This session

1 *Family based association analyses: Introduction*
(Camelia Minică)
2 *Genetic association test: Plink and R*
(Jenny van Dongen)
3 *Apply the biometrical model to the empirical results*
(Dorret Boomsma)

Exercises from this paper: *Effect of the IL6R gene on IL-6R concentration*

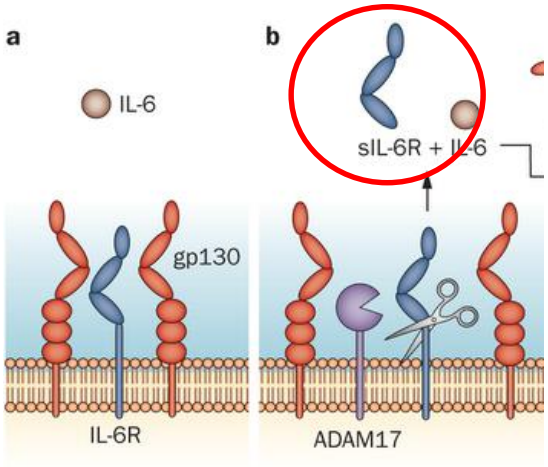
Behav Genet (2014) 44:368–382
DOI 10.1007/s10519-014-9656-8

ORIGINAL RESEARCH

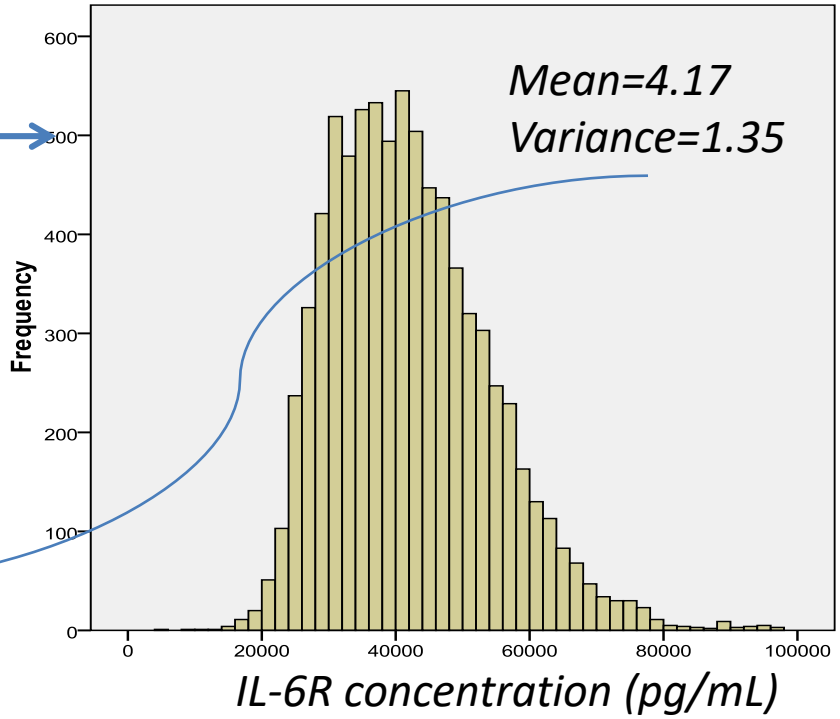
The Contribution of the Functional *IL6R* Polymorphism rs2228145, eQTLs and Other Genome-Wide SNPs to the Heritability of Plasma sIL-6R Levels

Jenny van Dongen · Rick Jansen · Dirk Smit · Jouke-Jan Hottenga · Hamdi Mbarek · Gonneke Willemsen · Cornelis Kluft · AAGC Collaborators · Brenda W. J. Penninx · Manuel A. Ferreira · Dorret I. Boomsma · Eco J. C. de Geus

- We measured soluble IL-6R concentration in blood in ~5000 individuals (from the Netherlands Twin Register)



- sIL-6R concentration in blood is a **quantitative trait**



Estimated in Mx

Genetics → IL-6R concentration → common disease

- IL-6R protein is encoded by the ***IL6R* gene** (chromosome 1)

- *IL6R* gene important for **several common diseases**
  - Asthma[1]
  - Coronary heart disease[2]
  - Type 1 diabetes[3]

[1]Ferreira M.A. *et al* Lancet 2011
[2]*IL6R* consortium Lancet 2012
[3]Ferreira R.C. *et al* PLoS Genetics 2013

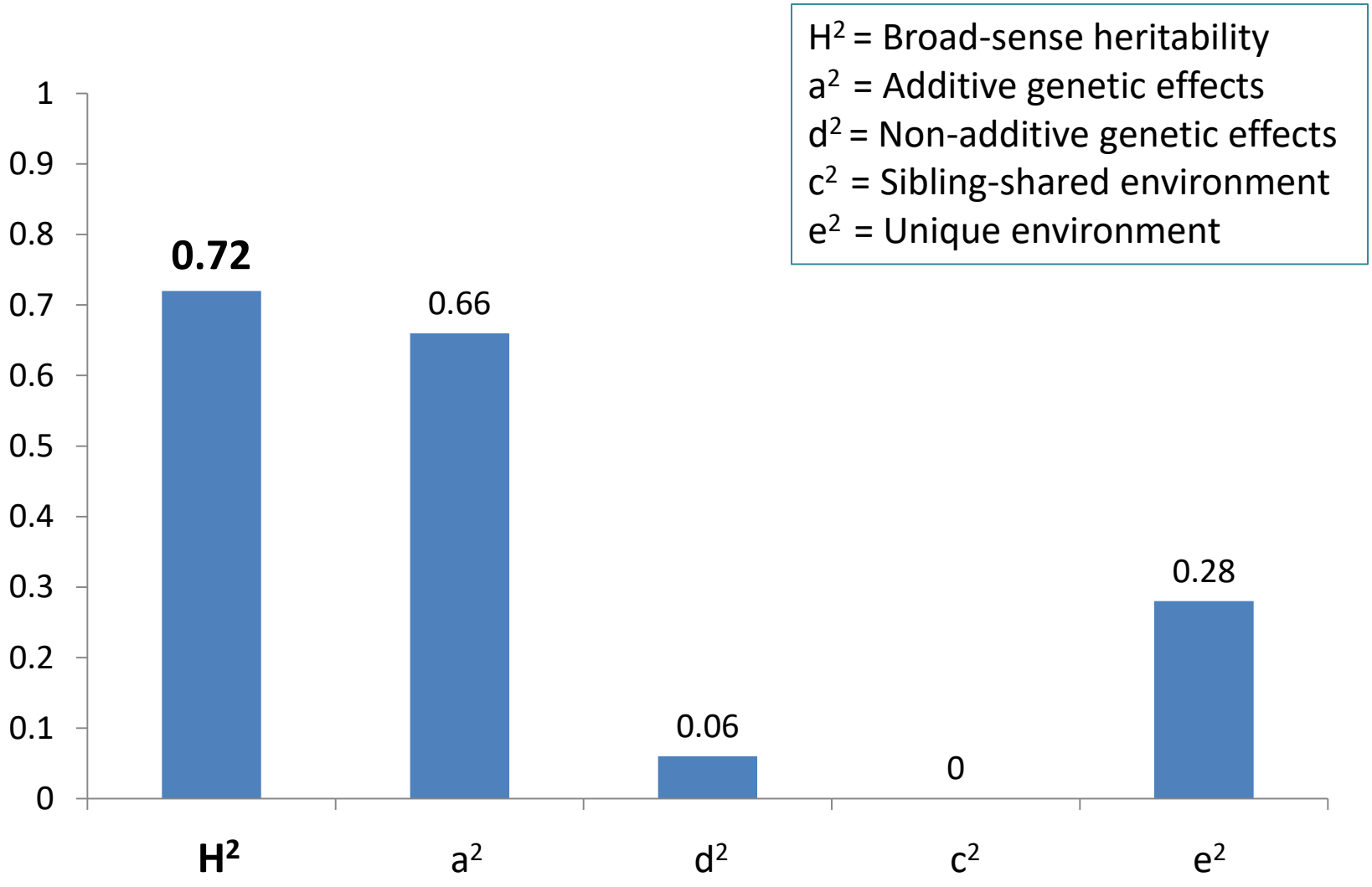
Analysis	N subjects	Mean age (SD), min-max	% Male	Cohort
Heritability analysis and biometrical model (MZ and DZ twins, siblings, and parents)	4980	42.7 (14.3), 18-89	36.2	NTR
GWA and GCTA (unrelated + related Ss)	4846	44.2 (14.4), 18-90	38.7	NTR
GCTA (unrelated Ss)	2875	46.5 (14.4), 18-89	38.8	NTR
Combined linkage and association analysis (Nuclear families)	1254	48.3 (15.7), 18-89	44.4	NTR
eQTL analysis (unrelated + related Ss)	4467	38.4 (13.0), 25-51	34.4	NTR + NESDA
Correlation between sIL-6R level and *IL6R* expression (unrelated + related Ss)	2727	37.5 (12.0), 18-79	34.5	NTR

Methods

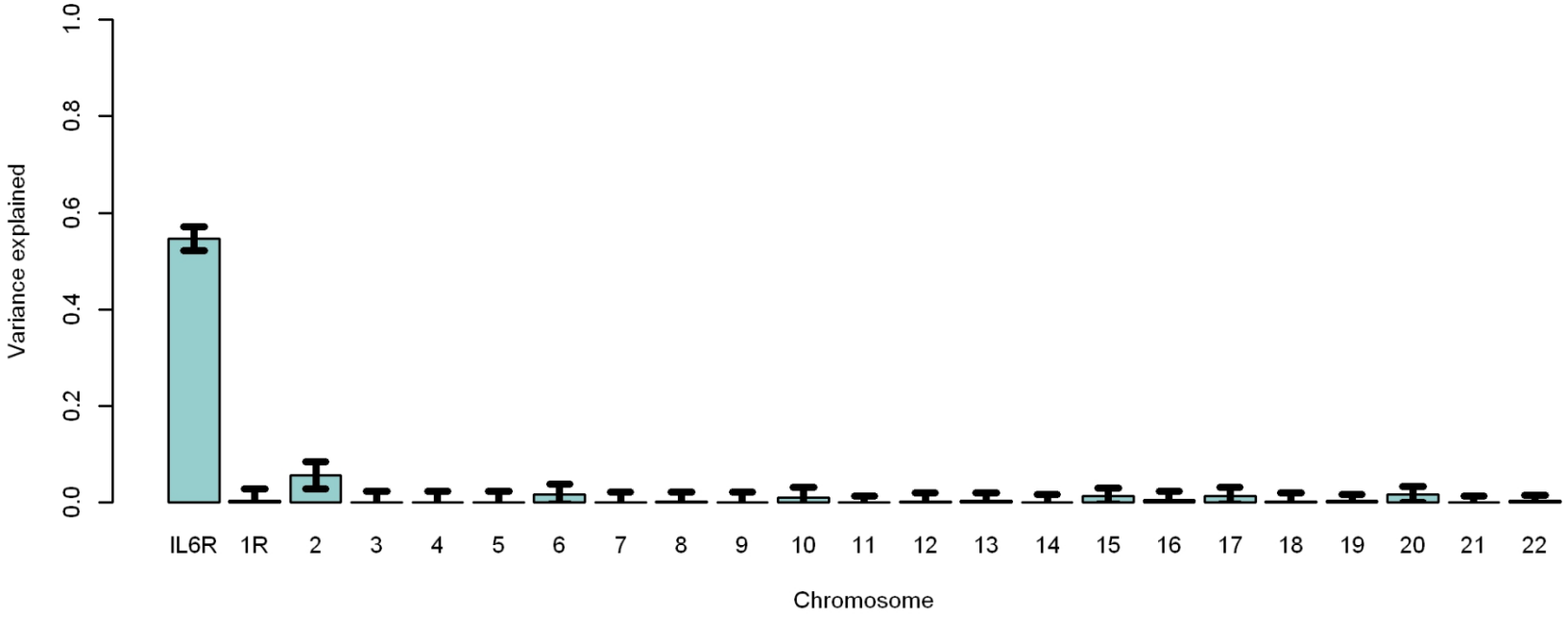
- We measured IL-6R concentration in ~5000 **twins & parents & siblings**
- We estimated **Heritability:** Variance of sIL-6R level explained by total genetic effects (Mx)

- We measured genome-wide SNP genotypes of the same subjects:
  - How much variance is explained by **all SNPs in the genome** (Genomewide-complex trait analysis, GCTA)
  - How much variance is explained by **all genetic variation in the *IL6R* gene** (linkage analysis)
  - How much variance is explained by **the SNP rs2228145**

Heritability of sIL-6R level (twin-family data)



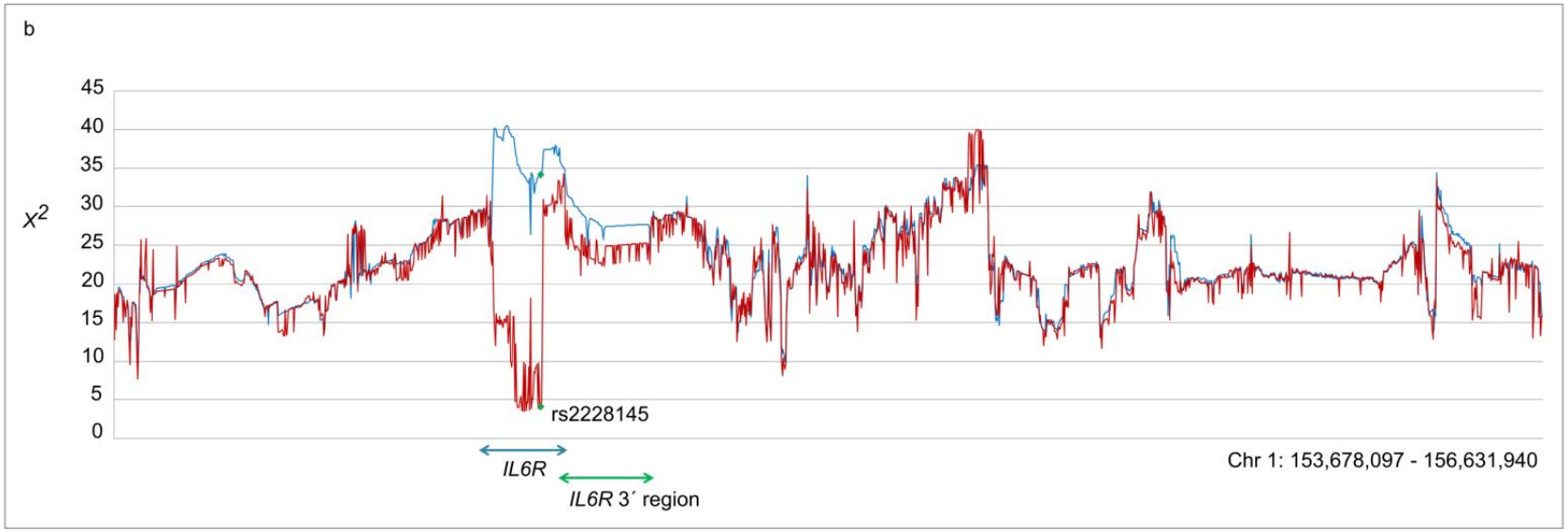
Variance explained by chromosome-wide SNPs (GCTA)



SNPs in the *IL6R* gene on Chromosome 1 (+/- 10MB): **54.7 %** (SE=2.5%)

Combined linkage and association analysis (qtdt)

- Chi-squared from linkage test
- Chi-squared from linkage test – while modeling association for individual SNPs



Chr1: (q21.3)

IL6R region:

1. Variance explained by linkage (V_A/V_{total}): **69 %**
2. Variance explained by linkage after correction for rs2228145: **19%**

Thus, we had twin – family data -> heritability
-> linkage

However, when looking at association, we need to adjust for clustering in the data.

Common Variant family-based GWAS (clustered data)

Camelia Minica

Conor Dolan **Dorret Boomsma**

VU VRIJE UNIVERSITEIT AMSTERDAM Faculteit der Psychologie en Pedagogiek

European Journal of Human Genetics (2014), 1–7

www.nature.com/ejhg

ARTICLE

Sandwich corrected standard errors in family-based genome-wide association studies

Camelia C Minică*,[1], Conor V Dolan[1], Maarten MD Kampert[2], Dorret I Boomsma[1] and Jacqueline M Vink[1]

Molecular Psychiatry (2014), 1–2

www.nature.com/mp

LETTER TO THE EDITOR

MZ twin pairs or MZ singletons in population family-based GWAS? More power in pairs

Why is this important?

Ignoring clustering in the data may lead to wrong conclusions (point estimates of effects OK, but SE too small)

Focus: family-based Genome-Wide Association Studies
However: these are regression based approaches, hence relevant for any analysis involving family data
Predictors: GV, polygenic score, other covariates

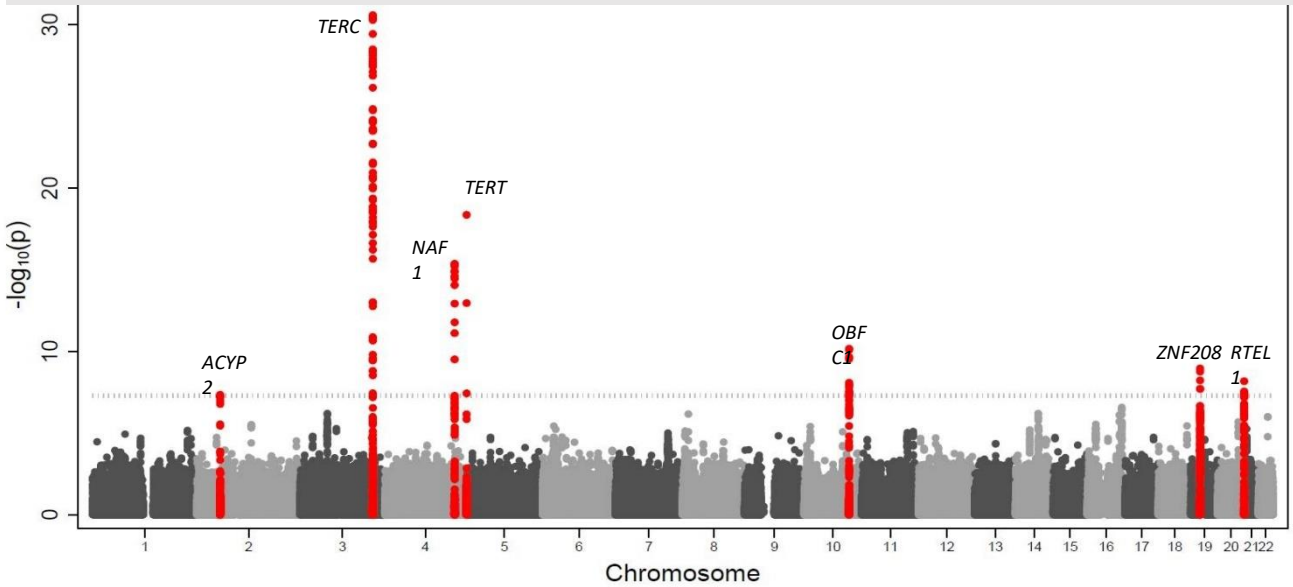
Why is this important?

- Many GWAS meta-analyses rely heavily on twin registries
- Twin registries have data collected in families readily available

Identification of seven loci affecting mean TELOMERE length and their association with disease

Veryan Codd et al. (ENGAGE consortium) *Nature Genetics*, 2013

Twin registries supplied 34% of samples



Nature genetics
Author Manuscript
Europe PMC Funders

Genome-wide meta-analysis identifies new susceptibility loci for migraine

Verneri Anttila, Bendik S. Winsvold, [...], and Aarno Palotie

Study	Cases	Controls
ALSPAC	3,134	5,103
Australia	1,683	2,383
B58C	1,165	4,141
deCODE	2,139	34,617
ERF	330	1,216
Finnish MA	1,032	3,513
FinnTwin	189	580
German MA	997	1,105
German MO	1,208	2,564
HUNT	1,608	1,097
LUMINA MA	820	4,774
LUMINA MO	1,118	2,016
NFBC1966	757	4,399
NTR&NESDA	282	2,260
Rotterdam	351	1,647
TWINS UK	972	3,837
WGHS	5,122	18,108
Young Finns	378	2,065

13% cases
9% controls

Sciencexpress Reports

GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment

There are 6 twin cohorts and total of 52 cohorts (11%)

- ***Finnish twin cohort***
- ***Netherlands twin register***
- ***QIMR (Australian twin register)***
- ***Swedish twin register***
- ***TwinsUK***
- ***Minnesota Twin – family study***

Twin registries supplied > 35% of total sample size

Some consortia protocols require discarding family members

Molecular psychiatry
Author Manuscript NIH Public Access

A mega-analysis of genome-wide association studies for major depressive disorder

Twin registries supplied 31% cases and 19% controls

UNRELATEDS

MZ pairs

or

MZ singletons?

MZ pairs or MZ singletons?

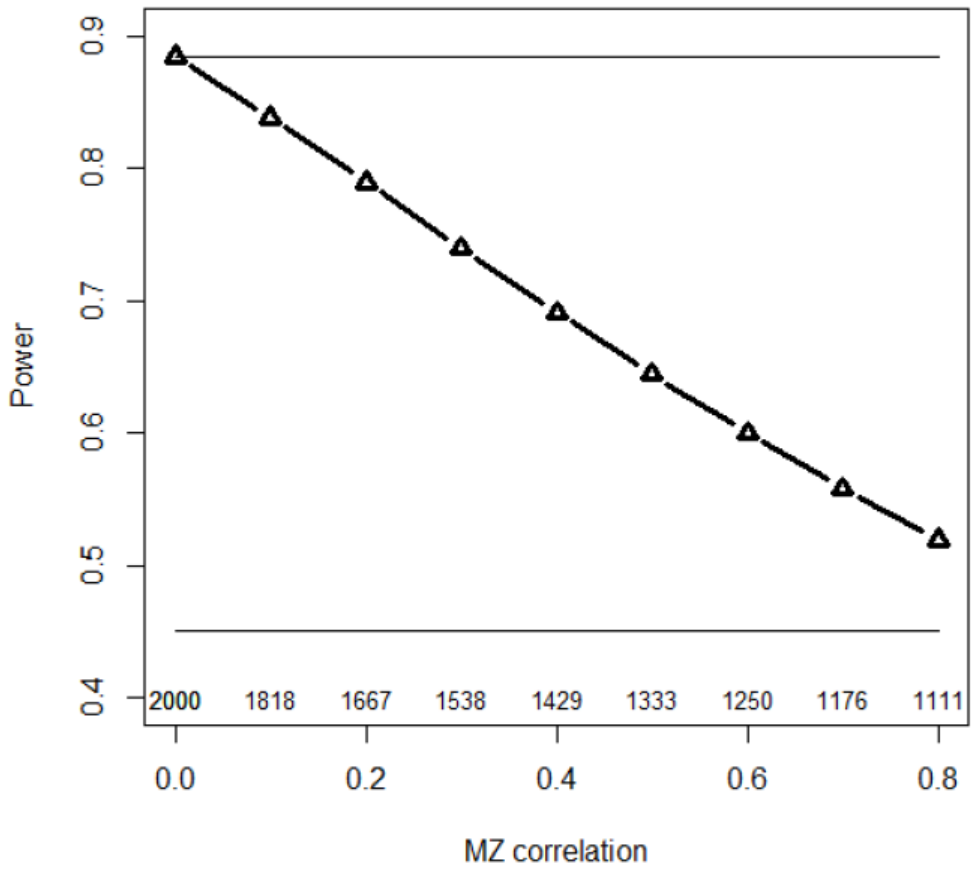
- Compute effective sample size:

$$N_E = (2*N) / (1+r)$$

intraclass correlation

ranges from **N** (**r** =1) to 2* **N** (**r**=0)

MZ pairs ~~or MZ singletons?~~



FAMILY-BASED GWAS: using efficiently correlated observations

Family-based GWAS
(continuous phenotype)

$$\mathbf{y}_{ij} = \mathbf{b}_0 + \mathbf{b}_1 * \mathbf{x}_{ij} + \boldsymbol{\varepsilon}_{ij}$$

where i is indicator of family (i=1..Nfam) and j is subjects (j=1..N)

y, **b** and **ε** are vectors

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} \mathbf{b}_0 \\ \mathbf{b}_1 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

Family-based GWAS

(model in matrix notation)

$$y = Xb + \varepsilon$$

$$\varepsilon = y - Xb$$

$$\varepsilon | X \sim N(0, V)$$

Family-based GWAS

$$\varepsilon | X \sim N(0, V)$$

$$V = \begin{bmatrix} V_1 & O & & O \\ O & V_2 & & O \\ & & \ddots & \\ O & O & & V_{N_{\text{fam}}} \end{bmatrix}$$

Family-based GWAS

$$\varepsilon | X \sim N(0, V)$$

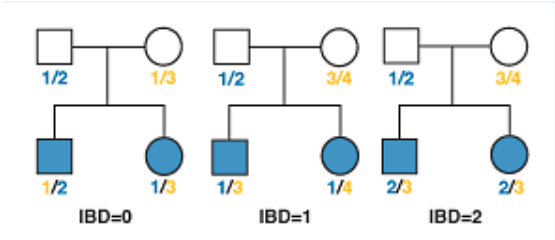
$$V(\Theta)$$

$$\Theta = [\sigma^2_A, \sigma^2_C, \sigma^2_E]$$

V modeled as an ACE

$$V(\Theta) = A \otimes \sigma^2_A + C \otimes \sigma^2_C + I \otimes \sigma^2_E$$

$$A = \begin{bmatrix} 1 & 0 & .5 & .5 & 0 & 0 & 0 & 0 \\ 0 & 1 & .5 & .5 & 0 & 0 & 0 & 0 \\ .5 & .5 & 1 & .5 & 0 & 0 & 0 & 0 \\ .5 & .5 & .5 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & .5 & .5 \\ 0 & 0 & 0 & 0 & 0 & 1 & .5 & .5 \\ 0 & 0 & 0 & 0 & .5 & .5 & 1 & .5 \\ 0 & 0 & 0 & 0 & .5 & .5 & .5 & 1 \end{bmatrix}$$



Expected proportion of the genome shared **IBD**

Genetic Relationship Matrix

e.g., 2 parents + 2 DZ twins

$$V(\Theta) = A \otimes \sigma^2_A + C \otimes \sigma^2_C + I \otimes \sigma^2_E$$

What other genetic information **A** could contain?

#1) The **actual** genome-wide relationship, defined as the **observed** proportion of the genome that two relatives share IBD, varies around its expectation because of Mendelian segregation, except for MZ twins and parent-offspring pairs. (Genotypic info: microsatellites).

Why bother?

OPEN ACCESS Freely available online PLoS GENETICS

Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings

Peter M. Visscher*, Sarah E. Medland, Manuel A. R. Ferreira, Katherine I. Morley, Gu Zhu, Belinda K. Cornes, Grant W. Montgomery, Nicholas G. Martin

2) GCTA (Yang et al 2011; Speed et al 2012) and variations (Zaitlen et al 2013)

$$V(\Theta) = A \otimes \sigma^2_A + C \otimes \sigma^2_C + I \otimes \sigma^2_E$$

What other genetic information could **A** contain?

GCTA: average allelic correlations between the individuals, where the alleles are observed in the measured SNPs

V modeled as an ACE

$$V(\Theta) = A \otimes \sigma^2_A + C \otimes \sigma^2_C + I \otimes \sigma^2_E$$

$$C = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

ESTIMATION?

Maximum Likelihood

$$\hat{\mathbf{b}}_{\mathrm{ML}} = \left(\mathbf{X}^{\mathrm{t}} \mathbf{V}\left(\hat{\boldsymbol{\Theta}}\right)^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{t}} \mathbf{V}\left(\hat{\boldsymbol{\Theta}}\right)^{-1} \mathbf{y}$$

$$\mathrm{var}(\hat{\mathbf{b}}_{\mathrm{ML}}) = (\mathbf{X}^{\mathrm{t}}\, \mathbf{V}(\hat{\boldsymbol{\Theta}})^{-1} \mathbf{X})^{-1}$$

Maximum Likelihood

$$\hat{\mathbf{b}}_{\mathrm{ML}} = \left(\mathbf{X}^{\mathrm{t}} \mathbf{V}\left(\hat{\boldsymbol{\Theta}}\right)^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{t}} \mathbf{V}\left(\hat{\boldsymbol{\Theta}}\right)^{-1} \mathbf{y}$$

correct model

$$\mathrm{var}(\hat{\mathbf{b}}_{\mathrm{ML}}) = (\mathbf{X}^{\mathrm{t}}\, \mathbf{V}(\hat{\boldsymbol{\Theta}})^{-1} \mathbf{X})^{-1}$$

What if my model for V is misspecified?

e.g.: model an ACE trait but ignore C

Maximum Likelihood

$$\hat{\mathbf{b}}_{\mathrm{ML}} = \left(\mathbf{X}^{\mathrm{t}} \mathbf{V}\left(\hat{\boldsymbol{\Theta}}\right)^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{t}} \mathbf{V}\left(\hat{\boldsymbol{\Theta}}\right)^{-1} \mathbf{y}$$

SANDWICH correction

misspecification?

$$\mathbf{V}(\hat{\boldsymbol{\Theta}}) = [\sigma^2_A,\ \sigma^2_E]$$

$$\mathrm{var}\left(\hat{\mathbf{b}}_{\mathrm{R-ML}}\right) = \left(\mathbf{X}^{\mathrm{t}} \mathbf{V}\left(\hat{\boldsymbol{\Theta}}_{\mathrm{m}}\right)^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{t}} \mathbf{V}\left(\hat{\boldsymbol{\Theta}}_{\mathrm{m}}\right)^{-1} (\mathbf{y} - \mathbf{X}\mathbf{b})(\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{t}} \mathbf{V}\left(\hat{\boldsymbol{\Theta}}_{\mathrm{m}}\right)^{-1} \mathbf{X} \left(\mathbf{X}^{\mathrm{t}} \mathbf{V}\left(\hat{\boldsymbol{\Theta}}_{\mathrm{m}}\right)^{-1} \mathbf{X} \right)^{-1}$$

What if the degree of misspecification is even larger?

e.g.: model an ACE trait but ignore AC

V modeled as an E

$$\mathbf{V}(\hat{\boldsymbol{\Theta}}) = \mathbf{I} \otimes \sigma^2_E$$

You assume there is no significant covariance between family members.

V modeled as an E

$$V(\hat{\Theta}) = I \otimes \sigma^2_E$$

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

ESTIMATION?

Unweighted Least Squares

$$\mathbf{b}_{ULS} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

$$\mathrm{var}(\mathbf{b})_{ULS} = (\mathbf{X'X})^{-1}\,\hat{\sigma}^2_E$$

$$\mathbf{V}(\hat{\boldsymbol{\Theta}}) = \hat{\sigma}^2_E\mathbf{I}$$

Unweighted Least Squares

$$\mathbf{b}_{ULS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\mathrm{var}(\mathbf{b})_{ULS} = (\mathbf{X}'\mathbf{X})^{-1}\,\hat{\sigma}^2_E$$

$$\mathbf{V}(\hat{\boldsymbol{\Theta}}) = \hat{\sigma}^2_E\mathbf{I}$$

misspecification

Unweighted Least Squares

$$\mathbf{b}_{ULS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\mathrm{var}(\mathbf{b})_{ULS} = (\mathbf{X}'\mathbf{X})^{-1}\,\hat{\sigma}^2_E$$

SANDWICH correction

misspecification

$$\mathbf{V}(\hat{\Theta}) = \hat{\sigma}^2_E\mathbf{I}$$

$$\mathrm{var}\left(\hat{\mathbf{b}}_{R-ULS}\right) = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\left(\mathbf{y}-\mathbf{X}\mathbf{b}\right)\left(\mathbf{y}-\mathbf{X}\mathbf{b}\right)^t\mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}$$

ML or ULS?

LEAST SQUARES: - non-iterative, very fast;
- correct standard errors;
- **E** model for the covariance matrix

misspecification

ML : - iterative;
- fast;
- **AE** model for the covariance matrix

misspecification for ACE traits

ML or ULS?

Two different estimators **may be consistent**, but they are not necessarily equally efficient.
so as N-> Large, b1-est tends to b1-true.

but **given N, one estimator may be more efficient: i.e., have a smaller standard error** (regardless whether the standard error is based on asymptotic theory or on a permutation test).

CONCLUSIONS
(quantitative traits)

- Full correct modeling (**RareMetal Worker (practical Sarah)**, OpenMx, Linear Mixed, Merlin, Mendel)

- **AE** type modeling standard (CGTA, FastLMM)
  (you probably can add the C coded matrix to GCTA if you are modeling close relateds)

- **CE/AE/E** type of modelling **with sandwich correction** (GEE)

- **E** type of modeling (Plink - - equivalent to GEE with independence correlation matrix) – low power (**generally not recommended**).

USEFUL SOFTWARE:

<u>PLINK1.7 + R-GEE+sandwich</u>:
http://pngu.mgh.Harvard.edu/~purcell/plink/rfunc.shtml
https://www.cog-genomics.org/plink2/
see EXAMPLE GEE: http://cameliaminica.nl/scripts.php

MERLIN and MERLIN-offline:
http://genepi.qimr.edu.au/staff/sarahMe/merlin-offline.html

GCTA-MLM-LOCO:
http://www.complextraitgenomics.com/software/gcta/mlmassoc.html

FAST-LMM:https://github.com/MicrosoftGenomics/FaST-LMM

PRACTICAL

Association analysis, family data

We will compare 3 options

- Plink1 --family
- gee, with option correlation structure="independence"
- gee, with option correlation structure="exchangeable"

E model

CE model

/faculty/jenny/2017/tuesday

```
mkdir practical_family
cp -r /faculty/jenny/2017/tuesday/* practical_family
cd practical_family
```

Note on --family in plink → use plink1!

- the option -- family is currently not implemented in plink2
- If you do use -- family in plink2, incorrect output is returned

Plink –association analysis

- Data

plink_covar.txt
rs2228145_plink.map
rs2228145_plink.ped

- Covariates (plink_covar.txt)

zage = z-score of age
PC1_NL PC2_NL PC3_NL = Dutch ancestry PCs
PC3_chip_effect PC5_chip_effect PC1_buccal = PCs to correct for chip and DNA source

- Run association test (1 SNP) - sIL6R, correcting for relatedness and 7 covariates
- We use plink version 1.07

```
plink1 --file rs2228145_plink --covar plink_covar.txt --linear --family --mperm 1000
```

The results are in **plink.assoc.linear** → have a look at this file

Output plink

- plink.assoc.linear

CHR	SNP	BP	A1	TEST	NMISS	BETA	STAT	P
1	**rs2228145**	**154426970**	**C**	**ADD**	**2572**	**1.226**	**47.22**	**0**
1	rs2228145	154426970	C	COV1	2572	0.171	10.16	8.626e-24
1	rs2228145	154426970	C	COV2	2572	-2.564	-1.307	0.1913
1	rs2228145	154426970	C	COV3	2572	-3.595	-1.393	0.1638
1	rs2228145	154426970	C	COV4	2572	0.2612	0.09378	0.9253
1	rs2228145	154426970	C	COV5	2572	-3.73	-2.457	0.01407
1	rs2228145	154426970	C	COV6	2572	-0.9417	-0.5481	0.5837
1	rs2228145	154426970	C	COV7	2572	9.234	0.943	0.3458

Gee – association analysis

- We will now use the R-package gee to test the association between our SNP and sIL-6R
- We are going to read in the plink ped file and covariate file in R.
- We will use gee, with 2 options:
  - Correlation structure= "independence“
  - Correlation structure= “exchangeable”

→ **Compare the results obtained with these 2 options – are they the same?**

- Open the R-script **association_rs2228145_gee.r** (click on it, it will open in R-studio)
- Run the script line by line

Output gee

Correlation structure “independence”

```
                   Estimate  Naive S.E.     Naive z Robust S.E.     Robust z
(Intercept)       3.2082952  0.03652159 87.84653889  0.03154167 101.71609306
genonum           1.2257800  0.02389002 51.30929293  0.02595836  47.22101667
zage              0.1710237  0.01675720 10.20598431  0.01683764  10.15722428
PC1_NL           -2.5636709  1.95036718 -1.31445552  1.96137938  -1.30707549
PC2_NL           -3.5952193  2.43177616 -1.47843347  2.58099455  -1.39295887
PC3_NL            0.2611566  2.62472212  0.09949875  2.78476265   0.09378055
PC3_chip_effect  -3.7295217  1.51768531 -2.45737482  1.51776706  -2.45724245
PC5_chip_effect  -0.9416699  1.54974148 -0.60763031  1.71813103  -0.54807804
PC1_buccal        9.2342455 10.99356713  0.83996809  9.79244696   0.94299674
>
```

Correlation structure “exchangeable”

- Identical estimates
- slightly larger Robust Z-statistics

```
> coeff
                   Estimate  Naive S.E.     Naive z Robust S.E.     Robust z
(Intercept)       3.20039476  0.03739698 85.57896946  0.03101160 103.19993373
genonum           1.22709542  0.02400555 51.11715551  0.02568134  47.78159138
zage              0.17856674  0.01659369 10.76112051  0.01682857  10.61092906
PC1_NL           -2.22687489  2.02532967 -1.09951230  1.91838437  -1.16080746
PC2_NL           -3.45332650  2.51218708 -1.37462951  2.57648882  -1.34032272
PC3_NL            0.03283707  2.71043312  0.01211506  2.74004517   0.01198414
PC3_chip_effect  -3.62243989  1.53635530 -2.35781390  1.49974450  -2.41537134
PC5_chip_effect  -1.04564505  1.58955286 -0.65782339  1.69688607  -0.61621406
PC1_buccal       11.61605341 11.21512815  1.03574861  9.75924743   1.19026119
>
```

- Compare the results obtained in gee to those obtained in plink1 (plink.assoc.linear)

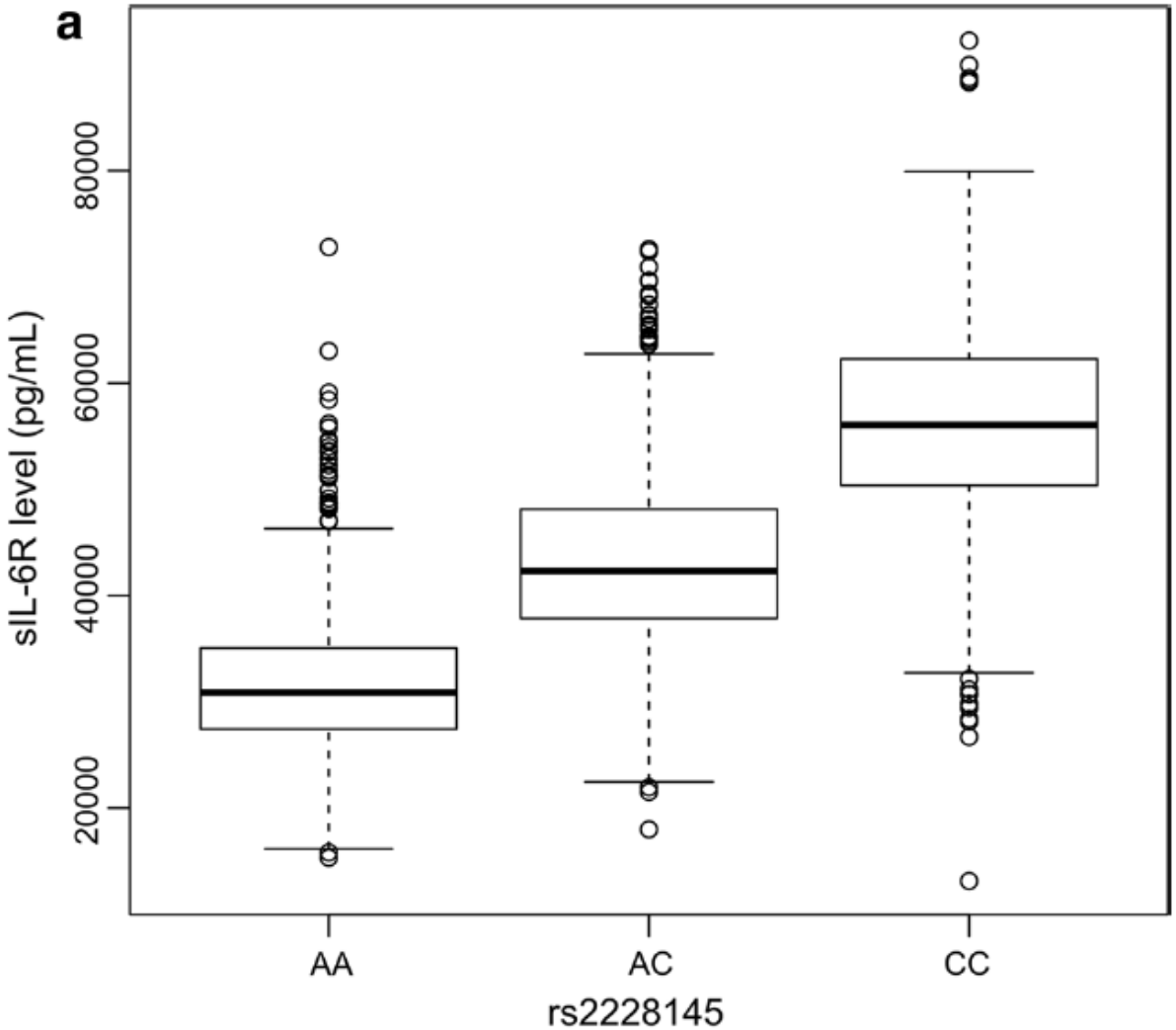
→ **Do you notice any difference?**

- Notice that the results (estimate and Robust Z) from gee with option “independence” are identical to those obtained in plink1
- → Gee with option corst=“independence” does the same as plink1 with option --family

Biometrical model

Rs2228145: **Large effect** on sIL-6R level (allele C increases sIL-6R concentration)

sIL-6R concentration



SNP rs2228145 g*enotype*

Exercise: *Effect of the IL6R gene on IL-6R concentration*

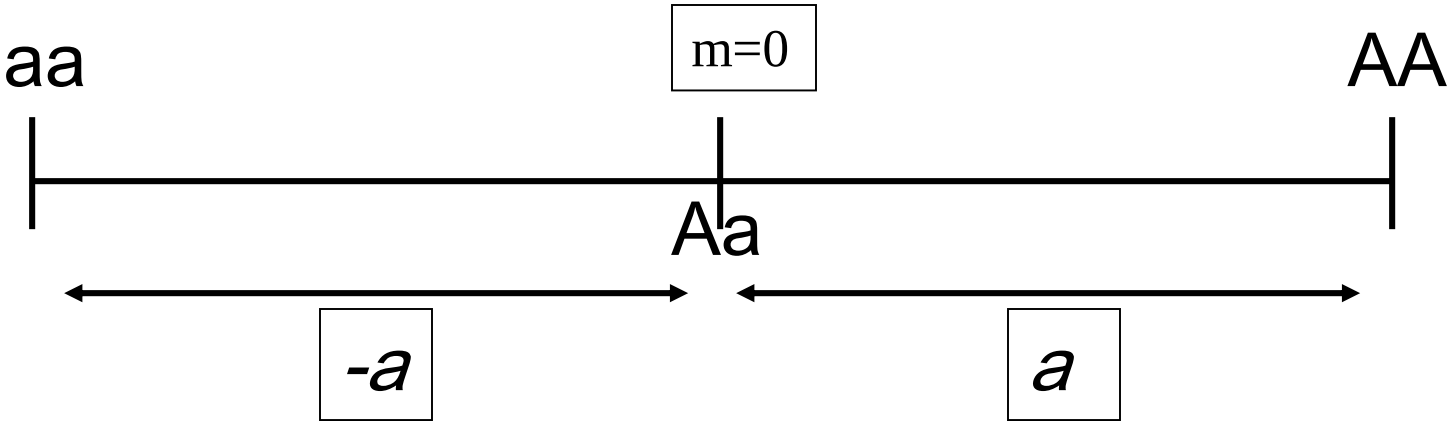
INFORMATION

- The SNP (single nucleotide polymorphism) has 2 alleles:
  - Minor allele: C, frequency: p=0.39
  - Major Allele: A, frequency: q =0.61
- Mean IL-6R concentration of each genotype:
  - CC: 5.698 (10^{-8} g/mL)
  - CA: 4.418 (10^{-8} g/mL)
  - AA: 3.238 (10^{-8} g/mL)
- Total Variance of IL-6R concentration=1.35

QUESTIONS (Falconer & MacKay; 1996: Introduction to quantitative genetics)

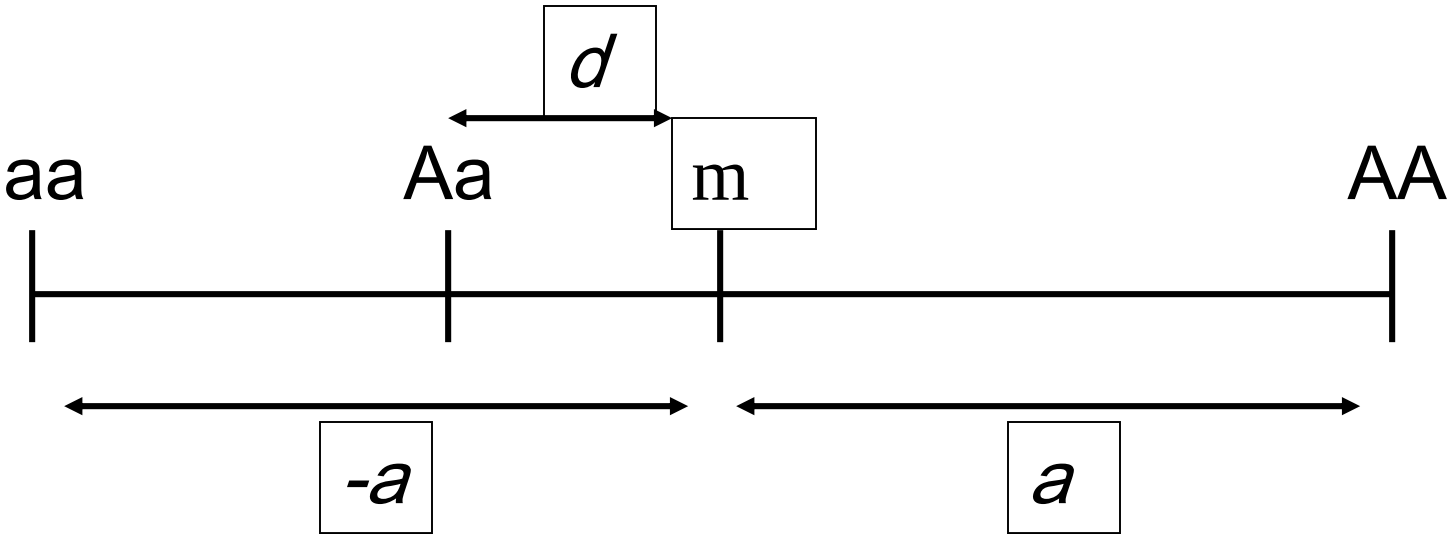
1. Calculate genotypic values (a and d) (page 109)
2. [Calculate the average effect of the alleles (page 113)]
3. Calculate the genotype frequencies (page 7)
4. Calculate the mean IL6-R concentration in the population (page 110)
5. Calculate how much of the variance is explained by this SNP (*Variance= Sum of squared deviations from the mean*)
6. Calculate heritability

Model: gene with 2 alleles A and a and 3 genotypes AA, Aa and aa



The difference on a quantitative scale between AA and aa is 2*a.*
The middle (m) is zero and the value of Aa is 0 (no dominance).

Model: gene with 2 alleles A and a and 3 genotypes AA, Aa and aa



The deviation from m (middle) of the heterozygote Aa is d: partial dominance.

Genotype (i)	AA	Aa	aa
Frequency (f)	p^2	2pq	q^2
Genotypic effect (x)	a	d	-a

Mean?

Genotype (i)	AA	Aa	aa
Frequency (f)	p^2	2pq	q^2
Genotypic effect (x)	a	d	-a
f * x	p^2 a	2pqd	- q^2 a

mean: $p^2 a + 2pqd - q^2 a =$

(recall p+q = 1)

$a(p^2 – q^2) + 2pqd =$

$a(p-q)(p+q) + 2pqd =$

Mean = a(p-q) + 2pqd

a(p-q) : attributable to homozygotes

2pqd : attributable to heterozygotes

Genotype (i)	AA	Aa	aa
Frequency (f)	p^2	2pq	q^2
Genotypic effect (x)	a	d	-a
f * x	p^2 a	2pqd	- q^2 a

mean: $p^2 a + 2pqd - q^2 a = a(p-q) + 2pqd$

Variation: $2pq[a+d(q-p)]^2 + (2pqd)^2$

Population variation depends on ‘a’ (difference between homozygote individuals), ‘d’ (deviation of heterozygote persons from zero) and on allele frequency (p & q).

Average effect
(associated with genes and not with genotypes)

The average effect of a gene (allele) is the mean deviation from the population mean of individuals which received that gene from one parent, the gene received from the other parent having come *at random* from the population.

Falconer (p112):The concept of average effect is not easy to grasp.

Average effect is related to genotypic values a and d

$q\,[a + d\,(q - p)] = \alpha_1$

$-p\,[a + d\,(q - p)] = \alpha_2$

Average effect of gene substitution is $\alpha_1 - \alpha_2 = \alpha$. This is the difference between the average effect of the 2 alleles:

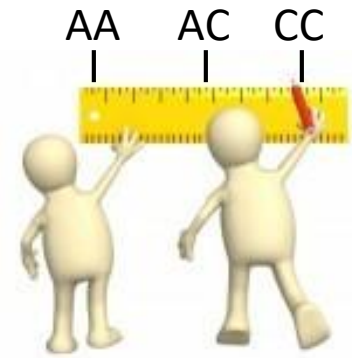
$\alpha = a + d(q-p)$

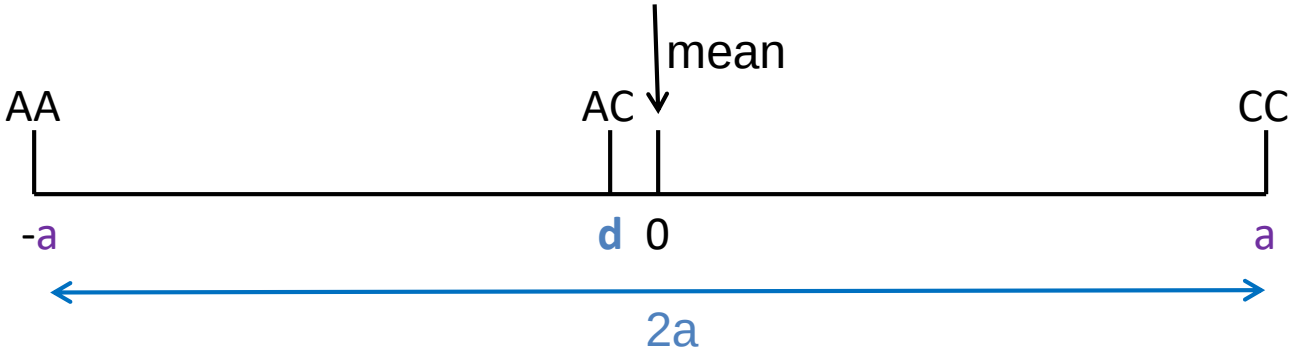
Mean IL-6R concentration of each genotype:

CC: 5.698 / CA: 4.418 / AA: 3.238 (10^{-8} g/mL)

Total Variance of IL-6R concentration=1.35

Frequencies: C, frequency: p=0.39 / A, frequency: q =0.61





QUESTIONS (Falconer & MacKay; 1996: Introduction to quantitative genetics)

1. Calculate genotypic values (a and d) (page 109)
2. Calculate the average effect of the alleles (page 113)
3. Calculate the genotype frequencies (page 7)
4. Calculate the mean IL6-R concentration in the population (page 110)
5. Calculate how much of the variance is explained by this SNP
   (*Variance= Sum of squared deviations from the mean*)
6. Calculate heritability