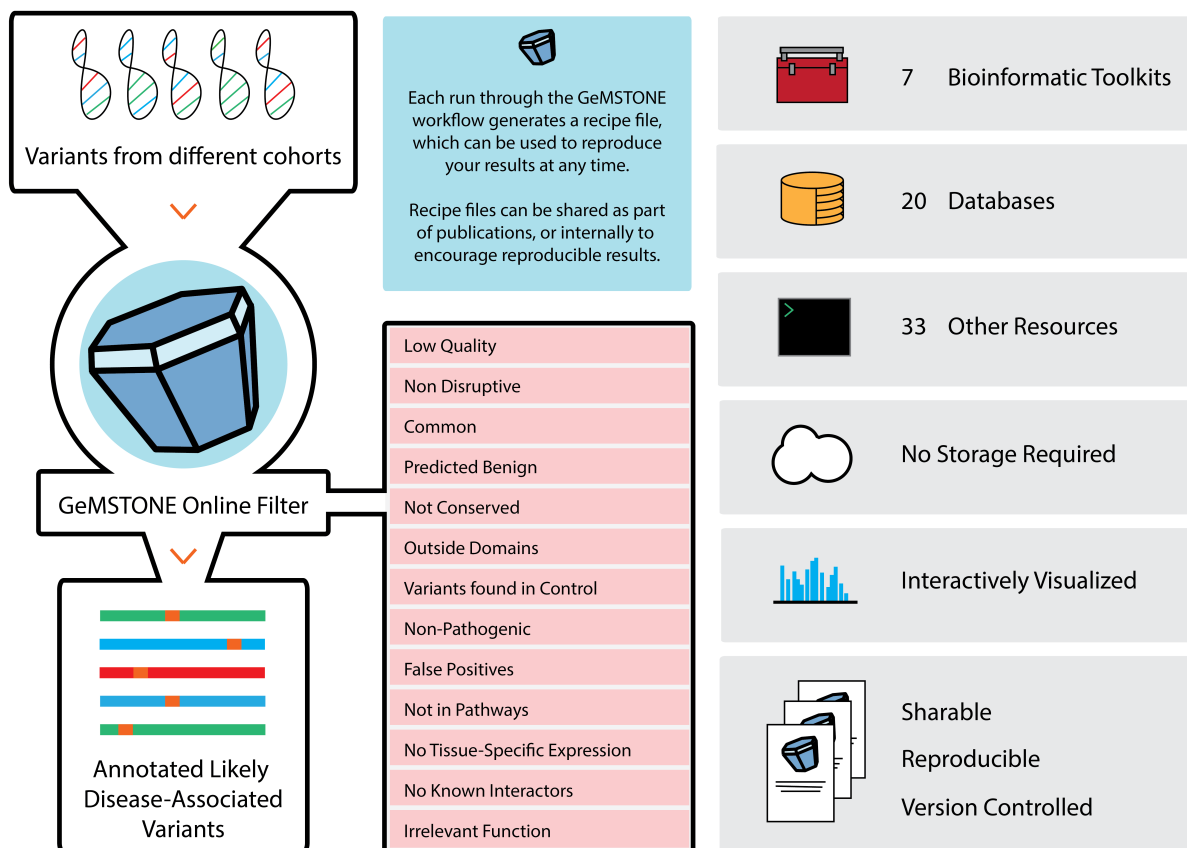


# GeMSTONE Manual



**Set Up a New Job P1-14**

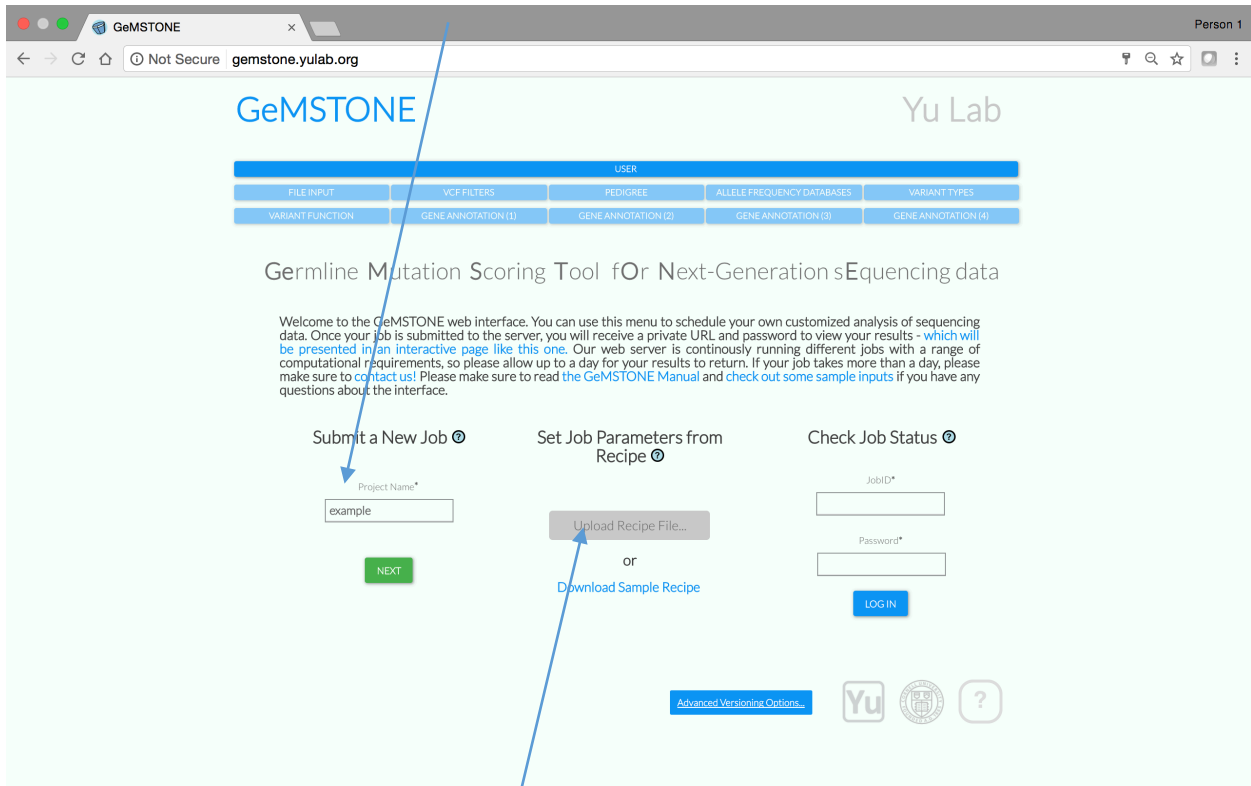
**Submit and Check Status P15**

**View and Download Results P16-17**

**Reproducibility and Version Controls P18-19**

**Benchmarks On Average Processing Time P20**

**Step1 @USER:** Start your job by giving it a name!



Or upload a recipe to use previous parameters to reproduce results!

**Step2 @FILE INPUT:** Upload your data – we always start with a *Variant Calling Format file (.vcf)*! Optionally, you could upload

- a *control file* (.vcf or .txt) to remove variants from the analysis
- a *pedigree file* (.ped) for co-segregation analysis.

Sample files are available for downloading under each entry!

\*\*\* Data submitted to GeMSTONE has a 2 months expiration date, after which it will be deleted from our servers. Data is not shared, or visible to any third parties. The Yu Lab does not keep any files submitted past the expiration date.

The screenshot shows the GeMSTONE web interface. At the top, there is a navigation menu with options: FILE INPUT, VCF FILTERS, PEDIGREE, ALLELE FREQUENCY DATABASES, and VARIANT TYPES. Below this, the 'File Upload' section is active, showing two sample files: 'sample\_1k\_genomes.vcf' and 'sample\_pedigree.ped'. The VCF file section includes a 'Download Sample VCF' button and a note that 'vcf.gz is required (max-size= 500M)'. The PED file section includes a 'Download Sample PED' button and a note about 'specifics on .ped format customization...'. To the right, there is a 'Human Genome Build' section with radio buttons for GRCh38 and GRCh37 (selected). Below this, there is a warning about the 2-month expiration date. A blue arrow points from the PED file section to a detailed explanation of PED file format requirements, including mandatory columns (Family ID, Individual ID, Paternal ID, Maternal ID, Sex, Phenotype) and an optional seventh column for ethnicity. A table lists available ethnicity identifiers from various databases (ExAC, 1000 Genomes, ESP6500, TAC) and their corresponding codes. The table is as follows:

ExAC	1000 Genomes	ESP6500	TAC
exac_ALL:Overall	kg_ALL:Overall	esp_ALL:Overall	tagc_A:Ashkenazi
exac_AFR:African/African American	kg_AFR:African	esp_AA:African American	
exac_AMR:Latino	kg_AMR:Ad Mixed American	esp_EA:European American	
exac_EAS:East Asian	kg_EAS:East Asian		
exac_FIN:Finnish	kg_EUR:European		
exac_NFE:Non-Finnish European	kg_SAS:South Asian		
exac_SAS:South Asian			
exac_OTH:Other			

Below the table, it states: 'If none of the above identifiers found in the seventh column, exac\_ALL will be used as default frequency database and ethnicity group for the allele frequency filtering. Less...'

Information from the .ped file is very important for co-segregation analysis:

- *Family ID* groups individuals with the same family ID into a family, which is the unit for co-segregation analysis.
- *Individual ID* uniquely identifies a sample in the VCF file (by exact matching, case sensitive); individual IDs indicated in the .ped file that do not match any sample in the VCF will be ignored.

- *Paternal and maternal IDs* identifies the familial relationship within a family. While parental sequence variants would be informative for co-segregation analysis (especially for recessive inheritance model), trios as well as any specific pedigree characteristics are required.
- *Sex information* will only be used for sex-linked inheritance model, if selected in co-segregation analysis.
- *Phenotype* identifies affection status, which is important in looking for variants that are co-inherited with the affection status within affected families. Under different inheritance models (**@PEDIGREE**), specific genotype criteria will be applied on the affected (with phenotype=2) and the unaffected (with phenotype=1); individuals with unknown phenotype will be ignored from the co-segregation analysis.
- *Ethnicity* (optional but unique) identifies the ethnicity of each sample using GeMSTONE-defined identifier (more details see *sepcifcs on .ped format customization @FILE INPUPT*), which is designed for rare variants filtering (**@VCF FILTERS**) in a sample-specific manner.

**Step3 @VCF FILTERS:** Remove low-quality variants (*using VCFtools*) and common variants.

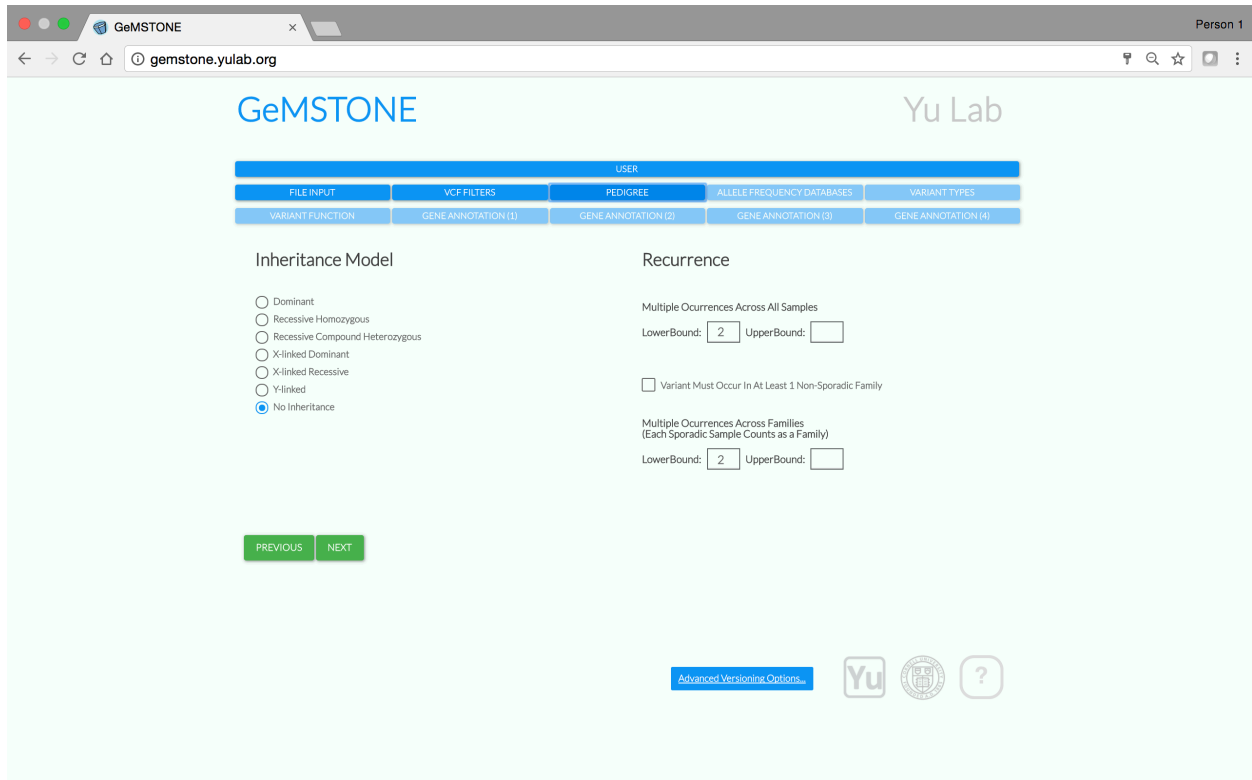
The screenshot shows the GeMSTONE web interface for configuring VCF filters. The page is titled 'GeMSTONE' and 'Yu Lab'. The main content area is divided into two columns: 'Site-basis' and 'Genotype-basis'. Under 'Site-basis', there are three configuration options: 'Phred-Scaled Quality Score Lowerbound' (input field empty, note: \*QUAL score in VCF file), 'Allele Frequency Upperbound' (input field '1.0', note: \*defaults to AF in ExAC unless population specified in PED file), and 'Ignore Variants Without PASS Flag' (checkbox). Under 'Genotype-basis', there are two configuration options: 'Genotype Quality Lowerbound' (input field '20', note: \*requires "GQ" FORMAT tag specified for all sites) and 'Individual Read Depth Lowerbound' (input field '10', note: \*requires "DP" FORMAT tag specified for all sites). At the bottom, there are 'PREVIOUS' and 'NEXT' buttons, and a link for 'Advanced Versioning Options...'. The browser address bar shows 'gemstone.yulab.org'.

- Phred-scaled quality score lowerbound: the minimum *QUAL* score for a variant to be included; *default*: None. (*vcftools* command: `--minQ <float>`)
- Ignore Variants Without PASS Flag: exclude variants of which the *FILTER* status is not PASS; *default*: OFF. (*vcftools* command: `--remove-filtered-all`)
- Genotype Quality Lowerbound: the minimum *GQ* score for a genotype to be considered; *default*: 20. (*vcftools* command: `--minGQ <float>`)
- Individual Read Depth Lowerbound: the minimum *DP* score for a genotype to be considered; *default*: 10. (*vcftools* command: `--minDP <float>`)

\*What are [QUAL score](#), [FILTER status](#), [GQ](#) and [DP score](#)?

- Allele Frequency Upperbound: the maximum *MAF* of the variant reported in general population, using the database and its sub-population matching the ethnicity of each sample as specified in the seventh column in the pedigree file; *default*: 1.0%, overall MAF reported in the ExAC database i.e. `exac_ALL`).

**Step4 @PEDIGREE:** Search for co-segregating and/or recurring variants.



- *Inheritance Model* analyzes a single family each time, looking for co-segregating events (see table below) or screening all variants shared by the affected (*No Inheritance*).

Inheritance model	<i>A variant will be kept if it is ...</i>	<i>A variant will be removed if it is...</i>
Dominant	HET in all the affected.	present in any the unaffected.
Recessive Homozygous	HOM_ALT in all the affected AND HET in the parents (if identified) of the unaffected.	HOM_ALT in any the unaffected other than the unaffected parents.
Recessive Compound Heterozygous ( <i>using comp_hets in GEMINI</i> )	HET at both sites in all the affected.	HOM_ALT at either site in any the unaffected OR HOM_REF in any parent (if identified) of the affected.
X-linked Dominant	HET in affected females AND present in affected males AND HET in mothers (if identified) of affected males.	present in any the unaffected .
X-linked Recessive	HOM_ALT in all affected females AND present in all affected males.	HOM_ALT in any unaffected females OR present in any affected males.
Y-linked	present exclusively on Y chromosome of affected males.	
No Inheritance	present exclusively in all the affected.	

\*HOM: homozygous; HET: heterozygous; ALT: alternate allele as opposed to reference allele

- *Recurrence* filter constrains to which degree a co-segregation event can happen across multiple families and/or the prevalence of the variant in sporadic samples.

Let’s take an example output file to understand how these options work! Suppose we have a pedigree file as below, in which there are

	1	2	3	4	5	6	7
1	F_GBR	HG00097	HG00101	HG00100		1	2 exac_NFE
2	F_GBR	HG00100	0	0		2	2 exac_NFE
3	F_GBR	HG00101	0	0		1	0 exac_NFE
4	NA19648	NA19648	0	0		2	2
5	NA19649	NA19649	0	0		1	2

- 5 samples: HG00097,HG00100,HG00101,NA19648,NA19649; all but HG00100 are known to be affected; the ethnicity of the first 3 samples was indicated to be close to non-finish European.
- One family: F\_GBR, consisting of 3 samples (kid HG00097, father HG00101, mother HG00100) and two sporadic samples: NA19648, NA19649.

Co-segregation analysis will be performed on each of the three ‘family units’ (F\_GBR, NA19648, NA19649) where each sporadic sample is considered as a family itself, then a union of the identified co-segregating variants will be summarized into a variant table (one of GeMSTONE’s outputs) as below, upon which the *Recurrence* filter will be applied.

	1	2	3	4	5	6	7	8	9	10	11	12
1	CHROM	POS	ID	REF	ALT	FILTER	INDIVIDUAL_ID	CONSEQL	PUTATIVE	GENE_NAME	ENTREZ	ENSEM
51	1	52499097	rs116535272	G	C	PASS	[NA19649]	missense_va	MODERATE	KTI12	112970	ENSGI
52	1	54605318	rs77544356	TG	T	PASS	[HG00097,HG00100]	frameshift_v	HIGH	CDCP2	200008	ENSGI
53	1	54605318	rs77544356	T	TGC	PASS	[HG00097,HG00100]	frameshift_v	HIGH	CDCP2	200008	ENSGI
54	1	54605318	rs77544356	T	TTG	PASS	[HG00097,HG00100]	frameshift_v	HIGH	CDCP2	200008	ENSGI
55	1	55223744	rs35201073	G	C	PASS	[NA19648]	missense_va	MODERATE	PARS2	25973	ENSGI
56	1	60520988	rs144671684	G	A	PASS	[NA19648],[NA19649]	missense_va	MODERATE	C1orf87	127795	ENSGI
57	1	62675673	rs200789118	G	T	PASS	[NA19648]	missense_va	MODERATE	L1TD1	54596	ENSGI
58	1	62676284	.	CAGA	C	PASS	[NA19648],[NA19649]	inframe_delc	MODERATE	L1TD1	54596	ENSGI
59	1	65129491	.	A	C	PASS	[NA19648]	missense_va	MODERATE	CACHD1	57685	ENSGI
60	1	67154849	.	G	C	PASS	[NA19648]	missense_va	MODERATE	SGIP1	84251	ENSGI
61	1	67390481	.	G	C	PASS	[NA19648],[NA19649]	missense_va	MODERATE	WDR78	79819	ENSGI
62	1	67447551	.	A	C	PASS	[NA19648],[NA19649]	missense_va	MODERATE	MIER1	57708	ENSGI
63	1	74670359	rs148933608	C	T	PASS	[NA19648],[NA19649]	missense_va	MODERATE	FPGT	8790	ENSGI
64	1	76345823	rs5745459	A	G	PASS	[NA19648]	missense_va	MODERATE	MSH4	4438	ENSGI
65	1	82456482	rs144339910	T	A	PASS	[NA19648]	missense_va	MODERATE	LPHN2	23266	ENSGI
66	1	86591837	rs11161747	G	T	PASS	[HG00097,HG00100],[NA19648],[NA19649]	missense_va	MODERATE	COL24A1	255631	ENSGI
67	1	87045799	rs201405115	C	T	PASS	[NA19648]	missense_va	MODERATE	CLCA4	22802	ENSGI
68	1	92200437	rs41286789	T	C	PASS	[NA19648]	missense_va	MODERATE	TGFB3	7049	ENSGI
69	1	93091349	rs200027454	A	C	PASS	[NA19648],[NA19649]	splice_donor	HIGH	EVIS	7813	ENSGI
70	1	94486816	rs200443984	C	A	PASS	[NA19648]	missense_va	MODERATE	ABCA4	24	ENSGI
71	1	109004611	.	C	A	PASS	[NA19649]	missense_va	MODERATE	NBPF6	653149	ENSGI
72	1	111957411	rs150120731	C	T	PASS	[NA19649]	missense_va	MODERATE	OVGP1	5016	ENSGI
73	1	111957517	rs3767609	T	C	PASS	[HG00097,HG00100],[NA19649]	missense_va	MODERATE	OVGP1	5016	ENSGI
74	1	111957570	rs45455292	G	C	PASS	[HG00097,HG00100],[NA19649]	missense_va	MODERATE	OVGP1	5016	ENSGI
75	1	111957592	rs56294468	A	G	PASS	[HG00097,HG00100],[NA19649]	missense_va	MODERATE	OVGP1	5016	ENSGI
76	1	114437834	rs201156296	A	C	PASS	[NA19648]	missense_va	MODERATE	AP4B1	10717	ENSGI
77	1	115124203	rs62621917	T	C	PASS	[NA19648]	missense_va	MODERATE	BCAS2	10286	ENSGI

Variants are listed by row with meta annotations (not shown in the screenshot). Focusing on the seventh column *INDIVIDUAL\_ID*, it reports the carriers of each variant with brackets indicating the same family unit, for examples, the #51 variant was observed in sample NA19649, the #52 variant was co-segregated in the two samples [HG00097,HG00101] in family F\_GBR (recall that the other sample in this family HG00100 was ignored as suggested in the pedigree file above), the #66 variant was observed in all 4 affected samples across 3 family units [HG00097,HG00101],[ NA19648],[ NA19649]. The *Reference* filter can then place constraints based on this column, options including

- *Multiple Occurrences Across All Samples*: consider each individual independently, and constrain the frequency a variant to be shared by all samples.

- *Variant Must Occur In At Least 1 Non-Sporadic Family*: if checked, requires the variant to be observed as a co-segregating event in at least one family with multiple family members, i.e. sporadic samples or families with only one individual sequencing available will not be considered.
- *Multiple Occurrences Across Families (Each Sporadic Sample Counts as a Family)*: similar with the first option, but consider each family (each sporadic sample counts as a family) as a unit.

<i>Multiple Occurrences Across All Samples</i>	<i>Variant Must Occur In At Least 1 Non-Sporadic Family</i>	<i>Multiple Occurrences Across Families (Each Sporadic Sample Counts as a Family)</i>	Variants will be selected in the example sheet (from #51-#77)
[2,-]	-	-	#52-54,#57,#59,#61-63,#66,#69,#73-75
[2,3]	-	-	Excluding #66 from above
[2,3]	checked		#73-75
-	-	[2,-]	Excluding #52-54 from the first result
	checked	[2,-]	#66,#73-75



**Step5 @ALLELE FREQUENCY DATABASES:** Annotate MAF of each variant from selected population database(s).

The screenshot shows the GemSTONE web interface at gemstone.yulab.org. The 'Allele Frequency Databases' section is active, displaying a grid of database options. The 'ExAC' option is selected with a checked checkbox. Below the grid are 'PREVIOUS' and 'NEXT' buttons.

USER			
FILE INPUT	VCF FILTERS	PEDIGREE	ALLELE FREQUENCY DATABASES
VARIANT FUNCTION	GENE ANNOTATION (1)	GENE ANNOTATION (2)	GENE ANNOTATION (3)
VARIANT TYPES			
GENE ANNOTATION (4)			

### Allele Frequency Databases

These options are only for annotation, not for filtering variants

<input checked="" type="checkbox"/> ExAC	<input type="checkbox"/> 1000 Genomes	<input type="checkbox"/> ESP6500	<input type="checkbox"/> TAGC
exac_ALL: Overall	kg_ALL: Overall	esp_ALL: Overall	tagc_AJ: Ashkenazi
exac_AFR: African/African American	kg_AFR: African	esp_AA: African American	
exac_AMR: Latino	kg_AMR: Ad Mixed American	esp_EA: European American	
exac_EAS: East Asian	kg_EAS: East Asian		
exac_FIN: Finnish	kg_EUR: European		
exac_NFE: Non-Finnish European	kg_SAS: South Asian		
exac_SAS: South Asian			
exac_OTH: Other			

**Step6 @VARIANT TYPES:** Select variant type(s) and transcript biotype(s) of interest based on SnpEff annotations. (`java -Xmx4g -jar snpEff.jar eff -v [GRCh37.75,GRCh38.86] -canon`)

The screenshot shows the GeMSTONE web interface. At the top, there is a navigation bar with 'GeMSTONE' and 'Yu Lab' logos. Below this is a menu with categories: USER, FILE INPUT, VCF FILTERS, PEDIGREE, ALLELE FREQUENCY DATABASES, and VARIANT TYPES. Under VARIANT TYPES, there are sub-categories: VARIANT FUNCTION, GENE ANNOTATION (1), GENE ANNOTATION (2), GENE ANNOTATION (3), and GENE ANNOTATION (4).

The main content area is titled 'Variant Consequence' and contains several sections of checkboxes:

- Coding Transcript Variant:**
  - Frameshift
  - Inframe Indel
  - Nonsynonymous (Missense, Start Loss, Stop Gained, Stop Lost)
  - Synonymous
- Splicing Variant:**
  - Exon Loss
  - Intron Gain
  - Splice Site
  - Splice Region
- Intergenic Variant:**
  - Up/Downstream
  - Other Intergenic Region
- Non-Coding Variant:**
  - Intron
  - UTR
  - Other Non-Coding Region
- Others:**
  - Regulatory Region Variant

Below this is the 'Transcript Biotype' section with more checkboxes:

- Protein Coding:**
  - Protein Coding (contains an Open Reading Frame (ORF))
  - Immunoglobulin (Ig) Variable Chain and T-cell Receptor (TCR) gene
  - Nonsense-mediated Decay
  - Non-translating CDS
  - Non-stop decay
  - Polymorphic Pseudogene
- Pseudogene:**
  - Pseudogene
  - Inactivated Immunoglobulin Gene
  - Disrupted Domain
- Short Noncoding:**
  - ncRNA
  - ncRNA Pseudogene
- Long Noncoding:**
  - Non-coding
  - Antisense
  - Sense Intronic
  - Sense Overlapping
  - Retained Intron
  - lincRNA
  - 3 overlapping ncRNA
  - Others (no ORFs / ambiguous # of ORFs)

At the bottom, there is a 'Custom Transcript File' section with a text input field and a 'PREVIOUS' button. A blue arrow points from the 'Advanced Versioning Options...' link to the 'Custom Transcript File' input field.

SnpEff by default uses canonical transcripts, you can upload your own here! Ensembl Transcript ID per line. (`java -Xmx4g -jar snpEff.jar eff -v [GRCh37.75,GRCh38.86] -onlyTr your_transcripts.txt`)

**Step7 @VARIANT FUNCTION:** Annotate and/or filter by *in silico* precisions on variant function.

Another very important informatic evidence for variant implication comes from *in silico* analysis, predicting a variant is likely to be deleterious in terms of biological function or in an evolutionary sense. You can Choose up to 23 different *in silico* predictors

- in terms of biological function (*Functional Predictions, Protein Stability Predictions*) or in an evolutionary sense (*Conservation Scores*);
- with customizable thresholds – check on any of the predictors, a slider for setting thresholds will show up!

PLUS! Use the '*global deleteriousness filter*' to set a threshold on the number of selected predictors needed in order for a variant to pass the filter. By default (or entering 0), GeMSTONE will annotate the number for you for future downstream decisions.

GeMSTONE Yu Lab

USER

FILE INPUT VCF FILTERS PEDIGREE ALLELE FREQUENCY DATABASES VARIANT TYPES

VARIANT FUNCTION GENE ANNOTATION (1) GENE ANNOTATION (2) GENE ANNOTATION (3) GENE ANNOTATION (4)

Functional Predictions Deleteriousness Thresholds

SIFT  
 PROVEAN  
 PolyPhen-2\_HDIV  
 PolyPhen-2\_HVAR  
 LRT  
 MutationTaster  
 MutationAssessor  
 FATHMM  
 FATHMMMKL  
 VEST3  
 CADD\_Pfired  
 DANN  
 MetaSVM  
 MetaLR  
 fitCons

Conservation Scores

GERP++  
 phyloP Vertebrate  
 phyloP Mammalian  
 phastCons Vertebrate  
 phastCons Mammalian  
 SiPhy

Protein Stability Prediction

Rosetta ddG

Deleteriousness Filter

Keep only variants that  + scores predict to be deleterious out of the 2 scores you've selected (only applicable to nsSNVs).

SIFT 0 to 0.05 PolyPhen-2\_HDIV 0.96 to 1

PREVIOUS NEXT

Advanced Versioning Options... Yu

**Step8 @GENE ANNOTATION (1)(2)(3):** Annotate and/or filter genes of biological function, disease implication, protein domain, protein-protein interaction, protein tissue expression, and initiate their 'crosstalk'!

*Pathway Enrichment Analysis* calculates enrichment of prioritized genes using a fisher exact test, one-sided  $p$ -value and FDR corrected  $q$ -value will be reported for each functional gene set in selected database(s). Enrichment calculation can be done with respect to the background to be either all genes before prioritization (*Genes from VCF Input*), or all human protein-coding genes (*Homo Sapiens Protein Coding Genes (Ensembl Biomart)*), or a gene list of your interest (*Custom Gene List*).

GeMSTONE Yu Lab

USER

FILE INPUT VCF FILTERS PEDIGREE ALLELE FREQUENCY DATABASES VARIANT TYPES

VARIANT FUNCTION GENE ANNOTATION (1) GENE ANNOTATION (2) GENE ANNOTATION (3) GENE ANNOTATION (4)

### Pathway Databases

Annotate genes using the following protein pathway databases.

Select pathway(s) from KEGG database  
1 selected

Select pathway(s) from BioCarta database  
Select options

Select pathway(s) from Reactome database  
Select options

Add Pathway Annotation For Interaction Partners  
 Filter Out Genes With No Pathway Annotation

### Pathway Enrichment Analysis

Calculate and report enriched pathways with q-value

From the following databases:

- KEGG
- BioCarta
- Reactome
- GO Biological Processes
- GO Cellular Components
- GO Molecular Function

Background Gene List

- Genes from VCF Input
- Homo Sapiens Protein Coding Genes (Ensembl Biomart)
- Custom Gene List

\* Requires Entrez Gene ID or HGNC Symbol. One gene per line.

[Download Sample Gene File](#)

PREVIOUS NEXT

Annotation of gene expression provides the level (or presence) of protein expressed in tissue(s) selected, and in addition, gene expression enrichment/preference in particular tissue(s), differential expression with respect to sex ( $\log_2\text{FoldChange}(\text{females/males})_{FDR}$ ), ethnicity ( $\log_2\text{FoldChange}(\text{AA/EA})_{FDR}$ ), age ( $\text{coefficient}_{FDR}$ ).

The screenshot shows the GeMSTONE web interface. At the top, there is a navigation bar with the GeMSTONE logo and 'Yu Lab' text. Below this is a menu bar with options like 'FILE INPUT', 'VCF FILTERS', 'PEDIGREE', 'ALLELE FREQUENCY DATABASES', and 'VARIANT TYPES'. The main content area is divided into two sections: 'GTEx (The Genotype-Tissue Expression Project)' and 'HPA (The Human Protein Atlas)'. Each section contains a grid of tissue names, with some cells highlighted in orange to indicate selection. The 'GTEx' grid has 'FIBROBLASTS', 'LIVER', and 'TESTIS' highlighted. The 'HPA' grid has 'APPENDIX' and 'LIVER' highlighted. At the bottom of the interface, there are 'PREVIOUS' and 'NEXT' buttons, and a footer with logos for 'Yu' and 'Yale University'.

**GTEx (The Genotype-Tissue Expression Project)**

ADRENAL GLAND	ANTERIOR CINGULATE CORTEX	AORTA	ATRIAL APPENDAGE	BLOOD
BREAST	CAUDATE (BASAL GANGLIA)	CEREBELLUM	COLON	CORONARY
CORTEX	FIBROBLASTS	HIPPOCAMPUS	HYPOTHALAMUS	LCL
LEFT VENTRICLE	LIVER	LUNG	MUCOSA	MUSCULARIS
NUCLEUS ACCUMBENS (BASAL GANGLIA)	OVARY	PANCREAS	PITUITARY	PROSTATE
PUTAMEN (BASAL GANGLIA)	SKELETAL MUSCLE	SKIN SUPRAPUBIC	SKIN LOWER LEG	STOMACH
SUBCUTANEOUS	TESTIS	THYROID GLAND	TIBIAL	TIBIAL NERVE
UTERUS	VAGINA	VISCERAL (OMENTUM)		

**HPA (The Human Protein Atlas)**

ADIPOSE TISSUE	ADRENAL GLAND	APPENDIX	BONE MARROW	CEREBRAL CORTEX
COLON	DUODENUM	ENDOMETRIUM	ESOPHAGUS	FALLOPIAN TUBE
GALLBLADDER	HEART MUSCLE	KIDNEY	LIVER	LUNG
LYMPH NODE	OVARY	PANCREAS	PLACENTA	PROSTATE
RECTUM	SALIVARY GLAND	SKELETAL MUSCLE	SKIN	SMALL INTESTINE
SMOOTH MUSCLE	SPLEEN	STOMACH	TESTIS	THYROID GLAND
TONSIL	URINARY BLADDER			

Navigation: PREVIOUS NEXT

Footer: [Advanced Versioning Options](#) | Yu | Yale University | ?

**Step8 @GENE ANNOTATION (4):** Investigate gene implication with disease.

- *GDI* (Gene Damage Index) and *RVIS* (Residual Variation Intolerance Score) quantitatively assess gene tolerance to variation in general population, predicting whether a gene is likely to harbor disease-causing mutations.
- *Genetic Association Test* tests phenotype-genotype association, seeking statistical evidence for prioritizing causal variants and/or genes. A single variant association test and four gene-based association tests are provided using PLINK/SEQ. Note that a control VCF file is required for association test.

GeMSTONE Yu Lab

USER

FILE INPUT VCF FILTERS PEDIGREE ALLELE FREQUENCY DATABASES VARIANT TYPES

VARIANT FUNCTION GENE ANNOTATION (1) GENE ANNOTATION (2) GENE ANNOTATION (3) GENE ANNOTATION (4)

GDI (Gene Damage Index) Disease Type

All diseases  
 Mendelian (general model)  
 Mendelian (autosomal dominant)  
 Mendelian (autosomal recessive)  
 Cancer (general model)  
 Cancer (autosomal dominant)  
 Cancer (autosomal recessive)  
 Primary immunodeficiency (general model)  
 Primary immunodeficiency (autosomal dominant)  
 Primary immunodeficiency (autosomal recessive)

Residual Variation Intolerance Score

Annotate RVIS Gene Score

Gene Burden Test

You can choose to upload an additional control to perform Gene Burden Tests.

BURDEN CONTROL VCF

Download Sample Burden Test Control VCF

BURDEN  
 calpha  
 vt  
 skat

Optional but **strongly recommended:**  
 Input your email to be notified when your job finishes running. A single job can take more than a day depending on the size of the dataset and the number of analyses scheduled in the workflow

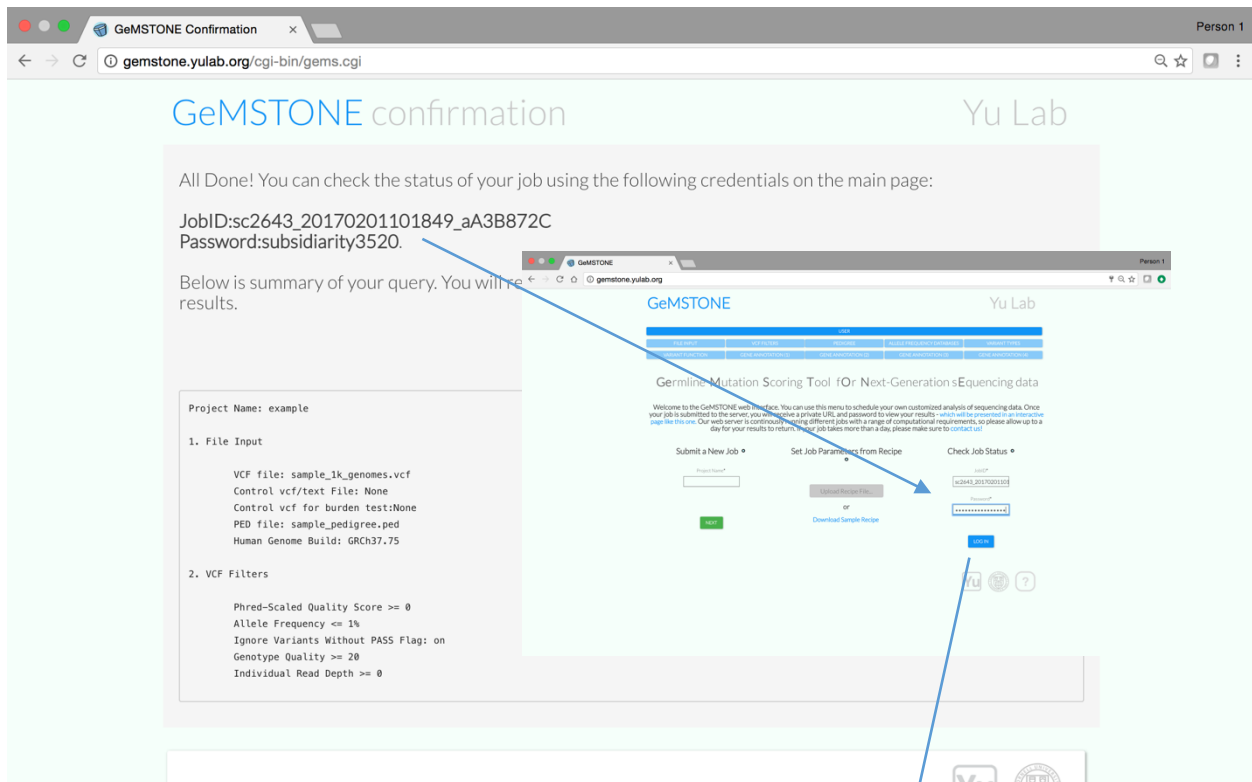
PREVIOUS SUBMIT

**SUBMIT!!!**

**@CONFIRMATION PAGE**

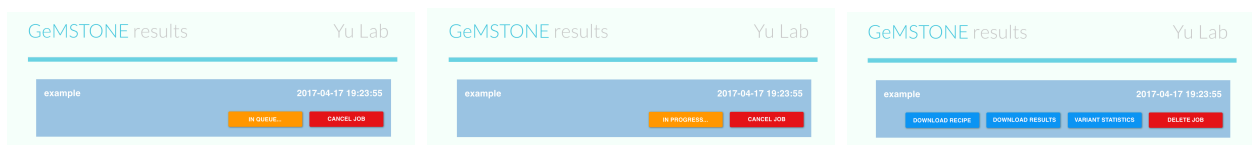
You will then be navigated to a *confirmation* page with

- a *JobID* with the *Password*, which you will need to login to your job interface @USER to check the process of your job and to download files, interactively view variant statistics after the job is finished. If you would like to provide an email address before submitting, we will send you a copy of the JobID and Password, and a notification once your job is finished!
- a summary listing all your selections and parameter settings for the submitted job and versions of all programs (with commands ran) and databases used. This file will be downloadable from the result page after the job is finished.



**@RESULTS PAGE**

Use your jobID and password to log in to the result page and view the job status – it either IN QUEUE, IN PROGRESS, or DONE with all your result files downloadable!





**@RESULTS PAGE**

- *DOWNLOAD RECIPE* provides a program readable recipe file that records all selections and parameter settings for this job and versions of all programs and databases used. You can upload the recipe @USER to re-use this workflow – all previous settings will be automatically loaded including all selections in all dropdown lists! Then you can readily apply this workflow to recapitulate a previous job, or to analyze new datasets with the identical or modified workflow.
- *DOWNLOAD RESULTS* will navigate you to a list of downloadable output files of the analysis, including main result tables as well as intermediate processed VCF files – see the schematic diagram in next page for details.
- *VARIANT STATISTICS* interactively shows general statistics of all variants before and after the analysis.

The screenshot displays the GemSTONE web interface for a job titled 'example' (ID: 00) dated 2017-04-17 19:23:55. The interface is divided into three main sections:

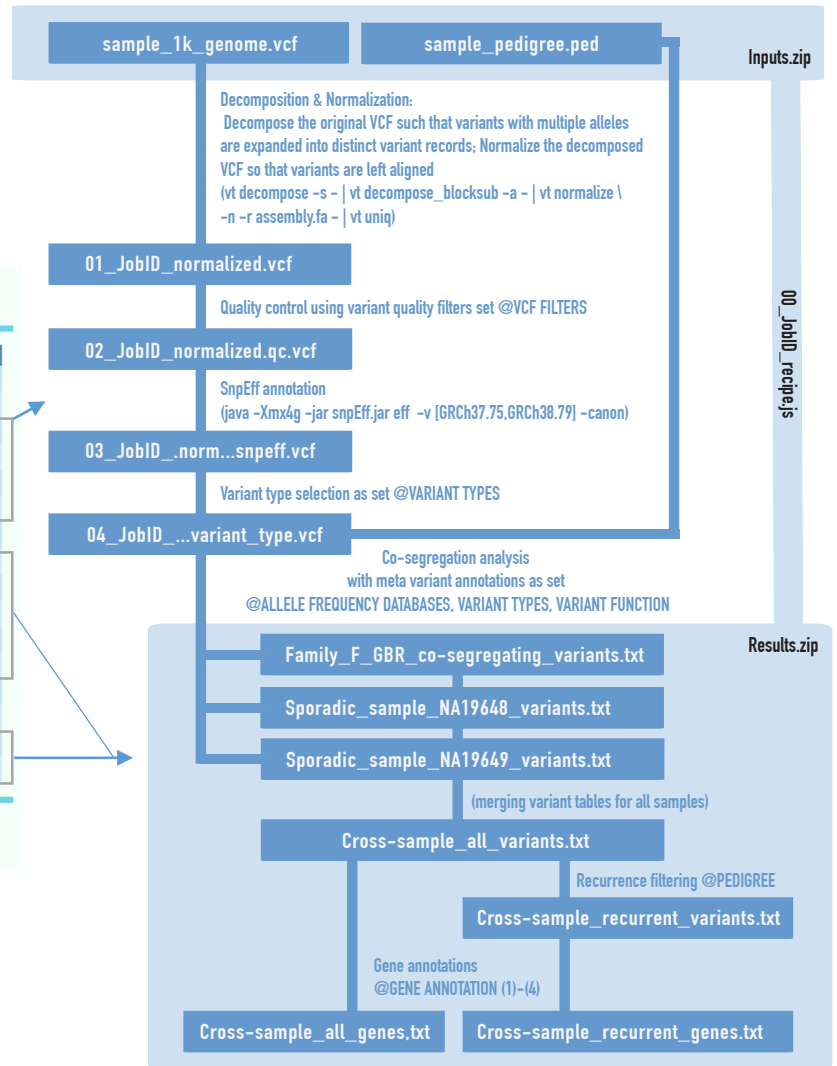
- Navigation Bar:** Contains buttons for 'DOWNLOAD RECIPE', 'DOWNLOAD RESULTS', 'VARIANT STATISTICS', and 'DELETE JOB'.
- GemSTONE (Left Panel):** Shows the job submission form with fields for 'Project Name' and 'Password', and a 'NEXT' button. A green notification box indicates that the job has been applied to the GemSTONE form.
- GemSTONE downloads (Middle Panel):** A table listing the output files and their sizes:
 

File	Size
00_sc2643_20170126103204_7ECDf368_recipe.js	1.0 kB
00_sc2643_20170126103204_7ECDf368_summary	1.0 kB
01_sc2643_20170126103204_7ECDf368.normalized.vcf	8.0 MB
02_sc2643_20170126103204_7ECDf368.normalized.qc.vcf	7.0 MB
03_sc2643_20170126103204_7ECDf368.normalized.qc.snpeff.vcf	34.0 MB
04_sc2643_20170126103204_7ECDf368.normalized.qc.snpeff.variant_type.vcf	6.0 MB
All_checksq4.zip	12.0 MB
Cross-sample_all_genes.bt	43.0 kB
Cross-sample_all_variants.bt	358.0 kB
Cross-sample_recurrent_genes.bt	9.0 kB
Cross-sample_recurrent_variants.bt	49.0 kB
Family_F_GBR_co-segregating_variants.bt	22.0 kB
Inputs.zip	1.0 MB
Results.zip	260.0 kB
Sporadic_sample_NA19648_variants.bt	194.0 kB
Sporadic_sample_NA19649_variants.bt	196.0 kB
- GemSTONE visualizer (Right Panel):** Displays various plots including 'Variant Density', 'Variant Quality', 'Mean Depth', 'Allele Frequency Spectrum', 'TSTV Ratio = 1.04', 'Insertion/Deletion Lengths', and 'Variant Type'.

GeMSTONE downloads Yu Lab

File	Size
00_sc2643_20170126103204_7ECDf368_recipe.js	1.0 kB
00_sc2643_20170126103204_7ECDf368_summary	1.0 kB
01_sc2643_20170126103204_7ECDf368.normalized.vcf	8.0 MB
02_sc2643_20170126103204_7ECDf368.normalized.qc.vcf	7.0 MB
03_sc2643_20170126103204_7ECDf368.normalized.qc.snpeff.vcf	34.0 MB
04_sc2643_20170126103204_7ECDf368.normalized.qc.snpeff.variant_type.vcf	6.0 MB
All_checksq4.zip	12.0 MB
Cross-sample_all_genes.txt	63.0 kB
Cross-sample_all_variants.txt	358.0 kB
Cross-sample_recurrent_genes.txt	9.0 kB
Cross-sample_recurrent_variants.txt	49.0 kB
Family_F_GBR_co-segregating_variants.txt	22.0 kB
Inputs.zip	1.0 MB
Results.zip	260.0 kB
Sporadic_sample_NA19648_variants.txt	194.0 kB
Sporadic_sample_NA19649_variants.txt	196.0 kB

Yu



@USER

Upload your recipe to use a workflow from previous analysis!

The screenshot shows the GemSTONE web interface. At the top, there is a navigation bar with the GemSTONE logo and 'Yu Lab'. Below this is a menu titled 'USER' with several options: FILE INPUT, VCF FILTERS, PEDIGREE, ALLELE FREQUENCY DATABASES, and VARIANT TYPES. Below the menu is a table with columns: VARIANT FUNCTION, GENE ANNOTATION (1), GENE ANNOTATION (2), GENE ANNOTATION (3), and GENE ANNOTATION (4). The main content area is titled 'Germline Mutation Scoring Tool fOr Next-Generation sEquencing data'. Below this is a welcome message and three main navigation buttons: 'Submit a New Job', 'Set Job Parameters from Recipe', and 'Check Job Status'. A green callout box highlights the 'Set Job Parameters from Recipe' button, stating that a recipe ID '00\_sc2643\_20170417185' has been applied to the form and that users should enter their project name and press NEXT to see parameters and upload data. Below the buttons are 'Advanced Versioning Options...' and logos for 'Yu' and a university.

After uploading the recipe file, all parameter settings and selections will be automatically loaded – check tabs like below! You can readily upload your inputs and submit or modify any of the options as usual!

A grid of seven screenshots showing different tabs in the GemSTONE web interface. The tabs shown are: 'File Input', 'VCF Filters', 'Pedigree', 'Allele Frequency Databases', 'Variant Types', 'Gene Annotation (1)', and 'Gene Annotation (2)'. Each screenshot shows a detailed configuration page for that specific tab, with various input fields, checkboxes, and buttons. A blue arrow points from the 'Set Job Parameters from Recipe' button in the previous screenshot to the first screenshot in this grid, indicating that the settings are automatically loaded.

### @Advanced Visioning Options

We keep in our system static versions of all the external resources where all the tools and datasets that we use for GeMSTONE are loaded onto our server so that we are able to ensure backwards-compatibility as we add updated versions of software or new tools.

The screenshot shows the GeMSTONE web interface with a 'Versioning Options' dialog box open. The dialog contains a grid of dropdown menus for various software packages and their versions. The ClinVar version dropdown is currently open, showing options for 2017-04 (selected), 2017-03, and 2017-02. Other visible options include dbNSFP (v3.4), SnpEff (4.3k), VT (v0.5), VCFtools (0.1.14), BCFtools (1.4), GEMINI (0.19.1), ExAC (0.3.1), ESP (v0.0.30), 1000 Genomes (Phase 3), TAGC (EGAD00001000781), Rosetta ddG (Rosetta 3), Pfam (31.0), MSigDB (v6.0), HGMD (2015.3), OMIM (2017-04), MGI (6.08), GDI (Itan\_et\_al\_2015), RVIS (Petrovsk\_et\_al\_2013), PLINKSEQ (0.10), GTEx (V6p), HPA (16.1), and HINT (Version 4).

If a recipe was uploaded whose workflow uses older software or datasets (latest versions are selected on site by default), there will be a prompt on the fly asking whether you want to use the legacy version or the latest version of the resources.

The screenshot shows the GeMSTONE web interface with a modal dialog box. The dialog box has a title bar that says 'gemstone.yulab.org says:'. The main text reads: 'This recipe was created using a different ClinVar\_version (2017-03). To undo this change check "Advanced Versioning Options" below.' There is an 'OK' button at the bottom right of the dialog. The background shows the GeMSTONE interface with a 'Submit a New Job' button and other navigation options.

Using a subset of our sample VCF file: it takes 11 minutes to process a 1MB VCF file (5 samples, 13,800 variants) under default settings and without being queued. Benchmarks on processing time in terms of # of variants and # of samples (as well as the composition of samples: # of sporadic samples and # of families) as shown below.

Job #	VCF Input File Size (MB)	# of Samples (S:sporadic,F:family)	# of Variants (#NS)*	GeMSTONE Default? (-:skip,+:implement)	Time (min)
1	1	5 (5S0F)	13,800 (2,166)	v	11
2	0.6	1 (1S0F)	13,800 (2,166)	v	11
3	5	5 (5S0F)	69,415 (11,633)	v	18
4	1	5 (5S0F)	13,800 (2,166)	- <i>in silico</i> predictions	3
5	5	5 (5S0F)	69,415 (11,633)	- <i>in silico</i> predictions	10
6	1	5 (2S1F)	13,800 (2,166)	+ Co-segregation (AD)	10
7	5	5 (2S1F)	69,415 (11,633)	+ Co-segregation (AD)	13
8	5	5 (1S2F)	69,415 (11,633)	+ Co-segregation (AD)	15

\*NS: Non-Synonymous variants selected by the default Variant Consequence filter for downstream annotations, filtering, and/or co-segregation analysis.

From the above sample jobs, the take-home messages are:

(i) by comparing jobs #1, #2, #3: under default settings, **processing time is *mainly* dependent on # of variants but not # of samples**. Explanation: under default settings GeMSTONE will screen all variants across all samples as a group (“No Inheritance”), so increasing # of variants will increase the site-by-site screening time whereas increase in # of samples will have much less effect (but significant increase in # of samples will affect the processing time when extracting genotypes for a much larger group of samples).

(ii) by comparing jobs #1 and #4, #3 and #5 respectively, ***in silico* predictions takes about 8 minutes and is independent of # of variants**, and it appeared to be the time-limiting step for the sample jobs. Explanation: this is due to the nature of dbNSFP searching tool where it sends queries by chromosome, i.e., the searching time depends on the # of chromosomes being affected by the variants but not the # of variants. The searching time is also independent of # of predictors selected. As shown it only took 3 minutes for a 1MB VCF job without deleterious predictions.

(iii) by comparing jobs #1 and #6, #3 and #7 respectively, **adding co-segregation analysis does not necessarily increase the processing time**. Explanation: when we group # of samples into a family and require variants to be co-segregating in this family, large fraction of the variants will be excluded in a more efficient way.

(iv) to extend on (iii) by comparing #7 and #8, **processing time of jobs with co-segregation analysis is dependent on the # of families**. Explanation: co-segregation analysis searches co-segregating variants in each family, so increase in the # of families will increase the overall co-segregation analysis time. **Thus the actual processing time will depend on a combined factor of sample composition and the mode of inheritance**.