Collapsing Analysis Workflow

Joe Hostyk, Ryan Dhindsa
j.hostyk@columbia.edu; rsd2135@cumc.columbia.edu
Created 2018-5-22; Edited 2020-2-11

The following workflow is based on previously published IGM collapsing analyses performed for idiopathic pulmonary fibrosis (Petrovski et al. 2017) and common epilepsies (Epi4K Consortium 2017).

Starting Steps

1. The handiest way to run all these steps is to create one bash file (files that have a ".sh" extension), put all the commands in it, and run them one at a time.

   The commands in the remainder of this workflow assume that these variables are defined in your bash script:

   ```
   PROJECT=/pathToProjectDirectory
   ATAV=/nfs/goldstein/software/sh/atav.sh
   ```

2. ATAV output will have a timestamp before filenames. Below, these are marked as `"TIMESTAMP"`. When running commands, change "TIMESTAMP" to the actual time in the ATAV-output files.

Step 1: Cohort Selection

In order to determine which samples you should use in your analysis, download a CSV that contains meta-information on all samples contained in Sequence:

- Go to the search page at https://sequence.igm.cumc.columbia.edu/search.php?action=searchGSS
  - If you do not have an account there, click "Request Account" and follow the steps there.
- For "Current Status", put "In DragenDB", and click "Search". You should be able to download a (large) file of all the available samples.
  - If you are working on exome samples, you can also put "Exome" in the "SeqType" field.

Then apply the following filtering criteria:

1) Ensure all of your control samples are approved for control use:
   `avaiContUsed == "yes" or "Yes"`

2) Choose controls with broad phenotypes that are not strongly genetic and do not have comorbidities with the phenotype of interest. The following broad phenotypes can almost always be included:
   `BroadPhenotype == "healthy family member", "control"`

   "Epilepsy" and "'kidney and urological disease" are the next biggest groups and can be used as long as they do not have comorbidities with your cohort's phenotype.

3) Keep samples with non-ambiguous sequencing gender (i.e. individuals whose sex can be determined from their sequencing data):
   `seqGender == "M" or "F"`

4) Remove samples with excess heterozygosity (determined by VerifyBamID), which suggests contamination:
   Keep samples where `ContaminationPercentage <= 2%`

5) Retain only well-covered samples:
   `CCDSBasesCov10X >= 85%`

6) Select exomes sequenced with well-attested kits. We typically retain individuals sampled with the following sequencing kits:
```
capture_kit == "65MB", "Roche", "RocheV2", "IDTERPv1", "IDTERPv1Plus",
"IDTxGEN", "IDTERPv1mtDNA", "AgilentCRE", "AgilentV4", "Agilentv5",
"AgilentV5", "AgilentV5UTR", "AgilentV6", "MedExome"
```

7) Remove IGM samples that are contained in the gnomAD database so that you can filter qualifying variants by their gnomAD allele frequencies later on:
   a. The filter --exclude-igm-gnomad-sample does this automatically. We only used this filter in Kinship because Kinship should be the first step, and the cohort will thus not contain any gnomAD samples from the get-go. As a sanity check, feel free to include the filter in all the subsequent steps as well.
   b. Additionally, all IGM samples contained in gnomAD can be found here: https://redmine.igm.cumc.columbia.edu/attachments/1832/igm_samples_to_exclude_when_using_gnomad_filter.txt

8) If you want to perform an ethnicity-specific run, you may use the SeqDB ethnicity PC prediction columns to retain individuals of your ancestry of interest. There are six possible ethnicities: Caucasian, Middle Eastern, Hispanic, East Asian, South Asian, and African. In order to be considered a member of that ethnicity, we require {Ethnicity}_prob > .95

Step 2: Create Sample File for ATAV

After determining the samples you are going to include in your collapsing analysis, make a tab-delimited file using the information from the SeqDB .csv file:

file format: Family ID, Individual ID, Paternal ID, Maternal ID, Sex, Phenotype, Sample Type, Capture Kit (tab-delimited)

1) Family ID: specify a family ID or use the same value as Individual ID to indicate this sample is being used as a non family sample
2) Individual ID: sample ID
3) Paternal ID: 0
4) Maternal ID: 0
5) Sex: 1=male, 2=female
6) Phenotype: 1=control, 2=case
7) Sample Type
   a. Can be Exome or Genome_As_Fake_Exome.
8) Capture Kit
   a. For Genome, put "Roche" as the kit.

Example: /nfs/goldstein/software/atav_home/data/sample/ALS_1424_DukeGr_ctrl.txt

Step 3: Kinship Pruning

At this stage, we need to remove all related samples from our cohort. We use KING to ensure that only unrelated (up to third-degree) individuals are retained in the sample list:

```
$ATAV \
    --ped-map \
    --kinship \
    --exclude-igm-gnomad-sample \
    --min-covered-case-percentage 95 \
    --min-covered-ctrl-percentage 95 \
    --min-coverage 10 \
    --variant
/nfs/goldstein/software/atav_home/data/variant/informative_snps.ld_pruned.37MB.txt \
    --sample $PROJECT/samples.txt \
```

Step 4: EIGENSTRAT

We then use the kinship-pruned sample list from Step 3 as input to EIGENSTRAT to remove ethnicity outliers from our sample list:

```
$ATAV \
    --ped-map \
    --eigenstrat \
    --min-covered-case-percentage 95 \
    --min-covered-ctrl-percentage 95 \
    --min-coverage 10 \
    --variant
/nfs/goldstein/software/atav_home/data/variant/informative_snps.ld_pruned.37MB.txt \
    --sample $PROJECT/Kinship/TIMESTAMP_Kinship_kinship_pruned_sample.txt \
   --out $PROJECT/Eigenstrat
```

To make things clearer later, set the following variable:

```
SAMPLES=$PROJECT/eigenstrat_pruned_sample.txt
```

Step 5: Site Coverage Harmonization

To alleviate issues coming from differential coverage between the case and control cohorts, we next prune out noisy nucleotide sites among the consensus coding sequence (CCDS release 15) public transcript or their 2 base pair intronic extensions:

```
$ATAV \
    --site-coverage-comparison \
    --gene-boundaries
/nfs/goldstein/software/atav_home/data/ccds/addjusted.CCDS.genes.index.r20.hg19.txt \
    --min-coverage 10 \
    --sample $SAMPLES \
    --out $PROJECT/Coverage/
```

We will use the output for our commands later.

Set these variables:

```
GENE_BOUNDARIES=$PROJECT/Coverage/TIMESTAMP_Coverage_site.clean.txt
COVERAGE_SUMMARY=$PROJECT/Coverage/TIMESTAMP_Coverage_coverage.summary.csv
```

Step 6: Synonymous Collapsing Analysis

We assume that synonymous variation in cases and controls should be drawn from the same random distribution. Therefore, we can use the exome-wide rate of rare synonymous variation to test whether there is significant inflation of neutral variation between the case and control cohorts.

```
$ATAV \
    --collapsing-dom \
    --mann-whitney-test
    --gene-boundaries $GENE_BOUNDARIES \
    --read-coverage-summary $COVERAGE_SUMMARY \
    --include-rvis \
    --include-known-var \
    --effect
LOW:start_retained,LOW:stop_retained_variant,LOW:synonymous_variant \
    --exclude-artifacts \
    --filter pass,likely,intermediate \
    --exclude-evs-qc-failed \
    --ccds-only \
    --min-coverage 10 \
    --include-qc-missing \
    --qd 5 --qual 50 --mq 40 --gq 20 --snv-sor 3 --indel-sor 10 --snv-fs
60 --indel-fs 200 --rprs -3 --mqrs -10 \
    --het-percent-alt-read 0.3-1 \
    --min-exac-vqslod-snv 5000 \
    --min-exac-vqslod-indel 5000 \
    --gnomad-exome-af 0 \
    --gnomad-exome-rf-tp-probability-snv 0 \
    --gnomad-exome-rf-tp-probability-indel 0 \
    --gnomad-exome-pop global \
    --exac-pop global \
    --exac-af 0 \
    --loo-af 0.0005 \
    --max-qc-fail-sample 0 \
    --sample $SAMPLES \
    --out $outputFolder/dominantSynonymous/
```

The QQ plot in the output should show no significant results.

The option --mann-whitney-test performs a Mann-Whitney U test to test for a significant difference in the number of rare synonymous variants per sample in the case versus control cohorts. The resulting p-value should *not* be significant. The results of the test are output in TIMESTAMP_qv_counts.txt. TIMESTAMP_qv_counts.png shows the distributions of the variants – the cases and controls should mostly overlap.

Step 8: Run Collapsing Analysis Models

Define the following variables in your bash script:

```
FUNCTIONAL_EFFECTS="HIGH:exon_loss_variant,HIGH:frameshift_variant,HIGH:rare_amino_acid_variant,HIGH:stop_gained,HIGH:start_lost,HIGH:stop_lost,HIGH:splice_acceptor_variant,HIGH:splice_donor_variant,HIGH:gene_fusion,HIGH:bidirectional_gene_fusion,MODERATE:3_prime_UTR_truncation+exon_loss_variant,MODERATE:5_prime_UTR_truncation+exon_loss_variant,MODERATE:coding_sequence_variant,MODERATE:disruptive_inframe_deletion,MODERATE:disruptive_inframe_insertion,MODERATE:conservative_inframe_deletion,MODERATE:conservative_inframe_insertion,MODERATE:missense_variant+splice_region_variant,MODERATE:missense_variant"

LOF_EFFECTS="HIGH:exon_loss_variant,HIGH:frameshift_variant,HIGH:rare_amino_acid_variant,HIGH:stop_gained,HIGH:stop_lost,HIGH:start_lost,HIGH:gene_fusion,HIGH:bidirectional_gene_fusion,HIGH:splice_acceptor_variant,HIGH:splice_donor_variant"
```

We use six models, designed and standardized by Slavé Petrovski:

	Dominant Ultra-Rare	**Dominant Rare**	**Dom Flex #1: Polyphen**	**Dom Flex #2: No Filter**	**Dom PTV**	**Recessive Autosomal**
LOO AF	.0005	.0005	0.001	0.001	0.001	.01
Gnomad AF	0	0.00002	0.001	0.001	0.001	.01
Exac AF	0	0.00005	0.001	0.001	0.001	.01
PP2 HumDiv	“probably”	“probably”	“probably”	None	None	None

Dominant Ultra-Rare:

```
$ATAV \
    --collapsing-dom \
    --mann-whitney-test
    --gene-boundaries $GENE_BOUNDARIES \
    --read-coverage-summary $COVERAGE_SUMMARY \
    --include-rvis \
    --include-known-var \
    --effect $FUNCTIONAL_EFFECTS \
    --polyphen probably \
    --exclude-artifacts \
    --filter pass,likely,intermediate \
    --exclude-evs-qc-failed \
    --ccds-only \
    --min-coverage 10 \
    --include-qc-missing \
    --qd 5 --qual 50 --mq 40 --gq 20 --snv-sor 3 --indel-sor 10 --snv-fs
60 --indel-fs 200 --rprs -3 --mqrs -10 \
    --het-percent-alt-read 0.3-1 \
    --min-exac-vqslod-snv 5000 \
    --min-exac-vqslod-indel 5000 \
    --gnomad-exome-af 0 \
    --gnomad-exome-rf-tp-probability-snv 0 \
    --gnomad-exome-rf-tp-probability-indel 0 \
    --gnomad-exome-pop global \
    --exac-pop global \
    --exac-af 0 \
    --loo-af 0.0005 \
    --max-qc-fail-sample 0 \
    --sample $SAMPLES \
    --out $outputFolder/dominantUltraRare/
```

Dominant Rare:

```
$ATAV \
    --collapsing-dom \
    --mann-whitney-test
    --gene-boundaries $GENE_BOUNDARIES \
    --read-coverage-summary $COVERAGE_SUMMARY \
    --include-rvis \
    --include-known-var \
    --effect $FUNCTIONAL_EFFECTS \
    --polyphen probably \
    --exclude-artifacts \
    --filter pass,likely,intermediate \
    --exclude-evs-qc-failed \
    --ccds-only \
    --min-coverage 10 \
    --include-qc-missing \
    --qd 5 --qual 50 --mq 40 --gq 20 --snv-sor 3 --indel-sor 10 --snv-fs
60 --indel-fs 200 --rprs -3 --mqrs -10 \
    --het-percent-alt-read 0.3-1 \
    --min-exac-vqslod-snv -2.632 \
    --min-exac-vqslod-indel 1.262 \
    --gnomad-exome-af 0.00002 \
    --gnomad-exome-rf-tp-probability-snv 0.01 \
    --gnomad-exome-rf-tp-probability-indel 0.02 \
    --gnomad-exome-pop global \
    --exac-pop global \
    --exac-af 0.00005 \
    --loo-af 0.0005 \
    --max-qc-fail-sample 0 \
    --sample $SAMPLES \
    --out $outputFolder/dominantRare/
```

Dominant Flexible #1: Polyphen Damaging

```
$ATAV \
    --collapsing-dom \
    --mann-whitney-test
    --gene-boundaries $GENE_BOUNDARIES \
    --read-coverage-summary $COVERAGE_SUMMARY \
    --include-rvis \
    --include-known-var \
    --effect $FUNCTIONAL_EFFECTS \
    --polyphen probably \
    --exclude-artifacts \
    --filter pass,likely,intermediate \
    --exclude-evs-qc-failed \
    --ccds-only \
    --min-coverage 10 \
    --include-qc-missing \
    --qd 5 --qual 50 --mq 40 --gq 20 --snv-sor 3 --indel-sor 10 --snv-fs
60 --indel-fs 200 --rprs -3 --mqrs -10 \
    --het-percent-alt-read 0.3-1 \
    --min-exac-vqslod-snv -2.632 \
    --min-exac-vqslod-indel 1.262 \
    --gnomad-exome-af 0.001 \
    --gnomad-exome-rf-tp-probability-snv 0.01 \
    --gnomad-exome-rf-tp-probability-indel 0.02 \
    --gnomad-exome-pop afr,amr,nfe,fin,eas,asj,sas \
    --exac-pop afr,amr,nfe,fin,eas,sas \
    --exac-af 0.001 \
    --loo-af 0.001 \
    --max-qc-fail-sample 2 \
    --sample $SAMPLES \
    --out $outputFolder/dominantFlexiblePolyphenDamaging/
```

Dominant Flexible #2: No Intolerance Filter

```
$ATAV \
    --collapsing-dom \
    --mann-whitney-test
    --gene-boundaries $GENE_BOUNDARIES \
    --read-coverage-summary $COVERAGE_SUMMARY \
    --include-rvis \
    --include-known-var \
    --effect $FUNCTIONAL_EFFECTS \
    --exclude-artifacts \
    --filter pass,likely,intermediate \
    --exclude-evs-qc-failed \
    --ccds-only \
    --min-coverage 10 \
    --include-qc-missing \
    --qd 5 --qual 50 --mq 40 --gq 20 --snv-sor 3 --indel-sor 10 --snv-fs
60 --indel-fs 200 --rprs -3 --mqrs -10 \
    --het-percent-alt-read 0.3-1 \
    --min-exac-vqslod-snv -2.632 \
    --min-exac-vqslod-indel 1.262 \
    --gnomad-exome-af 0.001 \
    --gnomad-exome-rf-tp-probability-snv 0.01 \
    --gnomad-exome-rf-tp-probability-indel 0.02 \
    --gnomad-exome-pop afr,amr,nfe,fin,eas,asj,sas \
    --exac-pop afr,amr,nfe,fin,eas,sas \
    --exac-af 0.001 \
    --loo-af 0.001 \
    --max-qc-fail-sample 2 \
    --sample $SAMPLES \
    --out $outputFolder/dominantFlexibleNoIntoleranceFilter/
```

Dominant PTV:

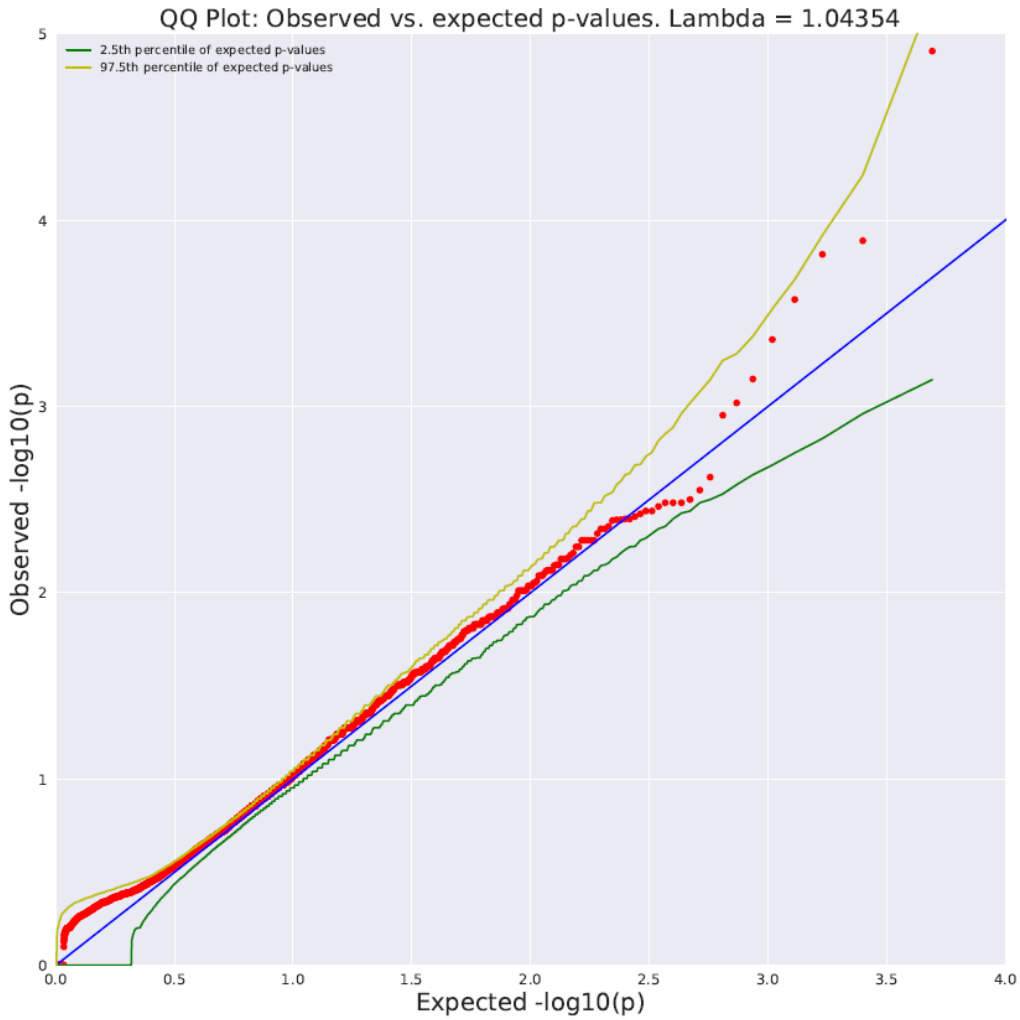
```
$ATAV \
    --collapsing-dom \
    --mann-whitney-test
    --gene-boundaries $GENE_BOUNDARIES \
    --read-coverage-summary $COVERAGE_SUMMARY \
    --include-rvis \
    --include-known-var \
    --effect $LOF_EFFECTS \
    --exclude-artifacts \
    --filter pass,likely,intermediate \
    --exclude-evs-qc-failed \
    --ccds-only \
    --min-coverage 10 \
    --include-qc-missing \
    --qd 5 --qual 50 --mq 40 --gq 20 --snv-sor 3 --indel-sor 10 --snv-fs
60 --indel-fs 200 --rprs -3 --mqrs -10 \
    --het-percent-alt-read 0.3-1 \
    --min-exac-vqslod-snv -2.632 \
    --min-exac-vqslod-indel 1.262 \
    --gnomad-exome-af 0.001 \
    --gnomad-exome-rf-tp-probability-snv 0.01 \
    --gnomad-exome-rf-tp-probability-indel 0.02 \
    --gnomad-exome-pop afr,amr,nfe,fin,eas,asj,sas \
    --exac-pop afr,amr,nfe,fin,eas,sas \
    --exac-af 0.001 \
    --loo-af 0.001 \
    --max-qc-fail-sample 2 \
    --sample $SAMPLES \
    --out $outputFolder/dominantPTV/
```

Recessive Autosomal:

```
$ATAV \
    --collapsing-comp-het \
    --mann-whitney-test
    --gene-boundaries $GENE_BOUNDARIES \
    --read-coverage-summary $COVERAGE_SUMMARY \
    --include-rvis \
    --effect $FUNCTIONAL_EFFECTS \
    --exclude-artifacts \
    --filter pass,likely,intermediate \
    --exclude-evs-qc-failed \
    --ccds-only \
    --min-coverage 10 \
    --include-qc-missing \
    --qd 5 --qual 50 --mq 40 --gq 20 --snv-sor 3 --indel-sor 10 --snv-fs
60 --indel-fs 200 --rprs -3 --mqrs -10 \
    --het-percent-alt-read 0.3-1 \
    --min-exac-vqslod-snv -2.632 \
    --min-exac-vqslod-indel 1.262 \
    --gnomad-exome-af 0.01 \
    --gnomad-exome-rf-tp-probability-snv 0.01 \
    --gnomad-exome-rf-tp-probability-indel 0.02 \
    --gnomad-exome-pop afr,amr,nfe,fin,eas,asj,sas \
    --exac-pop afr,amr,nfe,fin,eas,sas \
    --exac-af 0.01 \
    --loo-af 0.01 \
    --sample $SAMPLES \
    --out $outputFolder/recessiveAutosomal/
```

Validation

The QQ plot should not be over-/under-inflated.



An example of a clean QQ plot.

If there is inflation, possible solutions are:

- Different machines produce different QVs. (E.g. Samples sequenced on NovaSeq machines will show more QVs than those sequenced on HiSeq machines.) Only select cases/controls from similar machines.
- From the PED file from the Eigenstrat step, remove the sites (columns) that have a low genotyping rate, and then the samples (rows).

Optimizations

Collapsing requires a list of variants that fulfill the specified filters. The ATAV command collects and outputs those variants into a _genotypes.csv file, which takes a few hours.
If you realize you'd like to remove specific variants (e.g. from specific samples, or by using a new filter), you do not have to rerun the hours-long collapsing job with the new filter. You can filter that _genotypes.csv file directly, and use it as input to a new collapsing job. Simply collect all the variant IDs of interest, put them in a file (e.g. `variants.txt`), and put "`--variants variants.txt`" into the collapsing command.