GENOME-WIDE ASSOCIATION STUDIES

Timo Tõnis Sikka
18.11.2016

Outline

- Introduction
- Pipeline of a GWAS
- Samples and phenotypes
- Genotyping
- Imputation
- Association analysis
- Visualizing results
- Limitations of GWAS
- Findings from GWASes
- Conclusion
- Homework

Introduction

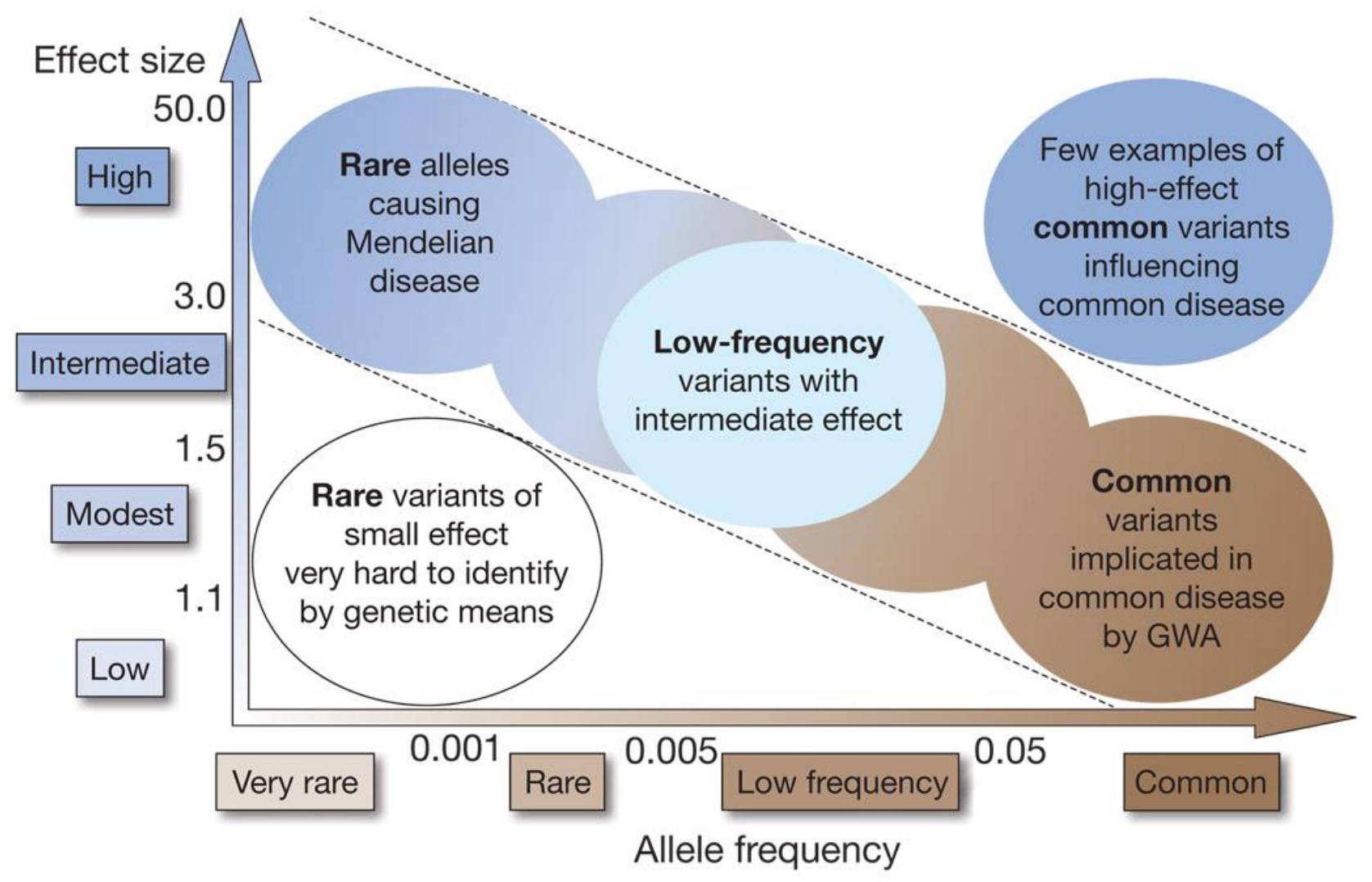
- What is a genome-wide association study?

= (bioinformatical) examination of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait (Wikipedia)

- Non-candidate approach

- Complex diseases and traits
- Single-nucleotide polymorphisms (SNPs) as markers

Introduction



Manolio *et al.,* Nature (2009)

Pipeline of a GWAS

Sample and phenotypes

↓

Genotyping

↓

Imputation

↓

Association analysis

↓

Preliminary results with best associations

↓

Replication

Samples and phenotype

- Cases and controls

- Large sample sizes
- Well-characterized phenotype

Genotyping

- Different microarrays (chips)
  - thousands of SNPs *(tagSNPs)*

Image: www.illumina.com

Imputation

= using statistics to predict unobserved genotypes

- Why is it useful?
  - brings down the cost of genotyping
- Reference panels:
  - HapMap2 (3 million SNPs)
  - 1000 Genomes phase 3 (over 84 million SNPs)
  - (population-based panels, Estonia!)
- SHAPEIT and IMPUTE2 (software for phasing and imputing SNPs)
- QUALITY CONTROL after genotyping and imputation

Quality control (QC)

- Sample QC
  - *Call rate (CR)*
  - *Hardy-Weinberg equilibrium (HWE)*
  - *Minor allele frequency (MAF)*
  - *Gender discordance*
  - *Heterozygosity*
  - *Identity by descent*
- Variant QC
  - *CR*
  - *HWE*
  - *MAF*

Association analysis

- SNPTEST https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html#introduction
- Input parameters

-pheno (continuous, binary, discrete)

-frequentist <number of test model>

-method score/expected

-covs <co-variants< #e.g. age, sex, PCs

- Output
  - *P-value*
  - *Effect size (beta)*
- Filtering results

Association analysis

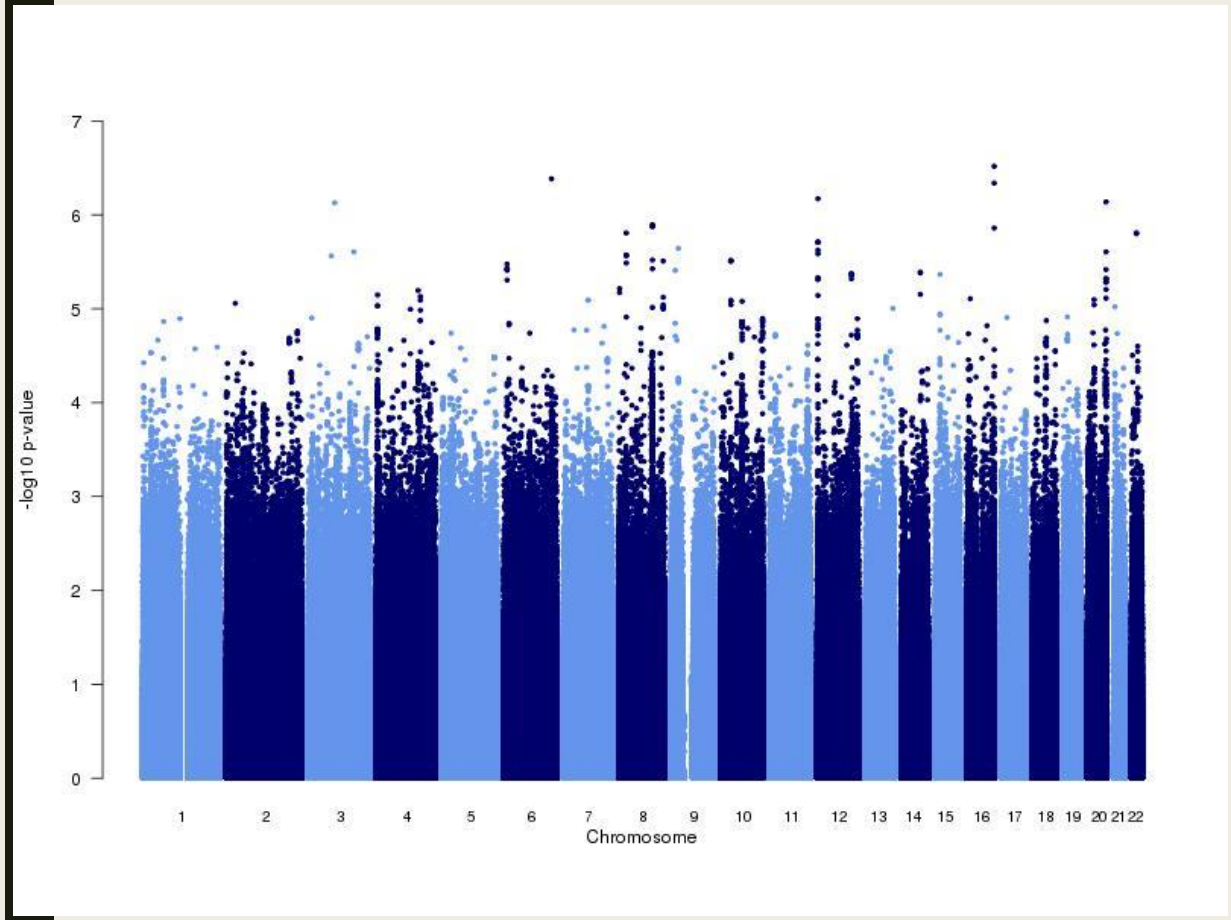
```
# Analysis: "SNPTEST analysis, started 2015-11-25 16:34:52"
#  started: 2015-11-25 16:34:54
#
# Analysis properties:
#   -data /gpfs/hpchome/tts/Glycans_GWAS/EGCUT_OMNI_imputed_06112015_chr20_ftga_ind_201115.impute.gz /gpfs/hpchome/tts/Glycans_GWAS/Glycans_OMNI_loplik_oopsline.sample (u
#   -exclude_samples /gpfs/hpchome/tts/Glycans_GWAS/eemaldada_need.txt (user-supplied)
#   -frequentist 1 (user-supplied)
#   -hwe  (user-supplied)
#   -method expected (user-supplied)
#   -o EGCUT_OMNI_HM2_IGP40_chr20.snptest (user-supplied)
#   -pheno IGP40 (user-supplied)
#   -use_raw_phenotypes  (user-supplied)
#
alternate_ids rsid chromosome position alleleA alleleB index average_maximum_posterior_call info cohort_1_AA cohort_1_AB cohort_1_BB cohort_1_NULL all_AA all_AB all_BB al
oportion cohort_1_hwe frequentist_add_pvalue frequentist_add_info frequentist_add_beta_1 frequentist_add_se_1 comment
--- rs4814683 NA 9795 G T 1 0.994921 0.989222 267.981 245.401 61.6189 0 267.981 245.401 61.6189 0 575 0.320555 -1.28516e-15 0.70103 0.600551 NA 0.0330308 0.063048 NA
--- rs6076506 NA 11231 G T 2 0.968674 0.509107 0.179018 32.5505 542.27 0 0.179018 32.5505 542.27 0 575 0.0286161 -7.90865e-16 1 0.440132 NA -0.195457 0.253016 NA
20 rs6139074 NA 11244 A C 3 1 1 363 191 21 0 363 191 21 0 575 0.202609 0 0.605359 0.956547 NA 0.0040676 0.0746202 NA
20 rs1418258 NA 11799 C T 4 1 1 268 245 62 0 268 245 62 0 575 0.32087 0 0.632177 0.617861 NA 0.0312645 0.0626347 NA
--- rs7274499 NA 12150 A C 5 0.983861 0.0711333 0.031011 9.24904 565.72 0 0.031011 9.24904 565.72 0 575 0.00809657 -4.94291e-16 1 NA NA NA NA frequentist_add:model_not_fi
it
--- rs6116610 NA 12934 A G 6 0.985745 0.20313 0.019998 9.63205 565.348 0 0.019998 9.63205 565.348 0 575 0.00841048 -3.95432e-16 1 0.976249 NA -0.0216295 0.726179 NA
--- rs13043000 NA 13288 G T 7 0.993952 0.972482 450.412 112.851 11.737 0 450.412 112.851 11.737 0 575 0.118543 2.17488e-15 0.220928 0.463422 NA 0.065995 0.0899462 NA
--- rs6054257 NA 14370 A G 8 0.993571 0.981512 362.848 190.887 21.265 0 362.848 190.887 21.265 0 575 0.202972 7.90865e-16 0.605089 0.991799 NA 0.000773367 0.0752121 NA
20 rs6086616 NA 16749 C T 9 1 1 251 255 69 0 251 255 69 0 575 0.341739 0 0.711837 0.711846 NA -0.0228748 0.061897 NA
20 rs6039403 NA 17094 A G 10 1 1 79 284 212 0 79 284 212 0 575 0.384348 0 0.332417 0.987449 NA 0.000978304 0.0621603 NA
--- rs17685809 NA 17408 C T 11 0.933492 0.69829 475.609 93.6222 5.76875 0 475.609 93.6222 5.76875 0 575 0.0914432 -1.97716e-16 0.797776 0.917079 NA 0.0126069 0.121035 NA
--- rs16987437 NA 20892 A G 12 0.996191 0.0399545 0 2.19015 572.81 0 0 2.19015 572.81 0 575 0.00190448 0 1 NA NA NA NA frequentist_add:model_not_fit:design_matrix_singula
20 rs6135141 NA 22347 A G 13 1 1 76 269 230 0 76 269 230 0 575 0.366087 0 0.92839 0.342473 NA -0.058505 0.0615791 NA
20 rs892665 NA 23254 A C 14 1 1 65 250 260 0 65 250 260 0 575 0.330435 0 0.706133 0.0215005 NA 0.14302 0.0620372 NA
--- rs6111385 NA 24962 C T 15 0.972959 0.93914 346.49 193.472 35.0382 0 346.49 193.472 35.0382 0 575 0.229172 8.89723e-16 0.239195 0.320942 NA 0.0702946 0.0707627 NA
20 rs2196239 NA 28655 A G 16 1 1 10 112 453 0 10 112 453 0 575 0.114783 0 0.304912 0.494539 NA 0.0621078 0.0908615 NA
--- rs7268260 NA 33116 C T 17 0.988518 0.892127 519.854 54.3089 0.836998 0 519.854 54.3089 0.836998 0 575 0.0486808 2.17488e-15 0.628873 0.160033 NA 0.206601 0.14686 NA
```

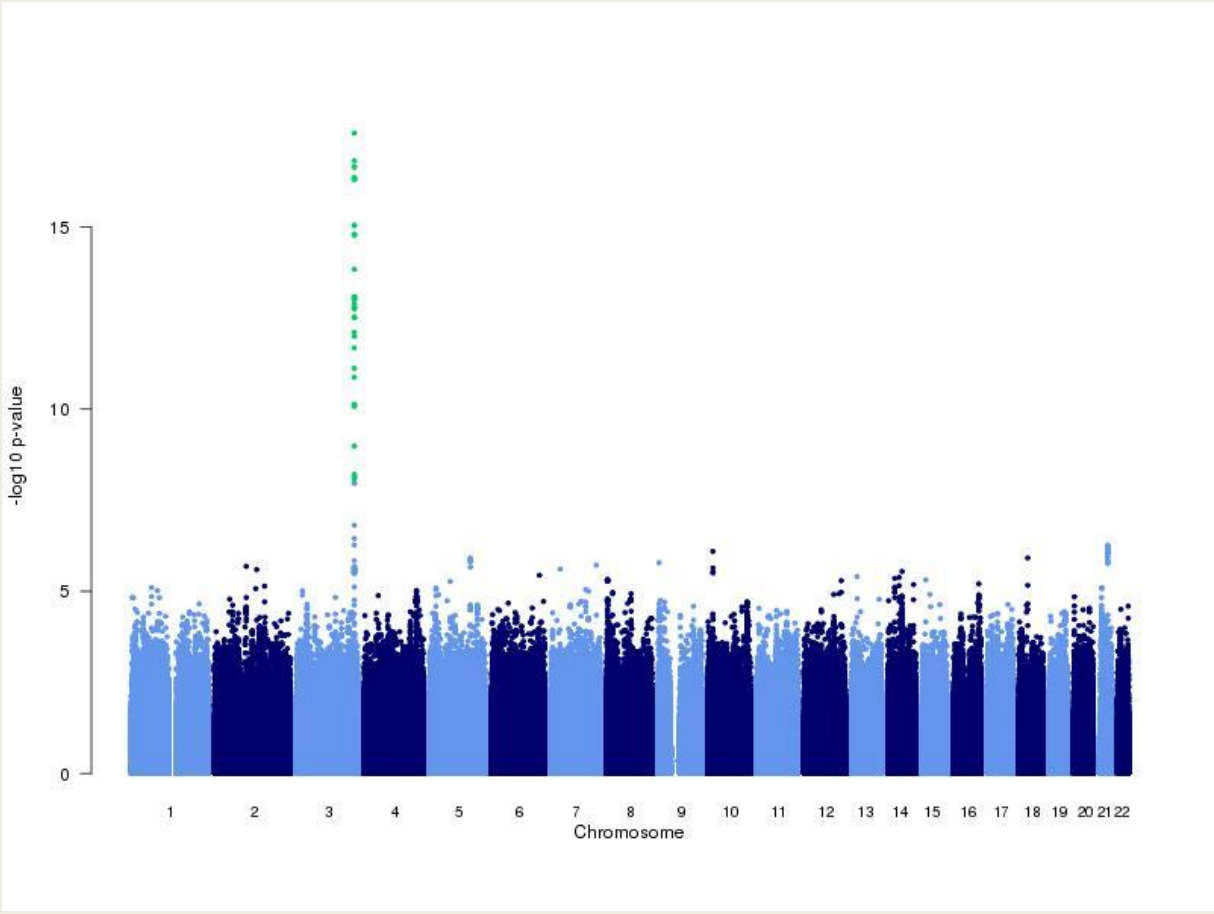
Visualizing GWAS results

- Post-filtering
- Manhattan plot
- Regional plot

Visualizing GWAS results

- Manhattan plot

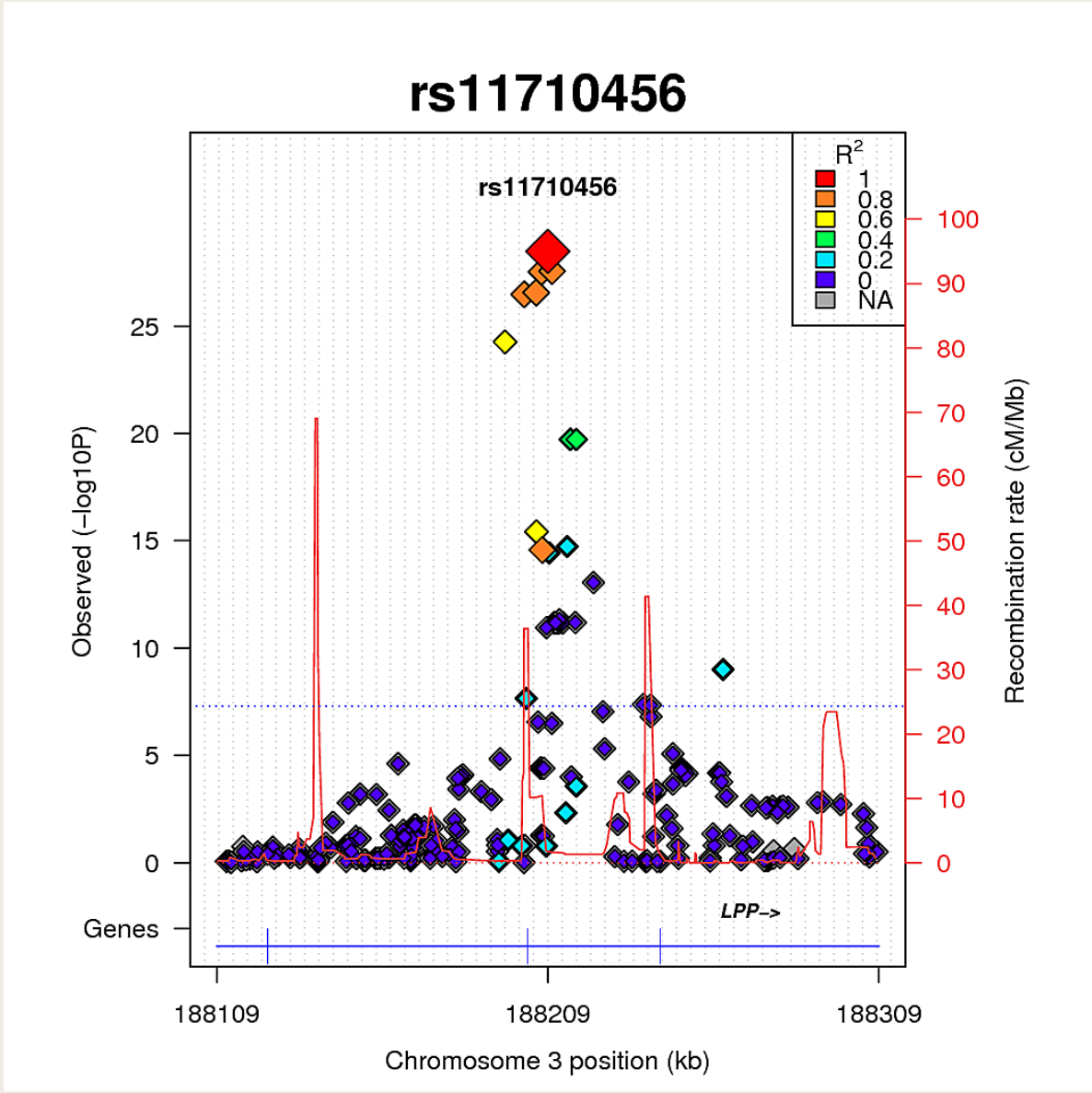




Visualizing GWAS results

- Regional plot
  - *LocusTrack*

http://gump.qimr.edu.au/general/gabrieC/LocusTrack/



Findings from GWASes

- GWAS catalog http://www.ebi.ac.uk/gwas/
- Over 29 000 trait associated SNPs (Nov 13, 2016)

Limitations of GWAS

- Dependent on microarray data
- Suitable for common variants with small effect
  - rare variants go missing
- Requires large sample sizes (which in turn asks for larger funds)
- Associated SNPs rarely in coding area
  - Requires further analysis
- Findings from studies with small samples often non-reliable (do not replicate)

Conclusion

- Robust method for scanning the genome for genotype-phenotype associations
- Has enabled to find thousands of associations between genotype and phenotype
- Effective for common variants, but not for rare variants
- Has its limitations that have to be taken into consideration

Homework

Please send your homework to

timotonis.sikka@gmail.com

by November 25

9:00 am

Homework

Task 1. What can we analyse using GWAS? Why is imputation used if we use micro-array data? What are the cons/limitations of GWASes?

Homework

Task 2. You want to analyse if there are any associations between body mass index (BMI) as phenotype and genetic variants. Please describe your analysis plan, starting from selecting the sample set and ending with visualizing the results after the association analysis. Which programs would you use at certain steps?

(Some ideas: would you pick all the people you can to your sample set or only very obese and very thin; which genotype data would you use (micro array or full sequenced genomes data); is imputation required?; what would the SNPTEST input parameters be (more information about SNPTEST parameters at https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html#frequentist_tests)

IF you have any questions, you can contact me by e-mail and I'll gladly explain.

All in all, there are no wrong answers to this task. You can have different approaches.

Example analysis about Alzheimer's GWAS study plan on next slide as guide!

Homework

Task 2. Example analysis plan for Alzheimer's GWAS.

1) I'd pick as many cases (people diagnosed with Alzheimer) and controls (healthy individuals) as I can. Number of controls should be equal or exceed the number of cases.
2) With limited funds I'd go for micro-array genotyping.
3) Would use imputation to have more genetic variants to analyse.
4) QC to remove bad samples and variants.
5) Association analysis (with SNPTEST) for a case-control study:

-pheno <Alzh> #binary phenotype – the individual either has Alzheimer or not

-frequentist 1 (additive model)

-method score (but could also be expected)

-covs age, sex and 5 first PCs

5) Visualize with Manhattan's plot to see if there are any hits (p-value <5x10E-8), if there are, then a regional plot (with LocusTrack).

(Optional: I would try to find out if the variant is actually causal and what is the pathway behind the SNP and trait association with further analysis.)

Homework

Task 3. You have data from a GWAS that examined possible associations between extraversion and genetic variants. Data can be found here:

https://www.dropbox.com/s/cj6fo5dhfwvzl97/testdata_chr10_Extraversion_for_LocusTrack.gz?dl=0

You want to see which SNP has the best association and if it's within a gene. For that, go to http://gump.qimr.edu.au/general/gabrieC/LocusTrack/index.html

1) Upload the previous file (can be as compressed file as it already is or unzipped) for the software: Manage Files > Results

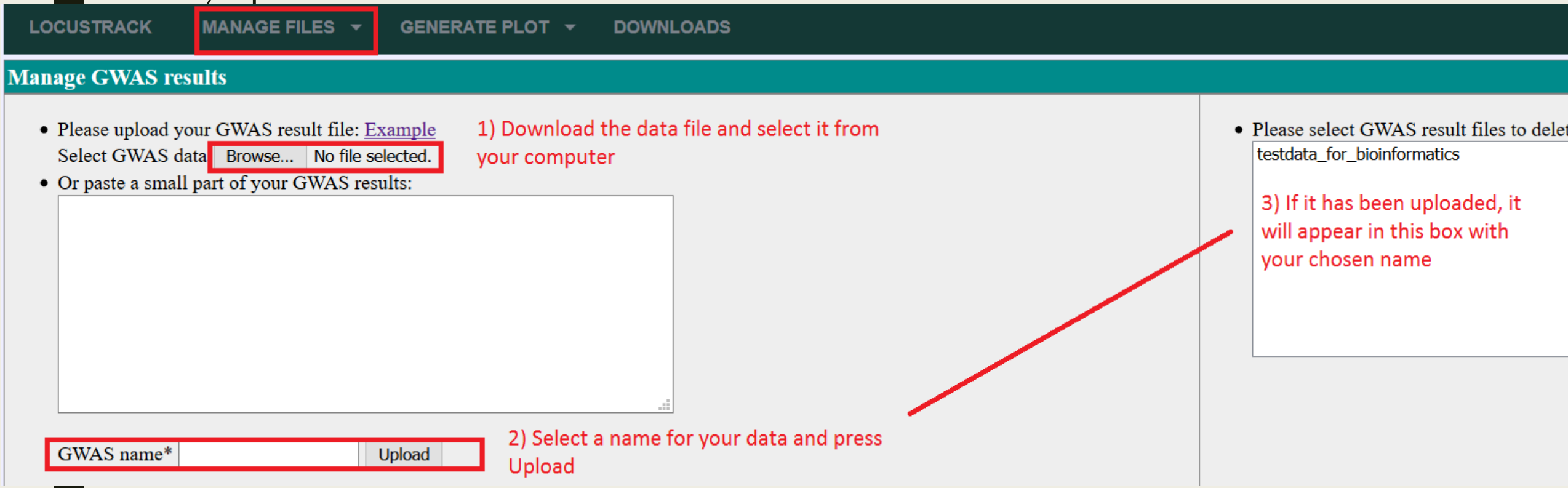
2) Select the parameters (under Generate plot > LocusTrack plot) you like and insert the e-mail where you'd like to receive the results (the creation for the plot will take a few minutes depending on the work load, should not take more than 10 minutes).
Explanation of options in the following slides!

Questions: Is the variant in a gene? If so, what gene? If not, what are the nearest genes?

Homework

Task 3. LocusTrack guide

1) Upload the data



Homework

Task 3. LocusTrack guide

2) Set parameters and generate plot.

More information about possible populations:

http://www.internationalgenome.org/category/population/