Genome wide association studies

Helga Westerlind, PhD

Outline

- About GWAS/Complex diseases
- How to GWAS
- Imputation

What is a genome wide association study?

Why are we doing them?

In what context?

How do we know there is genetics involved in the disease susceptibility?

Family studies!

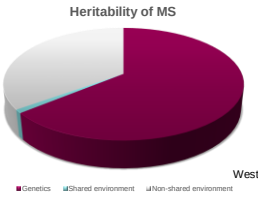
- Families with a higher number of cases than expected -> genetics are involved!

Family study – example from multiple sclerosis

	Risk Ratio (RR)
Monozygotic twin	23.62 (8.71-64.02)
Dizygotic twin	2.18 (0.71-6.68)
Sibling	7.13 (6.42-7.93)
Adopted sibling	1.87 (0.23-15.46)
Cousin	1.63 (1.36-1.97)

Westerlind et al, Brain 2014

But families share more than genetics!



Westerlind et al, Brain 2014

Multiple sclerosis is a complex disease!

Pic: stolen from insidegen.com

Multiple sclerosis risk factors

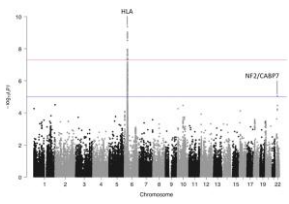
- Genetics – HLA DRB1*15:01
- Non-HLA genes >110 and counting…
- Environmental risk factors
  - → Smoking
  - → High BMI
  - → Working nightshift
  - → Low vitamin D-levels

Can you think of other complex diseases?

And why would you argue they're complex?

On to the genetics…

GWAS – the how to!



Gut. 2017 Mar;66(3):421-428. doi: 10.1136/gutjnl-2015-309934. Epub 2015 Nov 2.

Dense genotyping of immune-related loci identifies HLA variants associated with increased risk of collagenous colitis.

Westerlind H[1,2], Mellander MR[3,4], Bresso F[2,3,4], Munch A[5], Bonfiglio F[2], Assadi G[2], Ratter J[2], Höbenthal M[6], Lieb W[7], Källberg H[1], Brunedal B[1], Padyukov L[3], Halfvarson J[8], Törkvist L[4], Björk J[4], Andreasson A[9], Agreus L[9], Almer S[3,4], Miehlke S[10], Madisch A[11], Ohlsson B[12], Löfberg R[3,13], Hultcrantz R[14], Franke A[6], D'Amato M[2,15].

Collagenous colitis

- Inflammatory bowel disease
- Onset > 50 years
- Predominant in women

- Chronic watery diarrhea
- Diagnosis through biopsies showing deposition of collagen in lamina propria

- Etiology – unknown!

- Reports about familial clustering, no proper genetic study done

Study material

- 314 cases, 4,299 controls from Sweden and Germany

- Genotyped on the Immunochip

QUALITY CONTROL OF BOTH SAMPLES AND INDIVIDUALS!

Why quality control?

- Systematic bias leading to false results
- Unreliable assays
- Sample contamination
- Sample mix-up
- Low frequency of variants
- Population stratification and mixtures
- Batch effects

A cautious tale of failed QC

REPORT

Genetic Signatures of Exceptional Longevity in Humans

Paola Sebastiani[1,*], Nadia Solovieff[1], Annibale Puca[2], Stephen W. Hartley[1], Efthymia Melista[1], Stacy Andersen[2], Daniel A. Dworkis[1], Jemma B. Wilk[2], Richard H. Myers[2], Martin H. Steinberg[2], Monty Montano[3], Clinton T. Baldwin[1,*], Thomas T. Perls[2,*]

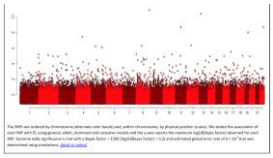
+ Author Affiliations

*To whom correspondence should be addressed. E-mail: sebas@bu.edu (P.S.); thperls@bu.edu (T.H.P.)

Science 28 Jul 2010:

DOI: 10.1126/science.1190532

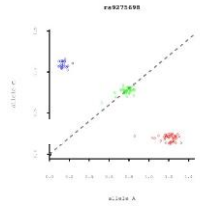
RETRACTED



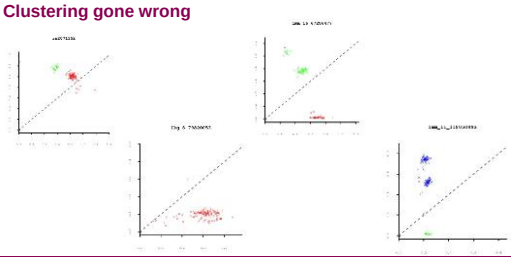
Genotyping arrays

- Oligonucleotide probes
  - →single nucleotide extension by labeled nucleotides
  - →allele specific probe + labeled ss DNA fragments
  - →etc

"Calling" of alleles



Clustering gone wrong



PLINK

Am J Hum Genet. 2007 Sep; 81(3): 559–575. PMCID: PMC1950838
Published online 2007 Jul 25. doi: 10.1086/519795

PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses

Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham

Citations on Web of Science per 2017-02-13: 8,666

QC of genotyping

Pipeline for QC SNP markers

Supplementary Table 1. QC markers pipeline (Immunochip discovery)

QC step	SE SNPs dropped #	SE SNPs left #	DE SNPs dropped #	DE SNPs left #
Initial number		196524		196524
Failing IIBDGC QC (including intensity outliers) *	52762	143762	52762	143762
Monomorphic	17174	126588	19495	124267
Success rate < 98%	142	126446	78	124189
Differential missing cases vs controls	1118	125328	4937	119252
Minor allele frequency < 0.01	6737	118591	6078	113174
HWE P < 1.0E-07	87	118504	275	112899
Passing QC in both datasets	787	110634	2265	110634

SE = Sweden; DE = Germany. * QC guidelines set by the International IBD Genetics Consortium (IIBDGC)

Hardy–Weinberg equilibrium

- Allele & genotype frequencies will remain constant in a population

  if f(A)=p & f(a)= q then:
- $f(AA)=p^2$
- f(Aa)=2pq
- $f(aa)=q^2$

$$p^2+2pq+q^2=1$$

Hardy–Weinberg equilibrium

- no mutation
- random mating
- no selection, genetic drift, etc
- Most populations are in HWE
  - → Robust
  - → one generation 'random mating' returns HW proportions
- When is HWE not held?
  - → Genotyping errors
  - → Batch effects
  - → Population stratification
  - → Association (very unlikely)

Be sure to check HWE in
- cases
- controls
- combined

Issues with Samples

- Quality
  - → poor DNA
- Identity
  - → contamination
  - → family (related samples)
  - → mix-up

QC of individuals

QC pipeline individuals

Supplementary Table 2. QC individuals pipeline (Immunochip discovery)

QC step	SE cases #	SE ctrls #	DE cases #	DE ctrls #
Initial number	163	2046	91	2018
Genotyping success rate < 95%	0	5	0	0
Genotype-phenotype sex mismatch	1	5	1	13
Genotype relatedness > 0.1	2	39	3	31
PCA outlier (> 6 SD)	0	2	0	0
Final sample size	160	1995	87	1974

SE = Sweden; DE = Germany; ctrls = controls.

Pruning to get "independent" markers

- Done using the software PLINK and the command –indep-pairwise
- GWAS: 50 5 0.5
- ICHIP: 50 5 0.2*
- "Problematic" genomic regions with high LD and/or known inversion* must also be removed

*Abraham G, PLoS One, 2014

Related samples - Genetic relationship

$$A_{jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(x_{ij} - 2p_i\right)\left(x_{ik} - 2p_i\right)}{2p_i\left(1 - p_i\right)}.$$

Samples from related individuals have more similar DNA.

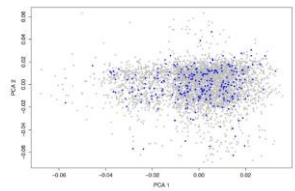
Filter away of member of pairs with high genetic relatedness (GWAS: > 0.025 / ICHIP: > 0.1)

Contamination (mixture of samples) leads to both more genetic relatedness and more heterozygotes.

Estimates might be biased and different software give different results.

Be skeptical of results and test different software!

PCA



- What do the results mean?
- What statistics do we use?
- How do we do the actual analysis?
- How do we interpret this?

Think about what data we have!

- Outcome?
- What are we investigating?
- Statistic of interest?
- Study design?
- Other things to take into account?
- What statistic could be used?

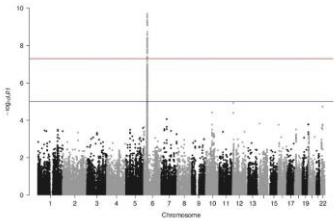
Odds ratio (recap I hope)

	Cases	Controls	Total
Exposed	a	c	a+c
Unexposed	b	d	b+d
Total	a+b	c+d	

$$\text{OR} = \frac{a/b}{c/d} = \frac{a*d}{c*d}$$

Why just not plot the risks per gene?

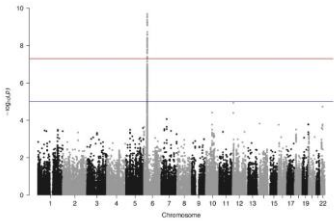
Results – but these can't be odds ratios!



A quick word about multiple testing

- What's the definition of a p-value?

Results!



Association != causation

So how determine if causal or not?

And how do we know if it's by random?

Imputation to investigate the HLA association

HLA = Human Leucocyte Antigen

So the major histocompatibility complec (MHC) in humans

Human Leucocyte Antigen

HLA class I:
Peptides from *inside* the cell to CD8+ T-cells (killer T-cells)

HLA class II:
Antigens from *outside* the cell to CD4+ T-cells (stimulating B-cells)

Most autoimmune disease have an HLA association

- Is collagenous collitis maybe an autoimmune disease?
- We want to go from SNP-data to HLA genotypes!

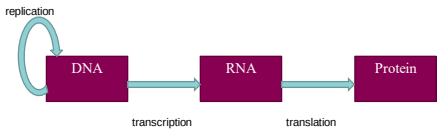
Imputation is the way to go!

A bit more about genetics first!

Outline

- The genome is not random
  → How the genome is structured
- Genetics differ between populations
  → A bit about migration and evolution of mankind
- Recombination creates diversity
  → More on inheritance
- Statistical modeling and software
  → The HowTo

The central dogma



A protein is not translated from a coherent stretch of DNA

Pic: news-medical.net

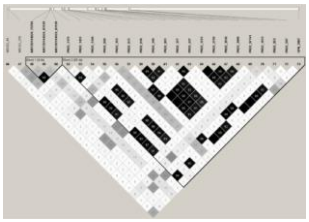
Linkage disequilibrium (LD)

- It's beneficial to preserve function, so there is high(er) correlation within a (functional) stretch of DNA.
- The LD differs across the genome.
- If we know the LD structure, we can infere the genotype at position B for an individual if we know the genotype at position A!

LD – more formal

- $D_{AB} = p_{AB} - p_A p_B$
- $D' = {}^{D}/_{D_{min}}$ $where\ D_{min} = \begin{cases} \max\{-p_{AB}, -(1-p_A)(1-p_B)\}\ when\ D < 0 \\ \min\{p_A(1-p_B), (1-p_A)p_B\}\ when\ D > 0 \end{cases}$
- $r = \frac{D}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}$

LD structure – an example



Pic: Olsen et al, BMC Genetics, 2007

The genetics differ between populations

Genetics differ across the world!

- Genetic variants differ in frequency between populations
- And certain variants might not even exist in some populations
- Addmixture, bottlenecks and selective pressure are some explanations for this

So somehow we need to take population into account…

Mitosis vs meiosis

- **Mitosis**: the cell copies itself. These are ***diploid*** (46 chromosomes, 23 pairs)
- **Meiosis**: creation of ***gametes*** (eggs/sperms), which are ***haploid*** (23 chromosomes)

Recombination creates diversity

- Is more common at certain position in the genome, so called **recombination hotspots**
- Is part in driving evolution (together with mutations, genetic drift, system of mating, population structure, selection and genetic linkage)
- But can also create dysfunctional genotypes that might be fatal
- By studying families we can learn more about recombination

In summary

- The correlation structure in the genome is called LD
- The genetics differ between populations
- There will be diversity in a population due to recombination

More on the statistics

- We can estimate
  → the correlation between different positions in the genome
  → And the frequencies of the genotypes in the population
- And we can use this to estimate the genotype at position i+1 conditional on position i

The genome is a Markov model!

(which is outside the scope of this course, but you should at least know there's a statistical way to model the genome)

Imputation using a Markov model

- Has been discussed extensively in the litterature
- A key publication is Lie & Stephens, Genetics 2003 (HOTSPOTTER)
- Lie and Stephens' approach is used by several software (BEAGLE, IMPUTE, MaCH)

Imputation continued

- By
  → assuming a Markov model
  → estimating the LD structure
  → and the genotypic frequencies

  we can infere the missing genotypes
- A key point is (of course) to have an accurate reference panel.
- Remember that mutations and recombinations introduce uncertatinty to the imputation.
- But most software also give a quality statistic of the imputed genotypes.

So by using imputation, we can go from SNP-data to genotypes

(there are other ways of doing this, this is just one version of inferring genotypes, but it's the gold standard for estimating the genotypes in the HLA region).

Table 2 rs2187668 (DQ2.5) genotypes and summary statistics

	TT	CT	CC	RAF	p Value	OR (95% CI)
Discovery						
Cases	8	94	145	0.222	1.6×10^{-8}	1.98 (1.56 to 2.51)
Controls	59	895	3015	0.128		
Replication						
Cases	3	24	40	0.224	1.3×10^{-4}	2.79 (1.65 to 4.73)
Controls	4	63	263	0.108		
Combined						
Cases	11	118	185	0.223	2.3×10^{-11}	2.06 (1.67 to 2.55)
Controls	63	958	3278	0.126		

CC, collagenous colitis; RAF, frequency of the risk allele (T=DQ2.5).

Westerlind et al, GUT 2017

In conclusion

- Collagenous colitis seem to have an association to HLA class II, implicating that it's an autoimmune disease!

Summary of the lecture

- Quality control of genetic data is important!
- We need to know about genetics and evolution to be able to use the correct statistics
- Statistics is just statistics, and inferring causality is not trivial. Association does not imply causation!!
- We need a plausible explanation, replication and preferably also functional studies to trust the results fully!