CCRaVAT (Case-Control Rare Variant Analysis Tool) & QuTie (Quantitative Trait)

Users Guide & Tutorial

Perl scripts to analyse low frequency and rare variants in case-control & quantitative trait genome-wide association studies

Written by:

AG Day-Williams
RW Lawrence
E Zeggini

Table of Contents

1) Introduction

Recent advances in high-throughput genotyping have made large-scale genetic association studies possible. Genome-wide association scans (GWAS) for complex disease have met with unprecedented success in identifying common susceptibility variants. However, the discovered common single nucleotide polymorphism (SNP) associations do not account for a large proportion of the genetic component of disease. The field is now focusing on the analysis of low frequency and rare variants (i.e. minor allele frequency (MAF) ≤0.05) to find the missing heritability in complex disease etiology (Bodmer and Bonilla, 2008; Manolio *et al*., 2009). While the sample sizes currently investigated are large enough for a well-powered GWAS of common variants, they are not large enough to provide sufficient power for the single-point analysis of rare variants with small to moderate effect sizes (Morris and Zeggini, 2009). We have developed rare variant analysis software, CCRaVAT and QuTie, which allow the large-scale analysis of low MAF polymorphisms by pooling rare variants within defined regions and treating them as a single "super-locus". This method helps identify regions that contain a significantly higher proportion of rare minor alleles in the disease cases or controls, or within groups of individuals with significantly different quantitative trait means. Collapsing multiple rare minor alleles into a single locus across pre-defined regions (either genes or sliding windows of defined sequence length) can substantially increase power for detecting association (Li and Leal, 2008; Morris and Zeggini, 2009). This approach, implemented in CCRaVAT and QuTie, can be applied to data arising from the targeted examination of specific regions or at the genome-wide scale.

2) Citing Software

All manuscripts with results from analyses performed using CCRaVAT or QuTie should cite the following reference:

Lawrence R, Day-Williams AG, Elliot KS, Morris AP, Zeggini E. CCRaVAT and QuTie - enabling analysis of rare variants in large-scale case control and quantitative trait association studies. BMC Bioinformatics 11, 527 (2010).

3) Downloading and Installing

The software is available from http://www.sanger.ac.uk/Software/rarevariant/. From the download tab of the webpage you will download a compressed tarred archive with all the software, the user's manual, and gene annotation files for genome builds 35 and 36. The file you download is named rare_variant_analysis_software.tar.gz. Execute the following commands from the Unix command line to set-up the software:

:> tar –xvzf rare_variant_analysis_software.tar.gz

:> unzip genes-b35.zip

:> unzip genes-b36.zip

All the software and gene annotation files are now ready to run rare variant association analysis. For a detailed description of the software and all the options read the user's manual.

4) Input Files

CCRaVAT and QuTie require two core input files, a pedigree file and a map file, per chromosome. An additional file, containing gene coordinates, is required if using genes rather than sliding windows to determine regions for collapsing rare variants. Below are detailed descriptions of the file formats.

PED File

The PED file is a white-space (tab or space) delimited file in standard pre-Makeped format. There are 6 mandatory columns:

Family ID

Individual ID

Paternal ID

Maternal ID

Sex (1 = Male, 2 = Female, 0 = unknown)

Phenotype

The family and individual ID should be unique per individual. Both programs require that

individuals in the input files are unrelated. The Mother ID and Father ID column entries can (should) be 0. The phenotypic data in the 6th column encode case-control status for CCRaVAT and quantitative trait values for QuTie. CCRaVAT requires controls to be coded as 1, cases coded as 2, and individuals of unknown disease status coded as -9. QuTie requires the phenotype values to be strictly numeric with the default missing value of -9. For quantitative trait values that contain a decimal point, the decimal point must be coded with a period (e.g. 1.234). QuTie allows the user to change the missing phenotype code from -9 to a user supplied code using the –miss option detailed below. Individuals with a missing phenotype code will not be analyzed.

Column 7 onwards holds the genotype data. The alleles of the markers can be represented in one of three ways: (1) A, C, G, T (2) 1, 2, 3, 4 or (3) 1, 2. The value for missing alleles is 0.

MAP file

The MAP file is a white-space (tab or space) delimited file. CCRaVAT and QuTie support two different file formats for the MAP file. Examples of the MAP file formats are shown below.

Three column MAP file:

Chr

Marker

Position (bp)

PLINK format, four-column, MAP file:

Chr

Marker

Genetic position (cM)

Physical position (bp)

The rsID or SNP identifier must be contained in the second column and the SNP base-pair position in column three or four (if using a plink format MAP file (genetic distance in column 3 and physical distance in column 4) the option "-plink" must be entered in the command line). If the MAP file has four columns of data including genetic distance (as column 3) only the SNP's physical position is used and the genetic distance is ignored.

Gene file (optional)

If you are using genes to determine analysis regions you will need a file listing the gene names and positions. The Gene file is a white-space (tab or space) delimited file that requires the first 5

columns to be:

Gene ID
Gene name
Chromosome
Gene start position (bp)
Gene stop position (bp)

If there are more than 5 columns, the software will ignore the information in the additional columns. The gene files for all 22 autosomes for both genome build 35 and build 36 are provided with the download of the software.

5) Running Analyses

Synopsis

perl Split_GW_PedMapFiles.pl [*option*]

perl CCRaVAT.pl [*options*]

perl QuTie.pl [*options*]

Split_GW_PedMapFiles

The first step in running any analysis is ensuring that CCRaVAT and QuTie run efficiently to provide you with the results in the least amount of time. Therefore, CCRaVAT and QuTie require that the analysis be run on a chromosome-by-chromosome basis. The Split_GW_PedMapFiles utility performs 3 tasks:

(1) Creates a directory for each chromosome in the data (e.g. Chr01, Chr02 etc);
(2) Splits a genome-wide ped and map file into chromosome-specific ped and map files (e.g. Chr01_file_name.map);
(3) Moves the chromosome-specific files into the appropriate directories.

The utility is a command line tool that prompts the user to input the file names of the genome-wide PED and MAP files to split. The chromosome-specific directories will be created as sub-directories of the directory from which the command is evoked.

COMMAND

perl Split_GW_PedMapFiles.pl

OPTIONS

-p The map file is in four-column PLINK format

CCRaVAT & QuTie

CCRaVAT and QuTie should be run from the same directory as the Split_GW_PedMapFile.pl utility. The programs will prompt the user to input the file names for the PED and MAP files. The options below work in both CCRaVAT and QuTie unless otherwise noted. For options that are assigned values the definition of the option uses three abbreviations for the values: *INT*, *FLOAT*, and *STR*. The *INT* abbreviation means that the option can only be assigned integer values (e.g. 1; 2; 3). The *FLOAT* abbreviation means that the option can only be assigned floating point values (e.g. 1.23;

5.678). The *STR* abbreviation means that the option can be assigned any string of characters.

OPTIONS

-help List the options available in CCRaVAT and QuTie

-plink MAP file is in four-column PLINK format

-miss=*STR* Designate the missing data value in the PED file. [Default = -9] [QuTie only]

-maf=*FLOAT* The maximum Minor Allele Frequency (MAF) to use in the collapsing algorithm [Default = 0.05; Acceptable range: 0 – 0.5]

-nchr Designates that analysis should be carried out on multiple chromosomes

-cstart=*INT* The chromosome to start analysis on [Default = 1; Acceptable range: 1- 22]

-cstop=*INT* The chromosome to stop analysis on [Default = 22; Acceptable range: 1-22]

-wind=*INT* The size of the window for sliding window analysis in kilobases (Kb) [Default = 100Kb; Acceptable range: 0 - 1000]

-gene Run the analysis based on gene regions

-ext=*INT* The number of kilobases (Kb) upstream and downstream of the gene's transcriptional start and stop positions to which analysis will be extended [Default = 50Kb; Acceptable range: 0 - 1000]

-ttest Perform a two-sample t-test to assess significance of the difference in means of individuals with and without rare variants [QuTie only]

-pout=*FLOAT* Pearson's chi squared [CCRaVAT] or Linear Regression [QuTie] p-value threshold for inclusion in the genome-wide significance output file. [Default 1 x 10^{-3}; Acceptable range: 0 – 1]

-cell=*INT* The minimum cell count in a contingency table below which a Fisher's exact test is used ***instead of*** chi squared test. [CCRaVAT only][Default = 20; Acceptable range: 0 – 100]

-pperm=*FLOAT* The p-value threshold a region must meet to initiate permutation testing. In CCRaVAT this is either the Pearson's chi squared or Fisher's exact p-value depending on the value of the –cell option. In QuTie this is based on the linear regression p-value. [Default 1 x 10^{-5}; Acceptable range: 0 – 1]

-nperm=*INT* The number of permutations to be run for each region meeting the significance threshold set by the –pperm option. The permutations are run using the method that generated the p-value for that region (i.e. Pearson's chi squared, Fisher's exact or linear regression) [Default = 100,000; Acceptable range: 1,000 – 1,000,000]

-qperm Force all permutation testing to perform the Pearson's chi squared method [CCRaVAT only]

-fout=*FLOAT* Fisher's exact p-value threshold for inclusion in the genome-wide significance output file for loci not analyzed by Pearson's chi-square. [Default = 1 x 10^{-3}; Acceptable range: 0 – 1] [CCRaVAT only]

-graph Produce a Manhattan plot of the –log10(p-value) of all regions [CCRaVAT and QuTie]. The p-value plotted in CCRaVAT is either the Person's chi-squared or Fisher's exact p-value, depending on the test performed for the locus. The p-value plotted in QuTie is the linear regression p-value. QuTie additionally produces histograms of quantitative trait values for : (1) All individuals, (2) For each significant region a single histogram of trait values comparing individuals with and without rare variant minor alleles.

-gYax=*INT* Set the height of the y-axis of the Manhattan plot. [Default = max(-log10(p-value)); Acceptable range: 1 – 50]

-gwid=*INT* Set the width of the Manhattan plot in pixels. Set the width of the mean comparison histograms [QuTie only]. [Default = 1200; Acceptable range: 100 – 10,000]

-ght=*INT* Set the height of the Manhattan plot in pixels. Set the height of the mean comparison histograms [QuTie only]. [Default = 400; Acceptable range: 50 – 2,000]

-gPnt=*INT* Set the size of the data points in the Manhattan plot in pixels. [Default = 4; Acceptable ranger: 1 – 10]

-glog=*INT* Highlight all regions with –log10(p-value) ≥*INT* in red. Produces the output file CCRaVAT_SigLOG10_gt_*INT*.txt which lists the highlighted regions. [Acceptable range: 1 – 10]

-gPval=*FLOAT* Highlight all regions with p-value ≤*FLOAT* in red. Produces the output file CCRaVAT_SigLOG10_gt_*INT*.txt, *INT* is the –log10 of the p-value, which lists the highlighted regions. [Acceptable range: 0 – 1]

-png Produce graphic output from pre-existing analysis results. This option can be combined with all the normal graphic manipulation options detailed above. If multiple chromosomes were analyzed combine with the –nchr , –cstart and –cstop options.

-gsing Run the –png option on a single output file.

EXAMPLES

- Run CCRaVAT genome-wide defining regions by genes with a MAF cut-off of 0.05

  :>perl CCRaVAT-0.1.pl –nchr –gene

- Run CCRaVAT for chromosomes 1 – 10 on windows of size 500Kb with a MAF cut-off of 0.02, a significance threshold of $1x10^{-7}$, performing Fisher's exact test for loci with a cell count below 40, running 500,000 permutations for significant regions with a threshold of $1x10^{-7}$, and producing a Manhattan plot

  :>perl CCRaVAT-0.1.pl –nchr –cstart=1 –cstop=10 –wind=500 –maf=0.02 –cell=40 –pout=0.0000001 –pperm=0.0000001 –nperm=500000 –graph

- Run QuTie genome-wide defining regions by genes with a MAF cut-off of 0.05

  :>perl QuTie-0.1.pl –nchr –gene

- Run QuTie for chromosomes 1-10 on windows of size 500Kb with a MAF cut-off of 0.02, a significance threshold of $1x10^{-7}$, performing a t-test, running 500,000 permutations for significant regions with a threshold of $1x10^{-7}$, and producing a Manhattan plot and histograms

  :>perl QuTie-0.1.pl –nchr –cstart=1 –cstop=10 –wind=500 –maf=0.02 –ttest –pout=0.0000001 –pperm=0.0000001 –nperm=500000 –graph

6) Output

CCRaVAT and QuTie produce 5 types of output:

(1) Summary of all significant regions genome-wide;

(2) Summary of all permutation results genome-wide;

(3) Chromosome-specific summary of all analyzed regions;

(4) A list of SNPs residing within each significant region;

(5) Graphic summary of all results[CCRaVAT and QuTie], and significant regions [QuTie only].

The first 4 types of output are contained in tab-delimited text files. The graphic outputs are PNG graphics. The following sections describe each type of output in more detail.

Summary of all significant regions genome-wide

The genome-wide summary file details the regions that have a p-value ≤ the significance threshold set with the -pout and -fout options. The summary file name is based on the chromosomal range analyzed, the type of analysis conducted (e.g. gene-based or sliding window), and the MAF cut-off used to define rare variants, and includes the wording 'WG_summary'. For example, for a genome-wide gene-based analysis with a MAF cut-off of 0.05, the summary output file names are:

CCRaVAT: chr1-22.WG_summary_CCRVgene_MAF0.05.txt

QuTie: chr1-22.WG_summary_QTRVgene_MAF0.05.txt.

For the same analysis run using a window of size 100 Kb, the file names are:

CCRaVAT: chr1-22.WG_summary_CCRVwin100kb_MAF0.05.txt

QuTie: chr1-22.WG_summary_QTRVwin100kb_MAF0.05.txt.

The summary file for CCRaVAT has 12 columns:

1) Region analysed;
2) Chromosome;
3) Region's start base-pair position;
4) Region's stop base-pair position;
5) Numbers of SNPs within the region. The first number represents the total number of SNPs and the second number represents the number of SNPs with MAF below the threshold set by the -maf option;
6) Number of cases with at least one rare variant minor allele;
7) Number of cases with no rare variant minor alleles;
8) Number of controls with at least one rare variant minor allele;

9) Number of controls with no rare variant minor alleles;
10) Pearson's chi-squared statistic;
11) P-value for statistic in column 10;
12) Fisher's exact p-value (optional). Any regions showing "No<20" means that the minimum cell count was higher than the threshold set, in this case 20.

The summary file for QuTie has 15 columns:

1) Region analysed;
2) Chromosome;
3) Region's start base-pair position;
4) Region's stop base-pair position;
5) Number of SNPs within the region. The first number represents the total number of SNPs and the second number represents the number of SNPs with MAF below the threshold set by the -maf option;
6) Number of individuals with a QT value with at least one rare variant minor allele;
7) Number of individuals with a QT value with no rare variant minor alleles;
8) The mean quantitative trait value for individuals with at least one rare variant minor allele;
9) The mean quantitative trait value for individuals with no rare variant minor alleles;
10) Linear regression p-value;
11) Linear regression beta coefficient;
12) Linear regression standard error;
13) The 95% confidence intervals;
14) The t-test statistic;
15) The t-test p-value.

Summary of all permutation results genome-wide

This summary file lists all regions that passed the significance threshold for running permutations. The file name contains the chromosome(s) analysed, followed by the wording '.GenWide_Signif_Pvals_Perms_MAF', followed by the MAF cut-off used, and the analysis run (e.g. gene-based or sliding window). For example, for a genome-wide, gene-based analysis with a MAF threshold of 0.05, the permutation summary output file name would be: chr1-22.GenWide_Signif_Pvals_Perms_MAF0.05gene.txt.

The permutation summary files for CCRaVAT and QuTie are formatted in the same way, and contain 8 columns:

1) Region Analyzed;
2) Chromosome;
3) Region's start base-pair position;
4) Region's stop base-pair position;
5) Number of cases and controls with and without rare variant minor alleles, or the number of individuals with QT values with and without rare variant minor alleles;
6) P-value from original analysis;
7) Number of permutations with p-value ≤ p-value in column 6, the total number of permutations run, and the permutation p-value.

Chromosome-specific summary of all analyzed regions

CCRaVAT and QuTie produce a chromosome-specific results summary file that includes all the analyzed regions, not just the significant regions. The CCRaVAT files have the following naming convention: Chr[#]_[PED file name]_[CCRVwin**kb/CCRVgene]_MAF[*].txt. If a window base analysis was run the "CCRVwin" is used and if a gene base analysis was run the "CCRVgene" is used. The files for QuTie have the same naming convention except it replaces "CCRV" with "QTRV" in the file name. The first line displays the command line used. The rest of the file is identical to the format of the summary file for significant regions for the different programs. Those regions where no rare variant minor alleles are present in the individuals studied are given a p-value of 1. In addition, the permutation results are listed after column 12 for CCRaVAT, and column 15 for QuTie.

List of SNPs residing within each significant region

For each region that produced a significant p-value, the programs produce a summary file with information for each SNP in the region. This file contains 4 columns:

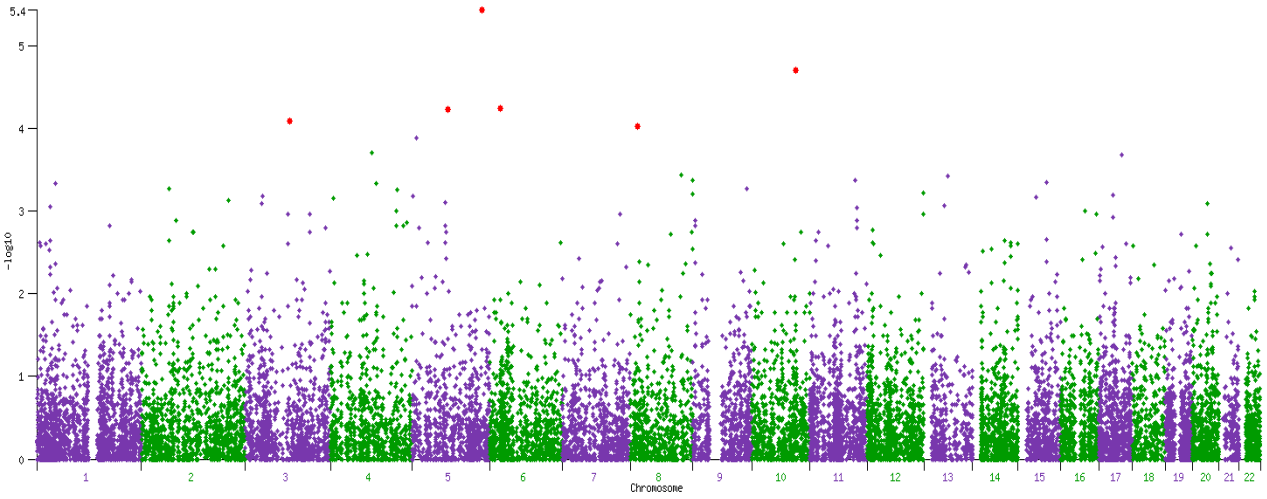
1) Marker name;
2) Chromosome;
3) Base-pair location;
4) MAF in all individuals.

The files have the following naming convention: RegionName_RegionStartPosition-RegionStopPosition-chr#_[CCRV/QTRV]_[gene/win*kb]_MAF*_SNPs.txt.

Graphic Output

Both CCRaVAT and QuTie produce a Manhattan plot (Figure 1) to allow users to easily visualize the genome-wide results of the analysis. Additionally, QuTie produces a summary histogram of the quantitative trait values of all individuals (Figure 2) and a histogram for each significant region comparing the distribution of the quantitative trait for individuals with and without rare variant minor alleles (Figure 3). All the graphic output files are PNG graphics. The naming convention for the Manhattan plot is the same as for the genome-wide summary file described above, simply with the .txt extension replaced by the .png extension. The summary histogram of all individuals follows the same naming convention as the summary file with the .txt extension replaced by _histogram.png. The histogram files for each significant region follow the same naming convention as the SNP list for the region, except that ' _SNPs.txt' is replaced with '_hist.png'.

Figure. 1. Genome-wide Manhattan plot for -log10 transformed p-values produced by CCRaVAT and QuTie. Regions with -log10(p) ≥ 4 are highlighted with larger red points.



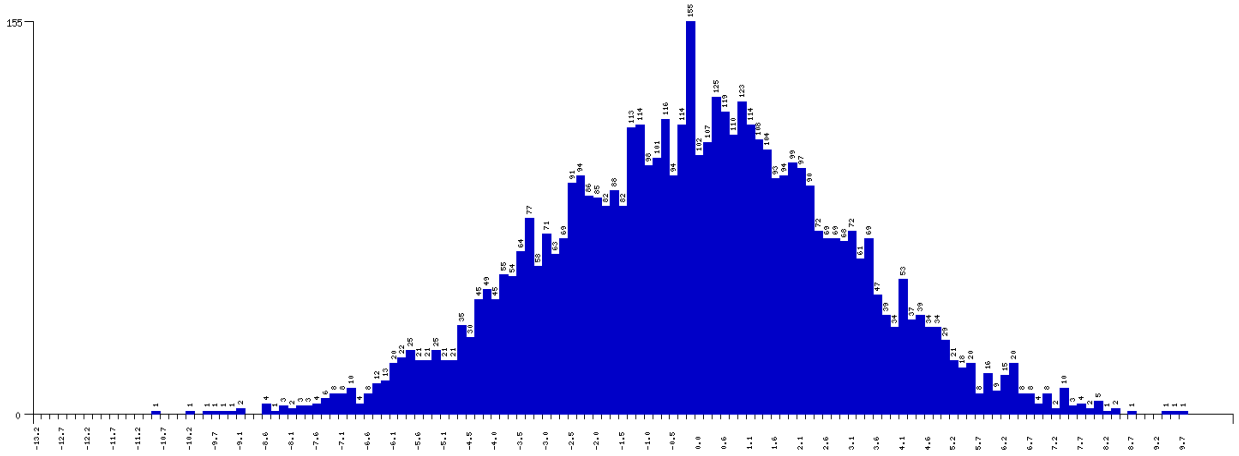


Figure. 2. Histogram displaying the distribution of QT values from all individuals analysed in QuTie.

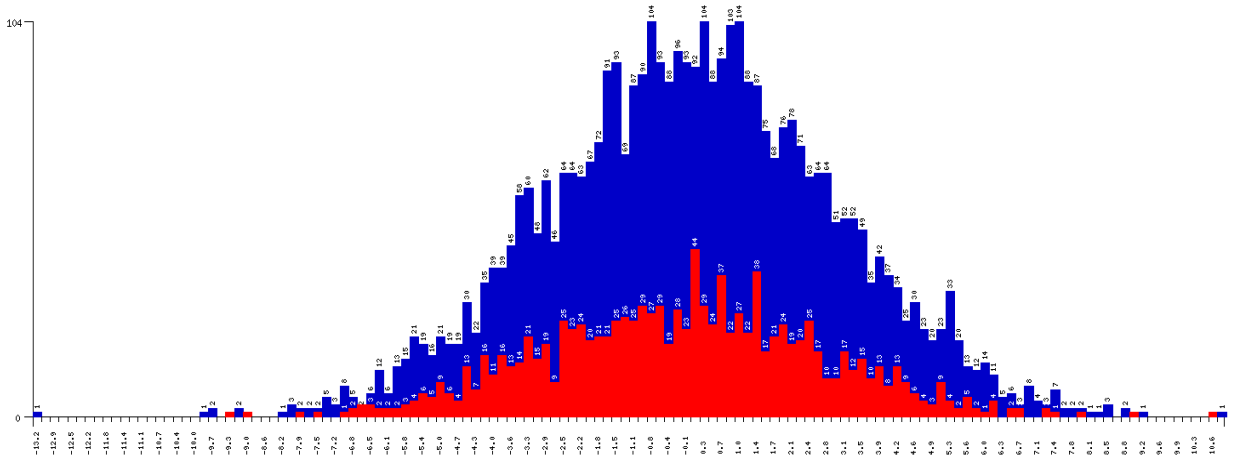


Figure. 3. Histogram displaying the distribution of QT values for individuals that either have (red) or do not have (blue) at least one rare variant minor allele within a defined region that passed the set threshold for significance in QuTie.

7) Methods and Implementation

The statistical properties of the rare variant super-locus or collapsing method that we have implemented have been described previously (Li and Leal, 2008; Morris and Zeggini, 2009). The first step in implementing this approach involves the definition of regions in which rare minor alleles are collapsed. These chromosomal regions can either be sliding windows of predefined length or genic regions defined by intervals either side of the transcriptional start and stop sites of genes. CCRaVAT and QuTie use the same method up to this point. The programs differ in the study designs analyzed and statistical techniques used. CCRaVAT analyzes case-control data and constructs a 2x2 contingency table of the presence or absence of rare variant minor alleles in cases and controls for each region. Differences in the proportion of cases and controls carrying rare minor alleles are tested using a Pearson's chi-squared test or a Fisher's exact test when cell counts are small. CCRaVAT also allows users to generate empirical p values by permuting case-control status a predefined number of times and repeating the analysis for each replicate. QuTie implements the analysis of quantitative traits and analyzes the differences in quantitative trait means for individuals carrying at least one rare variant minor allele and individuals carrying no rare variant minor alleles within the defined region. The quantitative trait values in the two groups are compared using linear regression and a Student's t-test. Figure 1 provides an overview of the analytical approach implemented in CCRaVAT and QuTie.

CCRaVAT and QuTie are command-line based utilities written in Perl. CCRaVAT and QuTie have been tested on a variety of GWAS datasets and the system requirements depend mainly on the size of the study (i.e. number of SNPs and individuals genotyped). CCRaVAT and QuTie require that the data be separated by chromosome for efficiency. For a genome-wide dataset separated by chromosome consisting of 450,000 SNPs typed in 5,000 individuals, CCRaVAT requires ~200Mb of RAM. The software development and testing of the applications was performed on machines with dual-core Athlon processors. The scripts can take a variable amount of time to run depending

on the options used. The run time for a typical gene-centric genome-wide analysis, using approximately 450,000 SNPs and 5,000 individuals separated by chromosome, is less than 24 hours (without permutations). Permutation testing can add considerably to the computing time depending on the number of regions analyzed and the numbers of permutations run.

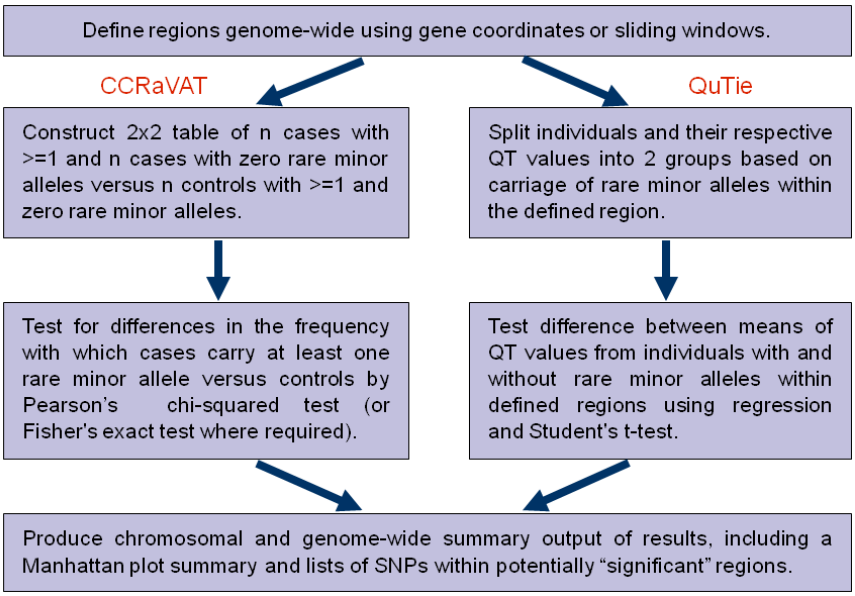


Fig. 4. Flowchart summarizing the rare variant analysis method implemented in CCRaVAT and QuTie.

8) Known Issues

- When running the genome-wide/multiple chromosomes option “-nchr”, the PED, MAP and gene information files must be in separate directories named “Chr01”, “Chr02”, …, Chr22. There should be only one PED file, one MAP file and one gene information file per directory. Multiple PED or MAP files per directory will cause errors.

- If running a gene-based analysis and not using the gene information files provided, make sure the files used have ‘chr’ (not case sensitive) in the filenames followed somewhere in the filename by ‘gene’ (not case sensitive). The script will otherwise not find the files and error messages will be produced.

- Re-running an analysis while there are results files from a previous analysis in the chromosome directories has given users problems in the past. Therefore we suggest that results are moved to a separate directory after each run. This will prevent results files being overwritten and any potential errors depending on how the input and output files are named.

- Do not cut and paste commands from Windows-based text editors (e.g. Word, WordPad) into a Unix terminal to run the programs. This will cause the programs to incorrectly parse the command line arguments because the two operating systems encode text differently.

9) Reporting bugs

CCRaVAT and QuTie have been extensively tested using genome-wide data.

If you have any problems running CCRaVAT or QuTie then please contact either: Robert Lawrence at: rlawrence@meddent.uwa.edu.au or Eleftheria Zeggini at: Eleftheria@sanger.ac.uk.

It is recommended to contact the author of CCRaVAT and QuTie (Robert Lawrence) regarding bugs or problems running the script.

10) References

Abecasis, GR *et al.* (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.*, **30**, 97-101.

Bodmer, W and Bonilla, C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet*., **40**, 695-71.

Li,B and Leal,SM. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83(3)**, 311-21 .

Manolio, TA *et al.*(2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747-753.

Morris,A.P. and Zeggini, E. (2009) An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol*, In press.

Purcell, S. *et al*. (2007) PLINK: a toolset for whole genome association and population-based linkage analysis. *Am. J. Hum Genet.*, **81**, 559-75.