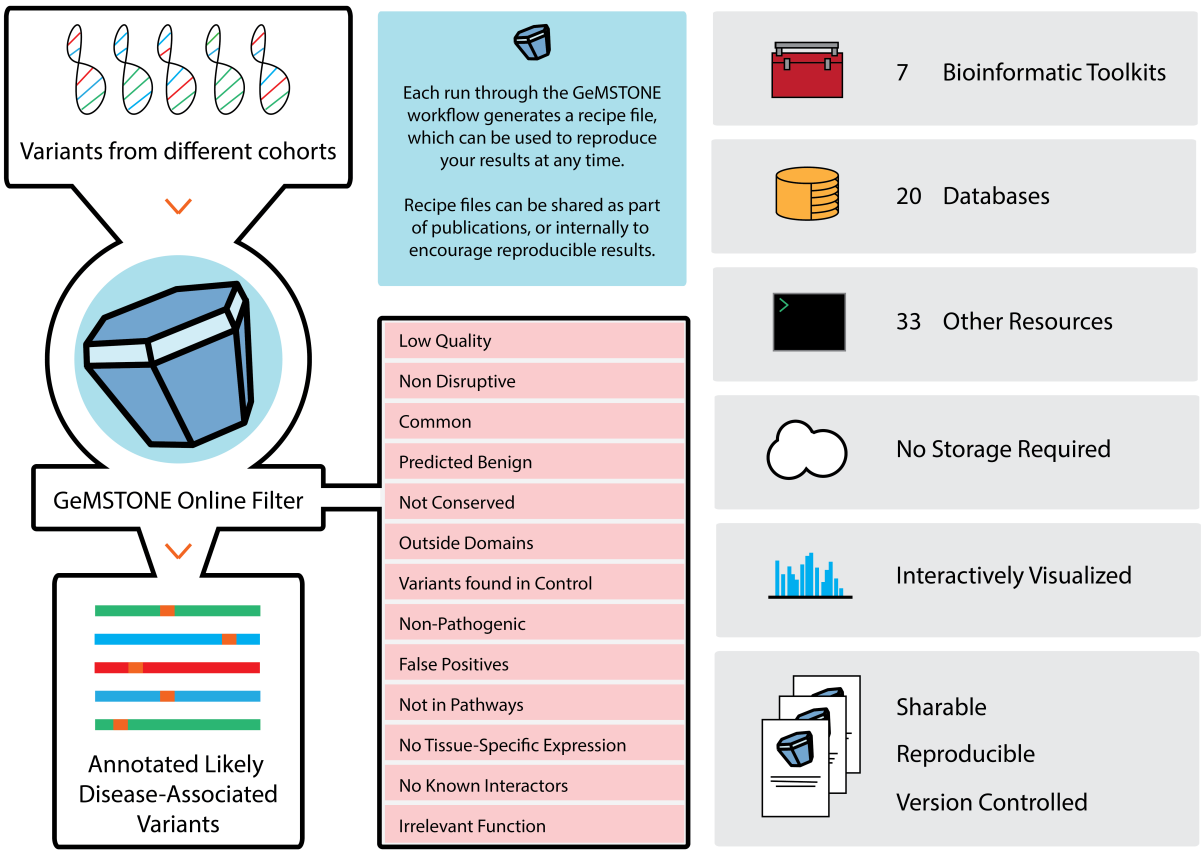
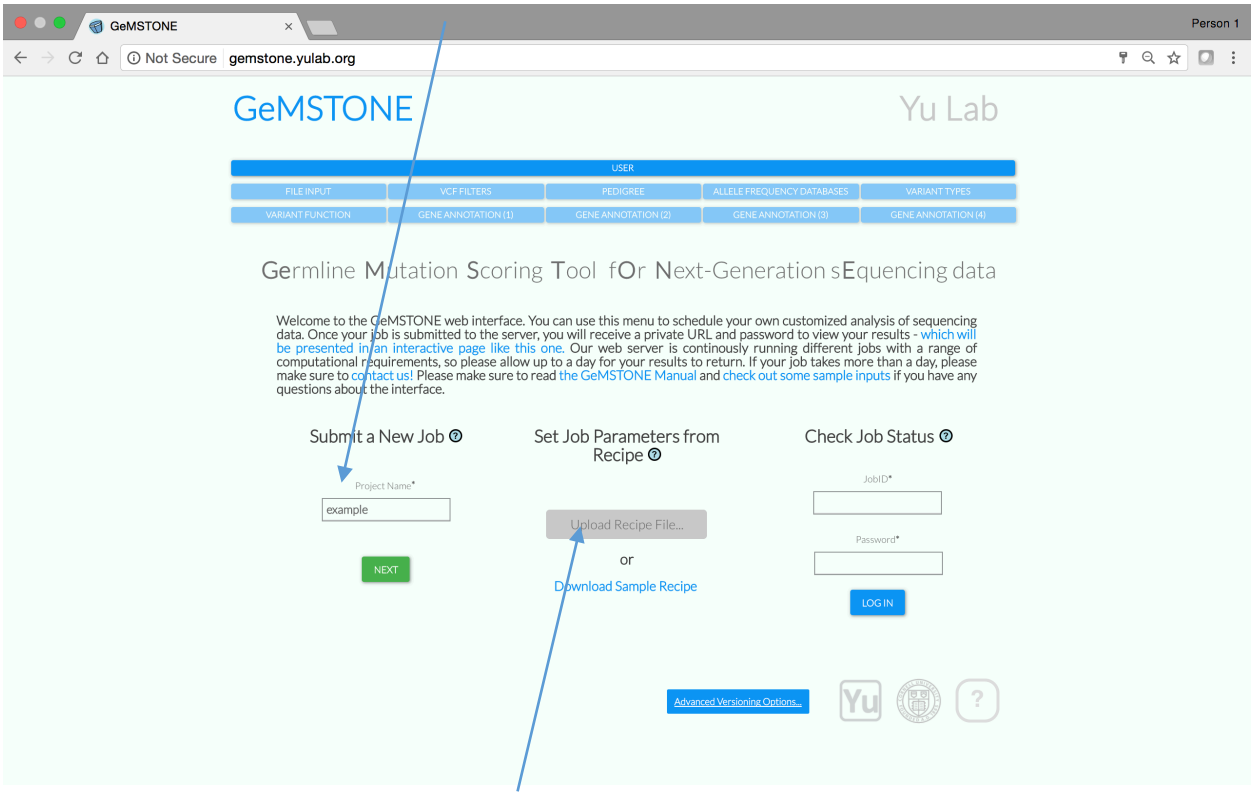
# GeMSTONE Manual



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



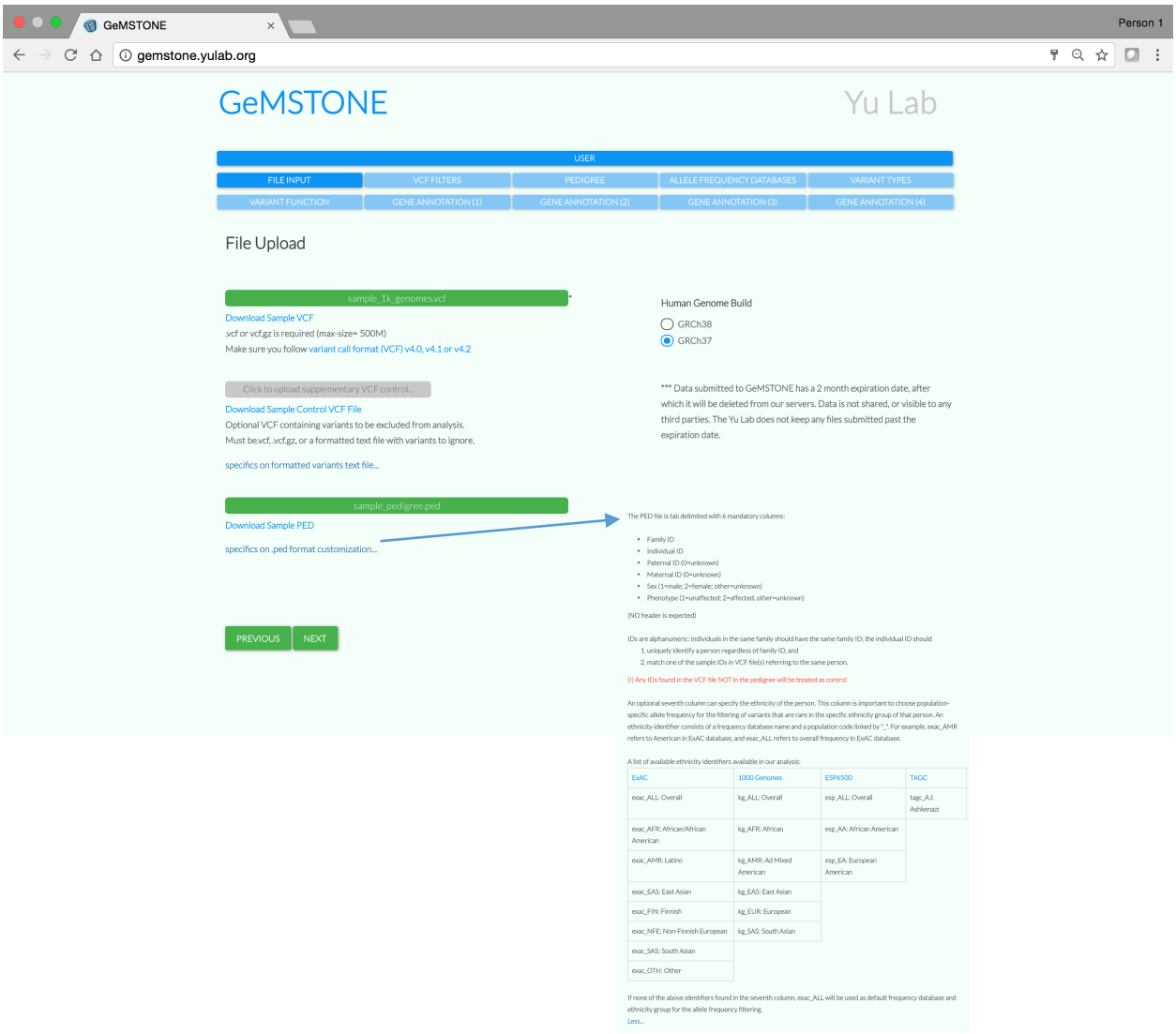

Or upload a recipe to use previous parameters to reproduce results!

**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.

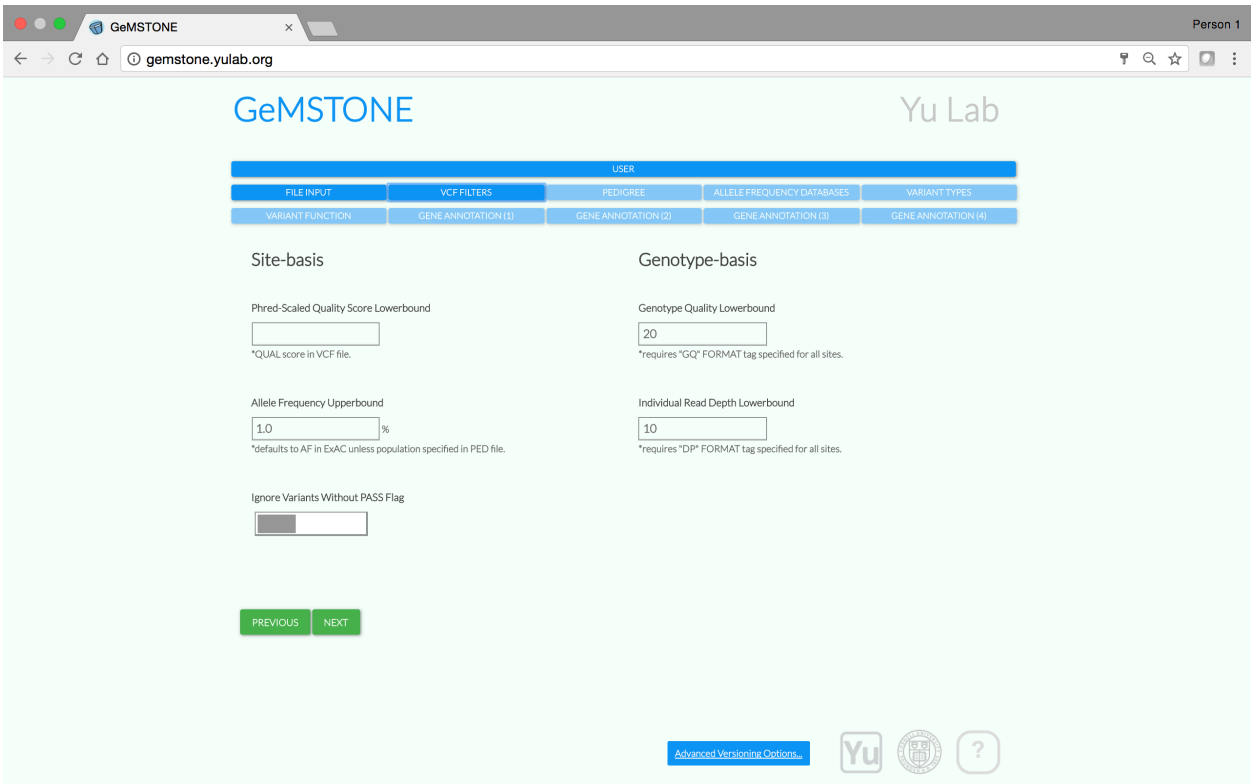



Information from the .ped file is very important for co-segregation analysis:

- *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.

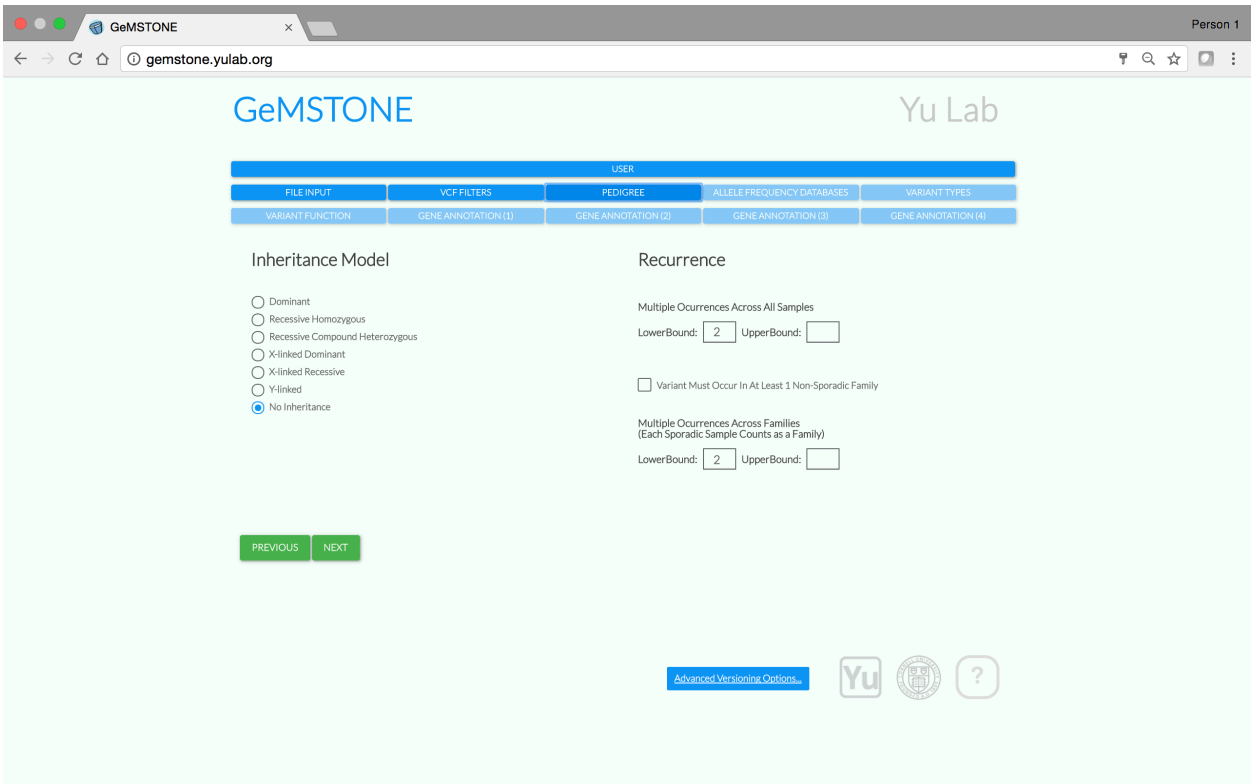



- Phred-scaled quality score lowerbound: the minimum *QUAL score* for a variant to be included; *default*: None. (*vcftools command: --minQ <float>*)
- Ignore Variants Without PASS Flag: exclude variants of which the *FILTER status* is not PASS; *default*: OFF. (*vcftools command: --remove-filtered-all*)
- 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** | ***A variant will be kept if it is …*** | ***A variant will be removed if it is…*** |
|---|---|---|
| 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 (*using comp_hets in GEMINI*) | 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 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CHROM | POS | ID | REF | ALT | FILTER | INDIVIDUAL_ID | CONSEQU | PUTATIVE | GENE_NA | ENTREZ_I | ENSE |
| 51 | 1 | 52499097 | rs116535272 | G | C | PASS | [NA19649] | missense_va | MODERATE | KTI12 | 112970 | ENSG |
| 52 | 1 | 54605318 | rs77544356 | TG | T | PASS | [HG00097,HG00100] | frameshift_v | HIGH | CDCP2 | 200008 | ENSG |
| 53 | 1 | 54605318 | rs77544356 | T | TGC | PASS | [HG00097,HG00100] | frameshift_v | HIGH | CDCP2 | 200008 | ENSG |
| 54 | 1 | 54605318 | rs77544356 | T | TTG | PASS | [HG00097,HG00100] | frameshift_v | HIGH | CDCP2 | 200008 | ENSG |
| 55 | 1 | 55223744 | rs35201073 | G | C | PASS | [NA19648] | missense_va | MODERATE | PARS2 | 25973 | ENSG |
| 56 | 1 | 60520988 | rs144671684 | G | A | PASS | [NA19648],[NA19649] | missense_va | MODERATE | C1orf87 | 127795 | ENSG |
| 57 | 1 | 62675673 | rs200789118 | G | T | PASS | [NA19648] | missense_va | MODERATE | L1TD1 | 54596 | ENSG |
| 58 | 1 | 62676284 | . | CAGA | C | PASS | [NA19648],[NA19649] | inframe_del | MODERATE | L1TD1 | 54596 | ENSG |
| 59 | 1 | 65129491 | . | A | C | PASS | [NA19648] | missense_va | MODERATE | CACHD1 | 57685 | ENSG |
| 60 | 1 | 67154849 | . | G | C | PASS | [NA19648] | missense_va | MODERATE | SGIP1 | 84251 | ENSG |
| 61 | 1 | 67390481 | . | G | C | PASS | [NA19648],[NA19649] | missense_va | MODERATE | WDR78 | 79819 | ENSG |
| 62 | 1 | 67447551 | . | A | C | PASS | [NA19648],[NA19649] | missense_va | MODERATE | MIER1 | 57708 | ENSG |
| 63 | 1 | 74670359 | rs148933608 | C | T | PASS | [NA19648],[NA19649] | missense_va | MODERATE | FPGT | 8790 | ENSG |
| 64 | 1 | 76345823 | rs5745459 | A | G | PASS | [NA19649] | missense_va | MODERATE | MSH4 | 4438 | ENSG |
| 65 | 1 | 82456482 | rs144339910 | T | A | PASS | [NA19648] | missense_va | MODERATE | LPHN2 | 23266 | ENSG |
| 66 | 1 | 86591837 | rs11161747 | G | T | PASS | [HG00097,HG00100],[NA19648],[NA19649] | missense_va | MODERATE | COL24A1 | 255631 | ENSG |
| 67 | 1 | 87045799 | rs201405115 | C | T | PASS | [NA19648] | missense_va | MODERATE | CLCA4 | 22802 | ENSG |
| 68 | 1 | 92200437 | rs41286789 | T | C | PASS | [NA19648] | missense_va | MODERATE | TGFBR3 | 7049 | ENSG |
| 69 | 1 | 93091349 | rs200027454 | A | C | PASS | [NA19648],[NA19649] | splice_donor | HIGH | EVI5 | 7813 | ENSG |
| 70 | 1 | 94486816 | rs200443984 | C | A | PASS | [NA19649] | missense_va | MODERATE | ABCA4 | 24 | ENSG |
| 71 | 1 | 109004611 | . | C | A | PASS | [NA19649] | missense_va | MODERATE | NBPF6 | 653149 | ENSG |
| 72 | 1 | 111957411 | rs150120731 | C | T | PASS | [NA19649] | missense_va | MODERATE | OVGP1 | 5016 | ENSG |
| 73 | 1 | 111957517 | rs3767609 | T | C | PASS | [HG00097,HG00100],[NA19649] | missense_va | MODERATE | OVGP1 | 5016 | ENSG |
| 74 | 1 | 111957570 | rs45455292 | G | C | PASS | [HG00097,HG00100],[NA19649] | missense_va | MODERATE | OVGP1 | 5016 | ENSG |
| 75 | 1 | 111957592 | rs56294468 | A | G | PASS | [HG00097,HG00100],[NA19649] | missense_va | MODERATE | OVGP1 | 5016 | ENSG |
| 76 | 1 | 114437834 | rs201156296 | A | C | PASS | [NA19648] | missense_va | MODERATE | AP4B1 | 10717 | ENSG |
| 77 | 1 | 115124203 | rs62621917 | T | C | PASS | [NA19648] | missense_va | MODERATE | BCAS2 | 10286 | ENSG |

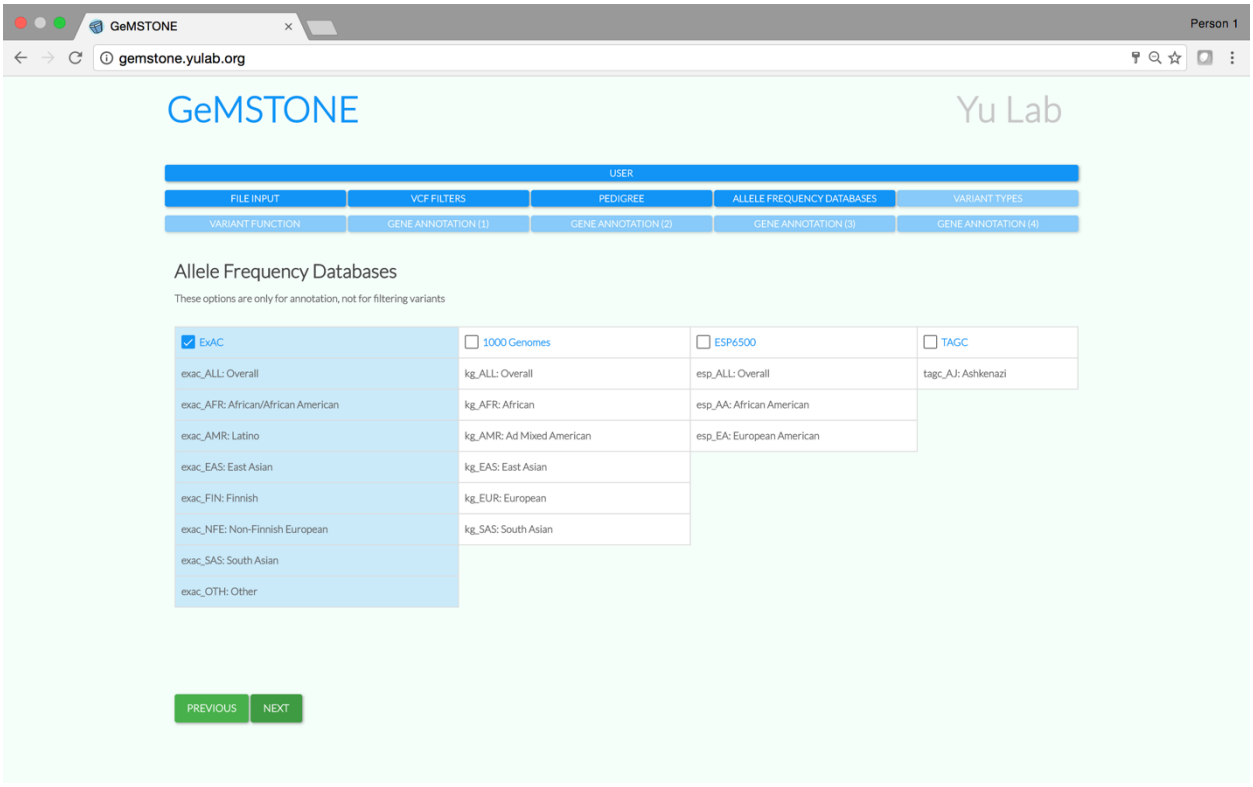
Variants are listed by row with meta annotations (not shown in the screenshot). Focusing on the seventh column *INDIVIDUAL_ID*, it reports the carriers of each variant with brackets indicating the same family unit, for examples, the #51 variant was observed in sample NA19649, the #52 variant was co-segregated 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.

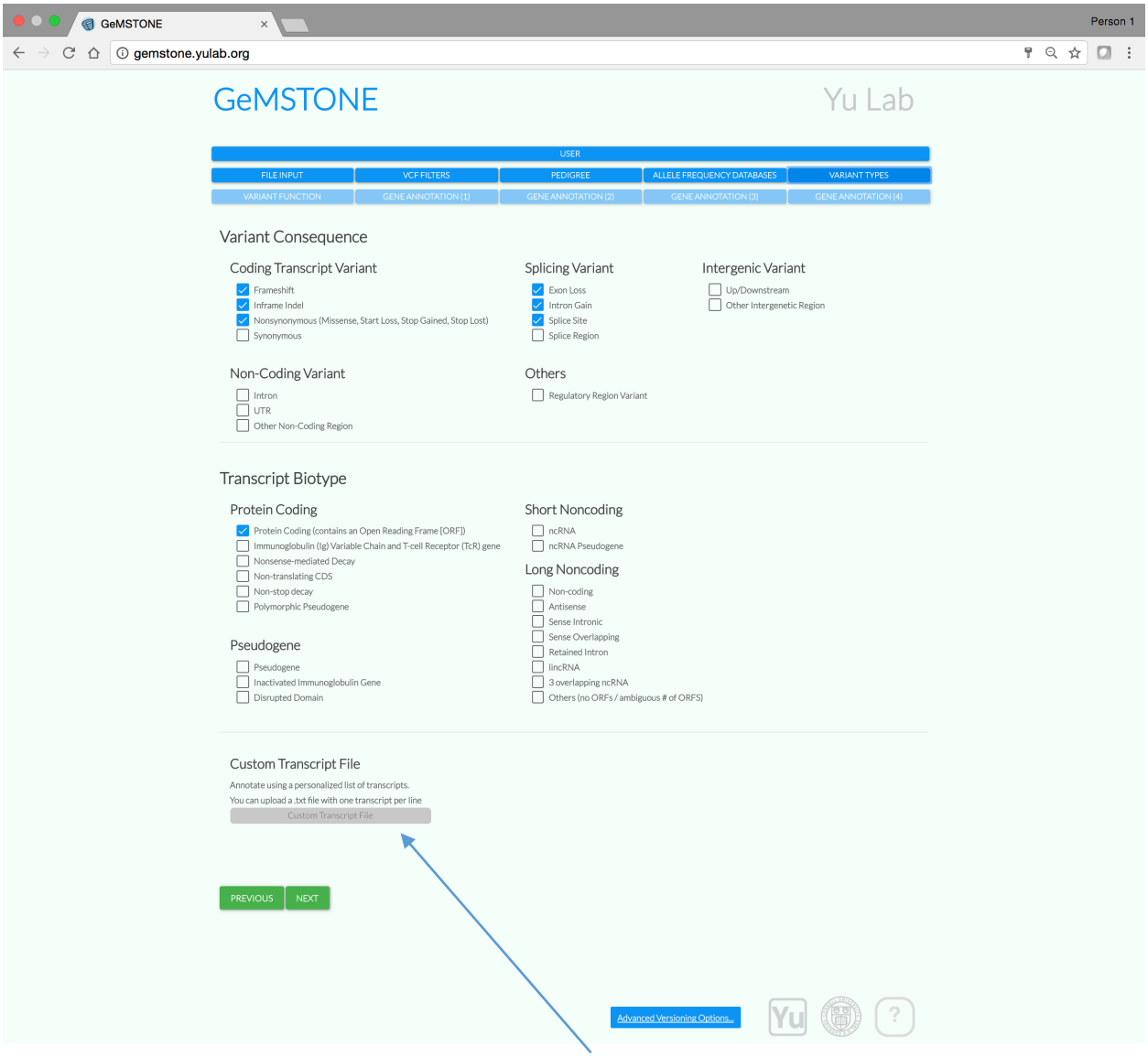
- *Variant Must Occur In At Least 1 Non-Sporadic Family*: if checked, requires the variant to be observed as a co-segregating event in at least one family with multiple family members, i.e. sporadic samples or families with only one individual sequencing available will not be considered.
- *Multiple 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.

| *Multiple Occurrences Across All Samples* | *Variant Must Occur In At Least 1 Non-Sporadic Family* | *Multiple Occurrences Across Families (Each Sporadic Sample Counts as a Family)* | 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).



**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*)

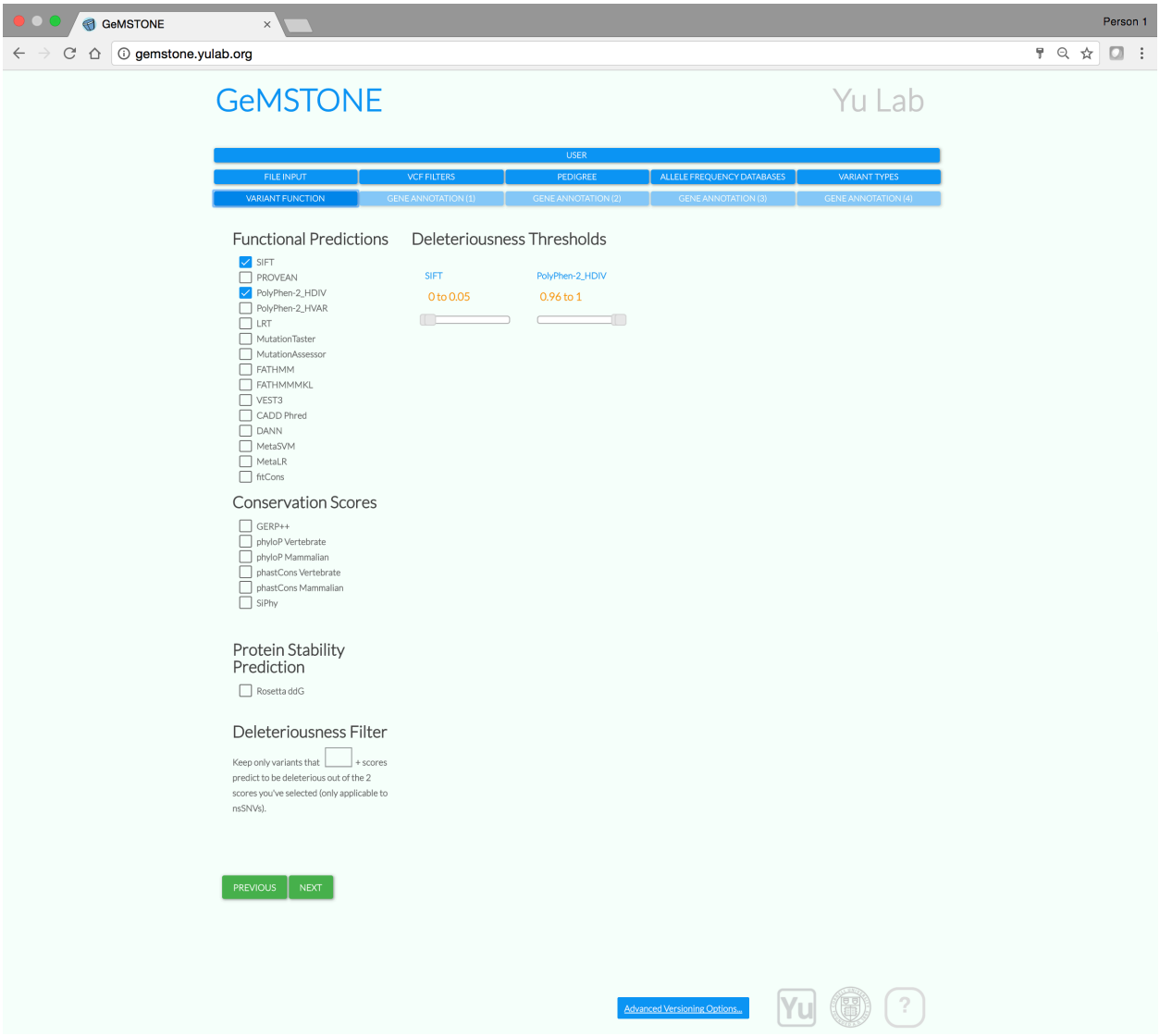



SnpEff by default uses canonical transcripts, you can upload your own here! Ensembl Transcript ID per line. (*java -Xmx4g -jar snpEff.jar eff -v [GRCh37.75,GRCh38.86] -onlyTr your_transcripts.txt*)

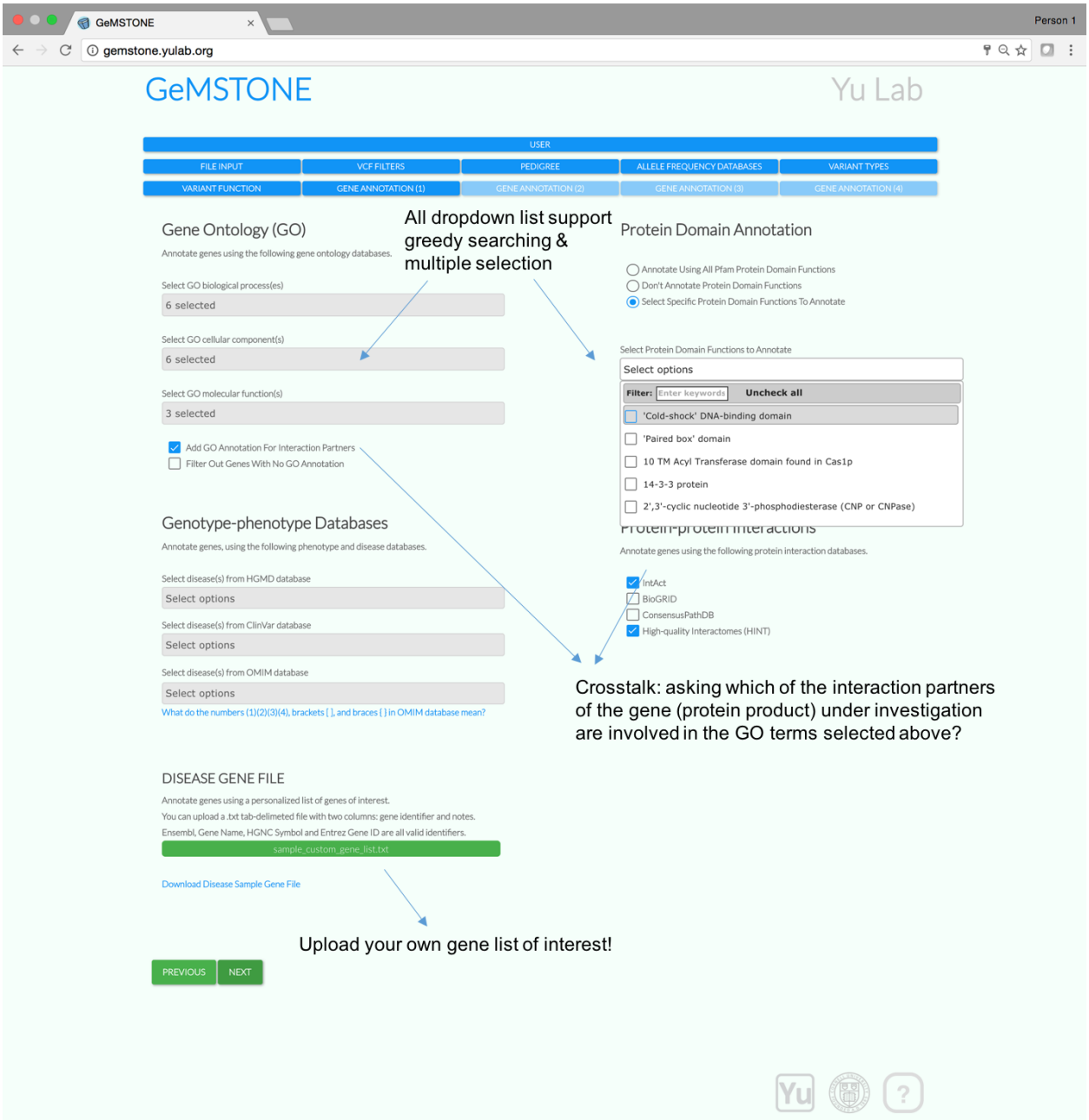
**Step7 @VARIANT FUNCTION**: Annotate and/or filter by *in silico* precisions on variant function. Another very important informatic evidence for variant implication comes from in silico analysis, predicting a variant is likely to be deleterious in terms of biological function or in an evolutionary sense. 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!

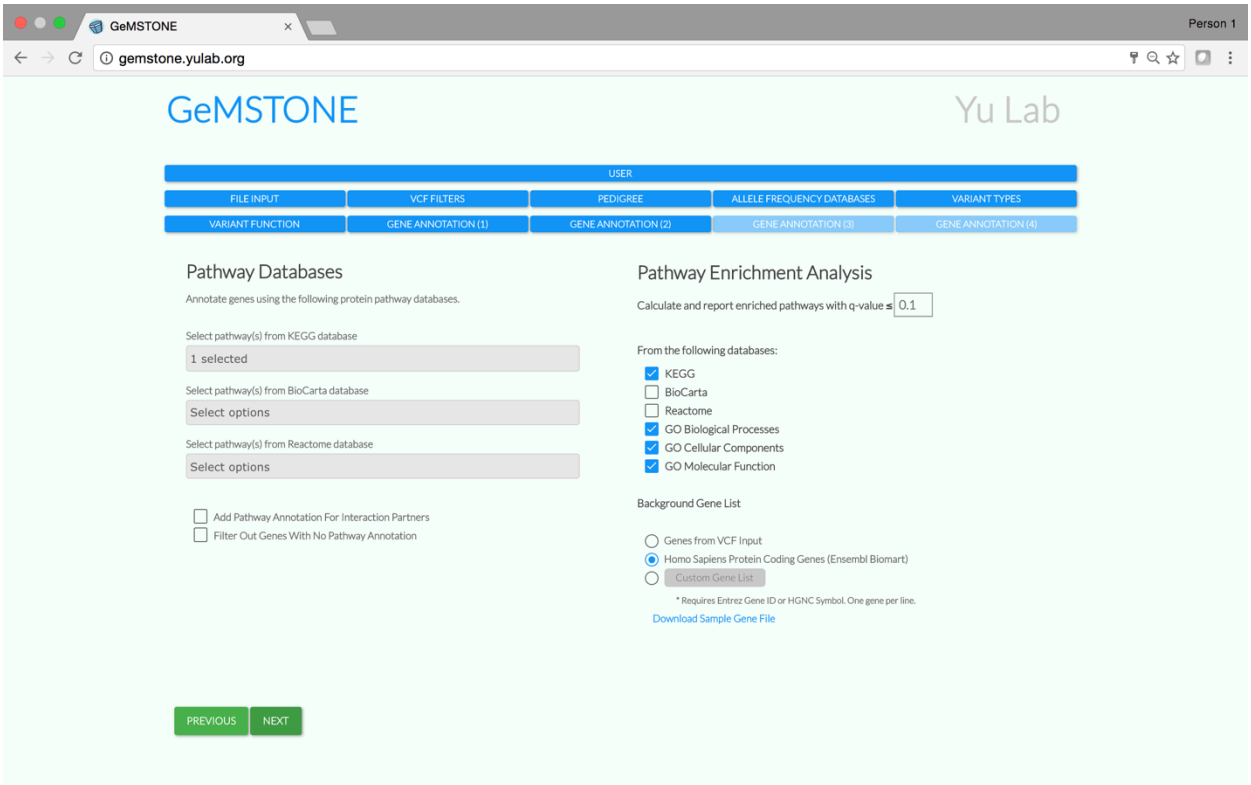
PLUS! Use the '*global deleteriousness filter*' to set a threshold on the number of selected predictors needed in order for a variant to pass the filter. By default (or entering 0), GeMSTONE will annotate the number for you for future downstream decisions.



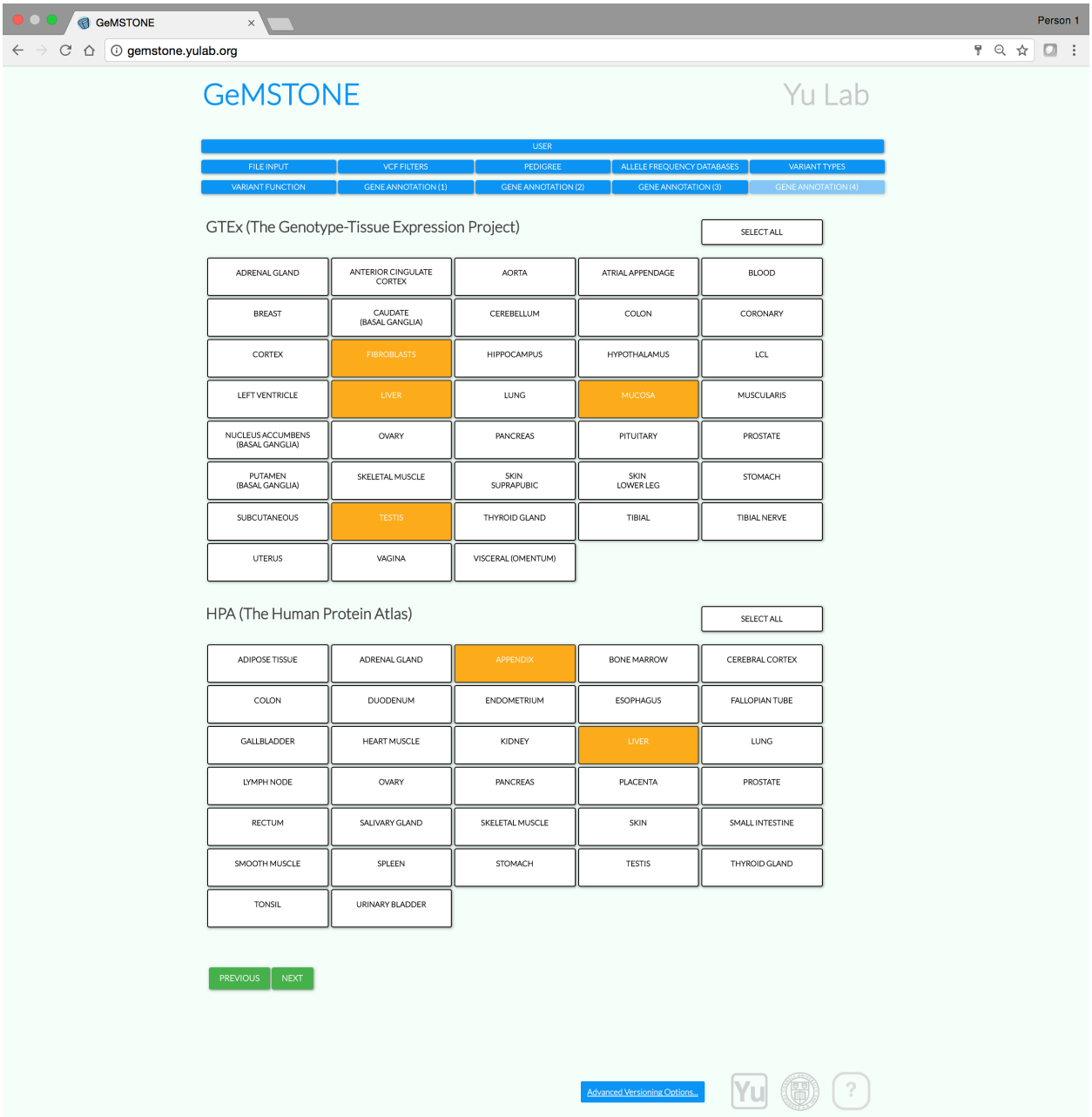
**Step8 @GENE ANNOTATION (1)(2)(3)**: Annotate and/or filter genes of biological function, disease implication, protein domain, protein-protein interaction, protein tissue expression, and initiate their 'crosstalk'!



*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*).

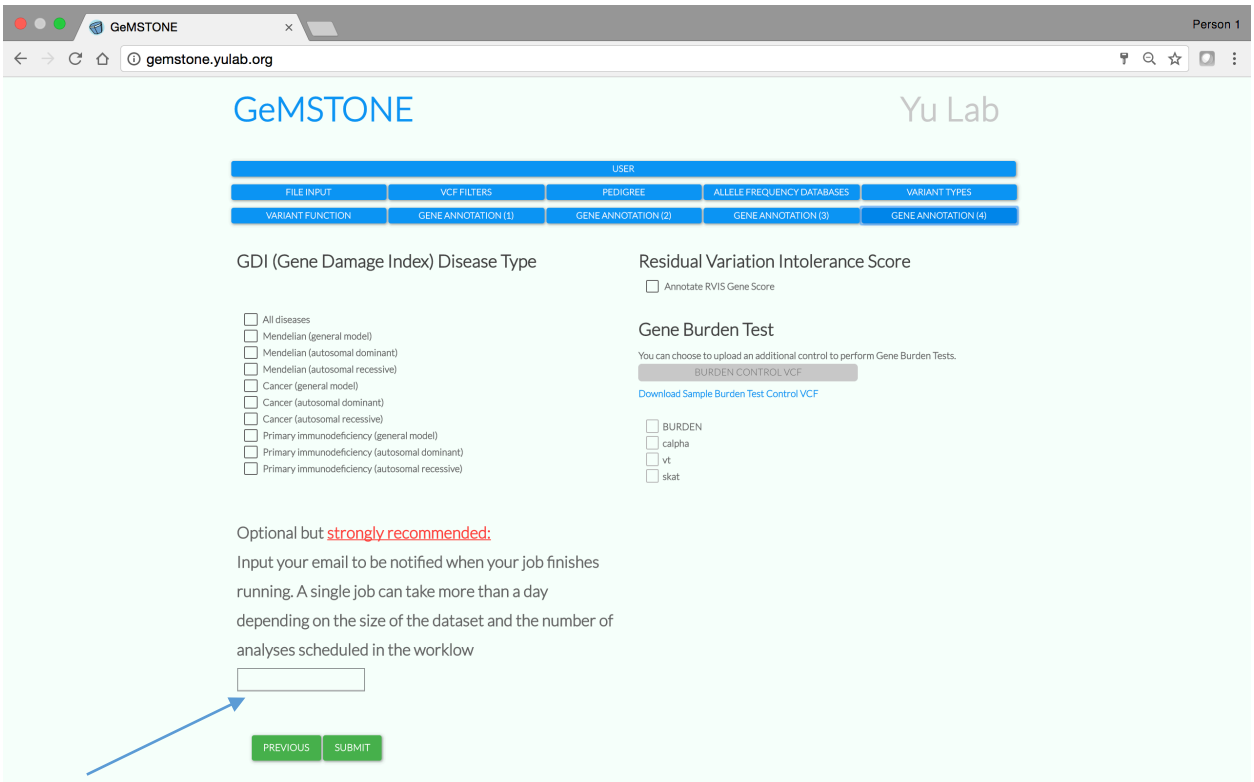


Annotation of gene expression provides the level (or presence) of protein expressed in tissue(s) selected, and in addition, gene expression enrichment/preference in particular tissue(s), differential expression with respect to sex (*log2FoldChange(females/males)_FDR*), ethnicity (*log2FoldChange(AA/EA)_FDR*), age (*coefficient_FDR*).



**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.

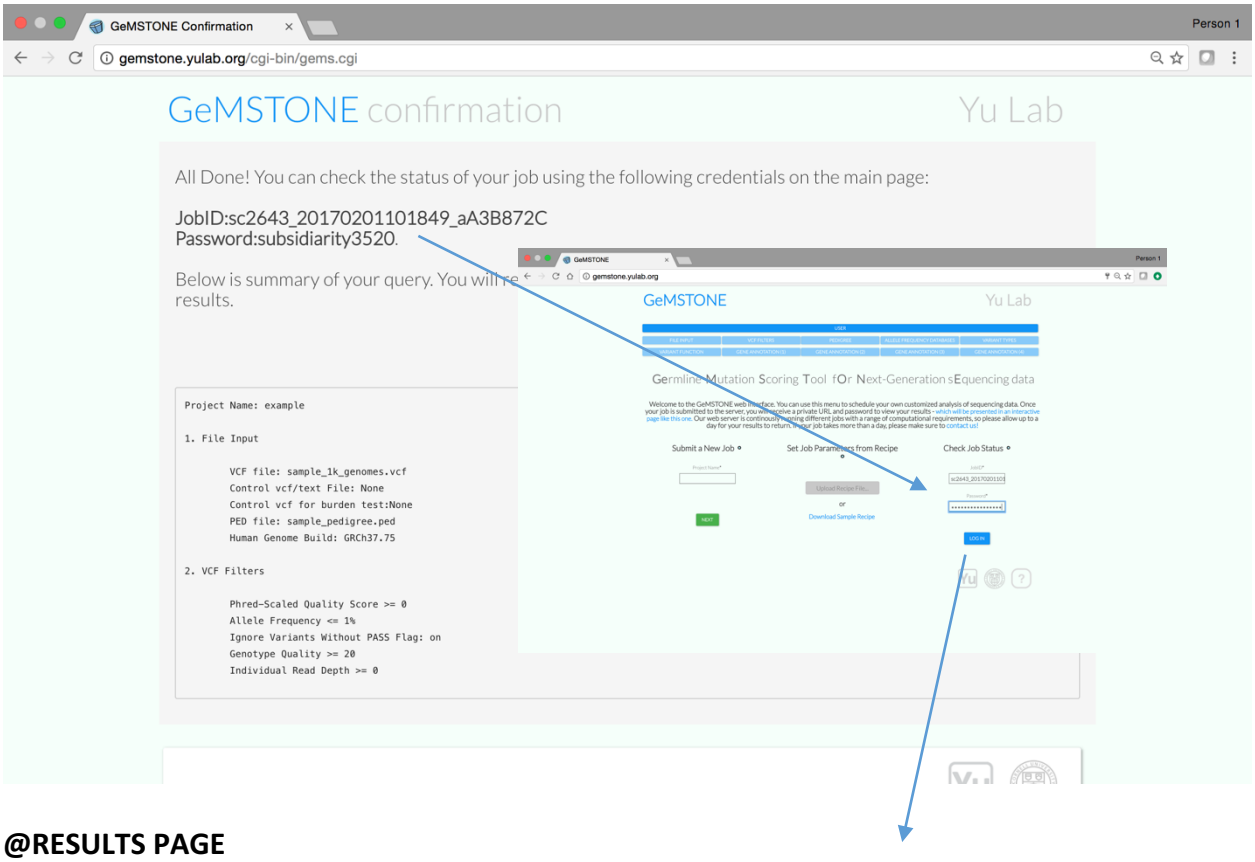



**SUBMIT!!!**

**@CONFIRMATION PAGE**

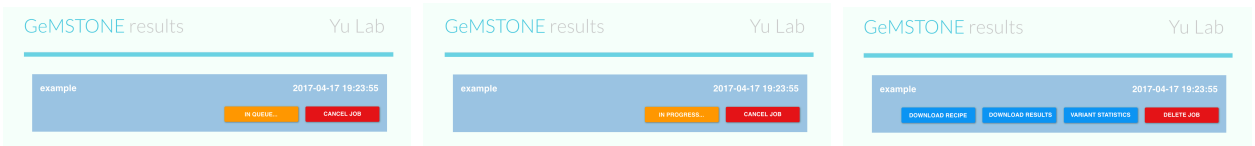
You will then be navigated to a *confirmation* page with

- 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.



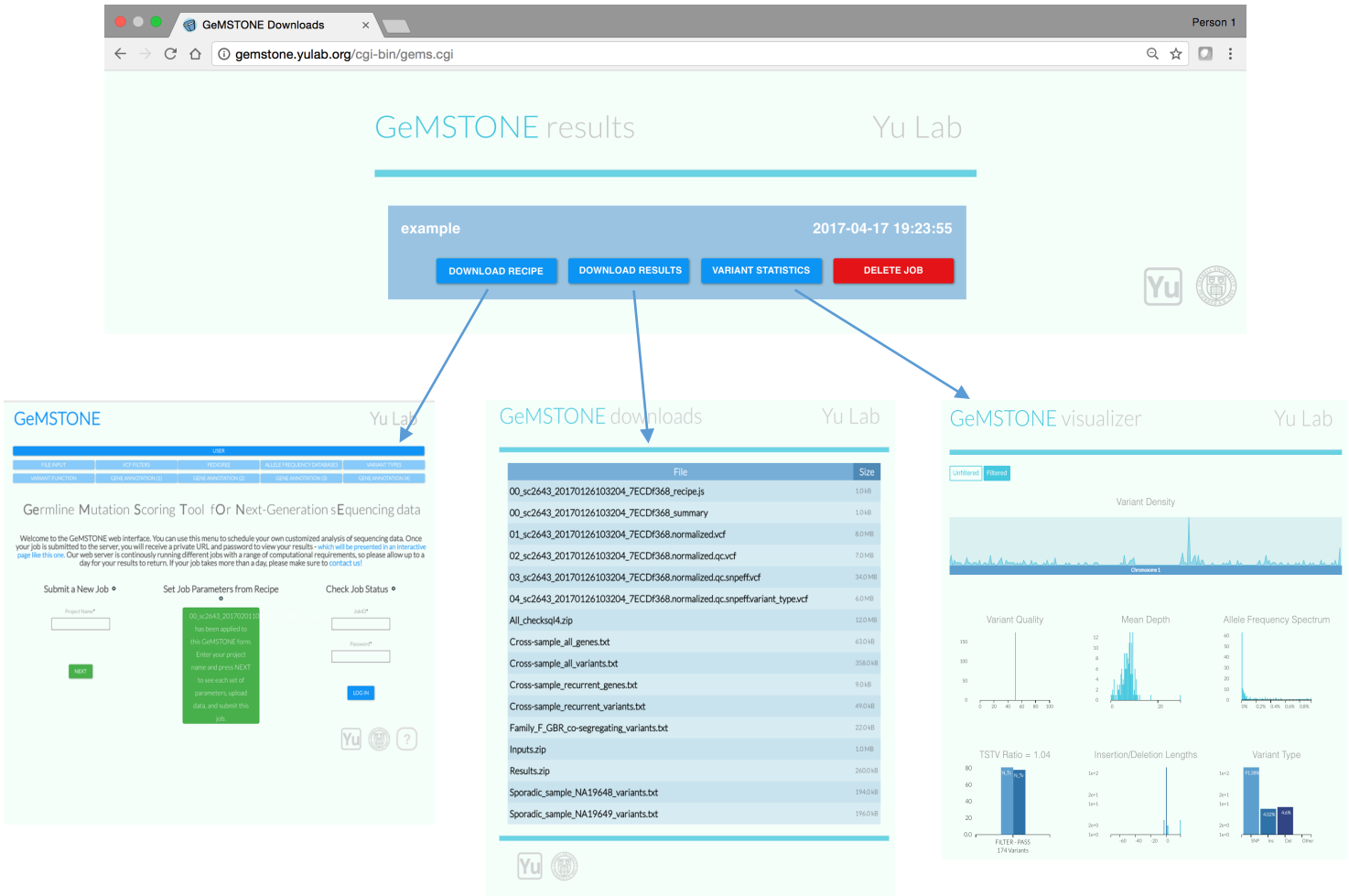

**@RESULTS PAGE**

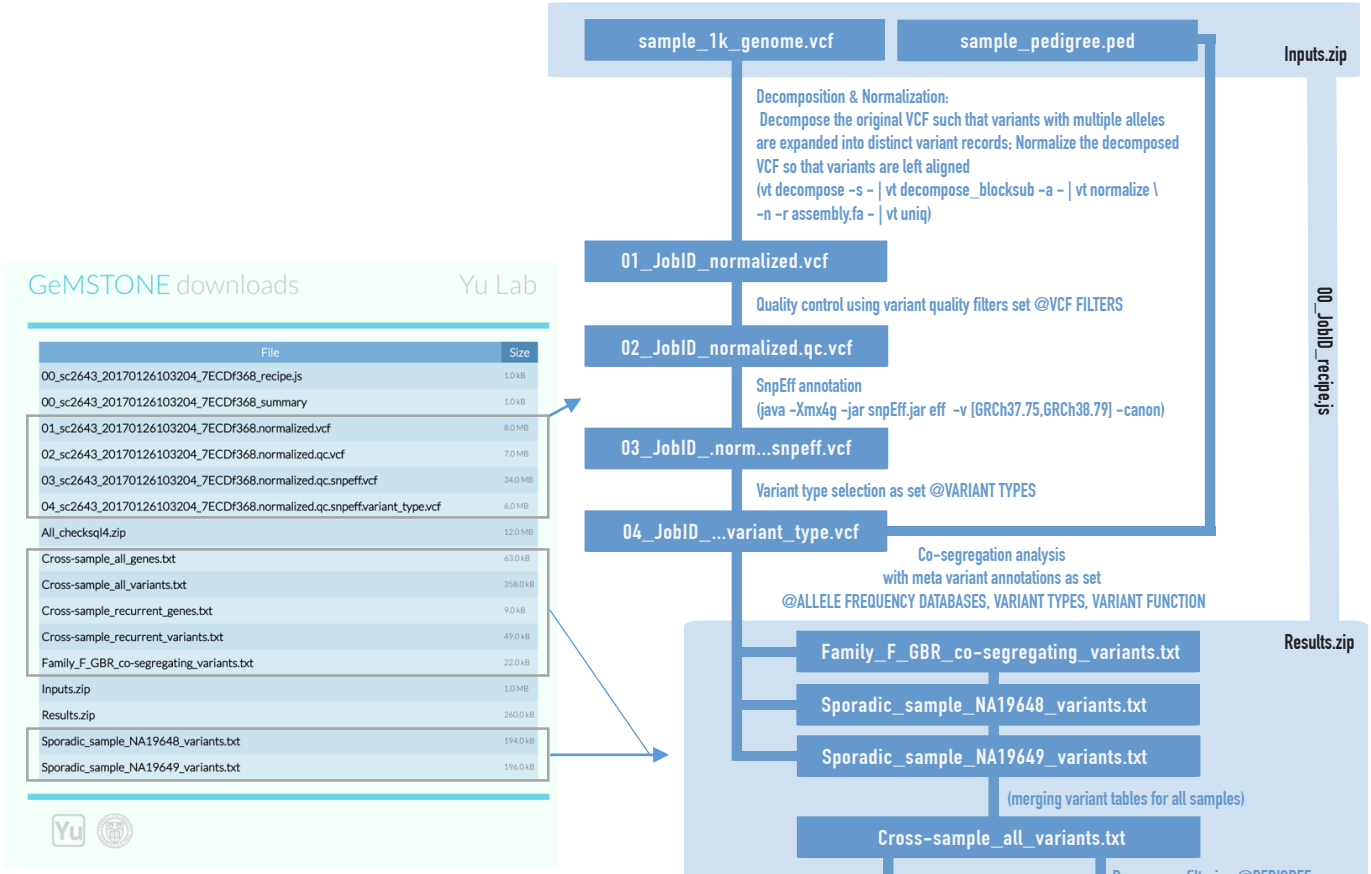
Use your jobID and password to log in to the result page and view the job status – it either IN QUEUE, IN PROGRESS, or DONE with all your result files downloadable!



**@RESULTS PAGE**

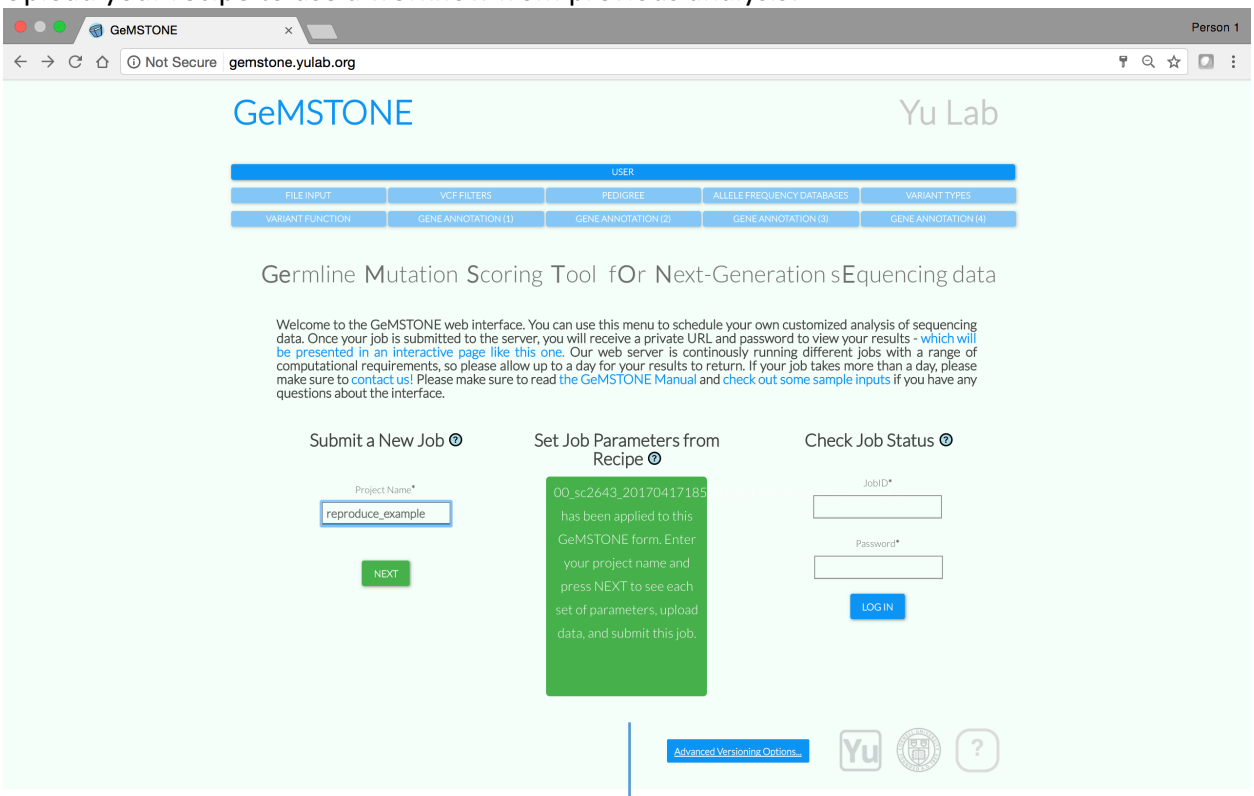
- *DOWNLOAD RECIPE* provides a program readable recipe file that records all selections and parameter settings for this job and versions of all programs and databases used. You can upload the recipe @USER to re-use this workflow – all previous settings will be automatically loaded including all selections 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.



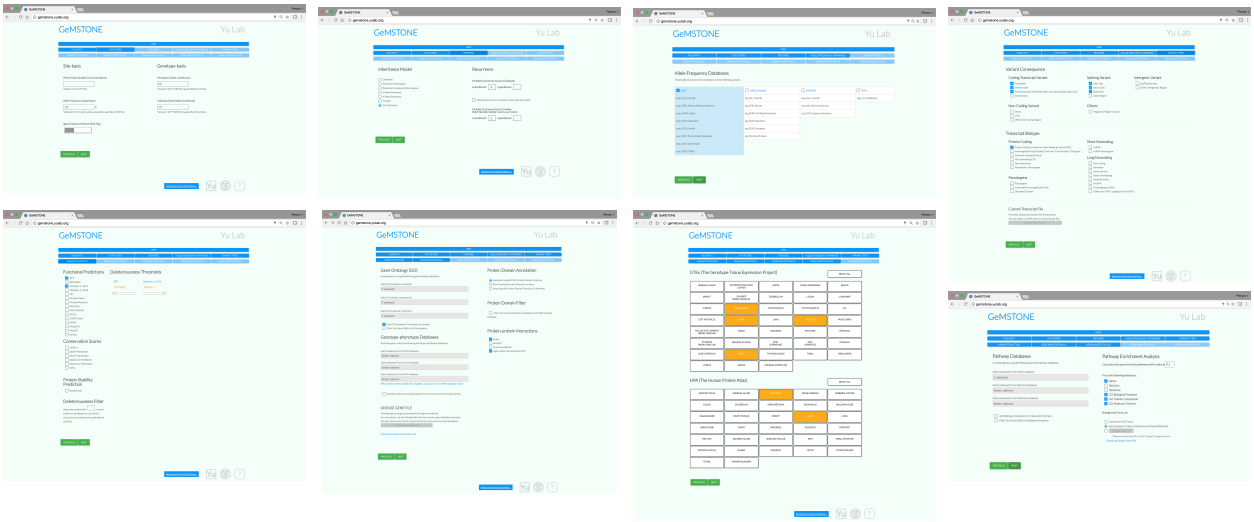

GeMSTONE downloads
Yu Lab
File
Size
00_sc2643_20170126103204_7ECDf368_recipe.js
00_sc2643_20170126103204_7ECDf368_summary
01_sc2643_20170126103204_7ECDf368.normalized.vcf
02_sc2643_20170126103204_7ECDf368.normalized.qc.vcf
03_sc2643_20170126103204_7ECDf368.normalized.qc.snpeff.vcf
04_sc2643_20170126103204_7ECDf368.normalized.qc.snpeff.variant_type.vcf
All_checksql4.zip
Cross-sample_all_genes.txt
Cross-sample_all_variants.txt
Cross-sample_recurrent_genes.txt
Cross-sample_recurrent_variants.txt
Family_F_GBR_co-segregating_variants.txt
Inputs.zip
Results.zip
Sporadic_sample_NA19648_variants.txt
Sporadic_sample_NA19649_variants.txt
sample_1k_genome.vcf
sample_pedigree.ped
Inputs.zip
Decomposition & Normalization:
Decompose the original VCF such that variants with multiple alleles are expanded into distinct variant records; Normalize the decomposed VCF so that variants are left aligned
(vt decompose -s - | vt decompose_blocksub -a - | vt normalize \
-n -r assembly.fa - | vt uniq)
01_JobID_normalized.vcf
Quality control using variant quality filters set @VCF FILTERS
02_JobID_normalized.qc.vcf
SnpEff annotation
(java -Xmx4g -jar snpEff.jar eff -v [GRCh37.75,GRCh38.79] -canon)
03_JobID_.norm...snpeff.vcf
Variant type selection as set @VARIANT TYPES
04_JobID_...variant_type.vcf
Co-segregation analysis
with meta variant annotations as set
@ALLELE FREQUENCY DATABASES, VARIANT TYPES, VARIANT FUNCTION
00_JobID_recipe.js
Results.zip
Family_F_GBR_co-segregating_variants.txt
Sporadic_sample_NA19648_variants.txt
Sporadic_sample_NA19649_variants.txt
(merging variant tables for all samples)
Cross-sample_all_variants.txt
Recurrence filtering @PEDIGREE
Cross-sample_recurrent_variants.txt
Gene annotations
@GENE ANNOTATION (1)-(4)
Cross-sample_all_genes.txt
Cross-sample_recurrent_genes.txt

**@USER**

Upload your recipe to use a workflow from previous analysis!

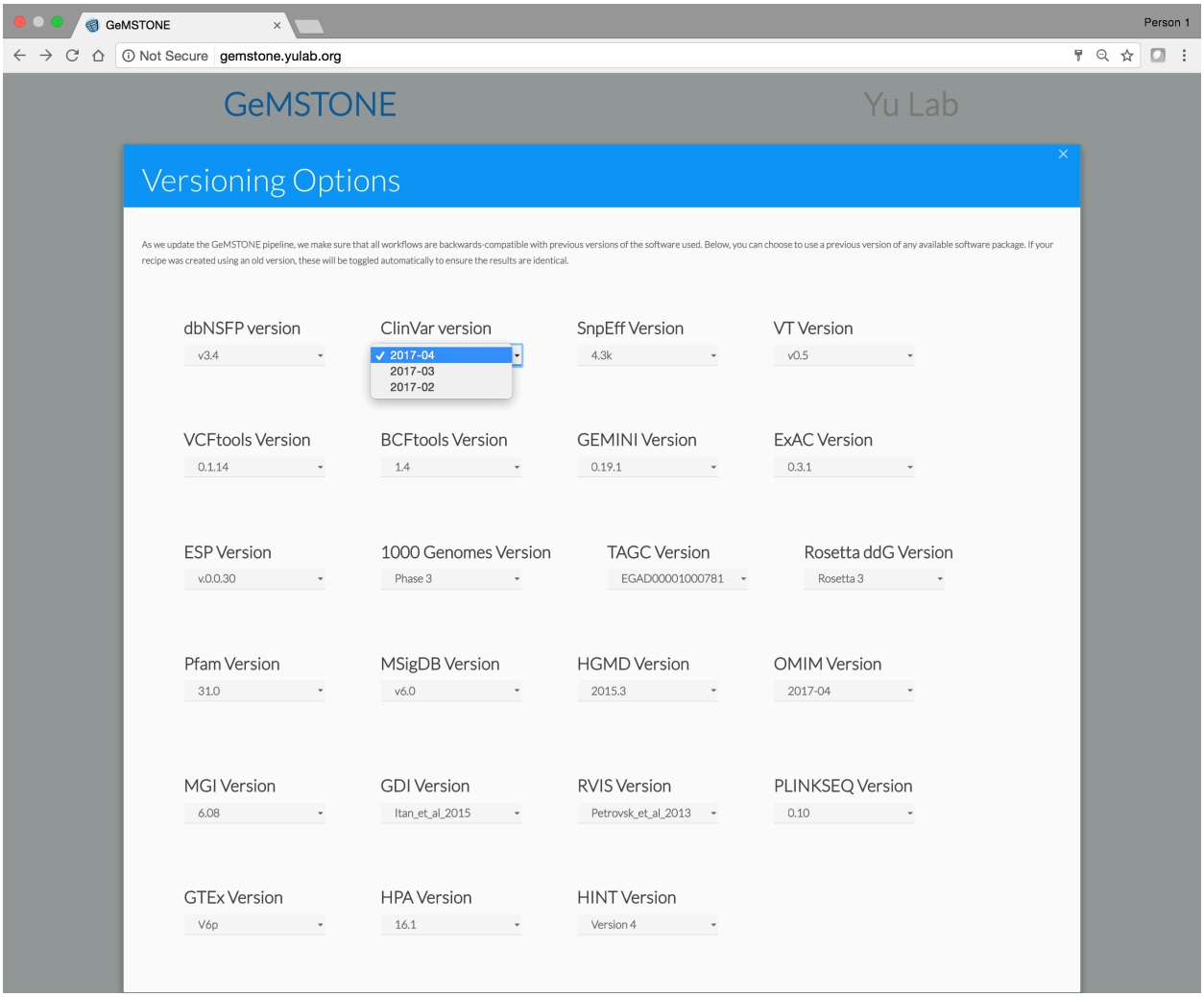



After uploading the recipe file, all parameter settings and selections will be automatically loaded – check tabs like below! You can readily upload your inputs and submit or modify any of the options as usual!

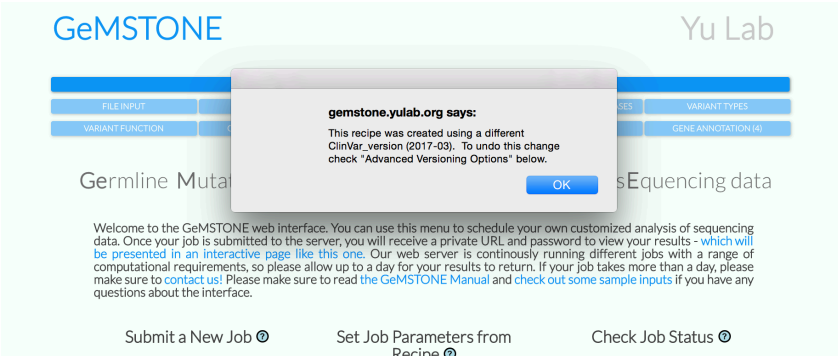


**@Advanced Visioning Options**

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.




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.



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) | √ | 11 |
| 2 | 0.6 | 1 (1S0F) | 13,800 (2,166) | √ | 11 |
| 3 | 5 | 5 (5S0F) | 69,415 (11,633) | √ | 18 |
| 4 | 1 | 5 (5S0F) | 13,800 (2,166) | - *in silico* predictions | 3 |
| 5 | 5 | 5 (5S0F) | 69,415 (11,633) | - *in silico* 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 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**.