on pedigree structure or size, e.g., requiring both parents to be present, or requiring the user to define genotypes for each sample (which is not feasible for large pedigrees).	Six inheritance models working for pedigree with multiple single samples or family members. "No Inheritance Model" allows variants screening with no co-segregation restrictions. "Recurrence Filter" further prioritizes co-segregating variants shared by multiple families.