Bioinformatics workshop I: Genome Wide Association studies

Cornelis Blauwendraat

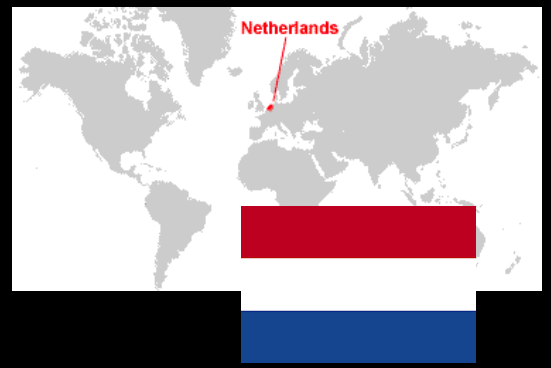
Sara Bandres-Ciga

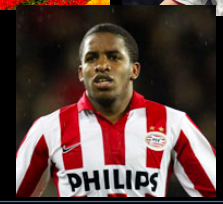
ISIA 2019

INTERNATIONAL SCHOOL ON INHERITED ATAXIAS:
FROM GENETICS TO CLINICS

LNG

Who are we?





NIH National Institutes of Health

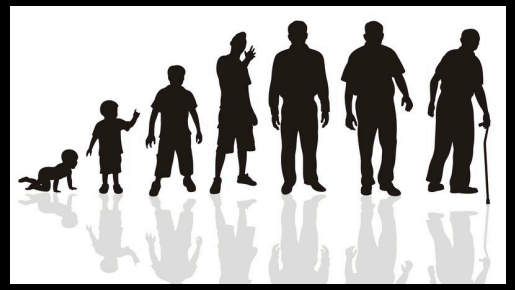
iPDGC
International Parkinson Disease Genomics Consortium

New $5 Million Program Will Explore Parkinson's Genetics in African, East Asian and Indian Populations

LARGE-PD
Latin American Research Consortium on the Genetics of Parkinsons Disease

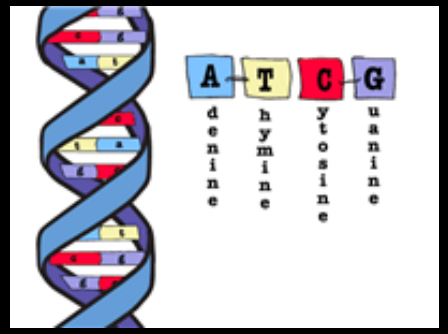
What causes diseases?

- Aging



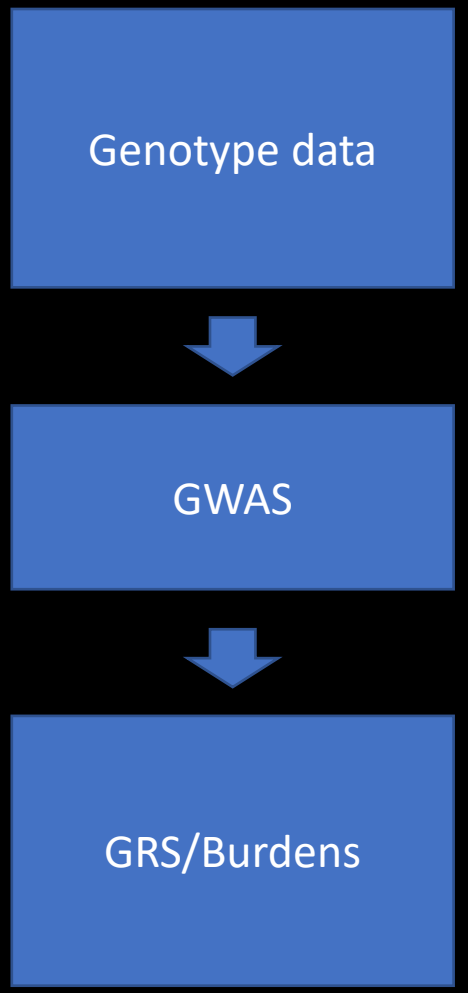
- Environment
  - Lifestyle (smoking, diet)
  - Pesticides

- Genetics

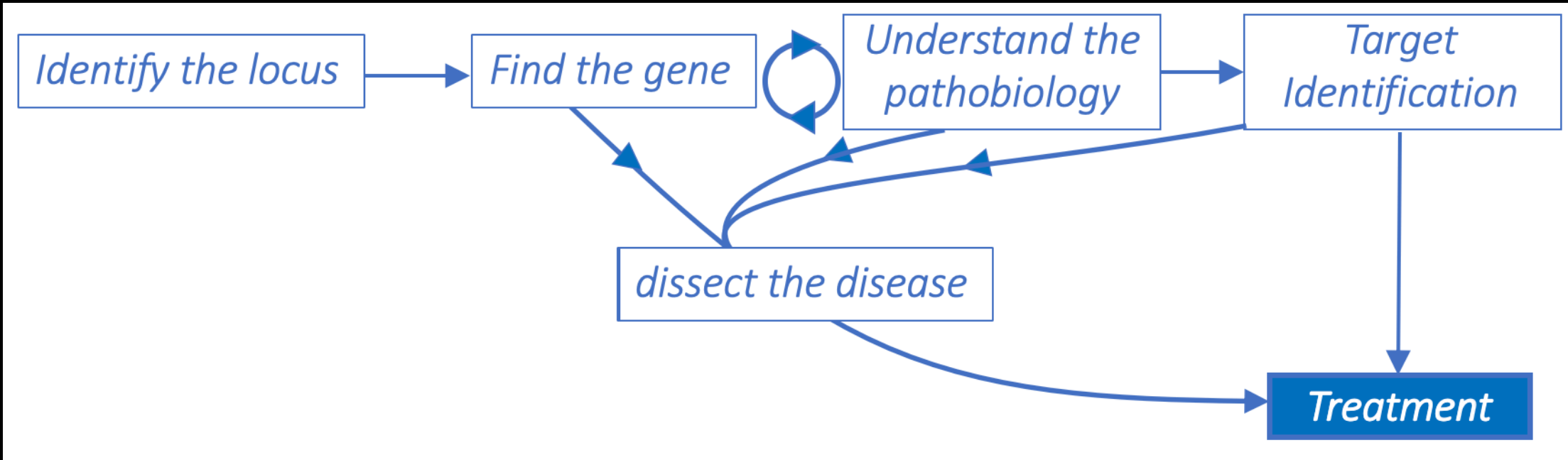


Overview

- Basic genetics
- Genotyping, filtering, QC and imputation
- Genome-wide association studies (GWAS)
- Genetic risk score (GRS)
- Burden testing



Why do we do genetics?



Basic genetics

- DNA -> A T C G

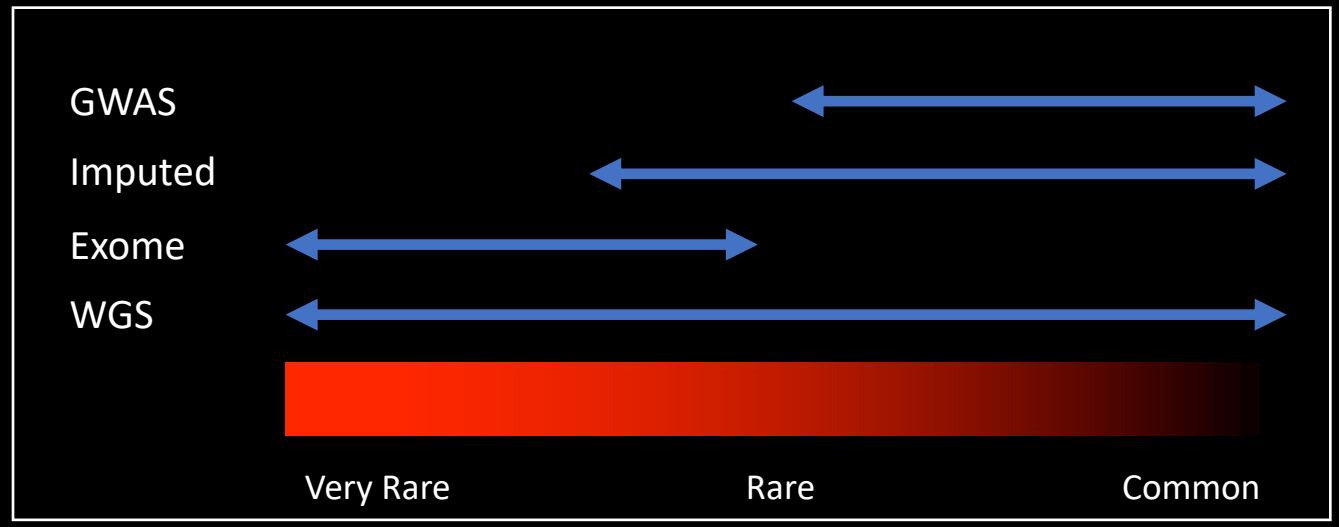
- Human genome, 3.0 $\times 10^9$ bp (base pairs)

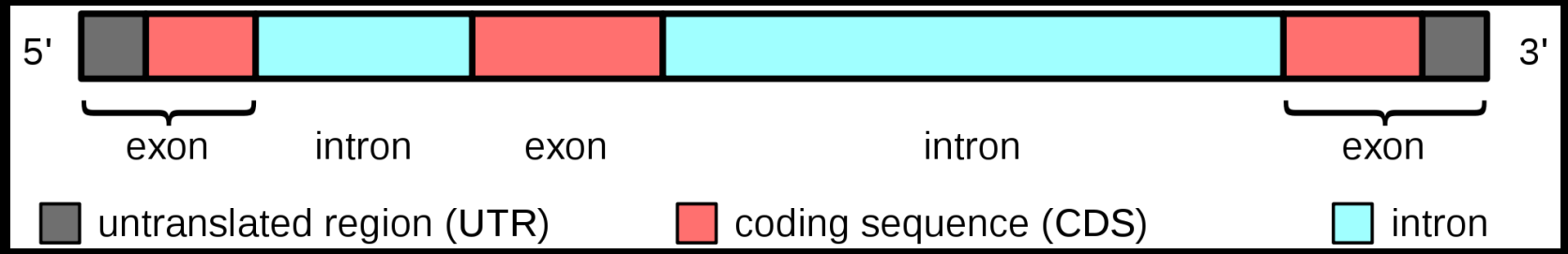
23 chromosomes, ~20,000 coding genes + >20,000 non-coding genes

- Diseases and DNA?
  - Causal mutations, Cystic Fibrosis, PD, AD, Breast cancer etc etc etc
  - Risk factors (GWAS), many common variant examples

DNA variant groups

- Frequency
  - Common variants
  - Rare variants
  - Ultra rare (private variants)
- Type
  - Coding
  - Non-coding
    - Intergenic
    - Intronic
    - UTR3 or UTR5

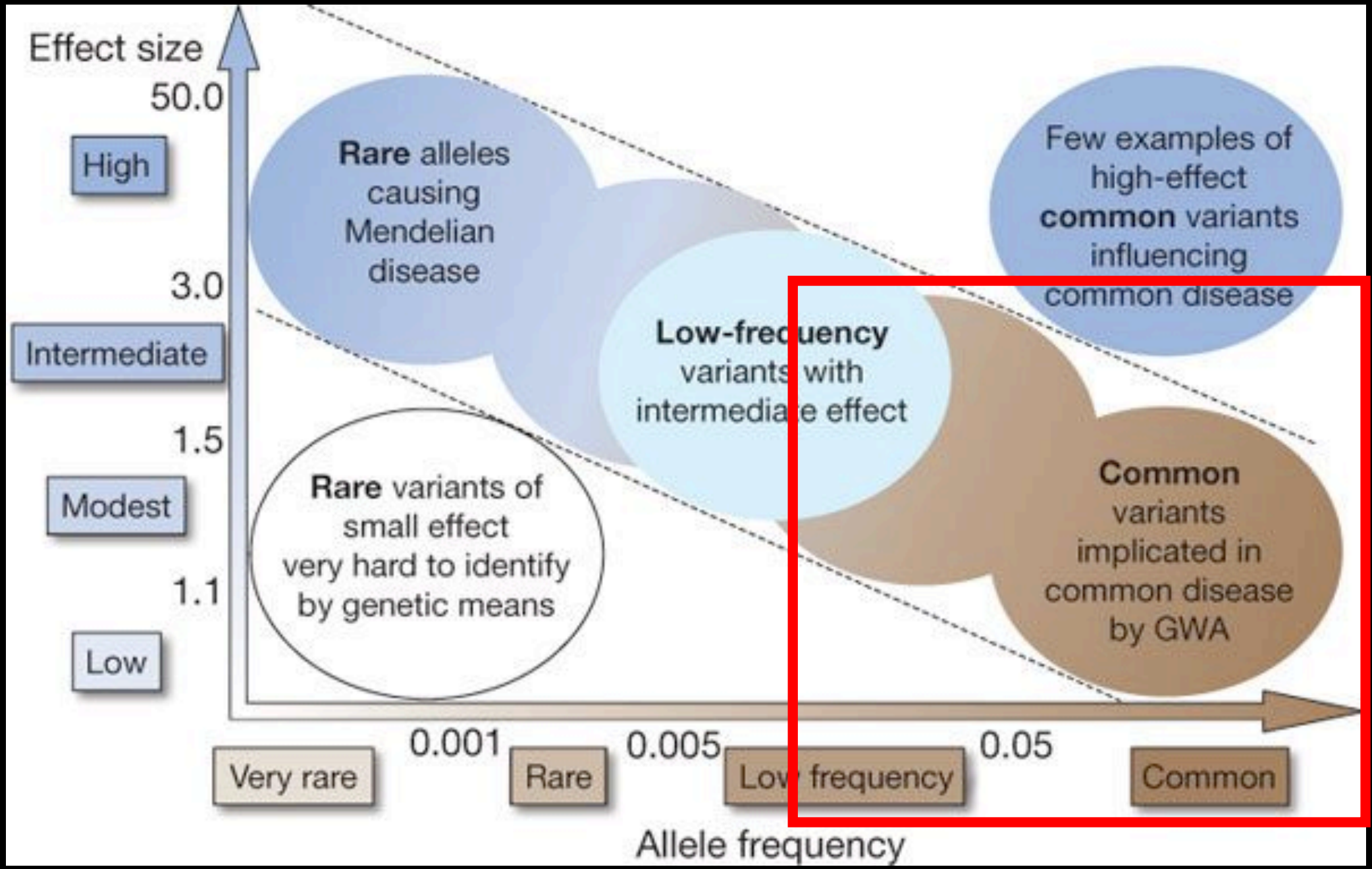




DNA variants in disease perspective

- Frequency
  - Common variants -----> can be risk factors
  - Rare variants -----> can be high risk factors
  - Ultra rare (private variants) -----> can be causal

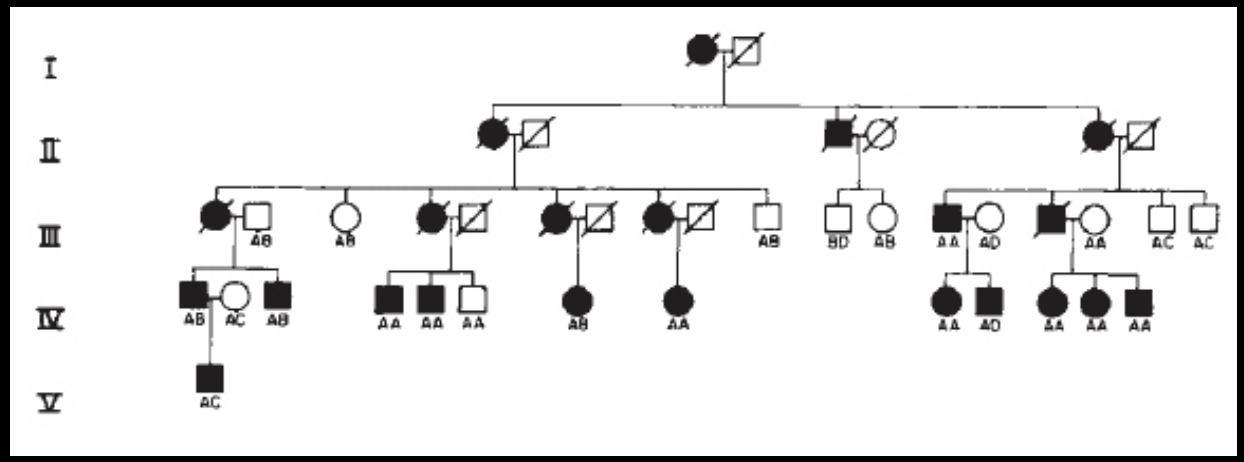
- Type
  - Coding -----> can be high risk or causal
  - Non-coding -----> can be risk factors
    - Intergenic
    - Intronic
    - UTR3 or UTR5


Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
0.001
0.005
0.05
Very rare
Rare
Low frequency
Common
Allele frequency

History of genetics

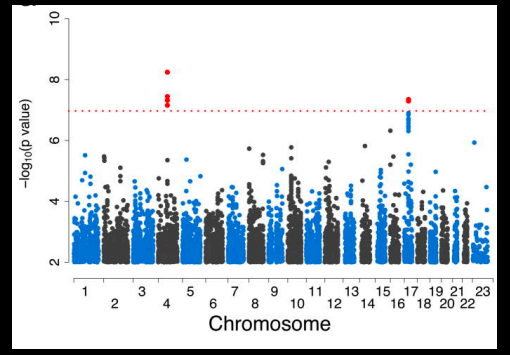
“Monogenic” forms:

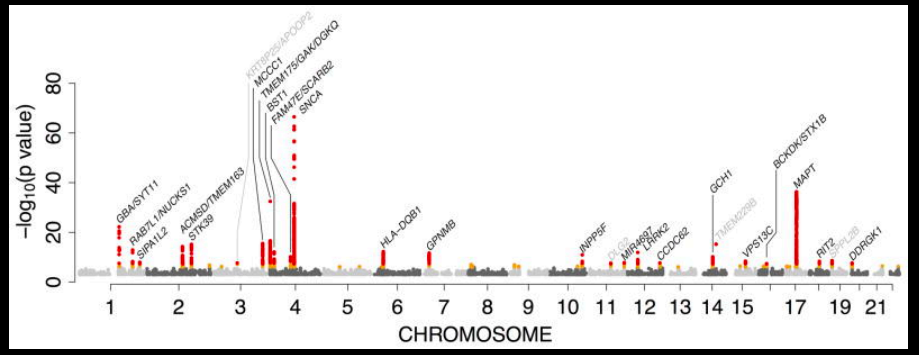
- Inheritance models:
  - Autosomal dominant
  - Autosomal recessive
  - X-linked



”Sporadic/Idiopathic” disease

- Started ~2009





From very small to very big

Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data

Hon-Chung Fung, Sonja Scholz, Mar Matarin, Javier Simón-Sánchez, Dena Hernandez, Angela Britton, J Raphael Gibbs, Carl Langefeld, Matt L Stiegert, Jennifer Schymick, Michael S Okun, Ronald J Mandel, Hubert H Fernandez, Kelly D Foote, Ramón L Rodríguez, Elizabeth Peckham, Fabienne Wavrant De Vrieze, Katrina Gwinn-Hardy, John A Hardy, Andrew Singleton

TITLE

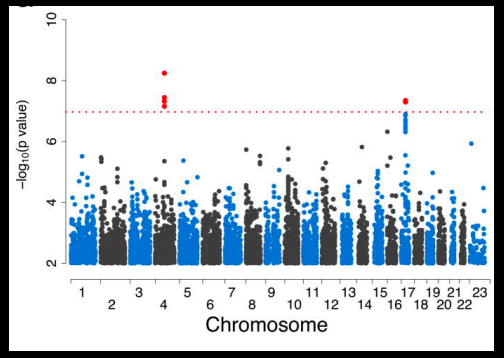
Parkinson's disease genetics: identifying novel risk loci, providing causal insights and improving estimates of heritable risk.

AUTHORS

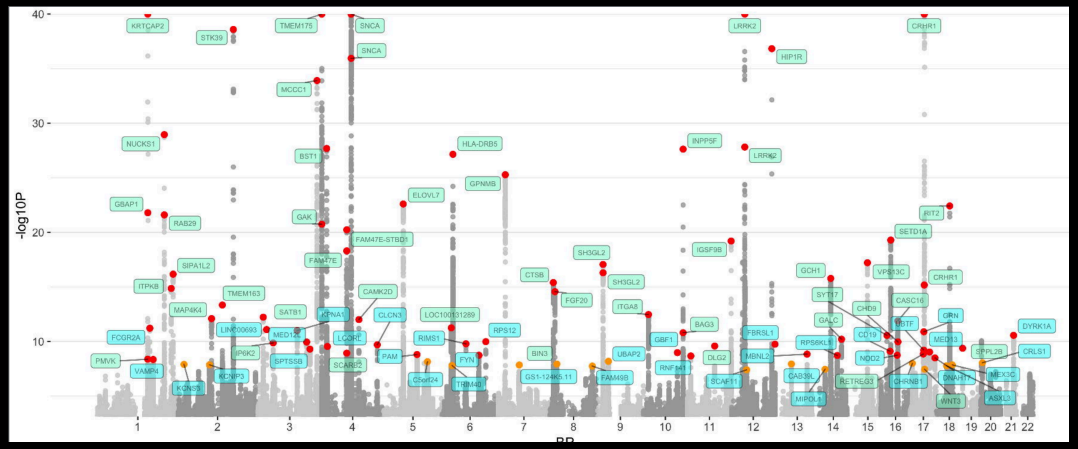
Mike A. Nalls[1,2,CA*], Cornelis Blauwendraat[1*], Costanza L. Vallerga[3,4*], Karl H
Bandres-Ciga[1*], Diana Chang[6], Manuela Tan[7], Demis A. Kia[7], Alastair J. No
Jose Bras[9,10], Emily Young[11], Rainer von Coelln[12], Javier Simón-Sánchez[13,1
Schulte[13,14], Manu Sharma[15], Lynne Krohn[16,17], Lasse Pihlstrom[18], Ari Siitonen[19,20], Hirotaka

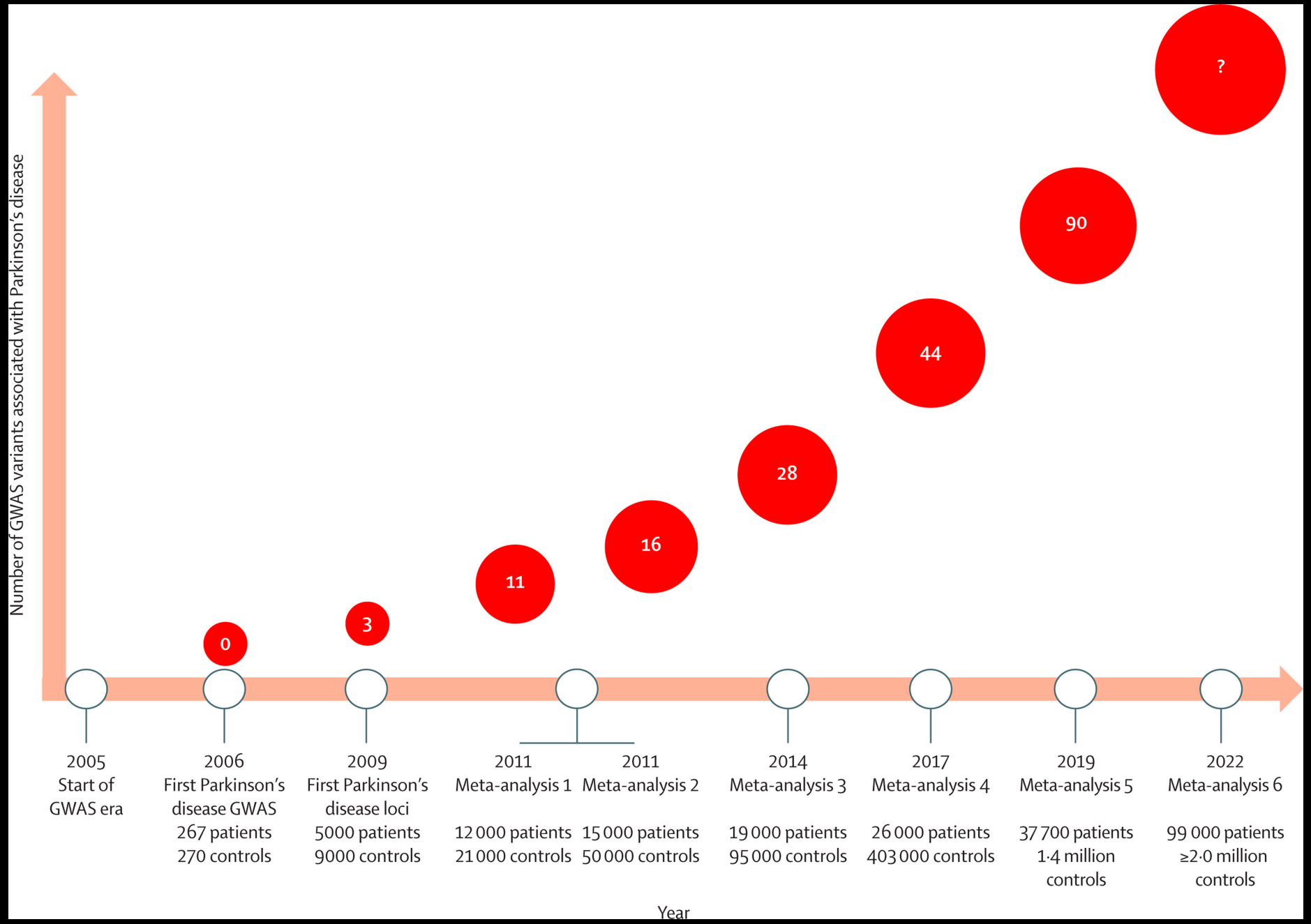
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

2006…
267 Parkinson's disease patients
270 Controls



2019…
37,700 cases Parkinson's disease patients
1,400,000 Controls



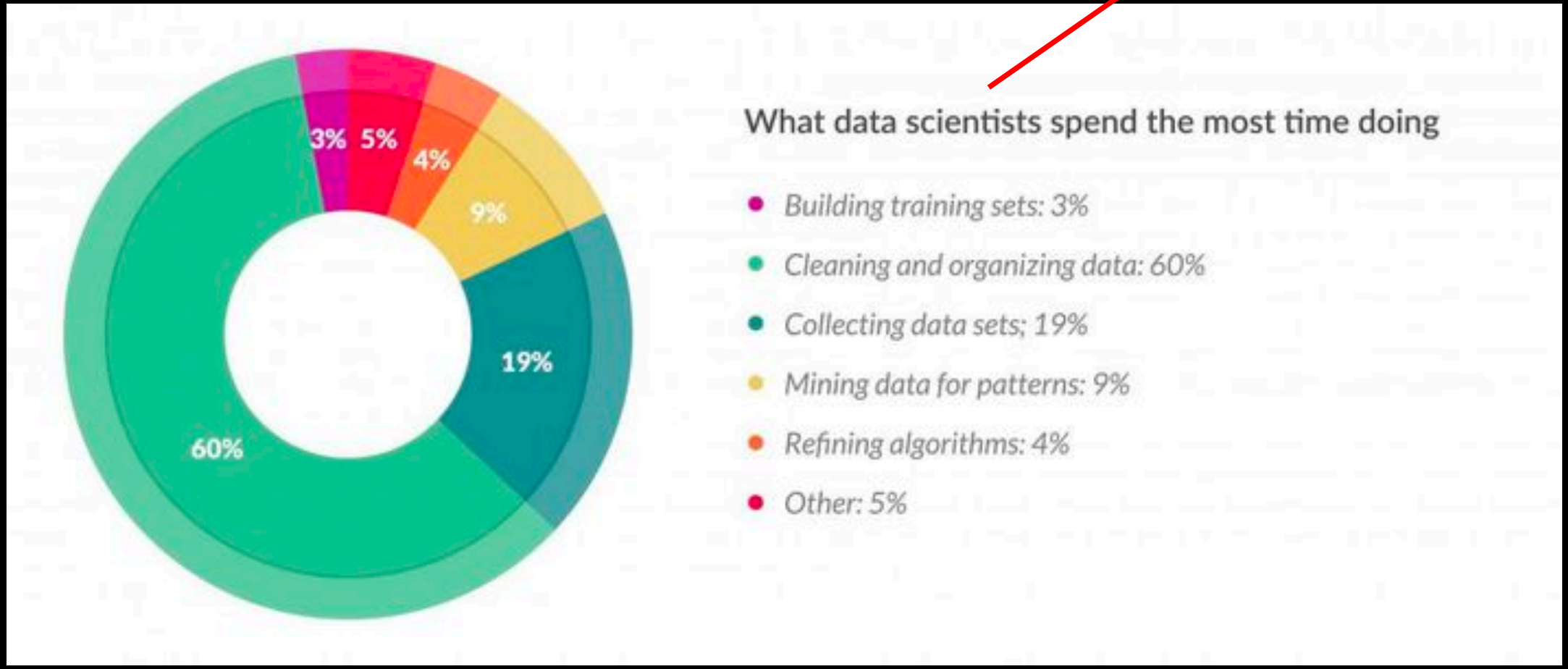

Number of GWAS variants associated with Parkinson's disease
0
3
11
16
28
44
90
?
2005
Start of GWAS era
2006
First Parkinson's disease GWAS
267 patients
270 controls
2009
First Parkinson's disease loci
5000 patients
9000 controls
2011
Meta-analysis 1
12 000 patients
21 000 controls
2011
Meta-analysis 2
15 000 patients
50 000 controls
2014
Meta-analysis 3
19 000 patients
95 000 controls
2017
Meta-analysis 4
26 000 patients
403 000 controls
2019
Meta-analysis 5
37 700 patients
1·4 million controls
2022
Meta-analysis 6
99 000 patients
≥2·0 million controls
Year

Important also to keep in mind

- How heritable do you expect your phenotype to be…
  - Some diseases might not be genetic
- The more common your disease the more likely to be a genetic factor to be involved?
  - Common diseases more genetics
- Also depends on how accurate your phenotype is…
  - Some diseases might have sub-groups?

Before we continue....

Or bioinformaticians



Genotyping, filtering, QC and imputation

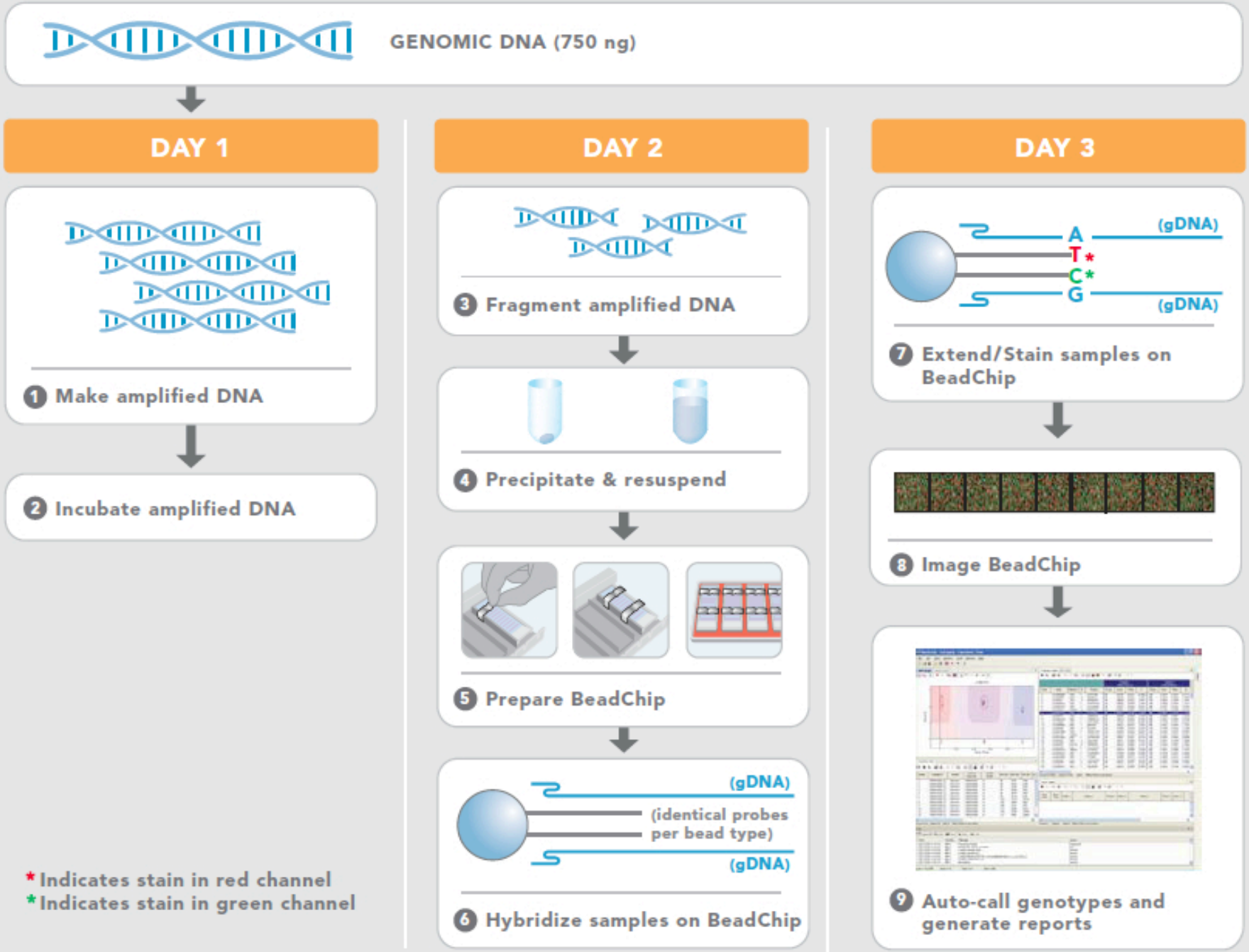
- Genotyping
  - Genotyping array, typically captures 500K-1Million variants
  - Why? Prize ~40$ much cheaper than genomes and exomes

- Filtering and QC
  - Identify low quality variants, samples and outliers
  - Why?

- Imputation
  - Impute variants based on reference panels
  - Why?

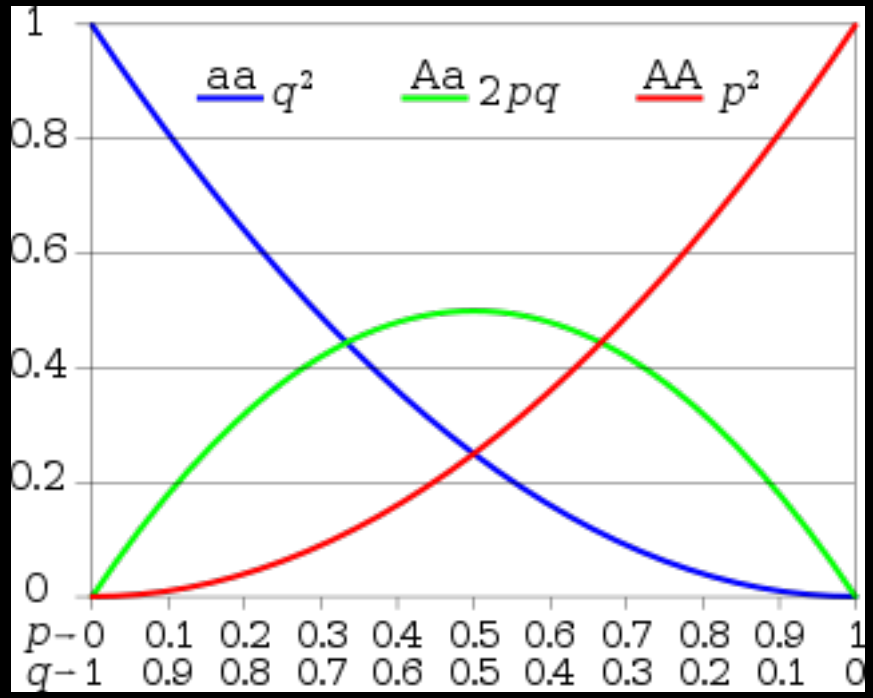
Data needed

FIGURE 1: INFINIUM II ASSAY PROTOCOL

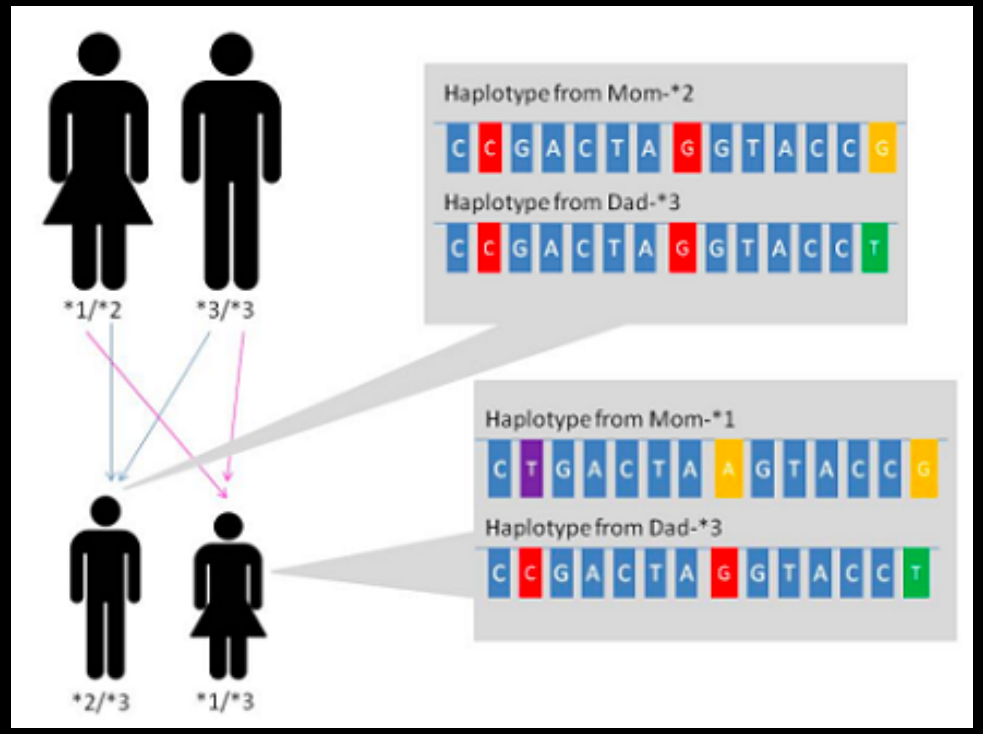


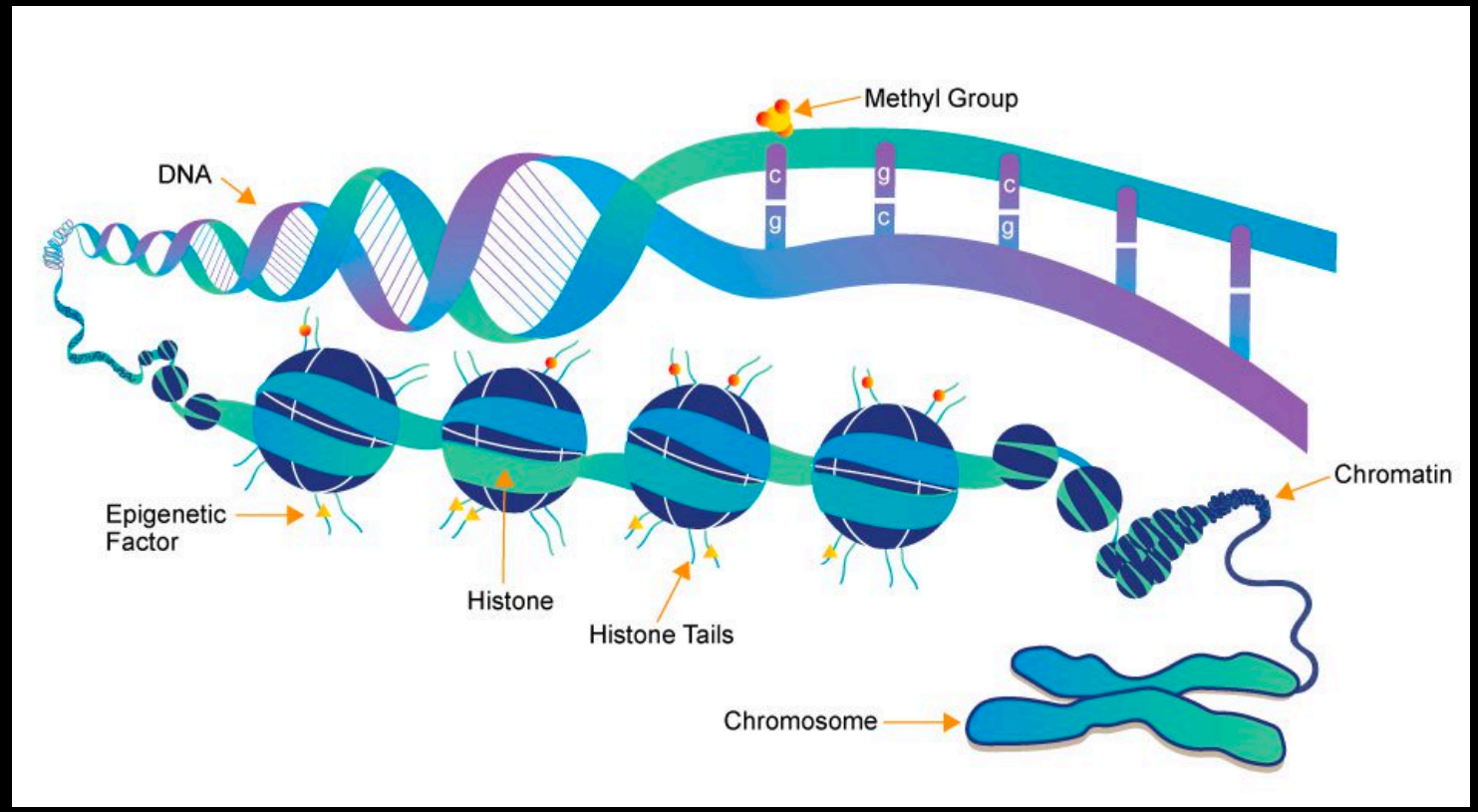
Filtering and QC

- Sample QC filtering
  - Sex mismatches
  - Heterozygosity rate
  - Call rate
  - (Optional Relatedness)
  - (Optional Ancestry)
- Variant QC filtering
  - Missingness between cases and controls
  - Hardy-Weinberg

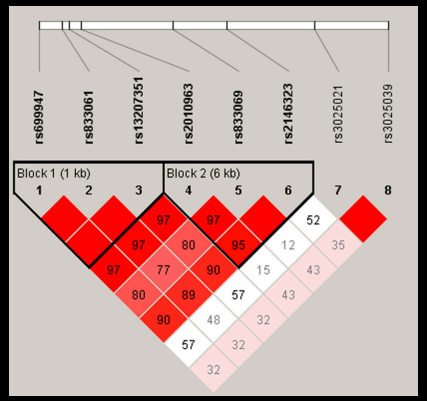


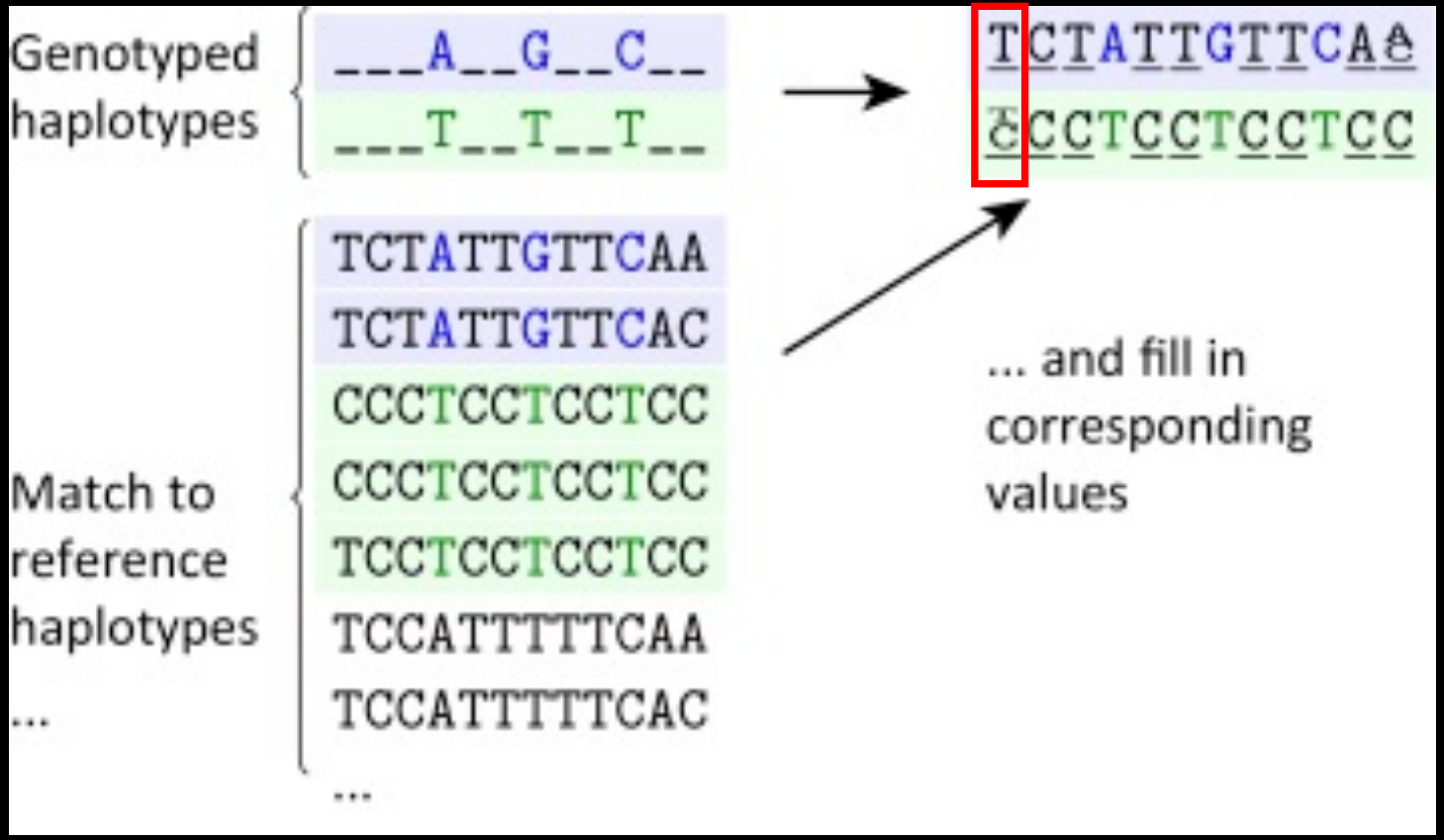
Imputation (filling in the gaps….)





Imputation (filling in the gaps….)





- Current reference panel
- The Haplotype Reference Consortium (HRC)
- Based on 36K European individuals

http://www.haplotype-reference-consortium.org/

Programs and Data formats

- Genome-studio (Illumina program to process genotypes)
- PLINK (Main genotype process/analysis software)
- GCTA (Related PLINK program)
- RVTESTS (Main GWAS/burden software)
- R (General plotting and figure language)

Programs and Data formats

- PLINK
  - Old format
    - .map (simple variant info file)
    - .ped (variant per individual file uncompressed)
  - Compressed format
    - .bim (simple variant info file)
    - .fam (simple phenotype info file)
    - .bed (compressed variant per individual file)
- RVTESTS
  - .vcf files (compressible variant per individual file)
    - Option to add more information like imputation quality

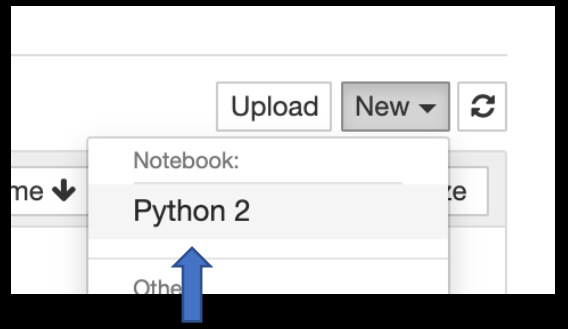
Lets try some exercises

- Go to:

https://notebook.hail.is/ and use Class key: isia2019

and click start… this is a Jupiter Notebooks session

Then go to right top



click

- Also go to: https://github.com/neurogenetics/Peru_course

Exercise 1

- Check “Part1: Clean GWAS data....”

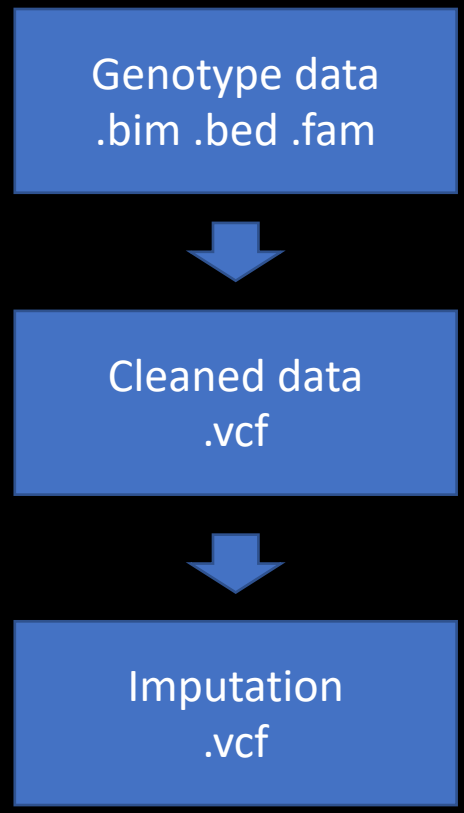
Part 1: Cleaning Data for GWAS

Preparing data for a genome-wide association study (GWAS) is a two step process

1. Clean the data
2. Harmonize the data with reference panel
    - No time for this today, but you can check the `STEP2_cleaning.sh` script if interested in how this is done

Step 1: Brief Overview of What the Scripts Do

1. `sh STEP1_cleaning.sh INPUTFILE` does the following:



Output

- Call rate outlier
- Heterozygosity rate outlier
- Sex mismatches (based on number of heterozygous SNPs on chrX)
- Hapmap plots
  - European
  - Asian
  - African
- Removes related individuals
- Final file = cleaned and QC’ed plink file (FILTERED.test_data.*)

Missing step2…

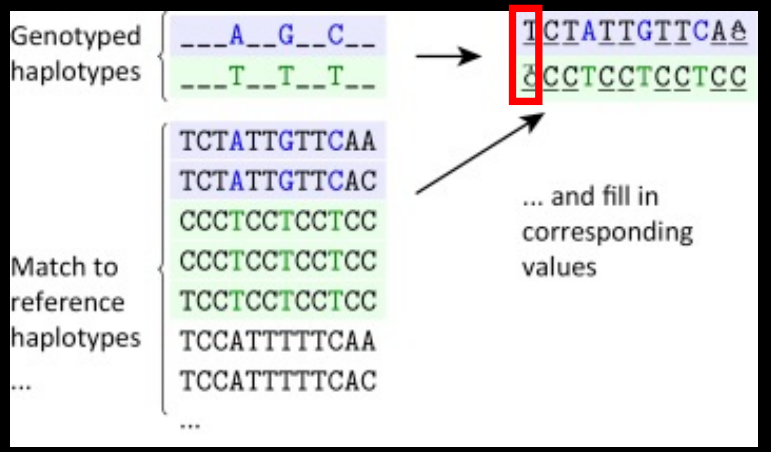
- Takes too long ~30 minutes

- Renames and reformats all the variants to the same format as the imputation panel…

Michigan Imputation Server

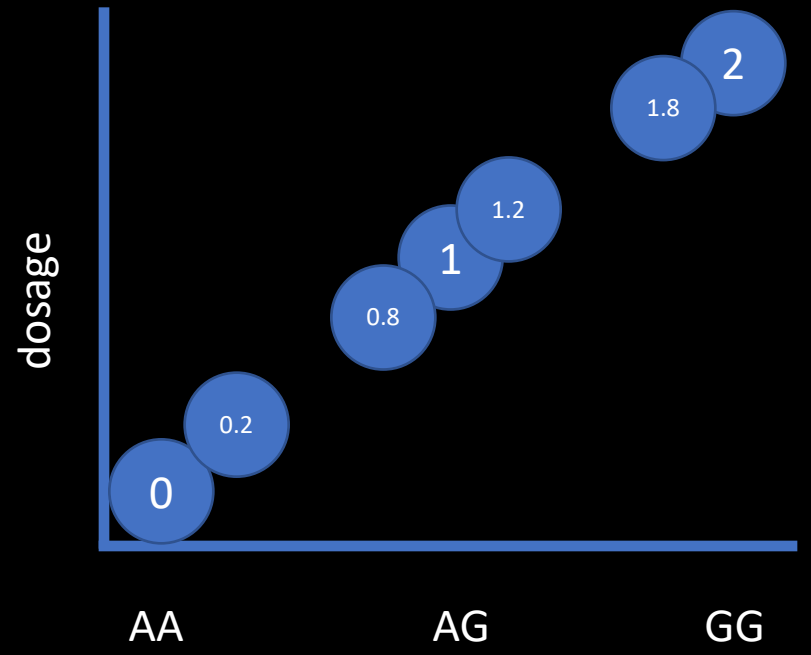
- https://imputationserver.sph.umich.edu/index.html

- Output:
  - dosage vcf file….
  - Imputation info file



Dosage vcf files??

- 0 = reference
- 1 = heterozygous
- 2 = homozygous alternative allele



Imputation quality measurement R2 stored in .info file

R2>0.8 high confidence (hardcall)

R2>0.3 moderate confidence (softcall)

.dose.vcf.gz file

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	3274_3274
7	14808	7:14808	T	C	.	PASS	AF=0.00011;MAF=0.00011;R2=0.00003	GT:DS:GP	0|0:0.000:1.000,0.000,0.000
7	15064	7:15064	T	C	.	PASS	AF=0.00124;MAF=0.00124;R2=0.00048	GT:DS:GP	0|1:1.002:0.000,1.000,0.000
7	20963	7:20963	C	G	.	PASS	AF=0.00038;MAF=0.00038;R2=0.00012	GT:DS:GP	1|1:2.001:0.000,0.001,1.000
7	20987	7:20987	T	C	.	PASS	AF=0.00050;MAF=0.00050;R2=0.00039	GT:DS:GP	0|0:0.001:0.999,0.001,0.000
7	21018	7:21018	G	A	.	PASS	AF=0.00018;MAF=0.00018;R2=0.00032	GT:DS:GP	0|1:0.501:0.999,0.001,0.000
7	30939	7:30939	T	C	.	PASS	AF=0.00055;MAF=0.00055;R2=0.00015	GT:DS:GP	0|0:0.001:0.999,0.001,0.000
7	31017	7:31017	G	C	.	PASS	AF=0.00020;MAF=0.00020;R2=0.00077	GT:DS:GP	0|0:0.000:1.000,0.000,0.000
7	31039	7:31039	G	T	.	PASS	AF=0.00046;MAF=0.00046;R2=0.00035	GT:DS:GP	0|0:0.001:0.999,0.001,0.000

GT = genotype
DS = dosage
GP = genotype posterior probabilities

.info.gz file

SNP	REF(0)	ALT(1)	ALT_Frq	MAF	AvgCall	Rsq	Genotyped	LooRsq	EmpR	EmpRsq	Dose0	Dose1
4:10229	T	C	0.00008	0.00008	0.99992	0.00103	Imputed	-	-	-	-	-
4:10408	T	C	0.00008	0.00008	0.99992	0.0007	Imputed	-	-	-	-	-
4:11689	C	T	0.00062	0.00062	0.99938	0.08511	Imputed	-	-	-	-	-
4:11720	G	A	0.00014	0.00014	0.99986	0.00005	Imputed	-	-	-	-	-
4:154894	A	G	0.00087	0.00087	0.99999	0.99126	Genotyped	0.494	0.999	0.9981	0.91938	0.00069
4:155415	G	A	0.00404	0.00404	0.99973	0.93652	Genotyped	0.098	0.998	0.99506	0.13028	0.00018
4:155655	C	G	0.00953	0.00953	0.99985	0.98432	Genotyped	0.335	0.917	0.84067	0.72925	0.01073
4:265547	A	G	0.01728	0.01728	0.99981	0.98886	Genotyped	0.433	0.942	0.88708	0.76483	0.01266
4:265955	T	C	0.0043	0.0043	0.99995	0.98868	Genotyped	0.518	0.918	0.84223	0.69849	0.0018

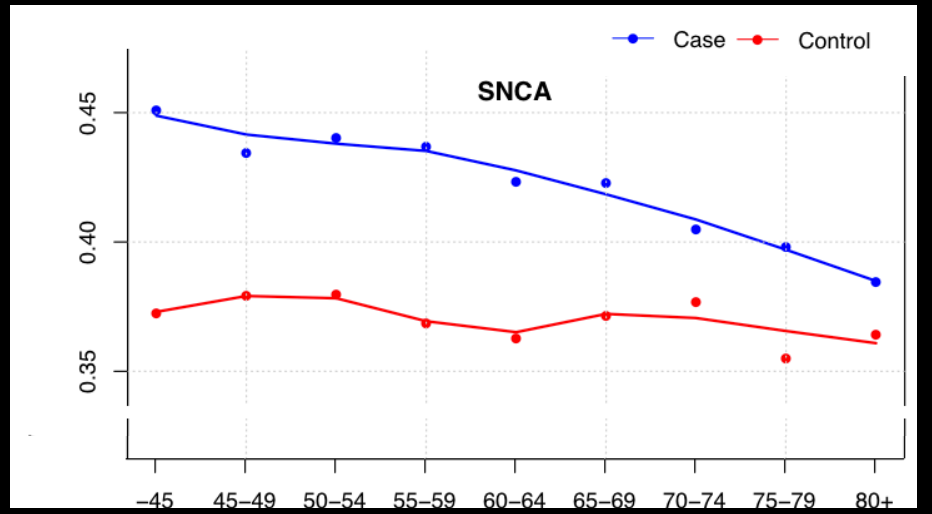
GWAS Simple concept

Formula = Disease ~ SNP + covariates

SNP = dosage of allele (0,1 or 2)

- Case-control example…
  - PD, SNCA SNP rs356203 frequency cases=0.4242, controls=0.3676
  - AD, APOE E4 allele SNP rs429358 frequency cases=~0.50, controls=~0.20

- Continuous phenotype examples
  - Age of onset PD, SNCA SNP rs356203
  - BMI value
  - Any blood level

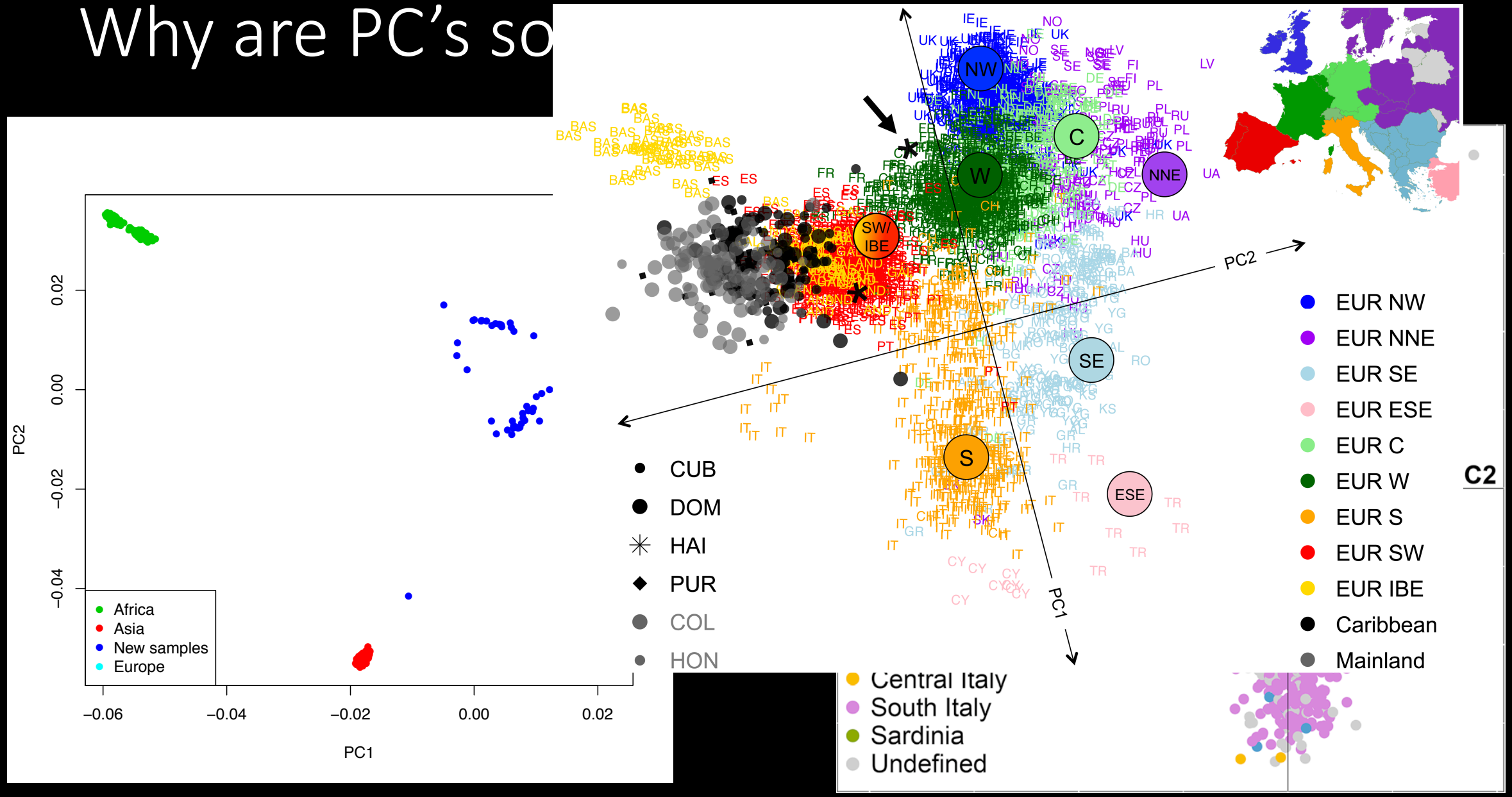


Covariates

- Very important….

- Sex
- Age (Age of onset, Age of recruitment, Birth year)
- Principal component (PC’s)
- Batch?
- Dataset?
- Site of sampling?
- Genotyping chip?

Why are PC's so



Exercise 2

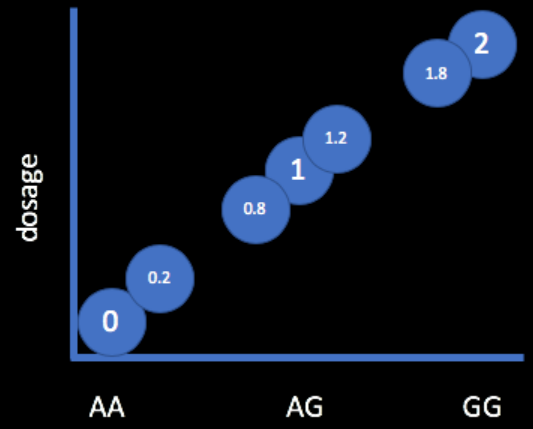
- Calculate principal components
- Check section:
- “Make principal components based on cleaned data...”

Step 6: Make Principle Components

Use PLINK to generate principle components on cleaned, pruned SNPs using the following script:

```
%%bash
# Go to the proper directory
cd GWAS_course_files/QC_PACKAGE/
plink --bfile FILTERED.EXAMPLE_DATA --maf 0.05 --geno 0.01 --hwe 5e-6 --autosome --exclude exclusion_reg
--make-bed --out pass1
```

Preparing for GWAS in RVTEST



- Prepare your phenotype and covariate files

FID	IID	Patid	Matid	Sex	Pheno	age	PC1	PC2	etc
sample1	sample1	0	0	1	2	65	0.5464	-1.454	..
sample2	sample2	0	0	2	1	48	0.4892	-1.351	..

- Check file “GWAS_course_files/GWAS/covariates.txt”

Exercise 3

- Check “Part2: Run GWAS....”

Part 2: Running a GWAS

- The GWAS will be run using the imputed data
  - We have no time to do the actual imputation during this demo
  - So we will use an example dataset
- ***Warning:*** Imputation imputes a lot of variants, many that might be not high quality

1. First, we create variant list R2>0.3
2. Run a GWAS on a small piece of the genome

Analyze output

```
%%bash
cd GWAS_course_files/GWAS/
sort -gk 9 EXAMPLE_DATA_GWAS.SingleWald.assoc | head
```

Inspect results file 4:90641340 is P=4.20E-06 = rs356220 which is one of the top variants for PD This variant has a beta of -0.29, which is on OR of ~1.3

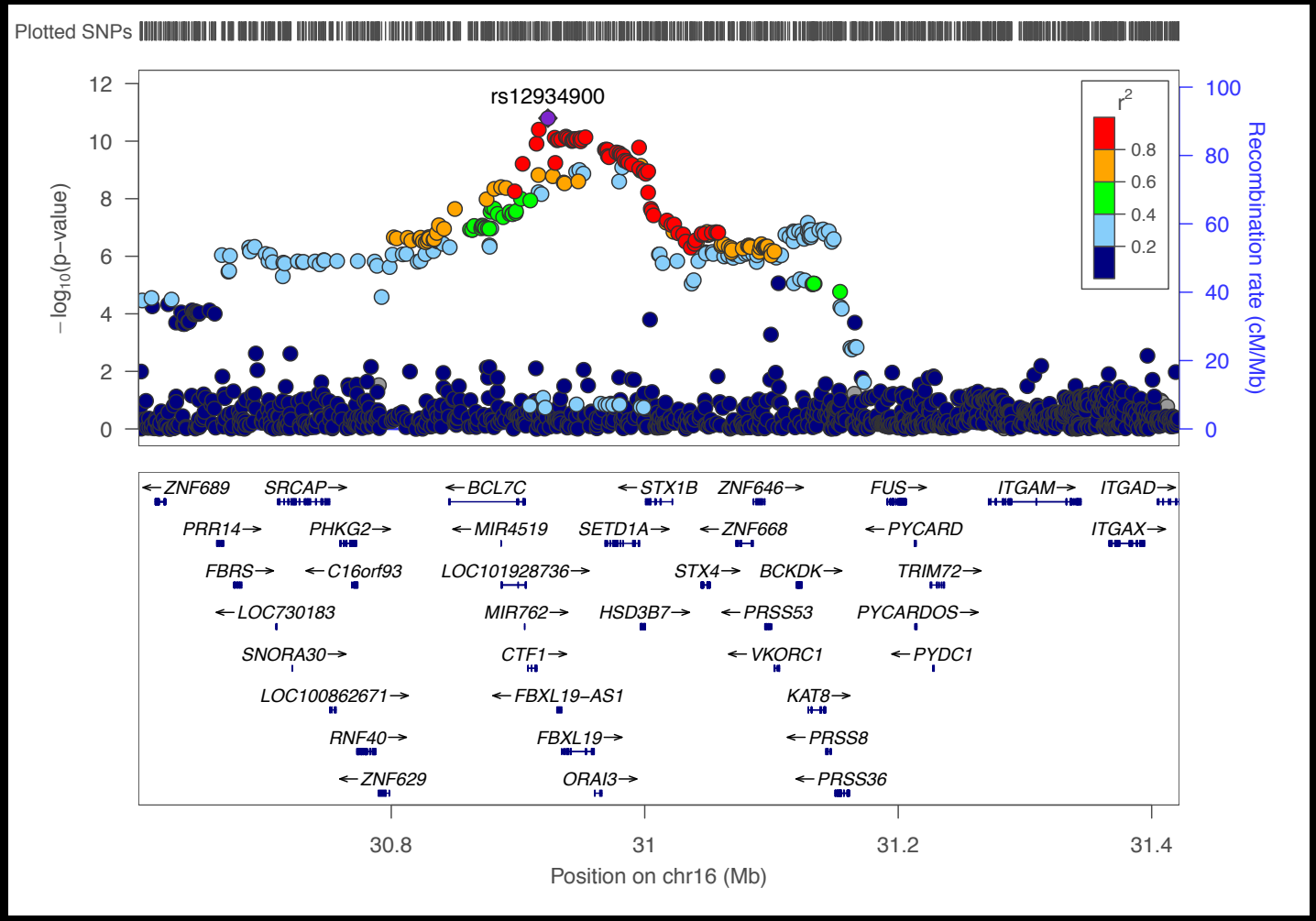
Locuszoom

http://locuszoom.org/

Very common way of plotting variants from a GWAS

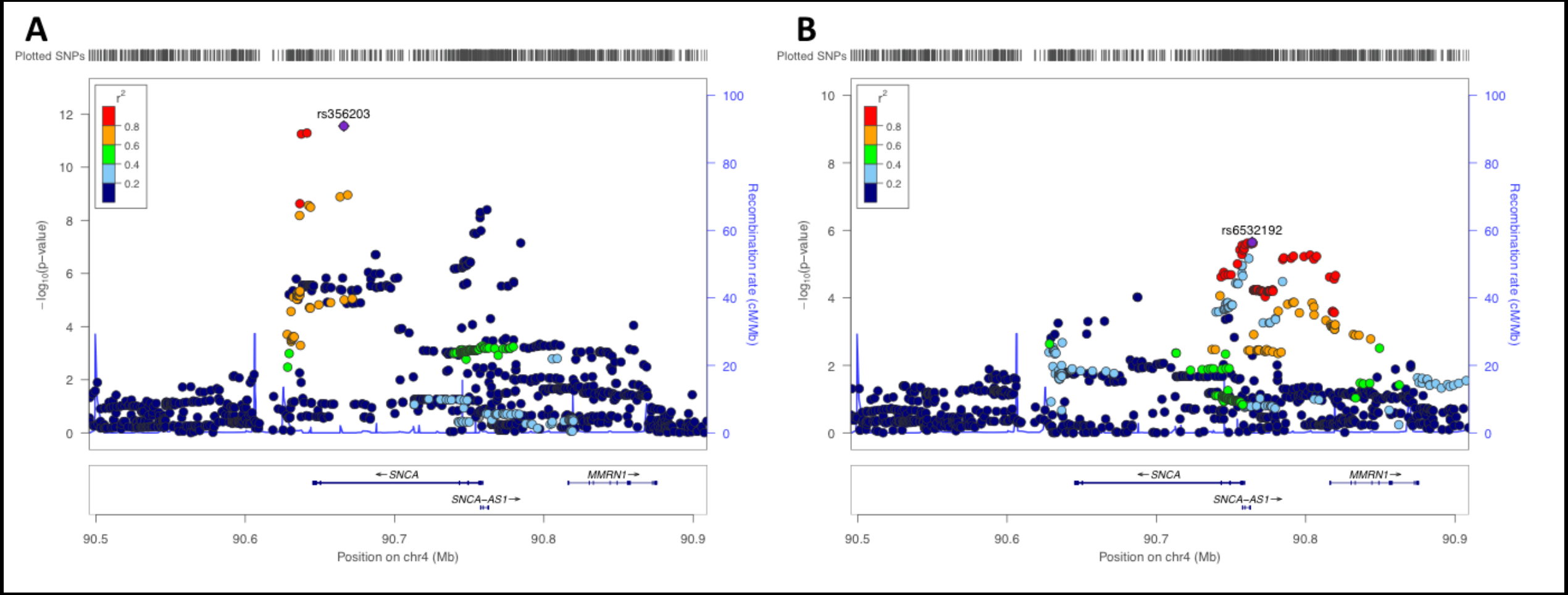
Only needed:

P-value and RS-ID (snp-name)



On GitHub you can find a file: locuszoom_example.txt.zip
Right click and “save link as” then upload to locuszoom, choose gene ”SNCA” and 200kb region

Independent hits in same region

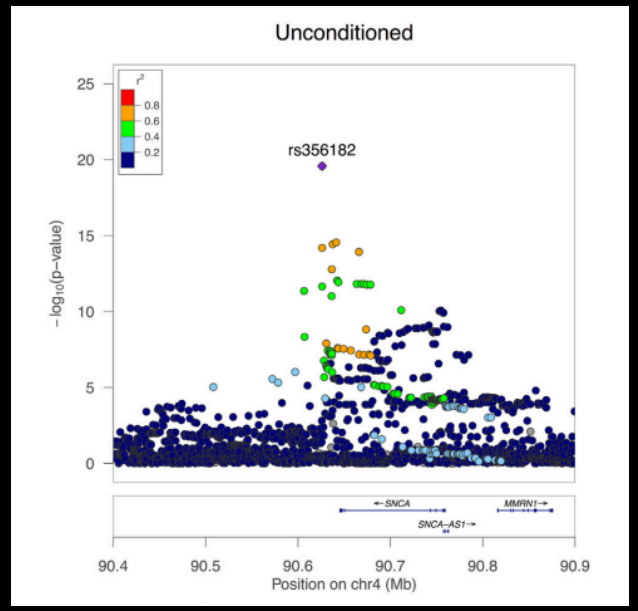


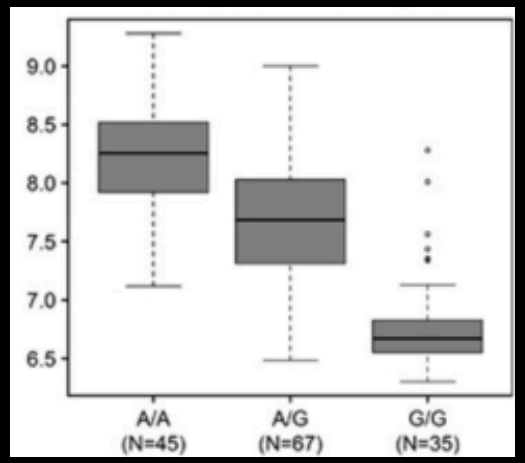
What to do with GWAS hits…

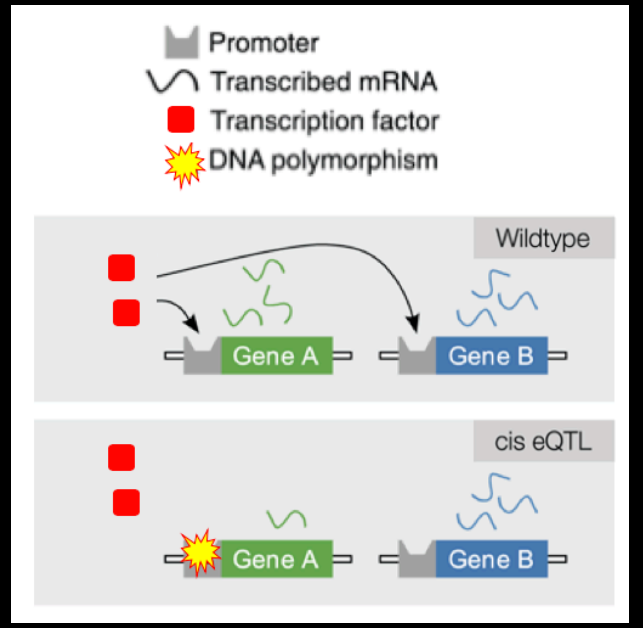
- Expression quantitative trait loci?
- Allele specific expression?
- Chromatin structure?
- Methylation?
- Structural variant?
- Repeat expansion?
- …

Example of GWAS locus 1

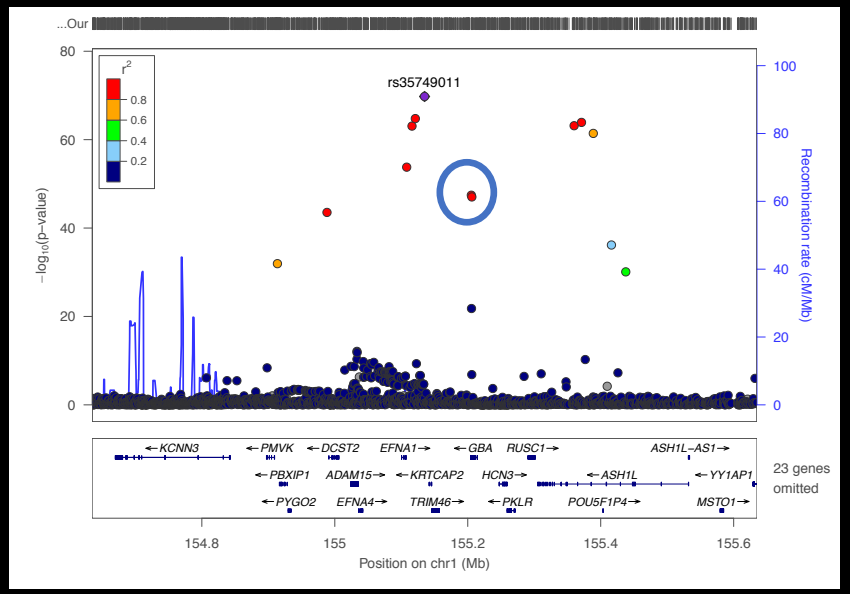
- eQTL



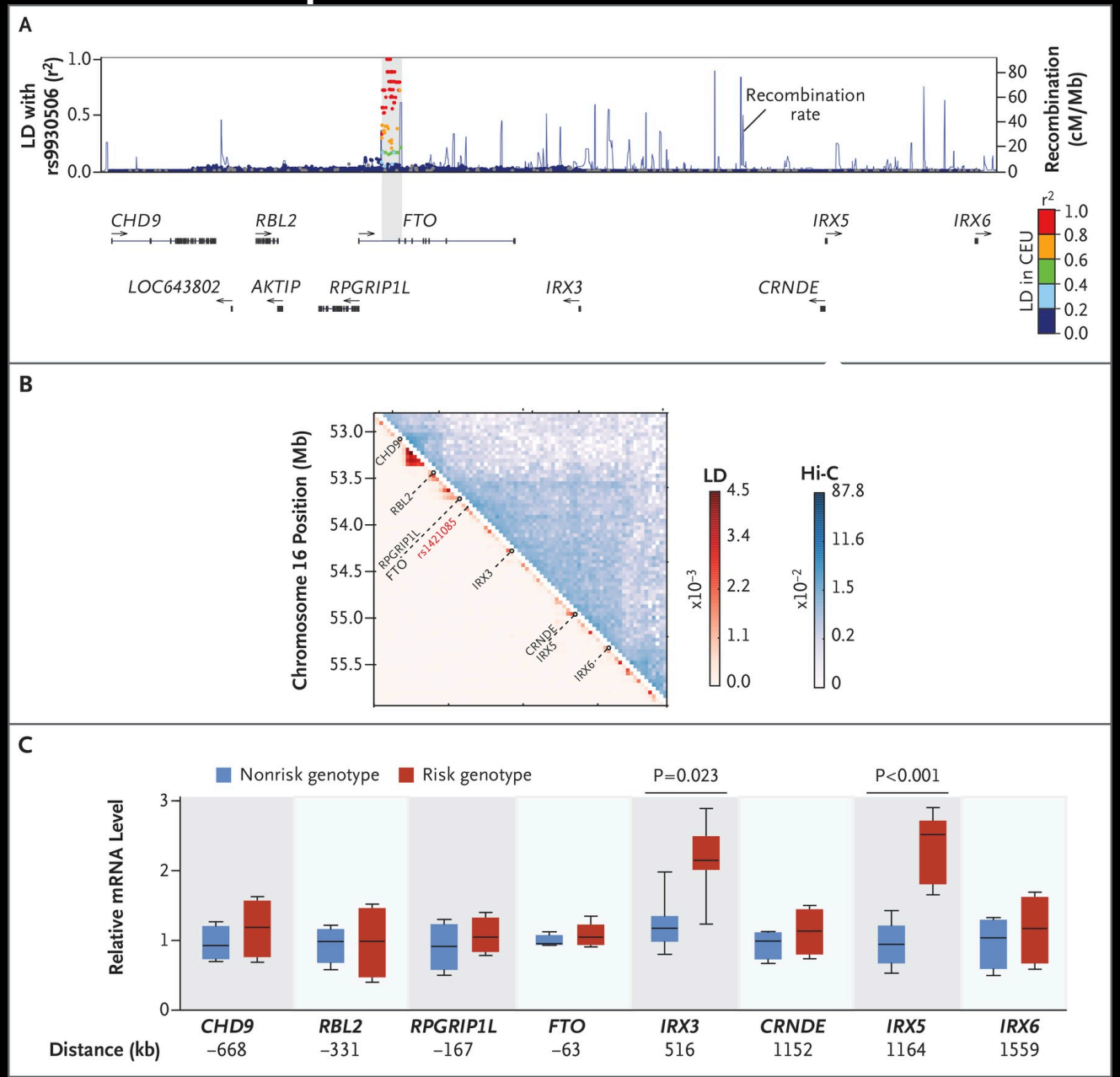


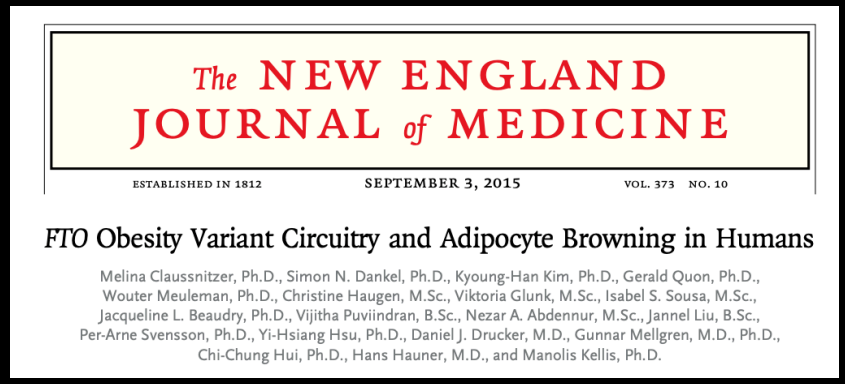


- Coding variant
  - Top snp in LD with *GBA* p.E326K
  - Coding variant is damaging



Example of GWAS locus 2




The NEW ENGLAND JOURNAL of MEDICINE
ESTABLISHED IN 1812 SEPTEMBER 3, 2015 VOL. 373 NO. 10

FTO Obesity Variant Circuitry and Adipocyte Browning in Humans

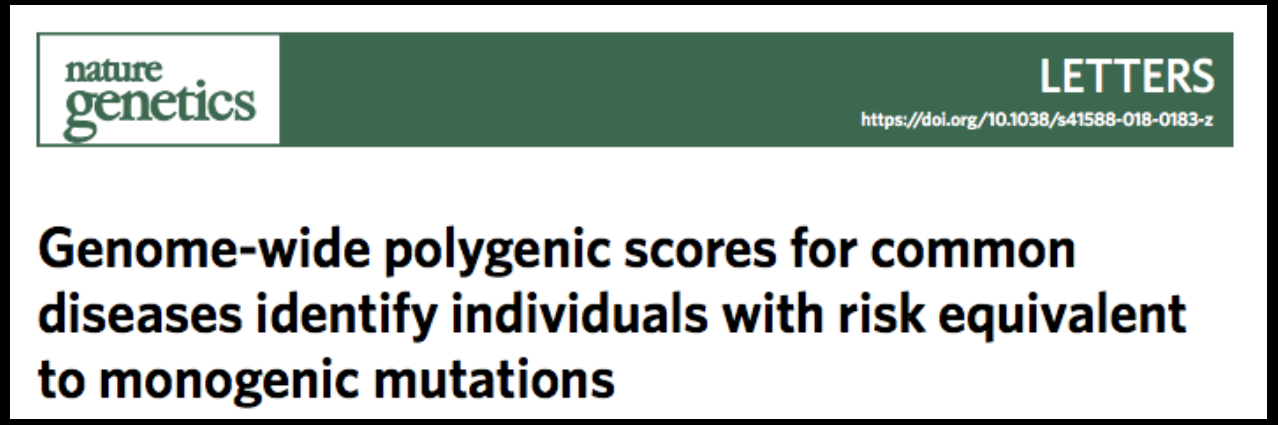
Melina Claussnitzer, Ph.D., Simon N. Dankel, Ph.D., Kyoung-Han Kim, Ph.D., Gerald Quon, Ph.D., Wouter Meuleman, Ph.D., Christine Haugen, M.Sc., Viktoria Glunk, M.Sc., Isabel S. Sousa, M.Sc., Jacqueline L. Beaudry, Ph.D., Vijitha Puviindran, B.Sc., Nezar A. Abdennur, M.Sc., Jannel Liu, B.Sc., Per-Arne Svensson, Ph.D., Yi-Hsiang Hsu, Ph.D., Daniel J. Drucker, M.D., Gunnar Mellgren, M.D., Ph.D., Chi-Chung Hui, Ph.D., Hans Hauner, M.D., and Manolis Kellis, Ph.D.

Claussnitzer - 2015

Genetic risk score (GRS)

- Also know as polygenic risk score (PRS)

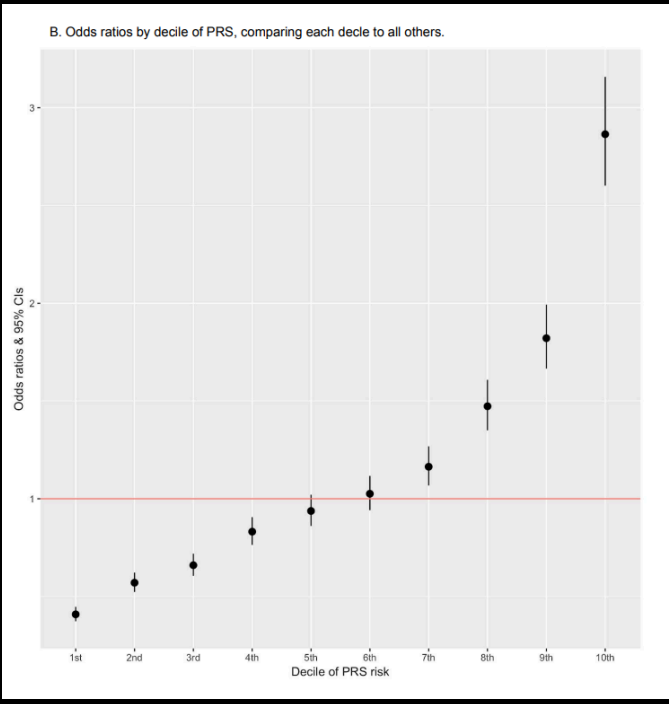
- Two general methods to do this
  - 1. simple way (we prefer the simple way ☺)
  - 2. complicated way



Parkinson's disease as example

- 90 independent loci…
- Each loci increase risk for PD with small amount
- Adding all of these increase risk alleles together creates a risk score

- Allele scoring, see --score in PLINK



Exercise 4

- Check “Part3: Run genetic risk score....”

Part 3: Generating a Genetic Risk Score (GRS)

In this section, you will calculate a GRS

Step 1: Going to the Proper Directory

Check overview of files that are in the folder:

```
%%bash
# Go to the proper directory
cd GWAS_course_files/GRS/
ls
```

What does the output mean?

- You get a score per individual….

- Check for example:

```
%%bash
cd GWAS_course_files/GRS/
head NEUROX_GRS.profile
```

- Logistic regression for association with covariates
- Boxplot to show differences

Burden testing (typically on rare coding variants)

- Very simple explanation…
  - Collapsing/merging of rare variants per gene
  - E.g. 5000 patients and 5000 controls
  - 100 rare (different) variants in cases and 5 in controls for gene A
  - Means that 2% of the patient carriers a rare variant in this gene and 0.1% of the controls, which would be significant.
- Commonly used in exome or genome sequencing data which mixed results due to typically limited power due to low sample sizes
- Note that there are many different ways/models and filtering steps for burden testing

Exercise 5

- Check “Part4: Run burden test...”

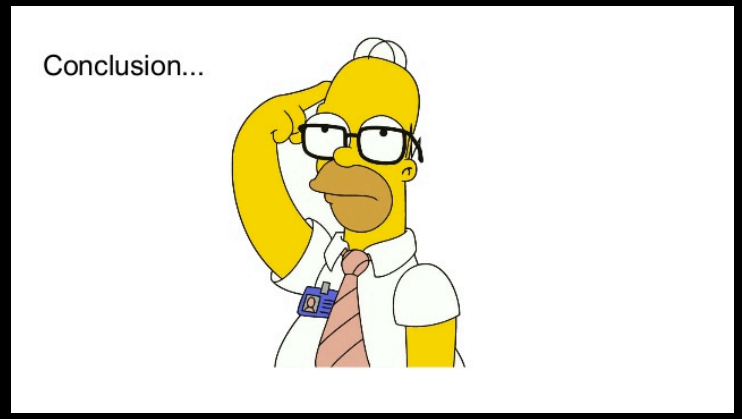
Part 4: Run a Burden Test

In preparation for this demo, we already created some files to make it easier, and this is how it was created:

```
## Subset only the variants of interest in this case 3 GBA variants that have been associated with Parki

# plink --bfile NEUROX --extract GBA_BURDEN_variants.txt --make-bed --out NEUROX_GBA
```

Questions...



https://github.com/neurogenetics

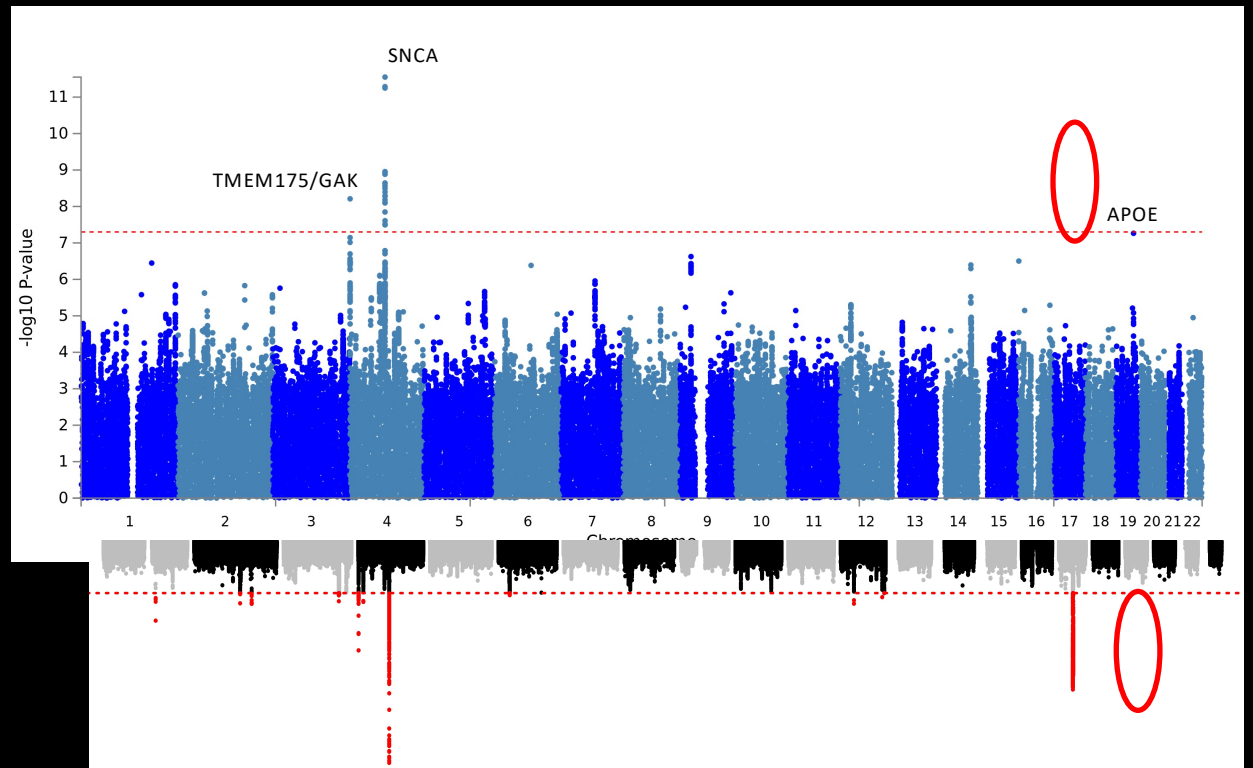
Feel free to contact us:

cornelis.blauwendraat@nih.gov

sara.bandresciga@nih.gov

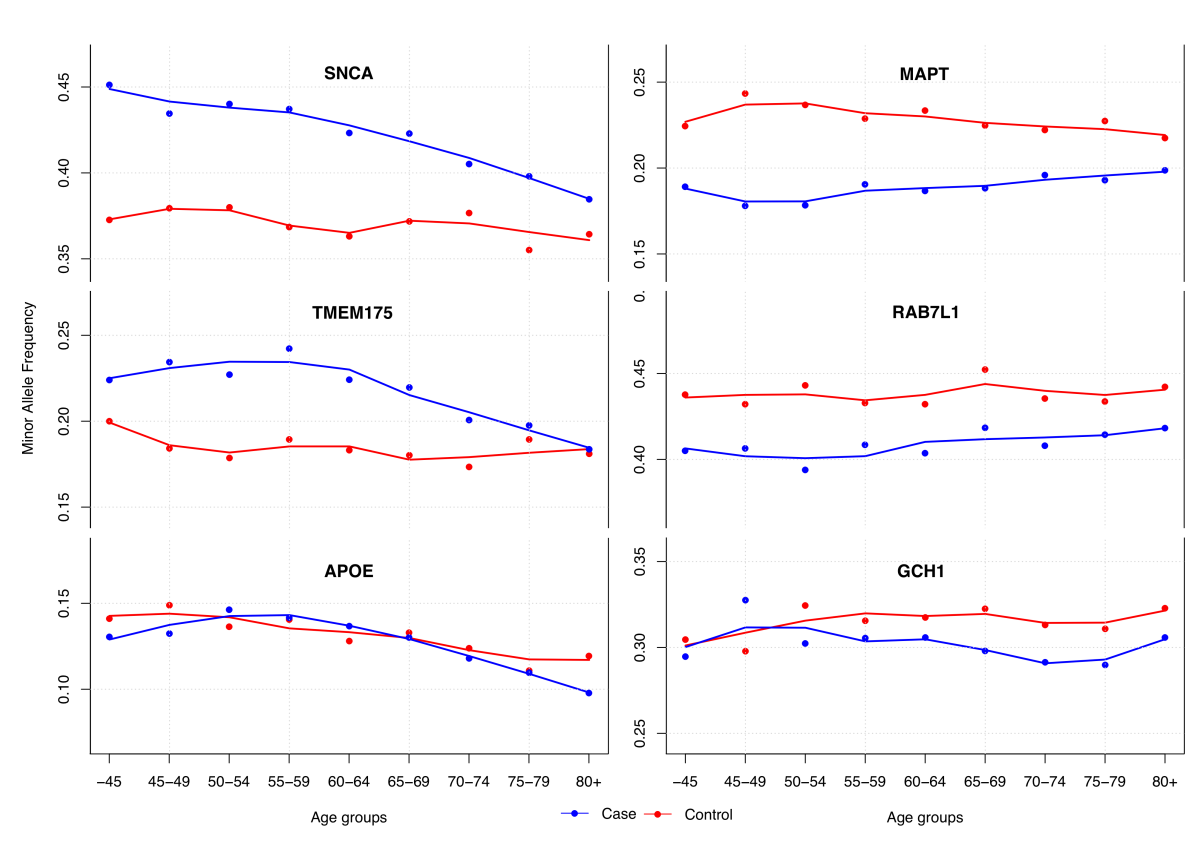
PD age of onset GWAS

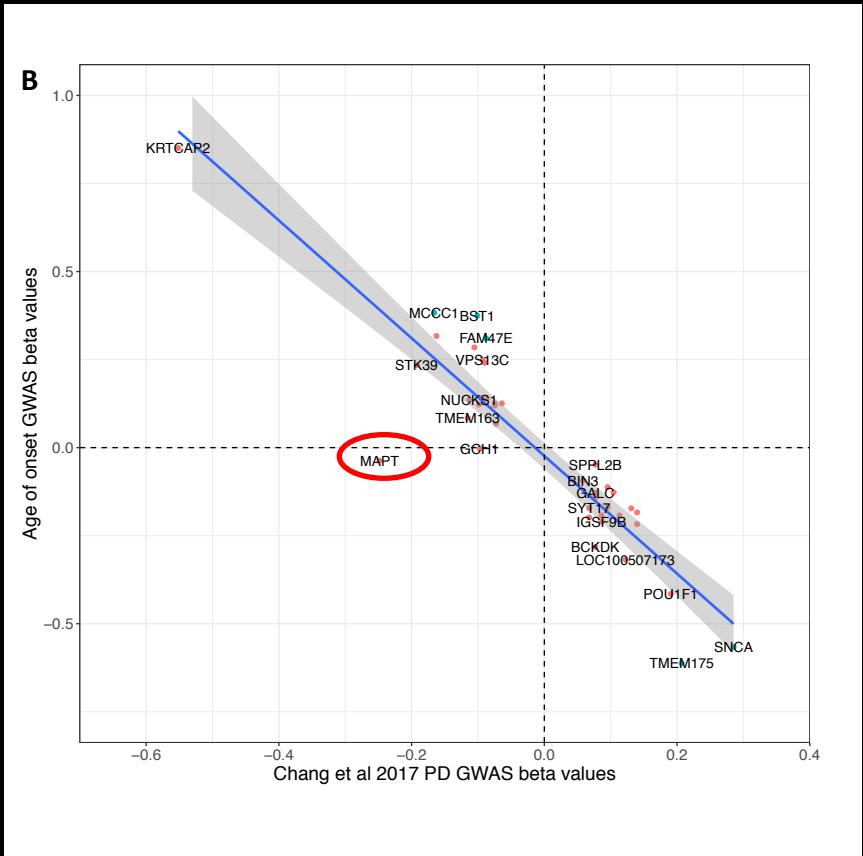
- 3 hits… less than expected?
- SNCA, kind of expected…
- TMEM175, known PD risk loci
- APOE, kind of expected as general aging marker



PD age of onset GWAS

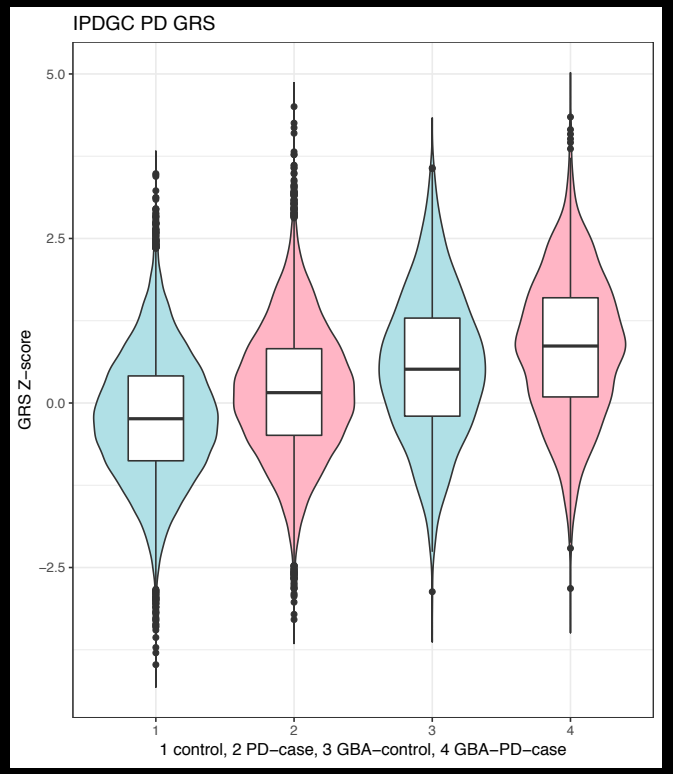
Some solid risk GWAS hits show no association with age of onset at all…

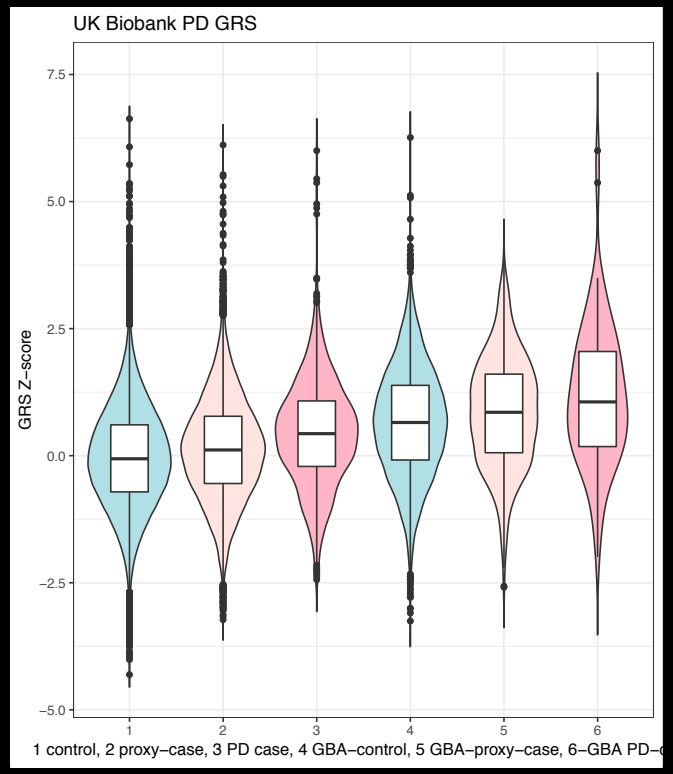




Genetic risk scoring

- Using the 90 GWAS variants
- Clear and highly significant differences between cases and controls





Prediction of PD?

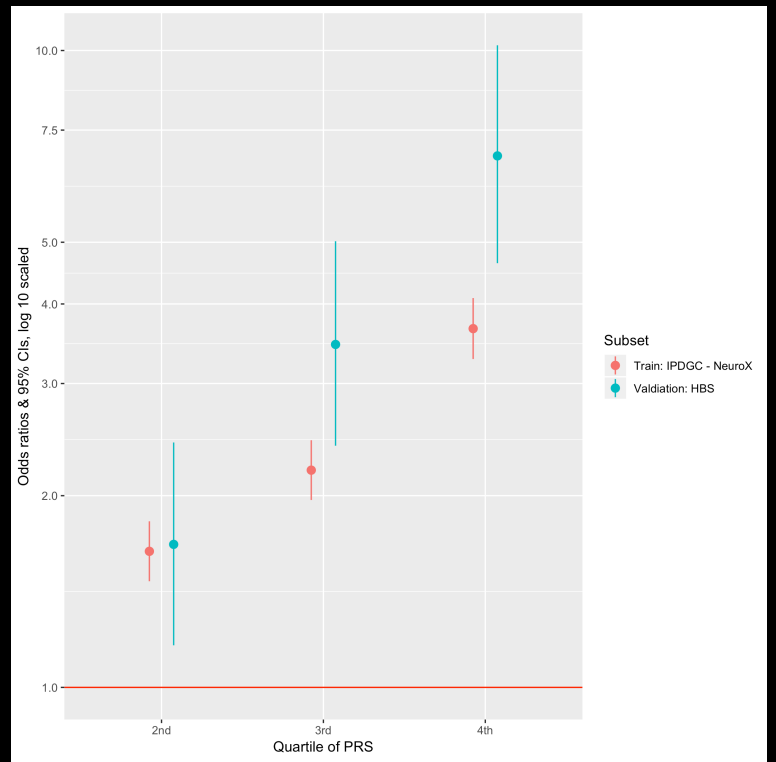
nature genetics

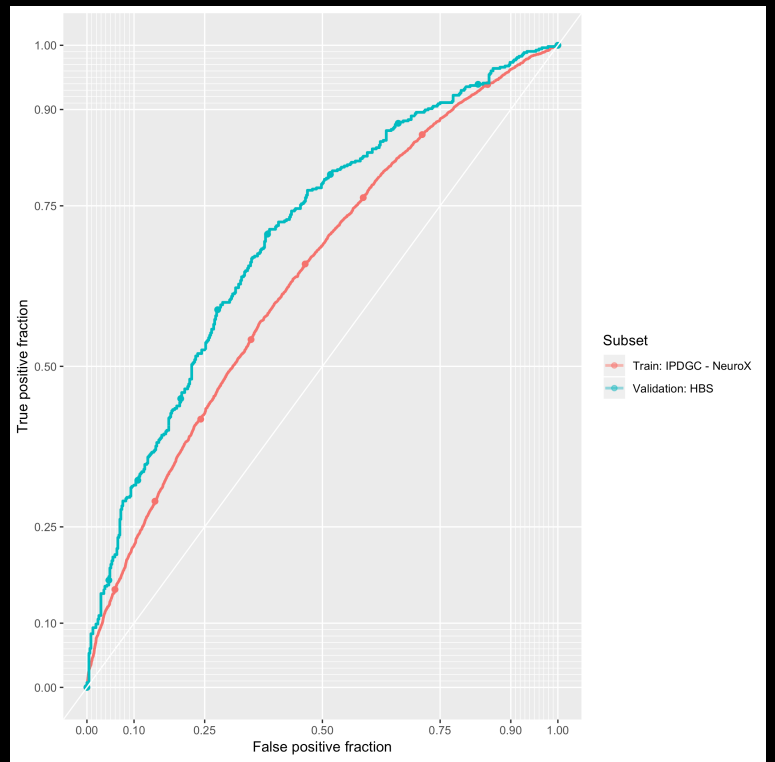
LETTERS

https://doi.org/10.1038/s41588-018-0183-z

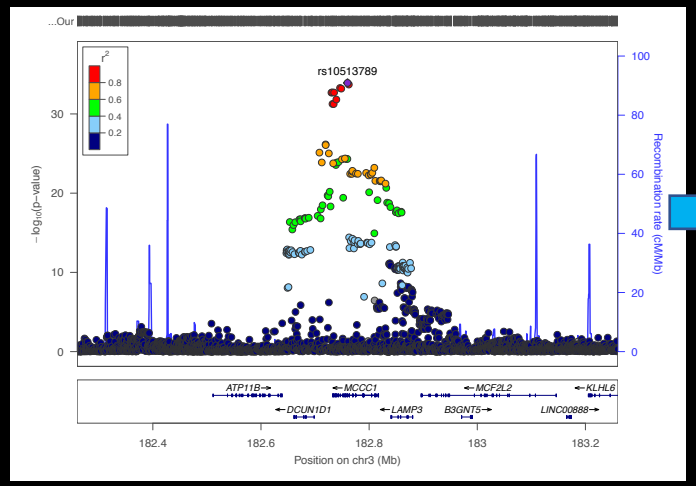
Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations

Amit V. Khera[1,2,3,4,5], Mark Chaffin[4,5], Krishna G. Aragam[1,2,3,4], Mary E. Haas[4], Carolina Roselli[4], Seung Hoan Choi[4], Pradeep Natarajan[2,3,4], Eric S. Lander[4], Steven A. Lubitz[2,3,4], Patrick T. Ellinor[2,3,4] and Sekar Kathiresan[1,2,3,4*]



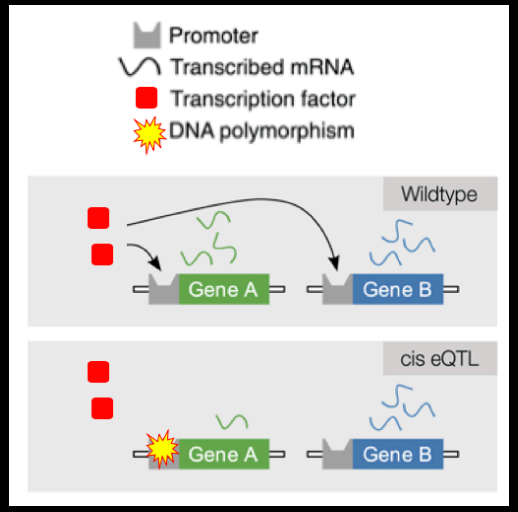


From GWAS loci to functional gene/variant



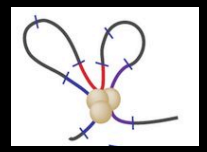
Functional effect?

- eQTL



Candidate gene(s)

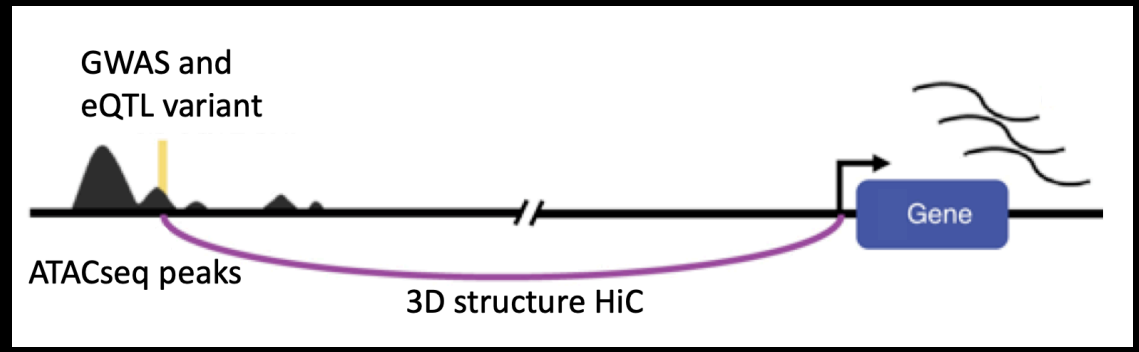
GWAS locus resolved



3D structure
HiC sequencing

\+

Open chromatin
ATACseq

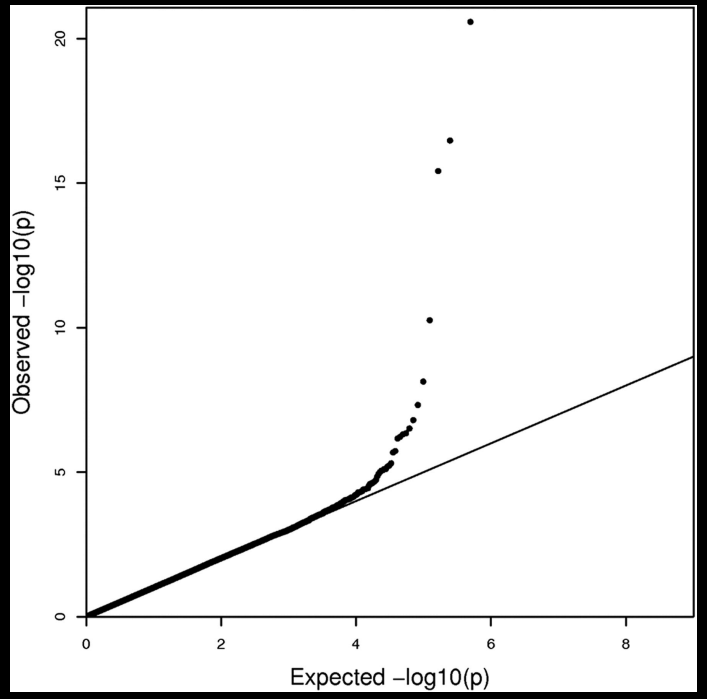


Candidate variant
CRISPR follow-up

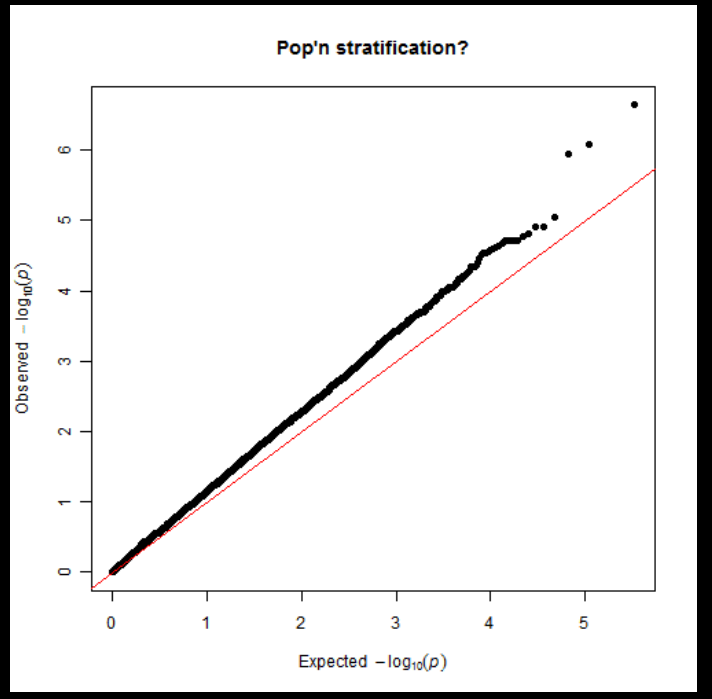
QQ-plot and lambda value

- See folder GWAS_more script QQplot.R

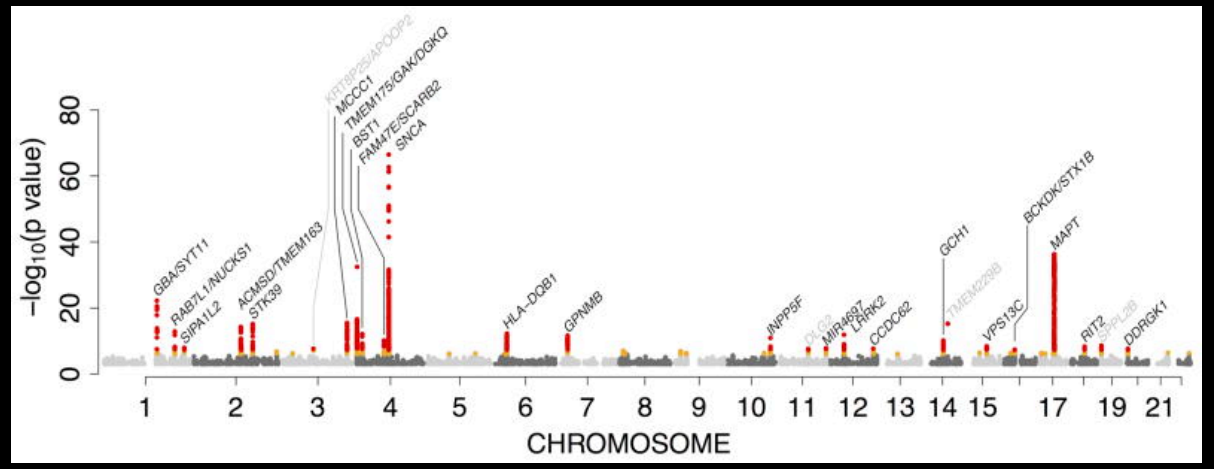
- Good example



Bad example



Manhattan plot



- See script Manhattan.R
- Or google manhattan plot in R, many many examples…
- Or go to http://fuma.ctglab.nl/ and they do everything for you…