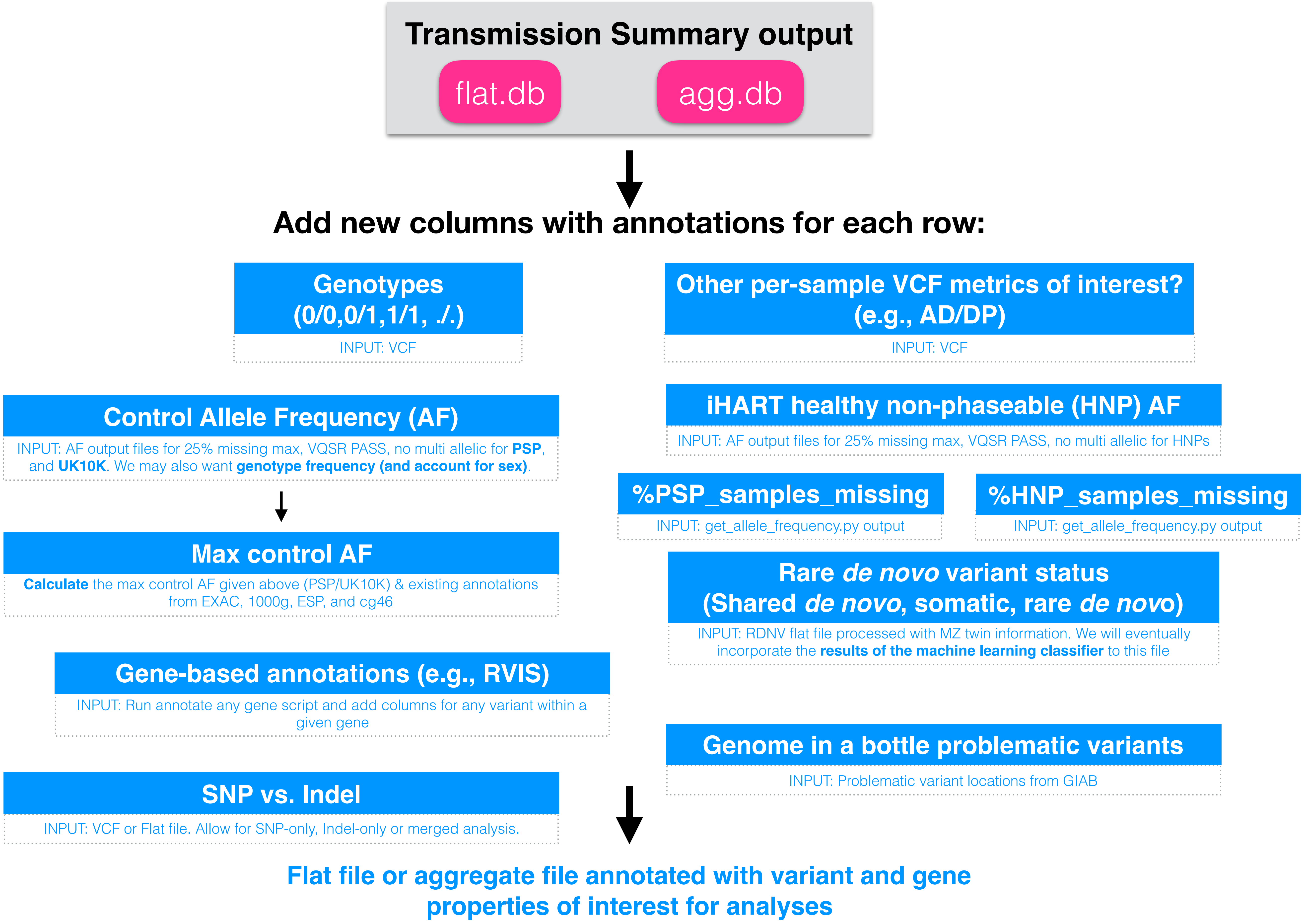
New analysis-based annotation and analysis-based filtering scripts

May 18th, 2016

Elizabeth K. Ruzzo, Laura Perez Cano, and Lee-Kai Wang

Analysis-based annotation



Transmission Summary output

flat.db

agg.db

Add new columns with annotations for each row:

Genotypes (0/0,0/1,1/1, ./.)

INPUT: VCF

Other per-sample VCF metrics of interest? (e.g., AD/DP)

INPUT: VCF

Control Allele Frequency (AF)

INPUT: AF output files for 25% missing max, VQSR PASS, no multi allelic for **PSP**, and **UK10K**. We may also want **genotype frequency (and account for sex)**.

Max control AF

Calculate the max control AF given above (PSP/UK10K) & existing annotations from EXAC, 1000g, ESP, and cg46

iHART healthy non-phaseable (HNP) AF

INPUT: AF output files for 25% missing max, VQSR PASS, no multi allelic for HNPs

%PSP_samples_missing

INPUT: get_allele_frequency.py output

%HNP_samples_missing

INPUT: get_allele_frequency.py output

Rare *de novo* variant status (Shared *de novo*, somatic, rare *de novo*)

INPUT: RDNV flat file processed with MZ twin information. We will eventually incorporate the **results of the machine learning classifier** to this file

Gene-based annotations (e.g., RVIS)

INPUT: Run annotate any gene script and add columns for any variant within a given gene

Genome in a bottle problematic variants

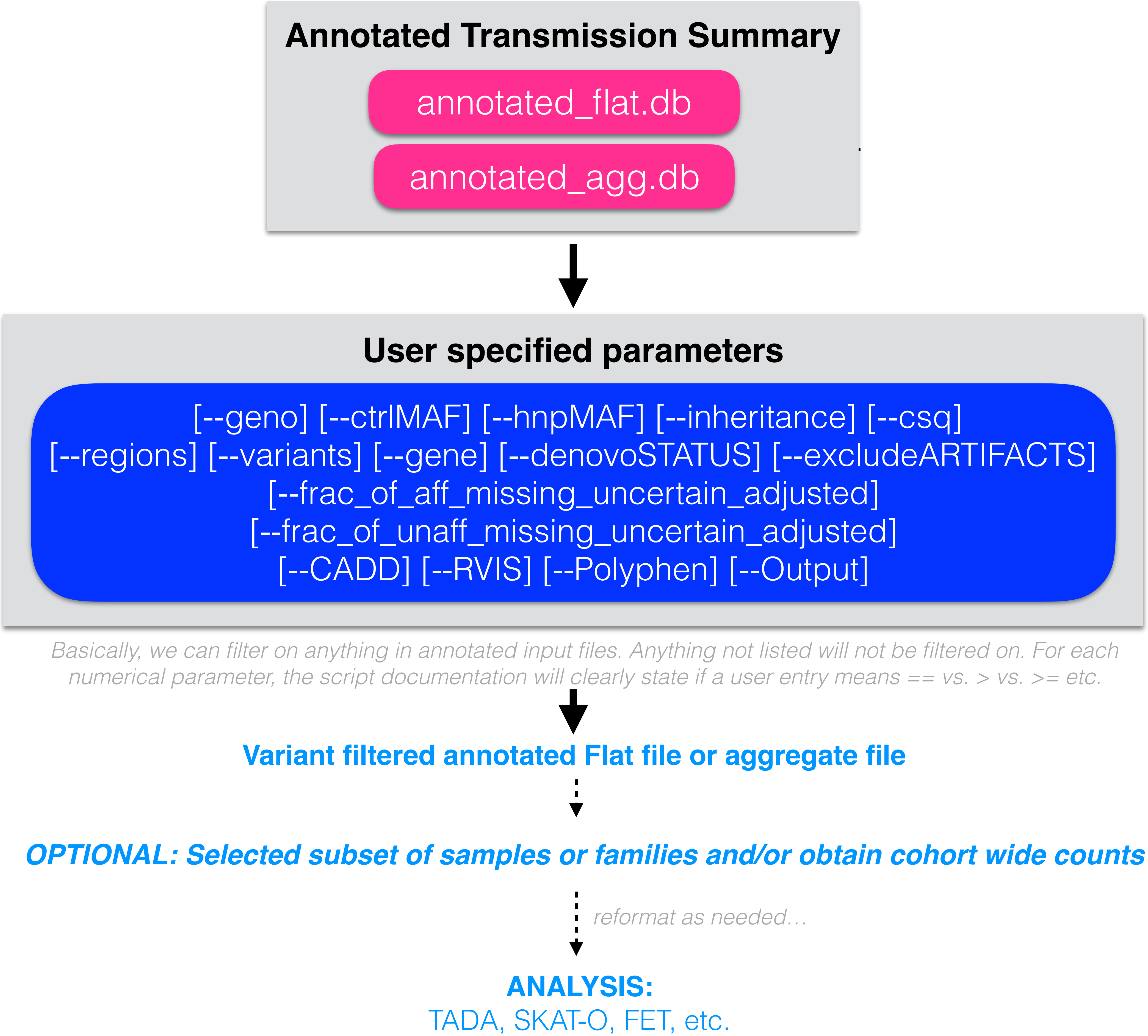
INPUT: Problematic variant locations from GIAB

SNP vs. Indel

INPUT: VCF or Flat file. Allow for SNP-only, Indel-only or merged analysis.

Flat file or aggregate file annotated with variant and gene properties of interest for analyses

Analysis-based VARIANT filtering



Annotated Transmission Summary

annotated_flat.db

annotated_agg.db

User specified parameters

[--geno] [--ctrlMAF] [--hnpMAF] [--inheritance] [--csq]
[--regions] [--variants] [--gene] [--denovoSTATUS] [--excludeARTIFACTS]
[--frac_of_aff_missing_uncertain_adjusted]
[--frac_of_unaff_missing_uncertain_adjusted]
[--CADD] [--RVIS] [--Polyphen] [--Output]

Basically, we can filter on anything in annotated input files. Anything not listed will not be filtered on. For each numerical parameter, the script documentation will clearly state if a user entry means == vs. > vs. >= etc.

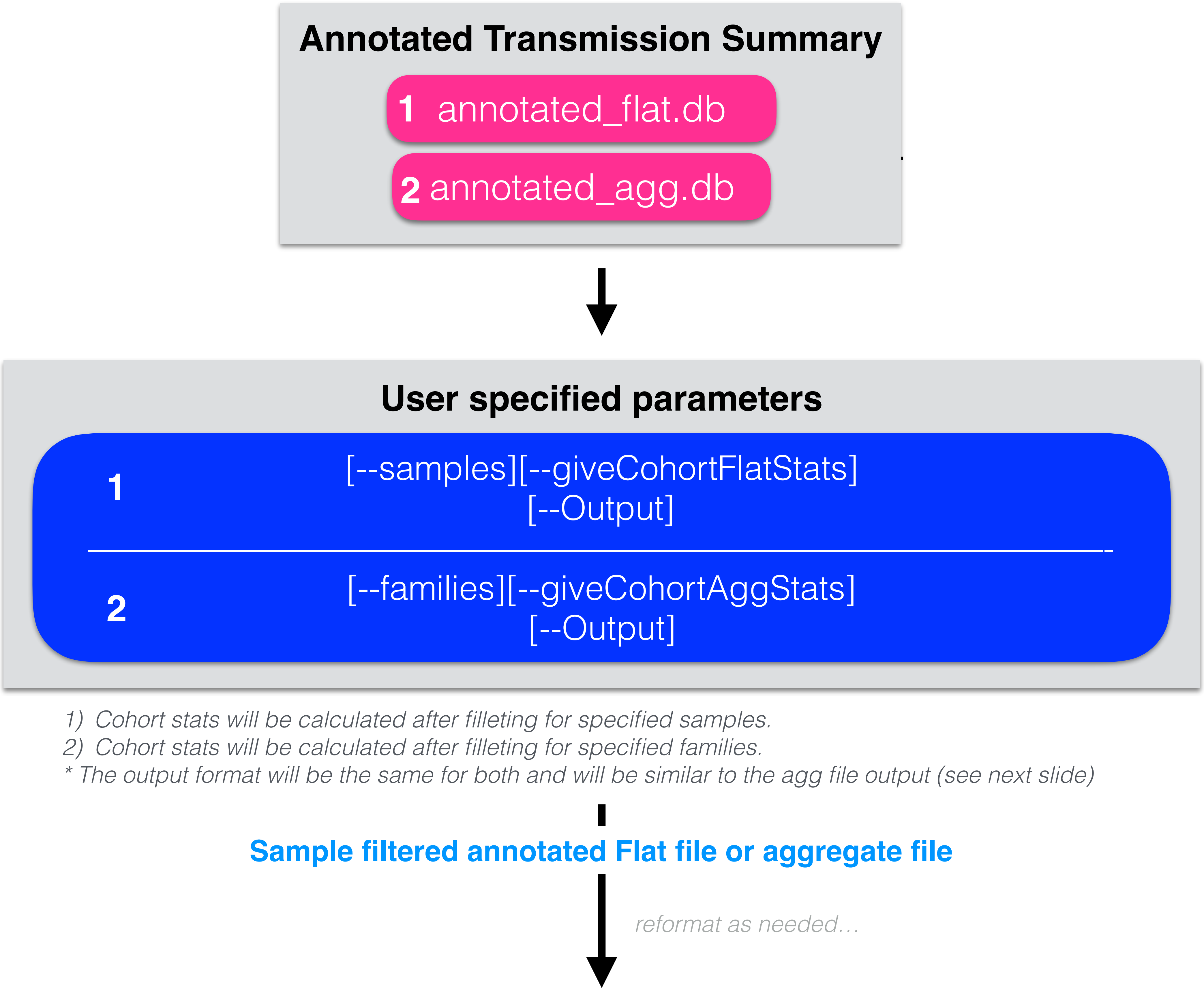
Variant filtered annotated Flat file or aggregate file

OPTIONAL: Selected subset of samples or families and/or obtain cohort wide counts

reformat as needed…

ANALYSIS:
TADA, SKAT-O, FET, etc.

Analysis-based SAMPLE filtering (Optional)



Annotated Transmission Summary

1 annotated_flat.db

2 annotated_agg.db

User specified parameters

1 [--samples][--giveCohortFlatStats]
[--Output]

2 [--families][--giveCohortAggStats]
[--Output]

1) Cohort stats will be calculated after filleting for specified samples.
2) Cohort stats will be calculated after filleting for specified families.
** The output format will be the same for both and will be similar to the agg file output (see next slide)*

Sample filtered annotated Flat file or aggregate file

reformat as needed…

ANALYSIS:
TADA, SKAT-O, FET, etc.

Filter flat file

- Input ped, filter on IID
- Adjust for missingness?
- Output will be one line per variant
- All sample annotations will be collapsed (distinct)

VCF

Chr	Position	Ref	Alt	(Parsed INFO)	inheritance-types	n_aff	n_unaff	n_carrier_aff	n_carrier_unaff	n_carrier_male_aff	n_carrier_male_unaff	n_carrier_female_aff	n_carrier_female_unaff
1	1232324	A	C	…	from_mother,from_father	422	173						
						422	173						
						422	173						
						422	173						
						422	173						

Cohort Variant Stats Output

- One line per variant (resulting from —giveCohortFlatStats or —giveCohortAggStats)
- Gives counts and fraction of affected and unaffected for each variant allowing adjustment for missingness or uncertainty
- Sample row shown below

VCF

Chr	Position	Ref	Alt	(Parsed INFO)	inheritance-type	families	fam_n	fam_n_aff	fam_n_unaff	n_missing_aff	n_missing_unaff	n_uncertain_aff	n_uncertain_unaff
1	1232324	A	C	…	from_mother, from_father	AU0965, AU0988	2	2	2	0	1	0	0
								n_aff_carriers	n_unaff_carriers	frac_of_aff	frac_of_unaff	frac_of_aff_missing_adjusted	frac_of_unaff_missing_adjusted
								2	1	1	0.5	1	1
								frac_of_aff_uncertain_adjusted	frac_of_unaff_uncertain_adjusted	frac_of_aff_missing_uncertain_adjusted	frac_of_unaff_missing_uncertain_adjusted		
								1	0.5	1	1		

*Add breakdown for each category by male/female

Comments

- We will want a log file for filtering which saves the command that was run
- It would be nice if the filtering script would automatically add the date to the output file (and maybe a few key filters like the max control allele frequency specified)
- We are still discussing certain feature ideas such as --unique which would output non-sample non-family specific columns and remove duplicate variant rows
- We are also still discussing the best way to deal with Cohort wide variables such as n_families_w_variant. One idea would be to add this as an annotation and then we can filter on them.
- Lee-Kai has already been working on a large matrix with all the gene-based annotations so this can be easily implemented for the annotation step. I am also considering adding gene-set lists like FMRP gene (0 or 1).