Statistical methods in GWATCH

Different tests for analysis of associations by different types of data are implemented into GWATCH. User having clinical and genotype data for one or several populations can select appropriate sample of tests for screening statistical associations of infection or disease progression with genotypes. Results of all selected statistical tests are subject to be visualized simultaneously on the Highway. Most of statistical tests are executed using R-project. A special option allows perform detailed statistical analysis (GWAS Trax report) for any selected SNP.

A.1. Statistical data

Data acceptable for the analysis consist of clinical data and genotype data. The genotype data contain information on all SNP to be analyzed and the corresponding genotype consists of two dummy (binary) variables specifying corresponding forms of two alleles or in one three levels categorical variable with whole genotype information associated with any individual. In the first case 1 is corresponds to minor SNP allele and 0 is corresponds to common SNP allele. In the second case common SNP homozygote is coded by -1, minor SNP homozygote is coded by 1 and heterozygote is coded by 0. The required SNP information is SNP-code and the corresponding coordinate. The input genotype data are expected to be sorted by ID and by SNP coordinate sequentially. For further analysis all individuals will be classified by genotype in different ways.

Genotype classification is used as an explanatory factor for all statistical tests. Four types of genotype classification are used: dominant (DOM) classification separates common homozygote from all others in two different groups; recessive (REC) classification separates minor homozygote from all others (common homozygote) and codominant (CDM) classification separates all individuals in three groups by their genotype; under allelic classification (ALC) two SNP alleles corresponding to any single individual are assumed as different observations with the same clinical data.

Two types of clinical data are acceptable for analysis: categorical and right-censored survival.

Categorical data consist of the ID variable and numeric categorical variable having two or more levels specifying disease status. In the first case it is recommended to correspond 1 to individuals who get symptoms of disease (or infection) and 0 otherwise. In the second case of more then two levels and ordinal categories it is recommended to use 0 for non symptoms disease (or infection) shown and to choose positive numbers corresponding to other levels in the same order as the original categories.

Right censored survival data contain information on exact time from baseline date (preferentially in days) to an event and type of the event is given by binary variable: 1 is corresponding to failure and 0 is corresponding to censoring, for any individual. For competing risks model it is possible to use several positive levels for different types of failures.

A.2 Statistical tools for associations

The GWATCH allows analyze associations of disease (or infection) features with genotype for all available SNP's. Four tests corresponding to different genotype classifications are produced for any clinical data by selected testing method. Stratified analysis is available if the input clinical data contains classification variable. In this case any selected group of individuals is analyzed separately, results of these tests are displayed in different lines at the highway.

Categorical tests (CT) are used for categorical statistical data analysis organized as m×k contingency table. The categorical data are required to perform categorical tests. Fisher's exact test (**R**-function "fisher.test()") for **2**×**2** contingency tables and its generalization in R are applied to produce P-value. The odds ratio for **2**×**2** contingency tables or the transformation $(1+\rho)/(1-\rho)$ of Pearson's correlation coefficient otherwise defines direction of the association and, therefore, color of the corresponding bar on the Highway.

Proportional hazards survival tests (PHST) are used for right-censored survival data analysis. The right-censored survival data are required to perform PHST. Cox proportional hazards model is used to produce *P*-value (**R**-function coxph(), package *survival*). The obtained hazard ratio in case of binary genotype classifications (ALC, DOM, REC) and exponentiated slope of Cox's regression line under CDM genotype classification define direction of the associations.

Categorical tests for survival data (*PCT*) are used to identify significant difference between groups of individuals grouped by survival data. The right-censored survival data are required to perform categorical survival tests. Baseline null hypothesis is formulated in terms of identity of cumulative distribution functions corresponding to different groups of individuals. Individuals involved into analysis are classified by observed failure or censoring times according to specified rules. All individuals having observed failure times are classified into k groups by the failure times, defined by k-1 breakpoints. All individuals censored before the maximal breakpoint are removed from the analysis.

Warning: PCT is not applicable for testing categorical null hypothesis formulated in terms of interval probabilities for failure times like it is in classical categorical analysis with continuous response variable. Target null hypothesis corresponding to PCT involves censoring and lag between infection time and starting follow up as well as rules of classification. On the other hand, under mild conditions on experimental design the baseline null hypothesis implies the target null hypothesis and, therefore, rejection of the target null hypothesis imply rejection of the baseline null hypothesis.

Hardy-Weinberg equilibrium (HWE) tests are used to evaluate significant deviation from Hardy–Weinberg equilibrium that is an indicator of genotyping errors. Haldan's exact test on Hardy–Weinberg equilibrium used to produce *P*-values. Sign of Hardy–Weinberg disequilibrium statistic is applied to specify direction of the disequilibrium. The **R**-function HWExact() of package *HardyWeingerg* is used to perform HWE test.

Other tests for longitudinal data analysis are ready to be installed.

A.3 Reports

After screening for associations of clinical traits and genotypes one may be interested closer review of certain SNPs. The report tool allows producing reports on extended statistical analysis for any single SNP. Important genotype information is given in the header on the front page: SNP identifier, SNP coordinate, chromosome, alleles and their frequencies. Header also contains information on populations involved into analysis. Besides the header, front page also lists summary for all tests with *P*-values and statistics on association (*SA*)-values classified for all tests in bar plot form. Following pages of report contain detailed information: contingency tables are produced in the form of corresponding bar plots for any categorical test (including progression categorical tests) and Kaplan–Meier survival curves are reported for all three genotypes for all survival tests.

A.4 Statistical tools for whole genome analysis

Several statistical tools addressing SNP compositions are available.

Polarization tool allows to inverse test results for minor and common SNP-alleles around some fixed SNP for better approximation of true associations. Polarization table is produced using linkage disequilibrium coefficients between closed SNP's. Linkage disequilibrium coefficients are calculated for 80 SNP's upstream of any fixed SNP and for 80 SNP downstream of the fixed SNP. In case of sufficiently large positive value of linkage disequilibrium with the fixed SNP the polarization mark is assigned to 1 and in case of sufficiently large negative linkage disequilibrium the polarization mark is assigned to -1. Otherwise, in case of sufficiently small linkage disequilibrium the polarization mark is assigned to 0. In the process of polarization around fixed SNP common and minor alleles of close SNP's are inverted if the polarization mark is -1 that imply inversion of direction of association with the disease of the SNP's with the polarization mark -1.

Manhattan plots of minus log *P*-values are producing for any single test.

Significant regions or regions of concentration of small *P*-values are selected by using kernel smoothing of minus log *P*-value

$$f_i = \sum_j (-\log p_j) * \exp((i-j)^2 / (2\sigma^2)).$$

Under larger σ role of any single *P*-value is smaller whereas under smaller σ the any single *P*-value plays important role in calculation the corresponding value of density. Moreover, density curve is smoother under larger σ and, therefore, larger σ allows identify wider regions with small and moderate *P*-values whereas smaller σ allows to identify peaks of small *P*-values. Densities obtained for single tests separately are averaged by all selected tests pointwise. Regions of concentration of small *P*-values are selected as sets of naighbour SNP's with corresponding density values over fixed threshold extended by σ SNP's to the right and σ SNP's to the left.

Bonferroni correction tool allows controlling for family-wise error rate by using Holm–Bonferroni method.* (work on this tool is in progress).

FDR correction tool allows controlling for family-wise error rate.* (work on this tool is in progress).

Table 1 Tests on associations and their capabilities

Test	Data type	Model	Genotype categorization				Phenotype		Test(s)	Data / Model
			ALC	CDM	DOM	REC	Class.	Covariate		
CT	Categorical, 2 categories	Multinomial, 2×2 table	YES	NO	YES	YES	YES	NO	Fisher's exact	Categorical/ multinomial
CT	Categorical	Multinomial, n × k-table	YES	YES	YES	YES	YES	NO	Chi-square, permutation chi-square	Categorical/ multinomial
PHST	Right-censored survival	Cox's regression	YES	YES	YES	YES	YES	YES	Likelihood ratio	Survival / Cox's proportional hazards
PCT	Right-censored Survival, 2-categories	Multinomial, with biased parameter	YES	NO	YES	YES	YES	NO	Fisher's exact	Survival transformed to categorical / multinomial
PCT	Right-censored Survival	Multinomial, with biased parameter	YES	YES	YES	YES	YES	NO	Chi-square, permutation chi-square	Survival transformed to categorical / multinomial