DASC6310 - Data Analysis in Biology and Life Science

Chapter 4: Genome Wide Association Studies

Yan Yan
yyan@tru.ca
Notes written by Yan Yan

Department of Computing Science
Thompson Rivers University

Section 1

What is Genome Wide Association Studies (GWAS)?

Why GWAS?

GWAS Analysis Overview

Data for GWAS

Running GWAS
- PLINK
- TASSEL Pipeline

Genetic Studies

- How do we know the genetic mechanisms of a trait? What causes it?
- Genomes drive the trait –study the relationships between genomes and trait
  - One major objective in genetic studies, e.g. human diseases

Genetic Studies

- How do we know the genetic mechanisms of a trait? What causes it?
- Genomes drive the trait –study the relationships between genomes and trait
  - One major objective in genetic studies, e.g. human diseases

Genetic Studies

Basic principle, if

- a region of the genome is associated with a trait and
- this genome region contains variations (e.g. coding differences) across different individuals impacting the phenotype

We can expect

- trait distribution of all those individuals having one particular form of that genomic variation will differ from those individuals have another.
- we can build statistical models to examine the statistical associations between the trait and genetic variations

Discussion: Is it only the genomes?

Genetic Studies

Basic principle, if

- a region of the genome is associated with a trait and
- this genome region contains variations (e.g. coding differences) across different individuals impacting the phenotype

We can expect

- trait distribution of all those individuals having one particular form of that genomic variation will differ from those individuals have another.
- we can build statistical models to examine the statistical associations between the trait and genetic variations

Discussion: Is it only the genomes?

Genetic Studies

Basic principle, if

- a region of the genome is associated with a trait and
- this genome region contains variations (e.g. coding differences) across different individuals impacting the phenotype

We can expect

- trait distribution of all those individuals having one particular form of that genomic variation will differ from those individuals have another.
- we can build statistical models to examine the statistical associations between the trait and genetic variations

Discussion: Is it only the genomes?

What is Genome Wide Association Studies (GWAS)?

GWAS is

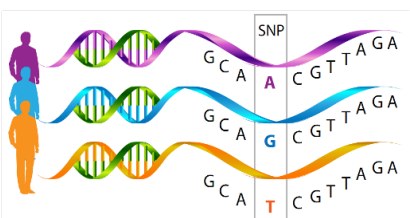
- genetic association studies
- between traits and genetic variations
  - what is a trait?
  - at the **whole genome** level

What is Genome Wide Association Studies (GWAS)?

In GWAS

- Genotype – genome, genetic variations, typically Single nucleotide polymorphism (SNP).
- Phenotype – traits, e.g. disease

Variations on genome is also called allele. A SNP typically has two alleles, major and minor. SNPs is also referred as markers.



	SNP1	SNP2	SNP3
individual 1	G	T	C
individual 2	A	C	C
individual 3	A	T	T

Table: An example of genotype data

What is Genome Wide Association Studies (GWAS)?

- SNP occurs $>1\%$ on a population.
- Major allele is also typically the reference.

Discussion:

- What is reference genome of a species? Do we have a it for all species?
- What's the difference between SNP and Mutation?

Note: Mutation includes deletion, insertion, duplications and substitution on DNA.

What is Genome Wide Association Studies (GWAS)?

- SNP occurs $>1\%$ on a population.
- Major allele is also typically the reference.

Discussion:

- What is reference genome of a species? Do we have a it for all species?
- What's the difference between SNP and Mutation?

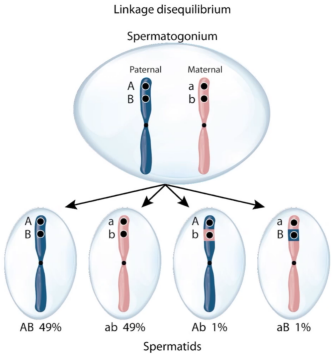
Note: Mutation includes deletion, insertion, duplications and substitution on DNA.

Linkage Disequilibrium (LD)

- Linkage: Two alleles located on the same chromosome
- Recombination of different segments on the same chromosome
- Linkage equilibrium: for frequency (F). we have F(AB)=F(A)*F(B)

Linkage Disequilibrium (LD)

- Non-random association of alleles at two or more loci in a general population.
- Usually seen by two genes located closely on one chromosome



From https://www.youtube.com/watch?v=DvrAuMyu4wU

How do we get the data?

- Phenotype – measure or record different traits
- Genotype – genotyping platforms (more on additional resources)
  - Illumina
  - Affymetrix
  - New trend – nanopore sequencing

Discussion: Scale of the data. Can you guess how many SNPs and how many samples typically we can get?

More on SNP genotyping see the "How SNP-typing works" Video (from Useful Genetics)
https://www.youtube.com/watch?v=NaonaIy_I2U&list=PLgh8WcYegg4fIyVjDUTw5oW2enNjBVtHF&index=9

How do we get the data?

- Phenotype – measure or record different traits
- Genotype – genotyping platforms (more on additional resources)
  - Illumina
  - Affymetrix
  - New trend – nanopore sequencing

Discussion: Scale of the data. Can you guess how many SNPs and how many samples typically we can get?

More on SNP genotyping see the "How SNP-typing works" Video (from Useful Genetics)
https://www.youtube.com/watch?v=Naona1y_I2U&list=PLgh8WcYegg47IyVjDUTw5oW2enMjBVtHF&index=9

Problems

Discussion: do you see any potential problems of the data?

Section 2

What is Genome Wide Association Studies (GWAS)?

Why GWAS?

GWAS Analysis Overview

Data for GWAS

Running GWAS
- PLINK
- TASSEL Pipeline

What we get from GWAS?

- Single SNP may have small to moderate effects, GWAS studies common variations and their associations to traits or disease.
- The significant SNPs can help locate the genes that are contributing to phenotype variations.
- It may not be the SNPs itself.

Discussions: Why not the SNPs itself?

What we get from GWAS?

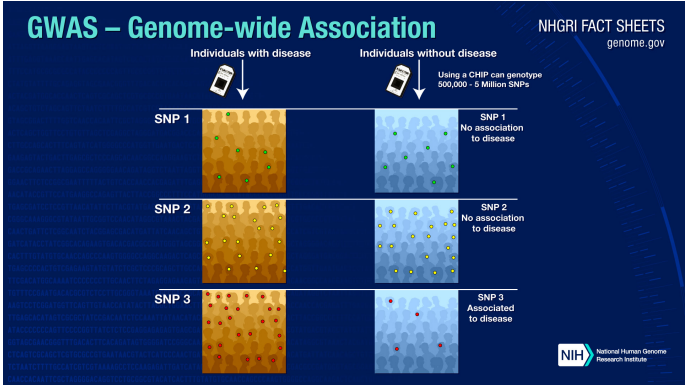
- Single SNP may have small to moderate effects, GWAS studies common variations and their associations to traits or disease.
- The significant SNPs can help locate the genes that are contributing to phenotype variations.
- It may not be the SNPs itself.

Discussions: Why not the SNPs itself?

A GWAS study example

In 2005, researcher found a common form of blindness is associated with variation in the gene for complement factor H, which produces a protein involved in regulating inflammation.

Other examples include studies finding risks of type 2 diabetes, heart disorders, cancers, etc. Recently also widely used in animal and plant studies.three independent studies found that



From https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet

Section 3

What is Genome Wide Association Studies (GWAS)?

Why GWAS?

GWAS Analysis Overview

Data for GWAS

Running GWAS
- PLINK
- TASSEL Pipeline

GWAS analysis pipeline

- Genotype calling (SNP calling) from all the markers typed on a given platform
- remove all low-quality samples and markers
- imputation for missing values
- and finally association analysis.

Here we focus on the last step. Note: There are follow-up analysis after association analysis such as map SNPs to genes, and identify gene functions, etc..

GWAS analysis pipeline

- Genotype calling (SNP calling) from all the markers typed on a given platform
- remove all low-quality samples and markers
- imputation for missing values
- and finally association analysis.

Here we focus on the last step. Note: There are follow-up analysis after association analysis such as map SNPs to genes, and identify gene functions, etc..

GWAS analysis models

- Statistical association studies
- Generalized linear model (GLM)
- Mixed linear model (MLM), also known as linear mixed model (LMM)

Other information in the model (as random effect)

- Population structure (K-matrix)
- Kinship (relation) matrix (Q-matrix)

Note: random effects control for non-independence from sample structure. Fixed effects are typically the parameters we're interested in testing.

More on MLM see https://www.youtube.com/watch?v=QCqF-2E86r0

GWAS analysis models

- Statistical association studies
- Generalized linear model (GLM)
- Mixed linear model (MLM), also known as linear mixed model (LMM)

Other information in the model (as random effect)

- Population structure (K-matrix)
- Kinship (relation) matrix (Q-matrix)

Note: random effects control for non-independence from sample structure. Fixed effects are typically the parameters we're interested in testing.

More on MLM see https://www.youtube.com/watch?v=QCqF-2E86r0

GWAS Software Overview

- PLINK – one of the earliest, human studies, a standard tool.
- TASSL and GAPIT – mainly for plant (can also do genome selection)
- FaST-LMM
- R packages, e.g. GWASTools, GenABEL, statgenGWAS

Input data and format vary.

GWAS Software Overview

- PLINK – one of the earliest, human studies, a standard tool.
- TASSL and GAPIT – mainly for plant (can also do genome selection)
- FaST-LMM
- R packages, e.g. GWASTools, GenABEL, statgenGWAS

Input data and format vary.

PLINK

- PLINK site https://zzz.bwh.harvard.edu/plink/, more explanations and easy to follow for beginners
- PLINK 1.9 https://www.cog-genomics.org/plink/, has GUI – gPlink
- Fast and efficient in computation
- Purely on analysis of genotype/phenotype data, no support for steps prior to this
- Required genotype in PLINK .ped & .map files; will convert to binary in computation
- Required phenotype in PLINK .pheno file, or can be embedded into .ped file if it is a single phenotype

PLINK

- PLINK site https://zzz.bwh.harvard.edu/plink/, more explanations and easy to follow for beginners
- PLINK 1.9 https://www.cog-genomics.org/plink/, has GUI – gPlink
- Fast and efficient in computation
- Purely on analysis of genotype/phenotype data, no support for steps prior to this
- Required genotype in PLINK .ped & .map files; will convert to binary in computation
- Required phenotype in PLINK .pheno file, or can be embedded into .ped file if it is a single phenotype

TASSEL

- TASSEK website https://tassel.bitbucket.io
- Java based, with GUI, good Youtube tutorials and forum for support
- R front end (rtassel)
- Tutorial data (the one we've used in our Lab 1)
- Accept different input data formats
- Associations GLM, LMM, and others, could include K and Q matrices in LMM
- More info check user manual
  https://bitbucket.org/tasseladmin/tassel-5-source/wiki/UserManual

TASSEL

- TASSEK website https://tassel.bitbucket.io
- Java based, with GUI, good Youtube tutorials and forum for support
- R front end (rtassel)
- Tutorial data (the one we've used in our Lab 1)
- Accept different input data formats
- Associations GLM, LMM, and others, could include K and Q matrices in LMM
- More info check user manual
  https://bitbucket.org/tasseladmin/tassel-5-source/wiki/UserManual

Section 4

What is Genome Wide Association Studies (GWAS)?

Why GWAS?

GWAS Analysis Overview

Data for GWAS

Running GWAS
- PLINK
- TASSEL Pipeline

Common Genotype (SNP) Data Formats

- PLINK ped & map files
- hmp (Hapmap text based file)
- vcf (Variant Call Format), a standardized format originally developed by the 1000 Genomes Project

PLINK Data Example

Arabidopsis Thaliana

- the thale cress, mouse-ear cress
- a model plant, diploid, 5 different chromosomes, one of the smallest genomes among plants.
- dataset from https://easygwas.ethz.ch/data/public/dataset/view/1/ (around 1.12GB .ped)
- PLINK ped file – rows are samples and columns are SNPs
- PLINK map file – rows are SNPs and columns are their locations
- Refer to the AT_intro.docx file

Hmp Data Example

TASSEL tutorial data

- The first row is the header labels, followed by one SNP in each row
- The first 11 columns describe attributes of the SNP, while the following columns describe the SNP value for a single germplasm line.
- Check the example at https://bitbucket.org/tasseladmin/tassel-5-source/wiki/UserManual/Load/Load#markdown-header-hapmap

VCF Data Example

TASSEL tutorial data

- A single 'fileformat' field is always required, must be the first line in the file, and details the VCF version number
- Contains meta information, after the ## string and has key=value pairs
- Could contain other information such as INFO fields and Filter field
- A single header line.
- Each row afterwards is information about one SNP.

VCF Data Example

- The VCF format is able to store a wide variety of different types of information including Reference Bases, Alternate Bases, Allele Frequency, Total Number of Alleles in the Genotype, Read Depth, Genotype Likelihoods, Genotype Quality and many other types.

- Check full info about VCF (and another example file) at
  http://samtools.github.io/hts-specs/VCFv4.2.pdf

Section 5

What is Genome Wide Association Studies (GWAS)?

Why GWAS?

GWAS Analysis Overview

Data for GWAS

Running GWAS
- PLINK
- TASSEL Pipeline

Running PLINK

- Check if PLINK is installed in your `PATH`
- General form `plink -flag1 -flag2 ...`
- e.g. `plink --file mydata`

Running PLINK

Make bed files (.bed + .bim + .fam) from plink text files (.map + .ped)

```
plink --file text_fileset --maf 0.05 --make-bed --out binary_fileset
```

- Make bed files `--make-bed`
- Set minor allele frequency `--maf`
- refer to plink files `--file`, .map + .ped shall have the **same** file basename and it will find both of them
- name the output file `--output file_name`

Running PLINK

Make plink text files from bed files

```
plink --bfile binary_fileset --recode --out new_text_fileset
```

Make vcf files from bed files

```
plink --bfile binary_fileset --recode vcf --out new_vcf_file
```

- `--recode` creates a new text fileset, after applying sample/variant filters and other operations. By default, the fileset includes a .ped and a .map file

Running PLINK

Run association analysis

```
plink --file mydata --pheno pheno.raw --assoc --maf 0.05 --out run1
```

Rerun the same with a different maf

```
plink --rerun run1.log --maf 0.1
```

- `--assoc`

Running PLINK

Run association analysis (no-sex)

```
plink --file mydata --pheno pheno.raw --assoc --maf 0.05 --allow-no-sex
--out run2
```

- `--allow-no-sex` option disables the automatic setting if the individual has an ambiguous sex code
- `----adjust` causes an .adjusted file to be generated with each association test report, containing several basic multiple testing corrections for the raw p-values. Entries in this file are sorted by significance value instead of genomic location.

Running TASSEL Pipeline

Manual is https://bytebucket.org/tasseladmin/tassel-5-source/wiki/docs/Tassel5PipelineCLI.pdf

- Check if TASSEL is installed in your `PATH`
- General form `run_pipeline.pl -flag1 -flag2 ...`
- Increasing Heap Size
  `run_pipeline.pl -Xmx10g`

Running TASSEL Pipeline

Examples:

```
run_pipeline.pl -fork1 -h genotype.hmp.txt -fork2 -t trait.txt -combine3
-input1 -input2 -intersect
```

- `-fork`<id> identifies the start of a pipeline segment that should be executed sequentially
- `-input`<id> specifies a pipeline segment as input to the plugin prior to this flag
- `-combine`<id> starts a new pipeline segment to combine data sets from multiple pipeline segments. It shall followed by `-input`<id>
- `-intersect` joins (intersect) input datasets based taxa (samples)
- Data controls
  - `-h` <hapmap file>
  - `-t` <trait file> TASSEL phenotype file

Running TASSEL Pipeline

Examples:

```
run_pipeline.pl -fork1 -h genotype.hmp.txt -fork2 -t trait.txt -combine3
-input1 -input2 -intersect -glm -export glm-file
```

- `-glm` run GWAS on GLM
- `-export`<file_name> Exports input dataset to specified filename
- Population matrix is optional with -q<Population_Matrix>
- Note: the newest version does NOT support `-glm`, but changed to `-FixedEffectLMPlugin -endPlugin`

See example
https://bitbucket.org/tasseladmin/tassel-5-source/wiki/UserManual/GLM/GLM

Running TASSEL Pipeline

Examples:

```
run_pipeline.pl -fork1 -h genotype.hmp.txt -fork2 -t trait.txt -fork3 -k
kamship.txt -combine4 -input1 -input2 -intersect -combine5 -input4 -input3
-mlm -export mlm-file
```

- `-mlm` run GWAS on MLM
- `-k` <kinship matrix file>
- `-q` <population matrix file>
- MLM takes longer time than GLM
- Kinship is **required** for MLM, and population is optional.

Running TASSEL Pipeline

Examples:
Another example of MLM
https://bitbucket.org/tasseladmin/tassel-5-source/wiki/UserManual/MLM/MLM

- Minor allele frequency `-siteMinAlleleFreq`, usually 0.05
- `-importGuess` <file> TASSEL will determine the file type
- `--mlmCompressionLevel` default is Optimum

Useful Commands

`screen`

- What if the task takes a long time while you are on a remote server, and your connection suddenly drop?
- SSH session is terminated, and your work is lost
- It is usually pre-installed. Check the version: `screen --version`
- To create a named session `screen -S session_name`

Useful Commands

`screen`

- What if the task takes a long time while you are on a remote server, and your connection suddenly drop?
- SSH session is terminated, and your work is lost
- It is usually pre-installed. Check the version: `screen --version`
- To create a named session `screen -S session_name`

Useful Commands

`screen`

- detach from the screen session at any time by typing: Ctrl+a d
- Reattach to a screen: `screen -r session_name`
- List all screens: `screen -ls`
- Kill a screen
  - in the screen: `exit`
  - out of the screen: `screen -X -S session_name quit`