Burden Testing Pipeline for Rare Variant Analysis

This pipeline performs **rare variant burden testing** using gene- or region-level aggregation in case-control cohorts. It supports variant filtering, genotype matrix creation, covariate computation, and association testing using multiple statistical models. Designed to handle diverse data types, family structures, and input formats.

---

1. Variant Filtering & Preprocessing

Run `variant_filtering_main.R` to prepare:

- Filtered variants using:
  - Variant impact (missense, frameshift, splice, etc.)
  - Frequency filters (gnomAD AF)
  - Functional scores (CADD, ClinVar)
  - Segregation rules (min carriers, outgroup filters)
- Optional mapping to custom **regions** (BED-style file or TSV)
- Outputs:
  - `filtered.tsv`: filtered variant table
  - `geno_matrix.tsv`: genotype matrix (gene or region by individual)
  - `covar.tsv`: principal components for population correction

Command example:

```
Rscript variant_filtering_main.R \
  --input raw_variants.tsv \
  --output filtered.tsv \
  --format subjects \
  --geno_matrix geno_matrix.tsv \
  --covar_file covar.tsv \
  --region_map region_map.tsv
```

---

2. Burden Testing

Run `burden_testing_pipeline.R` to test for association between aggregated rare variants and phenotype.

- Supported tests:
  - SKAT, SKAT-O (variance-component tests)
  - CMC, CAST, ACAT (burden tests)
  - Mixed models for covariate or kinship correction

Outputs:

- `Burden_Results.tsv`: Gene/Region | Test | PValue | EffectSize | NumVariants | FDR | Significant
- `Summary_Stats.tsv`: summary by method, number of hits, test diagnostics

---

3. QC Visualizations

- PCA plot from `covar.tsv`: PC1 vs PC2 colored by phenotype
- QQ plot of burden test p-values
- Variant count histograms per gene/region
- Volcano plot (effect size vs –log10(p))

Example Inputs

- `geno_matrix.tsv`: genotype matrix (grouped by gene or region)
- `pheno.tsv`: sample phenotype (1=case, 0=control)
- `covar.tsv`: PCs and optional variables (sex, age)
- `region_map.tsv` (optional): VariantID, RegionID pairs for region aggregation

---

Notes

- Region-based testing is fully supported using `--region_map`.
- Matrix generation and PCA are automated during filtering.

- Filtering logic and test parameters are customizable.

For full examples, see `/examples/` folder. This pipeline is modular and extensible for rare disease and familial cancer variant analysis.