Genome-Wide Association Study

Genome-Wide Association Study (GWAS) is a hypothesis-free method for identifying associations between genetic regions and traits. GWAS analysis are usually used to identify genes involved in human disease. By applying GWAS analysis to variant data we will be able to identify a given variant (or a set of variants) involved in a given phenotype or disorder.Based on a statistical test, GWAS analysis will provide a level of significance (or p-value) for each variant. OpenCGA implements GWAS analysis based on the statistical tests: chi-square and Fisher.

Useful Links

- GWAS Study

Implementation

OpenCGA GWAS analysis extends Oskar GWAS implementation. GWAS is implemented using Hadoop MapReduce over HBase.

Input

Parameters

OpenCGA support different input parameters:

- Variant data with sample genotypes
- Two list of samples (case-control study)
- Statistical test: chi square or fisher.

Instead of providing two lists of samples, users can provide:

- A phenotype, and
- A pedigree for those variant samples

Configuration

OpenCGA implementation supports different configuration variables. This can be setup in the OpenCGA installation folder or specified during execution.

Output

Files

GWAS analysis result includes a text file with score and plot image (WIP)

Text file that consists of a header line (starting with #), and then one line per variant with the following 12-13 columns:

chromosome	Chromosome code
start	Start base-pair coordinate
end	End base-pair coordinate
strand	Strand
reference	Reference allele
alternate	Alternate allele
dbSNP	Variant identifier
gene	Gene name
biotype	Bioytpe
conseq. types	List of consequence types
chi square	Allelic test chi-square statistic. ***Not present with 'fisher' test***.
p-value	Allelic test p-value
odd ratio	Odd ratio: odds(allele 1 | case) / odds(allele 1 | control)

Next, it shows the first lines of a result file after executing a GWAS analysis using the chi square test:

```
#chromosome        start        end        strand        reference
alternate        dbSNP        gene        biotype        conseq.
types        chi square        p-value        odd ratio
22        16054454        16054454        +        C        T
rs373998521                        intergenic_variant
2.4727272727272727        0.11583677431831574        0.0
22        16065809        16065809        +        T
C                ENSG00000233866        lincRNA
downstream_gene_variant        0.053968253968253915
0.8162967146689325        0.8
22        16065809        16065809        +        T
C                        regulatory_region_variant
0.053968253968253915        0.8162967146689325        0.8
22        16077310        16077310        +        T
A                ENSG00000229286        unprocessed_pseudogene
2KB_upstream_variant        0.9714285714285711
0.3243241555798487        3.0
22        16077310        16077310        +        T
A                        regulatory_region_variant
0.9714285714285711        0.3243241555798487        3.0
22        16080499        16080499        +        A        G
rs200119791        ENSG00000229286        unprocessed_pseudogene
upstream_gene_variant        1.8888888888888886
0.16932729721206297        Infinity
22        16080499        16080499        +        A        G
rs200119791        ENSG00000235265        unprocessed_pseudogene
downstream_gene_variant        1.8888888888888886
0.16932729721206297        Infinity
22        16084621        16084621        +        T
C                ENSG00000235265        unprocessed_pseudogene
non_coding_transcript_exon_variant,non_coding_transcript_variant
2.4425287356321843        0.11808572685033702        Infinity
```

Index

Score can be indexed.

Useful Links

- https://en.wikipedia.org/wiki/Genetic_association
- https://en.wikipedia.org/wiki/Genome-wide_association_study
- https://www.cog-genomics.org/plink/1.9/assoc