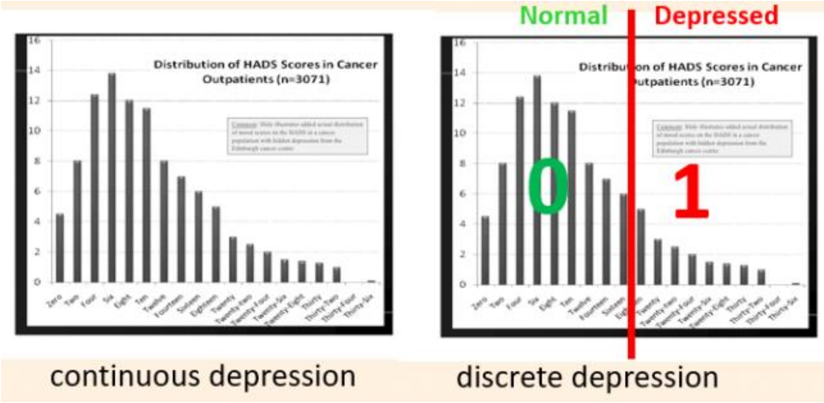
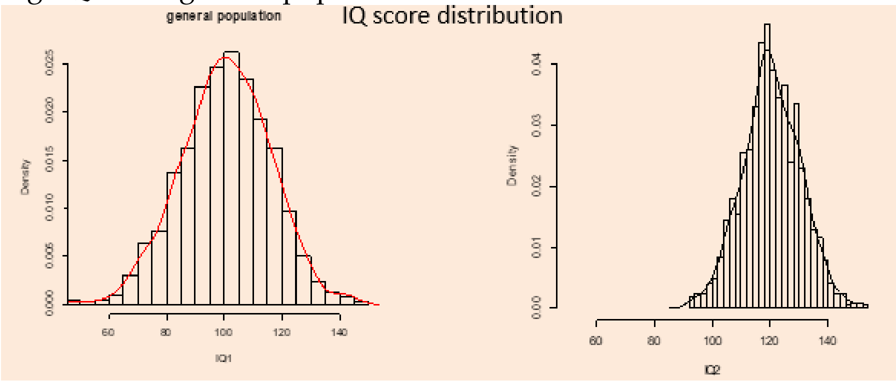
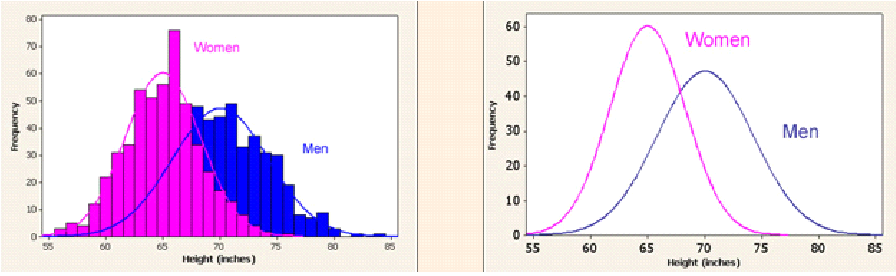
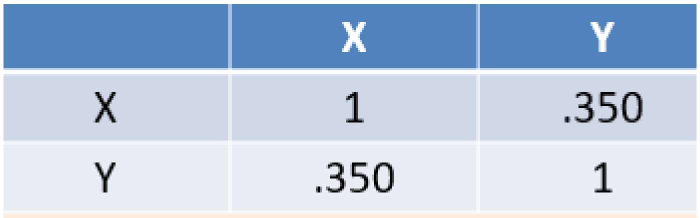
1a. From GWAS to the twin model via biometrical genetics (c.v.dolan@vu.nl)

What is the commonality between GWAS studies and twin studies?	Genome wide association studies (GWAS) The regression of a phenotype on measured variants Twin study Infering genetic effects from the phenotypic resemblance among twins Common biometrical underpinning - Relating genetic differences to phenotypic differences
Why is that 'having five fingers' is not in our genes?	GWAS is about the analysis of differences IQ genes - Genes that predicted differences in IQ Having five fingers is not in our genes - We can compare difference between people with five finger and people with more or less fingers
What is the science of biometrical genetics?	Biometrical genetics The science concerned with *inheritance of quantitative traits* Uses statistical analysis of the inheritance of difference phenotypes as related to plant or animal breeding
Can a continous phenotype (like depression) be discrete?	Statistics refresher Linear model - Random variables and parameters Y = b0 + b1*x + e Error is not observed - we do not know what the true values of b0 and b1 are Distribution of random variables - Expressed in a histogram RV are discrete or continuous ○ Discrete - Dice rolls, number of fingers ○ Continuous - Height (can be measured with infinitive precision) This varies according to your proposition - Depression could be discrete (you are or you are not) or continuous (there are hundreds of genes related to depression)

	
Why the distribution of a continuous phenotype is not always normal?	Distribution of a variable is a property of the different populations E.g. IQ in the general population and in students 
Why is it incorrect to say that 'height is normally distributed'?	Not all continuous variables are normally distributed 'A variable is normally distributed' - This is a statement about the population  Height is normally distributed is an incorrect statement, since it depends on gender
What is the difference between covariance and correlation?	Parameters b0 and b1 Mean(y) - b0 + b1*mean(x) Variance(y) = b1sqrt*variance(x)+ variance e Covariance between X and Y - Relationship between two random variables Pearson Product Moment - Linear association between two continuous variables -1 < r < +1 Covariance table - Standardized by the correlation coefficient



	X	Y
X	1	.350
Y	.350	1

Z-score - Linear association between X and Y in standardized units

What is the difference between regression and correlation?

Regression - Linear association between X and Y where X is the predictor and Y is the dependent

The value of Y is determined by two factors - b1sq*var(x) + var e

Are the individual differences of Y related to the individual differences of X?

R2 - Amount of variance explained from one variable to another

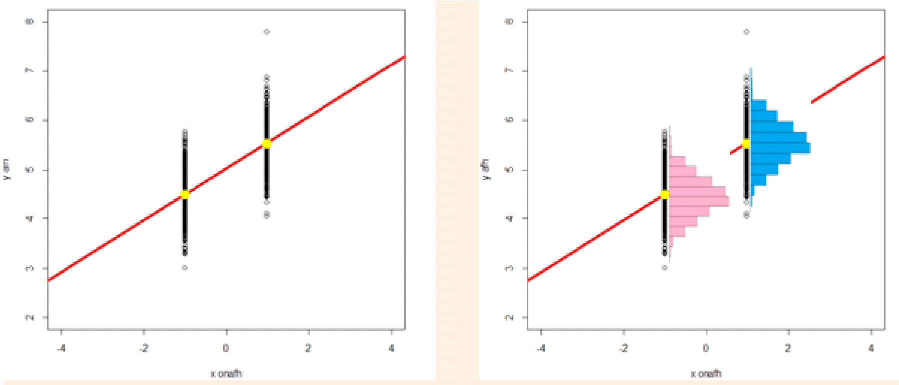
It is an effect size - It doesn't depend on which variables are used

Think about the individual components of the equation

What does it mean to say that a variable is normally distributed given another variable?

Distributional assumption of statistical inference?

Y I X - Normal (b0 + b1*x, stdev e

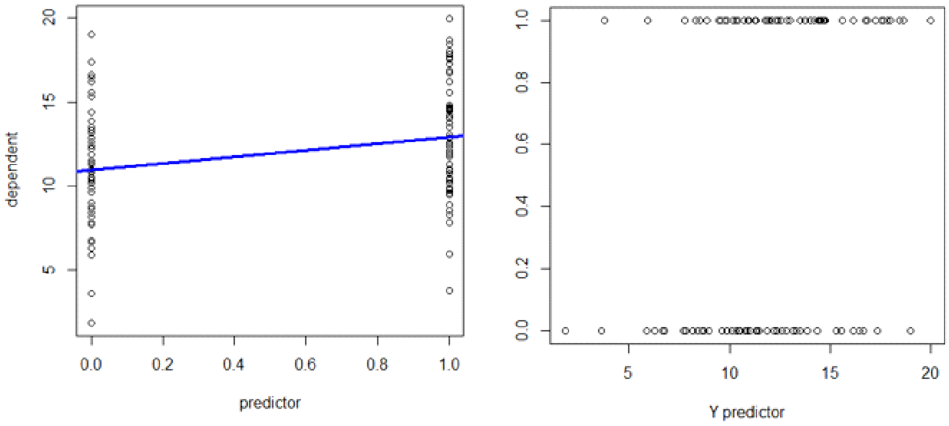


Y given X is normally distributed - **In each group of Y, the values of X are normally distributed**

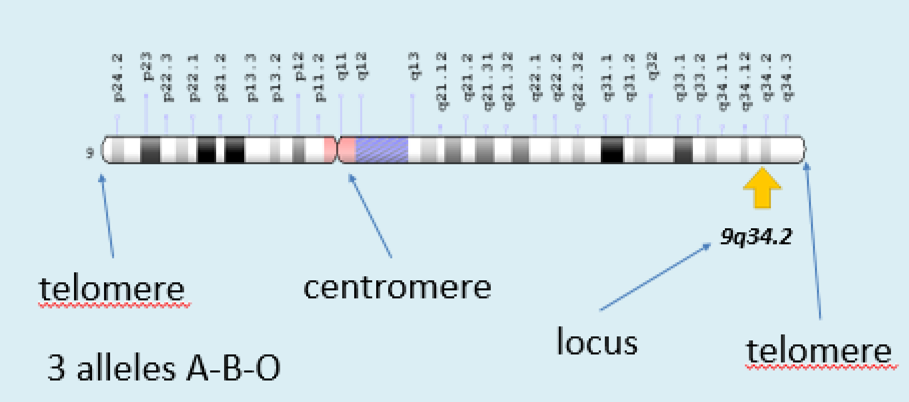
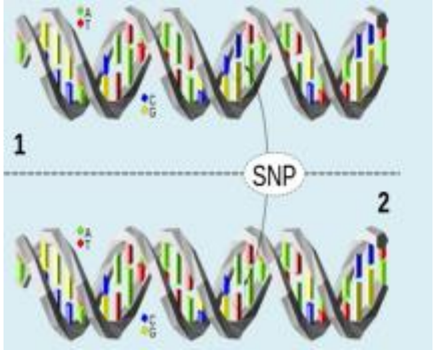
It does not mean that Y is normally distributed or that X is normally distributed

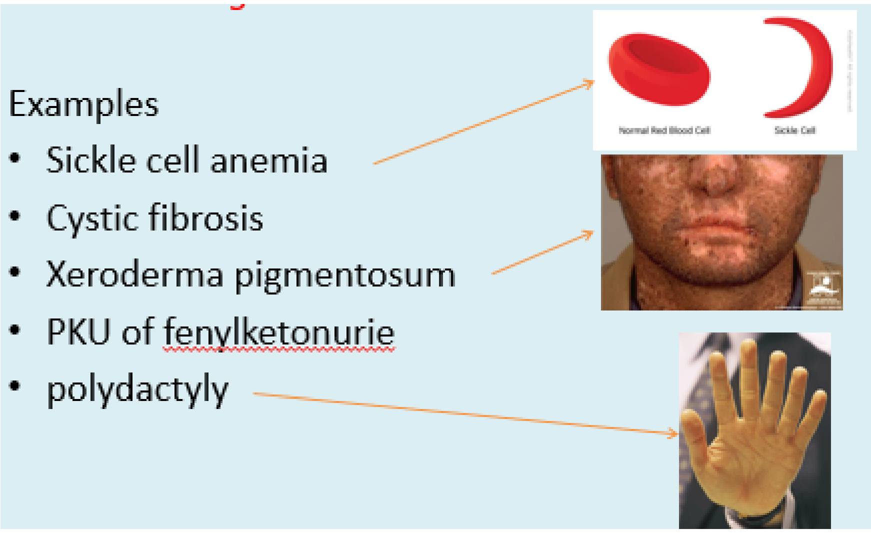
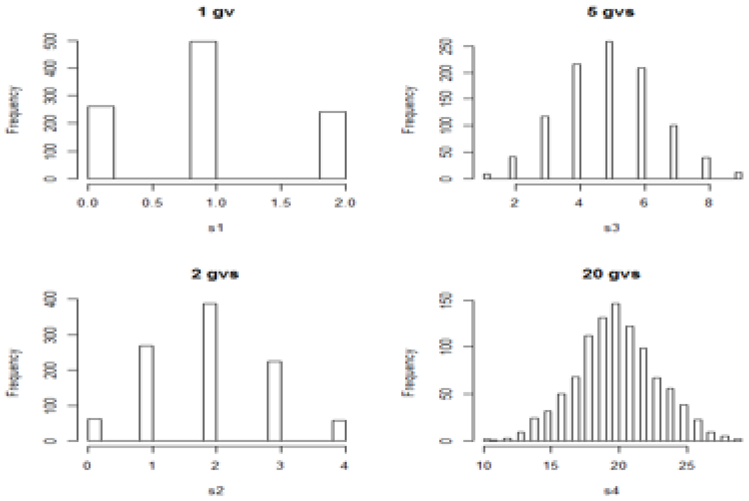
What statistical method should we used for

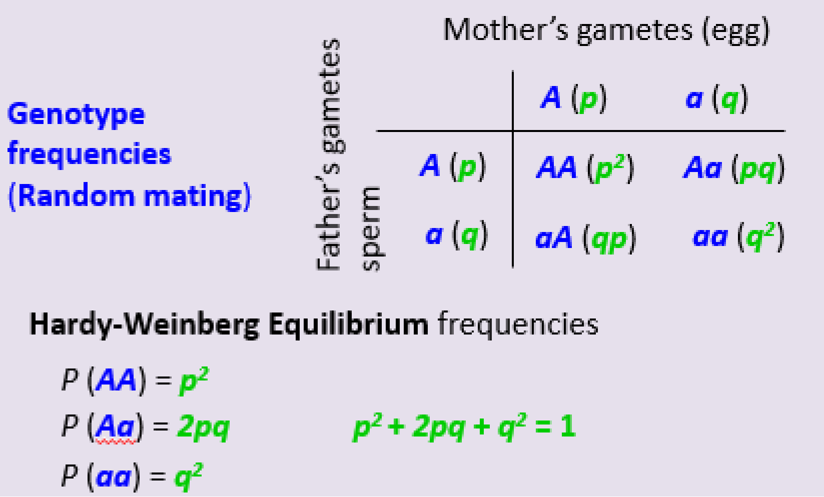
A binary dependent variable is not suited for linear regression - A logistic regression must be used

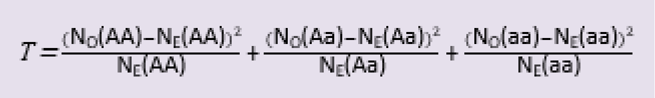
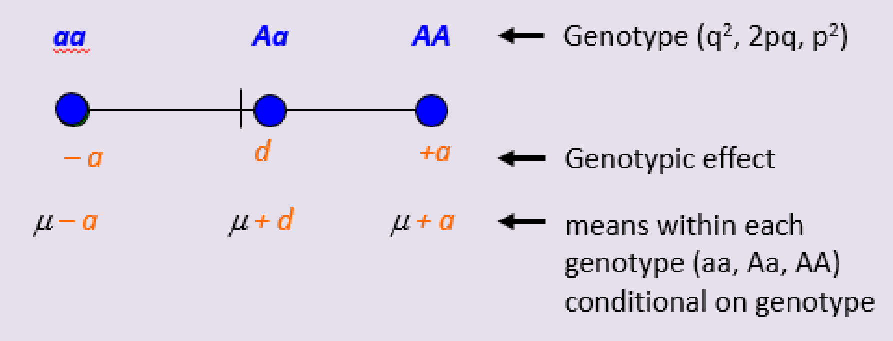
binary variables?	 This must be analysed with probabilites
What is the difference between conditional probabilities and unconditional probabilities?	Probabilites - Discrete random variable with a discrete outcome Unconditional probability Example: Hair and eye color How many people have light eyes and red hair - Unconditional How many people have dark eyes - Unconditional Conditional probabilities What is the probability of light eyes given red hair - Conditional probability

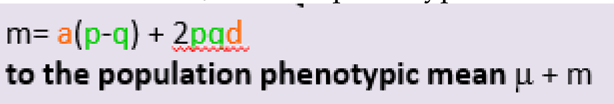
1b. How do people differ genetically?

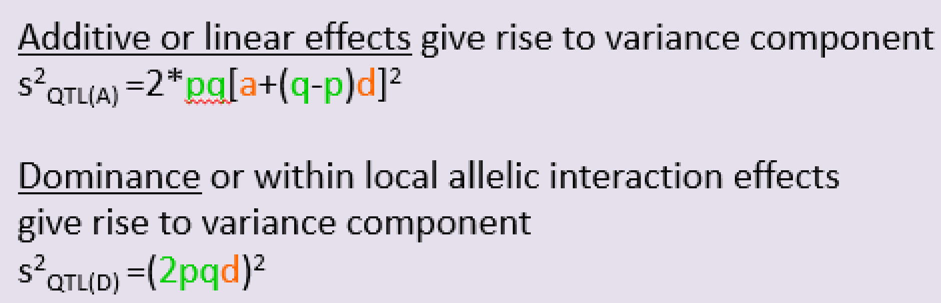
Define: **a. Gene** **b. Locus** **c. Allele**	Terminology Gene - Sequence of DNA that code for a particular product Locus - Site of a specific gene on a chromosome Allele - Alternative form of a gene at a locus Genotype - The combination of alleles at a particular locus Phenotype - Observed characteristic, trait
Where would a locus be if it was named "9q34.2"	Chromosome strucutre  p24.2 p23 p22.3 p22.1 p21.2 p13.3 p13.2 p12 p11.2 q11 q12 q13 q21.12 q21.2 q21.31 q21.32 q22.1 q22.2 q22.32 q31.1 q31.2 q32 q33.1 q33.2 q34.11 q34.12 q34.2 q34.3 9 telomere centromere 9q34.2 locus telomere 3 alleles A-B-O Each locus has 2 alleles - One paternal and one maternal Mendel's first law
What is a SNP	SNP (single nucleotide polymorphism) - Variant at a level of a single base pair  Variation of the human genome -> difference in protein structure -> phenotypic differences
What are Mendelian traits?	Mendelian traits - 1 to 1 phenotype-genotype relationship Variations in 1 gene causes variations in the trait

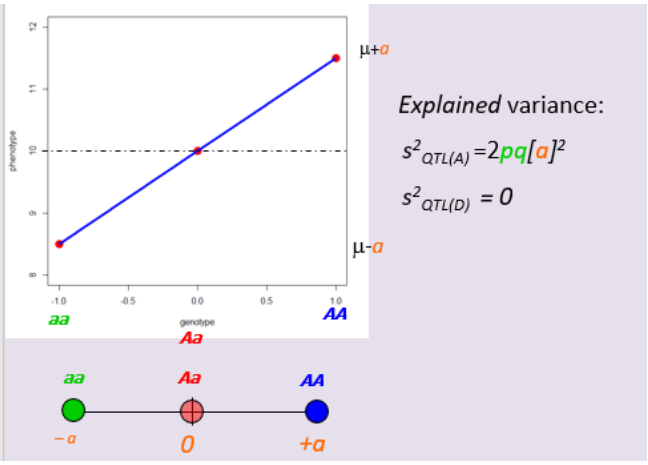
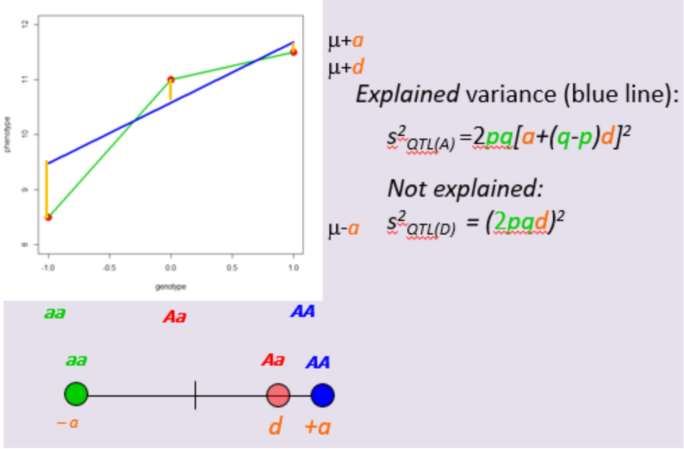
	 Examples • Sickle cell anemia • Cystic fibrosis • Xeroderma pigmentosum • PKU of fenylketonurie • polydactyly Sickle cell anemia Cystic fibrosis Xeroderma pigmentosum Fenylketonurie Polydactylyl
What are polygenic traits?	Polygenic traits - Quantitative/complex Accumulation of many single genes - Quantitative trait loci (QTL) The accumulation of many different gene give rise to a normal distribution 
Why is it difficult to have significant results in GWAS?	GWAS cohort study Y = b0 + b1*GV + e Var(y) = B1sqr * var(GV) + var e Predictive part depends on B1, the Genetic variance and the error

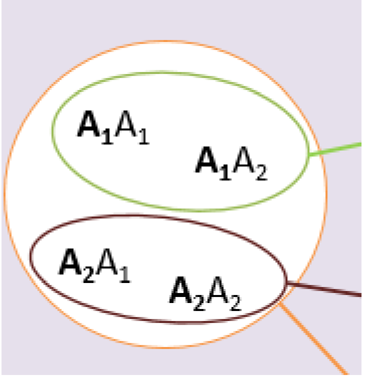
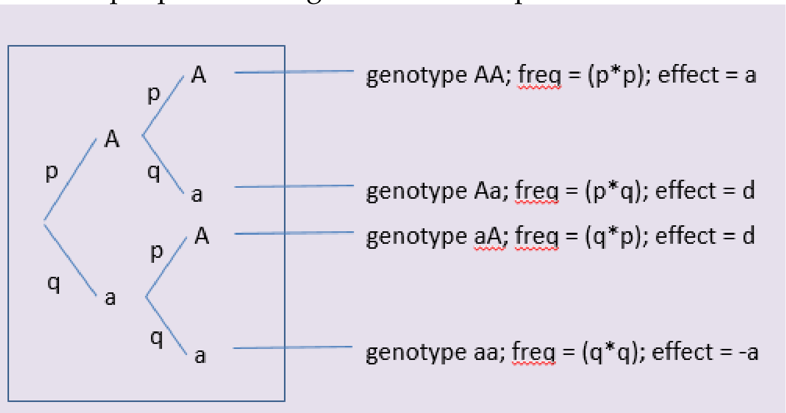
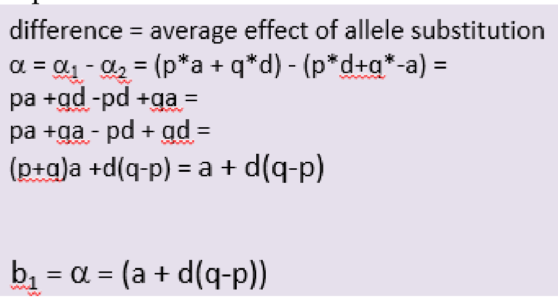
	H-null: b1 = 0 Bonferroni correction - With a million tests you are bound to find many false positives
What is Hardy Weinberg equilibrium?	Define what is the distribution of alleles in the population Biallelic: A and a In GWAS: SNPs Frequency of A is p Frequency of a is q What are the genotype frequencies (predicted from the the allele frequency)?  Genotype frequencies (Random mating) Mother's gametes (egg) Father's gametes sperm $\begin{array}{ccc} & A\ (p) & a\ (q) \\ \hline A\ (p) & AA\ (p^2) & Aa\ (pq) \\ a\ (q) & aA\ (qp) & aa\ (q^2) \end{array}$ **Hardy-Weinberg Equilibrium** frequencies $P(AA) = p^2$ $P(Aa) = 2pq$ $p^2 + 2pq + q^2 = 1$ $P(aa) = q^2$ Hardy Weinberg equilibrium
Why is HW useful for GWAS?	Observed genotype frequencies - Observed empirically P(AA) = N(AA) / Ntotal P(Aa) = N(Aa) / Ntotal P(aa) = N(aa) / Ntotal Estimates allele frequency: P = p(AA) + 1/2 *p(Aa) P = p(aa) + 1/2*p(Aa) GWAS studies - Need to establish that what they observe is close to Hardy Weinberg equilibrium This is important because it the genotype testing is not 100% reliable - Some loci are easy to genotype, others are difficult HW assumes that all genotypes have the same fit/ability to reproduce T(1) - Chi-square with one degree of freedom distribution (p = 3.84 is

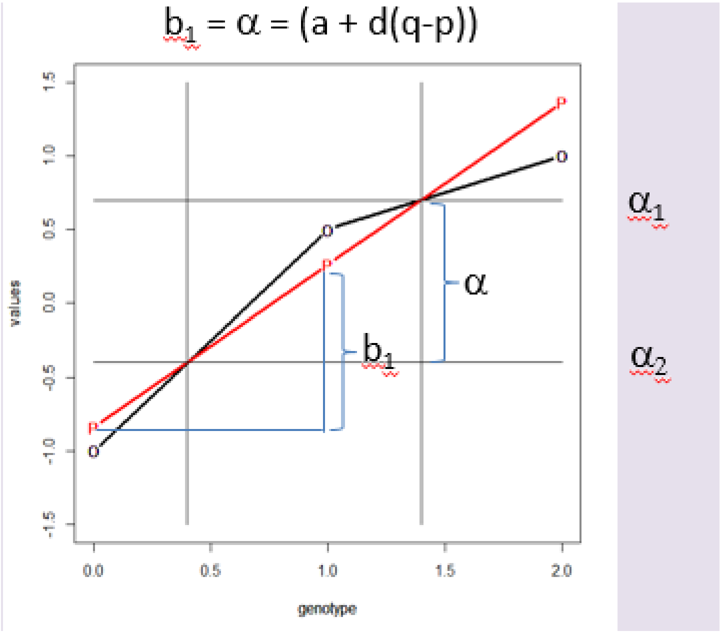
	associated with alpha <0.05)  $T = \frac{(N_O(AA)-N_E(AA))^2}{N_E(AA)} + \frac{(N_O(Aa)-N_E(Aa))^2}{N_E(Aa)} + \frac{(N_O(aa)-N_E(aa))^2}{N_E(aa)}$ Degrees of freedom - 2x2 pq table; once you know p, you know q
How is the "mean phenotype" measured in a population?	We know frequencies, we now assign effects to the genotypes M - a - Effect of genotype aa M + d - Effect of genotype Aa M + a - Effect of genotype AA  aa Aa AA ← Genotype (q^2, 2pq, p^2) $-a$ d $+a$ ← Genotypic effect $\mu-a$ $\mu+d$ $\mu+a$ ← means within each genotype (aa, Aa, AA) conditional on genotype Take all individuals with a genotype and calculate their mean phenotypes

Describe the formula for the contribution of QTL to the phenotype mean?	Contribution of the QTL to the phenotype mean
	
	m= a(p-q) + 2pqd
	to the population phenotypic mean μ + m
	
	M = a(p-q) + 2pqd
	
	A = Homozygous effect
	D = Heterozygous effect
	If a and d equal 0 - There is no effect of the genetic variant on the phenotype

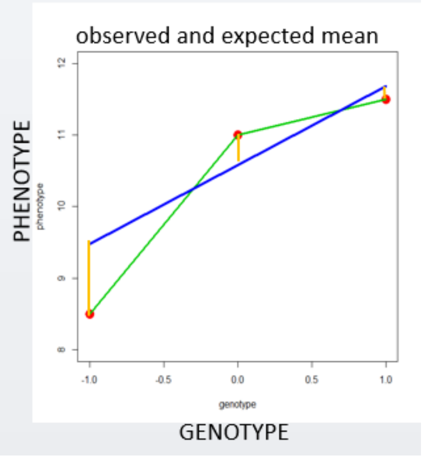
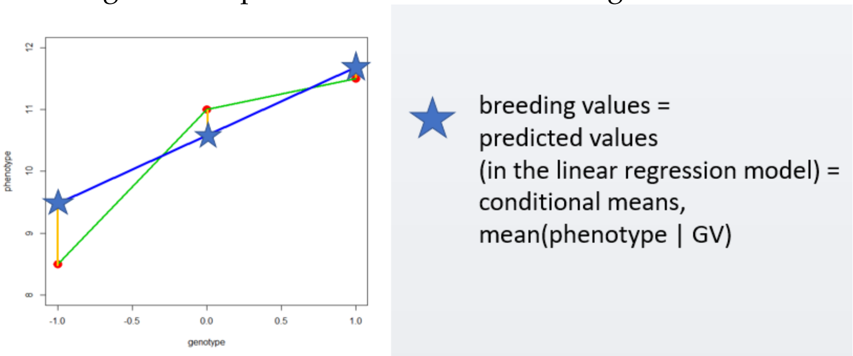
What is the difference between additive and dominant effects?	 Additive or linear effects give rise to variance component $s^2_{QTL(A)} = 2^*pq[a+(q-p)d]^2$ Dominance or within local allelic interaction effects give rise to variance component $s^2_{QTL(D)} = (2pqd)^2$ Additive effects - d equals 0

	Dominance effects - d is different from zero
How to derive allele distribution from phenotype?	Deriving the distribution of alleles in a population: Take all aa individuals and calculate their mean phenotypic value: mu - a (conditional mean is an arbitrary variable)
In a graph of genotype plotted against genotype, what is the visual difference between additive effects and dominant effects?	When d equals zero, the relationship is dominant - Linear model When d is not equal to zero, the relationship is not dominant - Non-linear model Additive/Linear effects  *Explained* variance: $s^2_{QTL(A)}=2pq[a]^2$ $s^2_{QTL(D)} = 0$ Explained variance: ssqr = 2pq(a)sqr If it equals zero - There is no dominance or recessiveness Dominance effects - d is in the middle  *Explained* variance (blue line): $s^2_{QTL(A)}=2pq[a+(q-p)d]^2$ *Not explained:* $s^2_{QTL(D)} = (2pqd)^2$ Explained variance: ssqr = 2pq(a+(q-p)d)sqr Not explained: ssqr = (2pqd)sqr Dominance becomes part of the residual

What is alpha (in the context of GWAS)? **How to calculate alpha?**	Alpha - Average effect of allele substitution (same as b1) Derivation: Image of population (aA and Aa is different)  Take all people with large A in the first position  p A q a p A q a p A q a genotype AA; freq = (p*p); effect = a genotype Aa; freq = (p*q); effect = d genotype aA; freq = (q*p); effect = d genotype aa; freq = (q*q); effect = -a What is the probability that the second allele is A? P (effect a) What is the probability that the second allele is a? Q (effect d) Conditional mean1 = p*a + q*d Take all people with small a in the first position What is the probability that the second allele is A? P (effect d) What is the probability that the second allele is a? Q (effect -a) Conditional mean2 = p*d + q*(-a) Alpha = Conditional mean1 - Conditional mean2  difference = average effect of allele substitution $\alpha = \alpha_1 - \alpha_2 = (p*a + q*d) - (p*d+q*-a) =$ $pa + qd - pd + qa =$ $pa + qa - pd + qd =$ $(p+q)a + d(q-p) = a + d(q-p)$ $b_1 = \alpha = (a + d(q-p))$

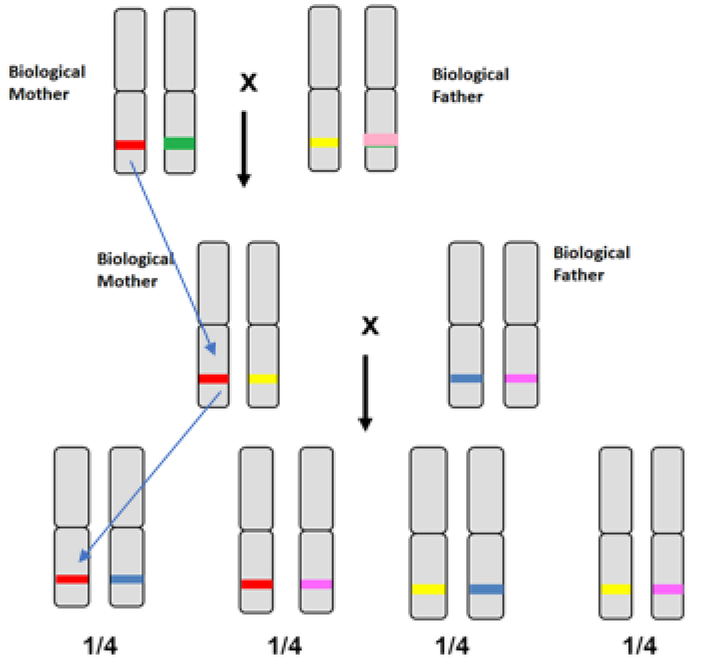
Why is alpha the same as b1 in an additive model?	Predicitive values  Genotype AA = 2qalpha Genotype Aa = (q-p)alpha Genotype aa = 2palpha

2. Additive + Dominance Component

How to observe additive or dominant effects in a graph?	 Dots - Conditional means (calculated from phenotypic data) Blue line - Regression model Yellow line - Residuals
What are breeding values?	Breeding values = predicted values from the regression model  ★ breeding values = predicted values (in the linear regression model) = conditional means, mean(phenotype | GV) Why is it called 'breeding values'? They are predictive of the offspring phenotype and therefore can be used in animal and plant breeding programs
What is the main variable that predicts the success of a breeding program?	Example: Breeding sheeps Goal: Extend fertility period Identify genes that have an effect on extended fertility Choose sheep with highest breeding values The higher the effect of GV on phenotype (higher r squared), the more sucessful the breeding program is

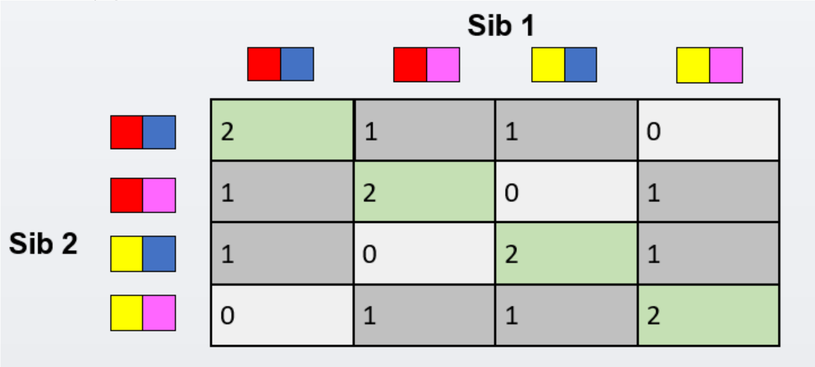
What is identity by descent? How many alleles IBD do monozygotic twins have? And dyzygotic twins?

Identity by descent - When the alleles can be traced from the parents



Each parent provides one allele for every allele possible

Monozygotic twins = 2 alleles IBD



		Sib 1			
		🟥🟦	🟥🟪	🟨🟦	🟨🟪
	🟥🟦	2	1	1	0
	🟥🟪	1	2	0	1
Sib 2	🟨🟦	1	0	2	1
	🟨🟪	0	1	1	2

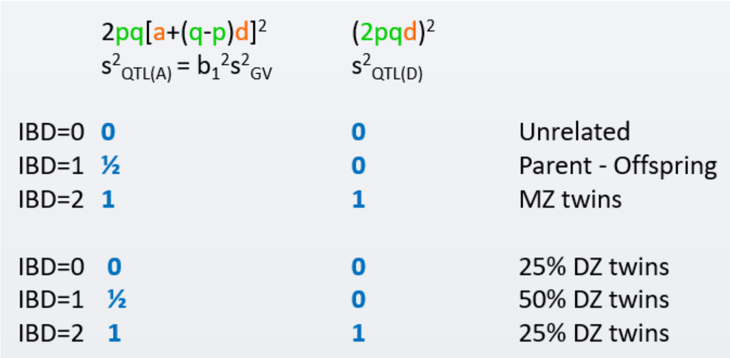
Dizygotic twins = Can have 0, 1 or 2 alleles IBD

Non twins = Always 1 IBD (one from each parent)

What is the proportion of additive variance shared by

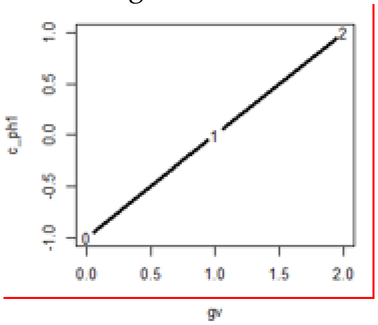
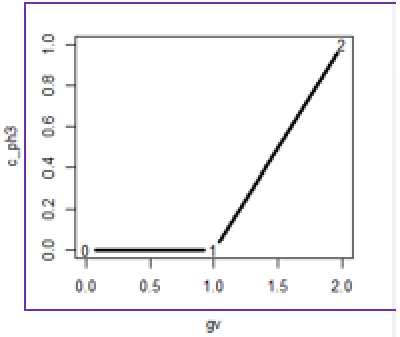
a. **unrelated people**
b. **Parent-offspring**
c. **MZ twins**

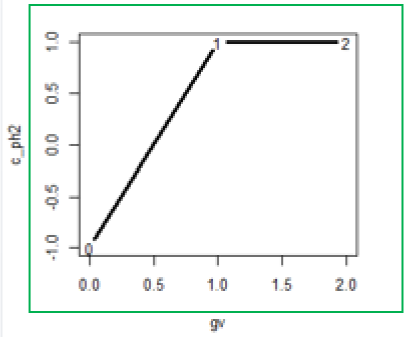
Proportion of shared variance

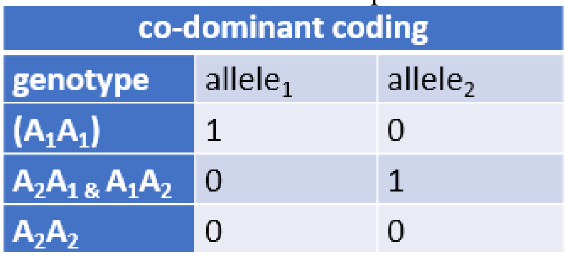
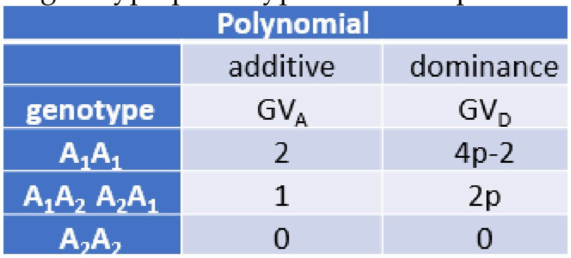
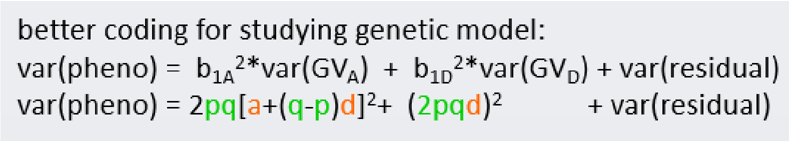


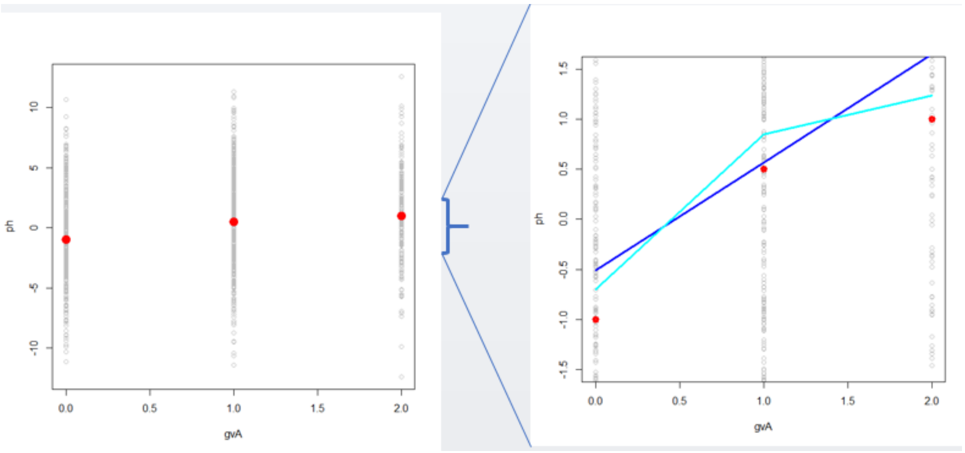
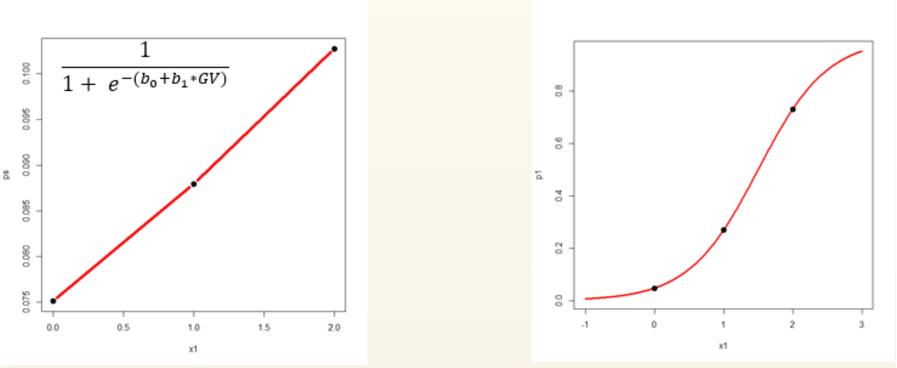
	$2pq[a+(q-p)d]^2$ $s^2_{QTL(A)} = b_1^2 s^2_{GV}$	$(2pqd)^2$ $s^2_{QTL(D)}$	
IBD=0	0	0	Unrelated
IBD=1	½	0	Parent - Offspring
IBD=2	1	1	MZ twins
IBD=0	0	0	25% DZ twins
IBD=1	½	0	50% DZ twins
IBD=2	1	1	25% DZ twins

What is the proportion of dominance variance shared by **a. unrelated people** **b. Parent-offspring** **c. MZ twins**	Additive genetic components - Not shared by unrelated people, shared 50% by parent and offspring Dominant genetic components - Not shared by parents and offspring

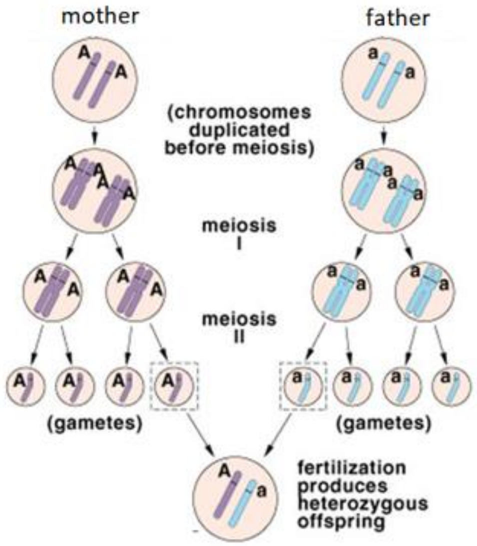
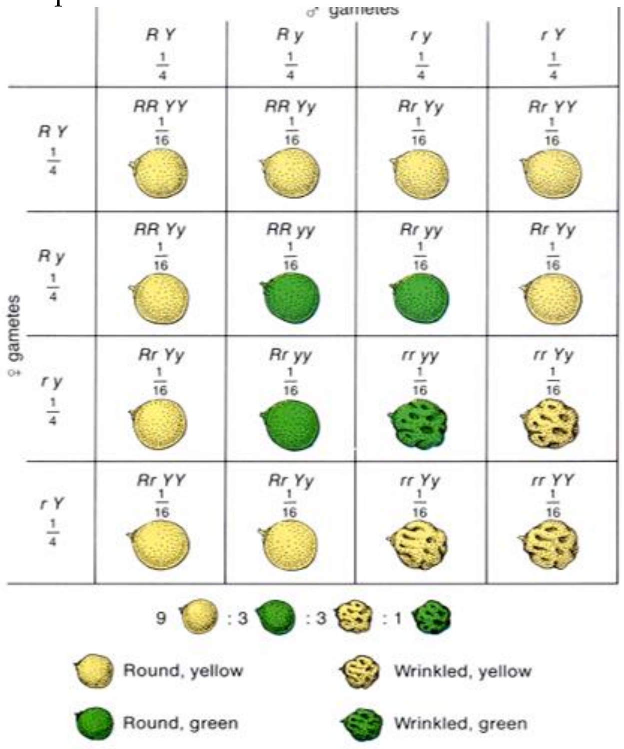
What is the difference between additive coding, dominant coding and recessive coding?	Additive coding
	
	AA = 2
	Aa = 1
	aa = 0
	
	Dominant coding
	
	AA = 1
	Aa = 1
	aa = 0
	
	Recessive coding

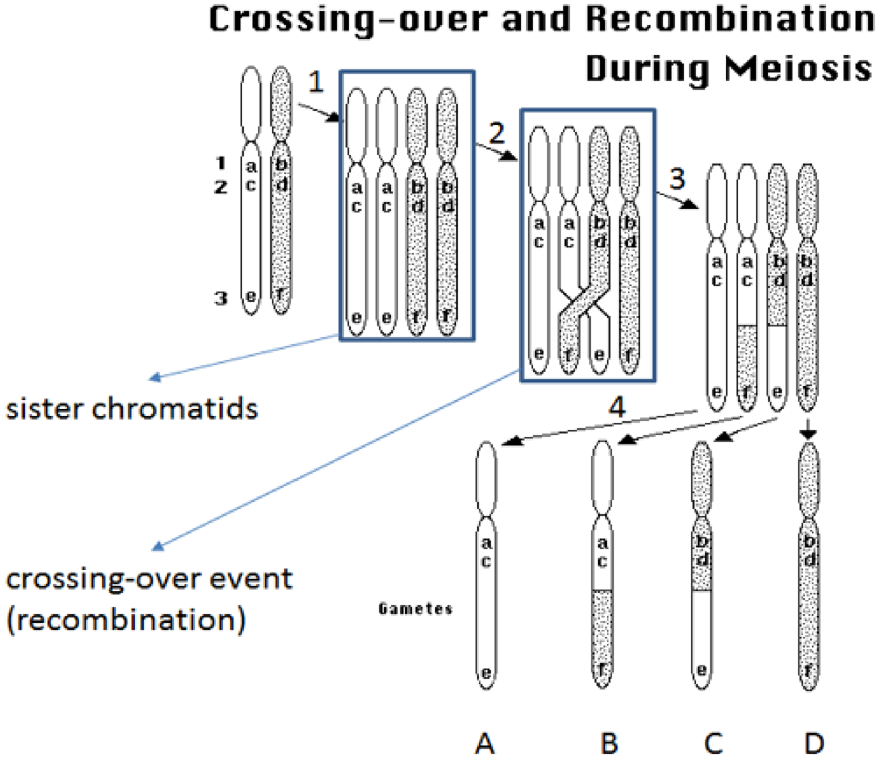
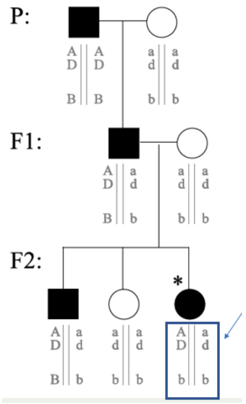
<table>
<tr><td></td><td>


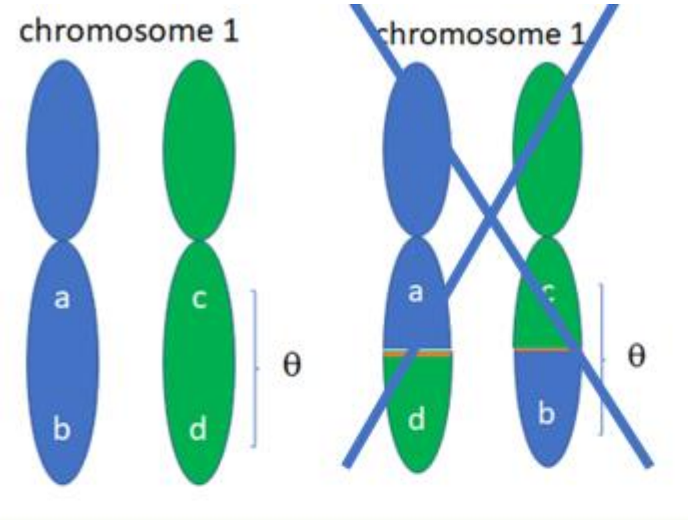
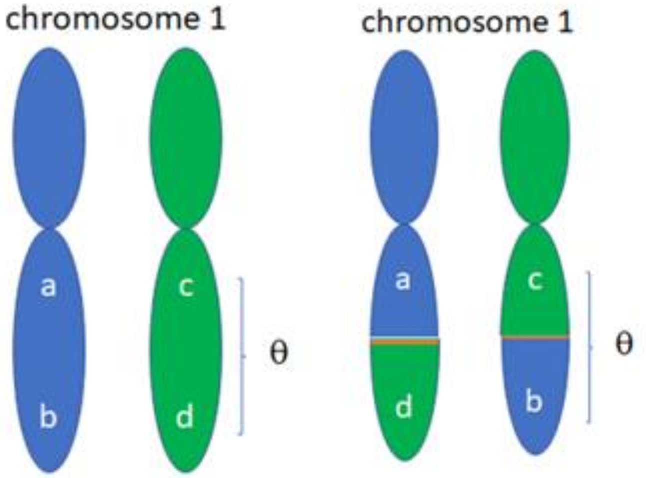
AA = 1
Aa = 0
aa = 0
Figure - Dominant coding and recessive coding
</td></tr>
<tr><td><b>How to represent non-linear relationships in a regression model?</b></td><td>
What about non-linear relationships?

<table>
<tr><th colspan="3">co-dominant coding</th></tr>
<tr><td>genotype</td><td>allele$_1$</td><td>allele$_2$</td></tr>
<tr><td>(A_1A_1)</td><td>1</td><td>0</td></tr>
<tr><td>A_2A_1 &amp; A_1A_2</td><td>0</td><td>1</td></tr>
<tr><td>A_2A_2</td><td>0</td><td>0</td></tr>
</table>
Dummy coding - Allelic association test
</td></tr>
<tr><td><b>How is the additive and dominant effects of genotype described in model to explain variance of phenotype?</b></td><td>
Explore genotype-phenotype relationship

<table>
<tr><th colspan="3">Polynomial</th></tr>
<tr><td></td><td>additive</td><td>dominance</td></tr>
<tr><td>genotype</td><td>GV_A</td><td>GV_D</td></tr>
<tr><td>A_1A_1</td><td>2</td><td>4p-2</td></tr>
<tr><td>A_1A_2 A_2A_1</td><td>1</td><td>2p</td></tr>
<tr><td>A_2A_2</td><td>0</td><td>0</td></tr>
</table>

better coding for studying genetic model:
var(pheno) = b_{1A}^2*var(GV_A) + b_{1D}^2*var(GV_D) + var(residual)
var(pheno) = $2pq[a+(q-p)d]^2 + (2pqd)^2$ + var(residual)
Reject null hypothesis - Dominance is present
Coding guarantees that the correlation between Gva and Gvd is zero (avoid multicollinearity in the model)
</td></tr>
<tr><td><b>How can dominance be visually identified from a graph?</b></td><td>Graph - a big difference between the lines indicates dominance</td></tr>
</table>

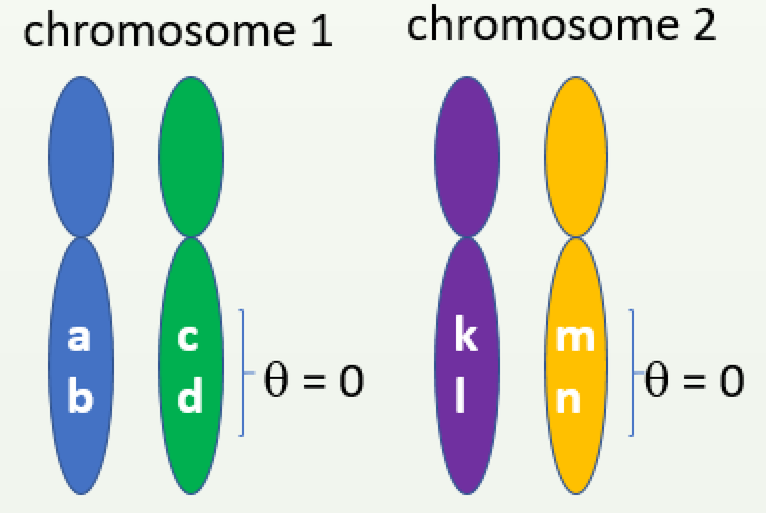
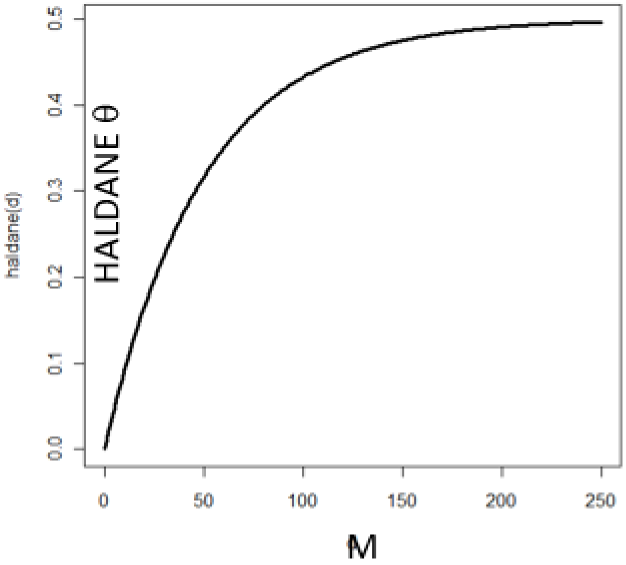
	
What is the difference between a simple linear model and a logistic regression?	Dichotomous outcome variable - A logistic regression model must be used Probability of scoring 1 (having the disease) based on the genotype variant logistic function $prob(1\|GV) = \frac{1}{1+ e^{-(b_0+b_1*GV)}}$ Exp (x) = e raised to the power of x
What happens when there is a big effect of genetic relative risk in a logistic regression model?	Genetic relative risk - Ratio of risk based on a reference (condGV=1/condGV=0)  1 = no effect Large = big effect (the regression line becomes S-shaped)

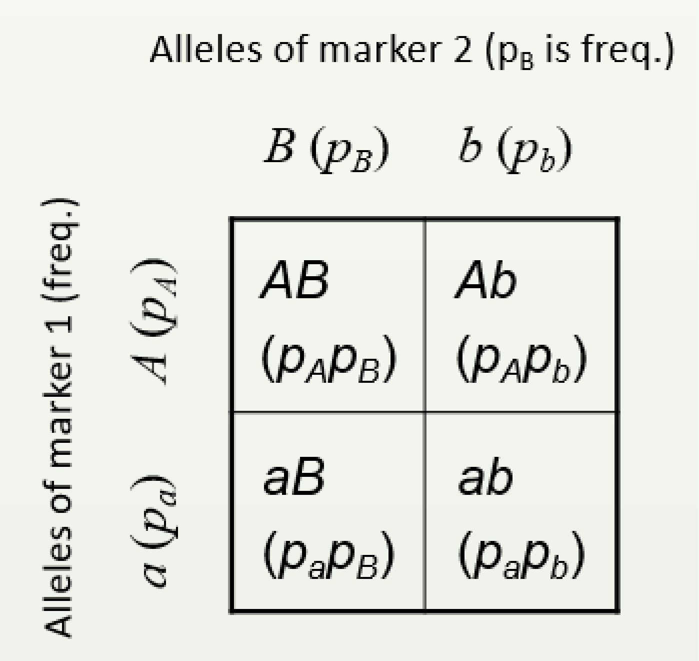
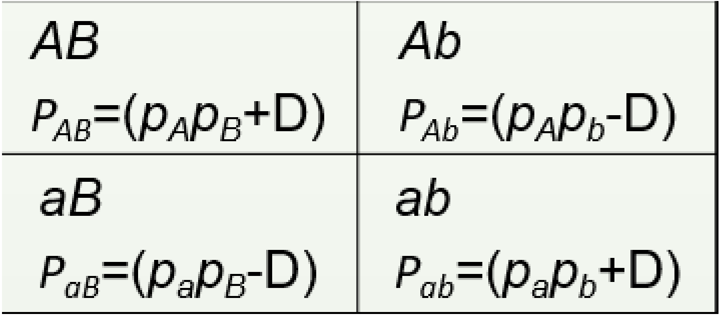
3a. Linkage equilibrium and disequilibrium

What is Mendel's first law of segregation?	Mendel’s first law of segregation - Parents alleles are chosen at random Random selection - 1 out 4 for each parent  mother father A A a a (chromosomes duplicated before meiosis) meiosis I meiosis II (gametes) (gametes) fertilization produces heterozygous offspring
What is Mendel's second law of segregation? **When is Mendel's second law of segregation violated?**	Mendel’s second law of segregation - The assortment for each allele is independent from one another  ♂ gametes R Y 1/4 R y 1/4 r y 1/4 r Y 1/4 ♀ gametes R Y 1/4 R y 1/4 r y 1/4 r Y 1/4 RR YY 1/16 RR Yy 1/16 Rr Yy 1/16 Rr YY 1/16 RR Yy 1/16 RR yy 1/16 Rr yy 1/16 Rr Yy 1/16 Rr Yy 1/16 Rr yy 1/16 rr yy 1/16 rr Yy 1/16 Rr YY 1/16 Rr Yy 1/16 rr Yy 1/16 rr YY 1/16 9 : 3 : 3 : 1 Round, yellow Wrinkled, yellow Round, green Wrinkled, green Exception; Linkage - Proximity of genetic loci in a chromosome

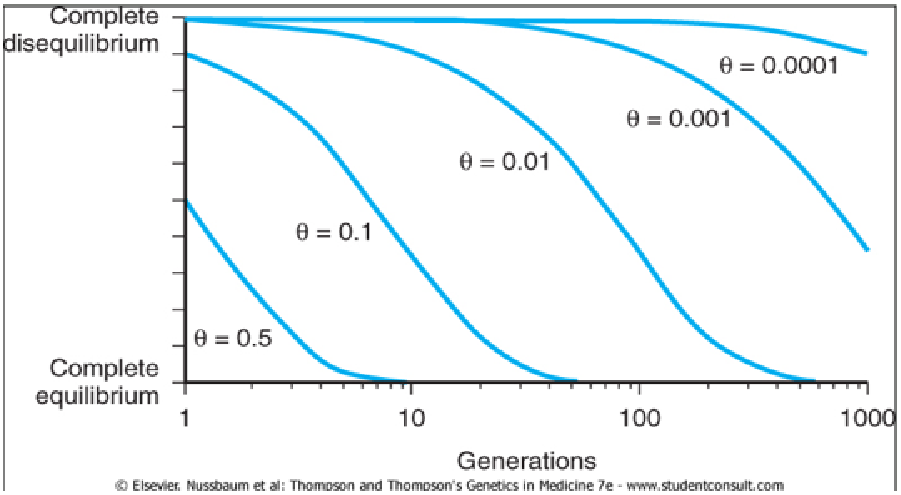
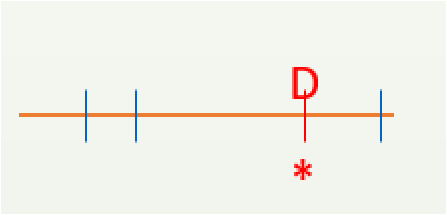
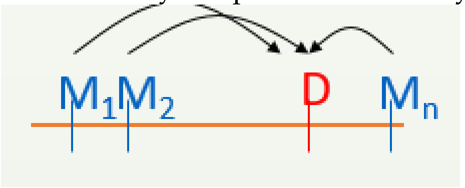
	Creates gametic phase disequilibrium
What is recombination?	Crossing over and recombination Crossing over happens during second phase of meiosis  Crossing-over and Recombination During Meiosis 1 2 3 4 1 2 3 a c b d e f sister chromatids crossing-over event (recombination) Gametes A B C D Chromosomes A and D - No crossing over Alleles a and c - Violates Mendel's second law because they are transmitted together Chromosomes B and C - Only happens due to cross over
What happens when two loci are very close together regarding Mendel's second law? And when they are in different chromosomes or very far apart in the same chromosome?	Probability of recombination depends of the distance Large distance - Probability is 0.5 (Mendel's second law holds) Short distance - Probability is very close to 0  ABD is a haplotype ABd is a new haplotype

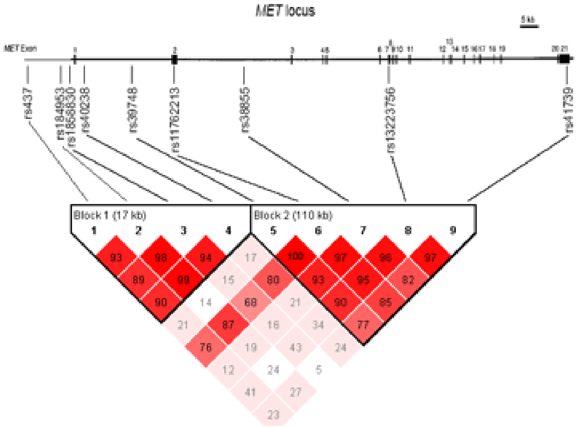
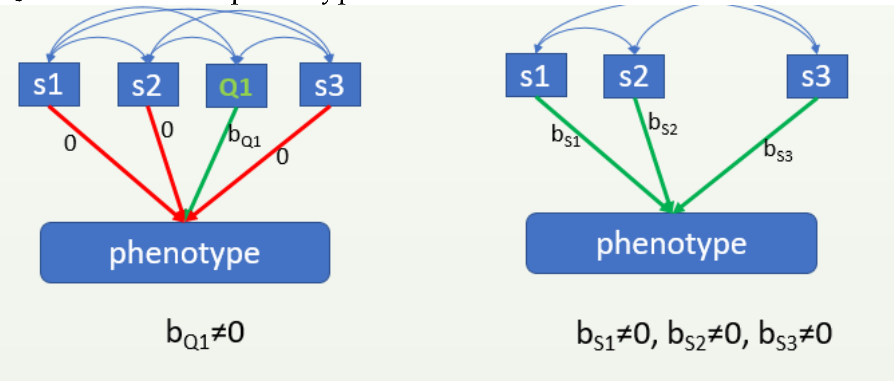
What is theta?	Same chromosomes - Theta is the recombination probability Theta is a function of distance - Base pairs of DNA  chromosome 1 a b c d θ chromosome 1 a d c b θ If theta is 0 - Recombination is unlikely Once you observe one allele, you are likely to observe the other (markers can be used to observe causal genes)  chromosome 1 a b c d θ chromosome 1 a d c b θ If theta is 0.5 - Recombination is as likely as non-recombination
What is the value of theta when the loci are in different chromosomes?	Different chromosomes - Theta is 0.5 (chromosomes are chosen at random, there is no dependency)

	 chromosome 1 chromosome 2 a b c d θ = 0 k l m n θ = 0
What is the unit of theta?	Distance in chromosome - Expressed in centimorgans 1 Morgan = 1 million basepairs 1 centiMorgan = 10 thousand basepairs  Theta - Depends on centimorgans (distance between loci)
If the haplotypes depend on allele probability, what is the value of theta?	Quantifying close: Linkage equilibrium Haplotypes depend on the allele probability Loci are independent - Either very far apart in the same chromosome or in different chromosomes (theta = 0.5)

<table>
<tr><td></td><td>
<table>
<tr><td colspan="2"></td><td colspan="2">Alleles of marker 2 (p_B is freq.)</td></tr>
<tr><td></td><td></td><td>$B\ (p_B)$</td><td>$b\ (p_b)$</td></tr>
<tr><td rowspan="2">Alleles of marker 1 (freq.)</td><td>$A\ (p_A)$</td><td>AB
(p_Ap_B)</td><td>Ab
(p_Ap_b)</td></tr>
<tr><td>$a\ (p_a)$</td><td>aB
(p_ap_B)</td><td>ab
(p_ap_b)</td></tr>
</table></td></tr>
<tr><td><b>What is the influence of linkage disequilibrium and HW disequilibrium?</b></td><td>Linkage disequilibrium between two markers
Has no bearing on Hardy Weinberg equilibrium (genotypes at each marker do not have to be randomly selected)</td></tr>
<tr><td><b>What is D'? What happens when D' equals 0?</b></td><td>Measure of LD: D′ (standardization value that indicates linkage disequibrilium)

<table>
<tr><td>AB
P_{AB}=(p_Ap_B+D)</td><td>Ab
P_{Ab}=(p_Ap_b-D)</td></tr>
<tr><td>aB
P_{aB}=(p_ap_B-D)</td><td>ab
P_{ab}=(p_ap_b+D)</td></tr>
</table>
If D′ is not 0, the loci are correlated (Linkage equilibrium)
If D′ is different from 0, loci are not perfectly correlated (Linkage disequilibrium)

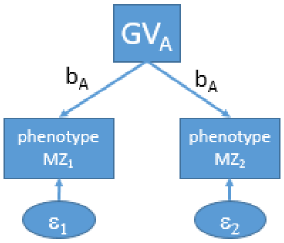
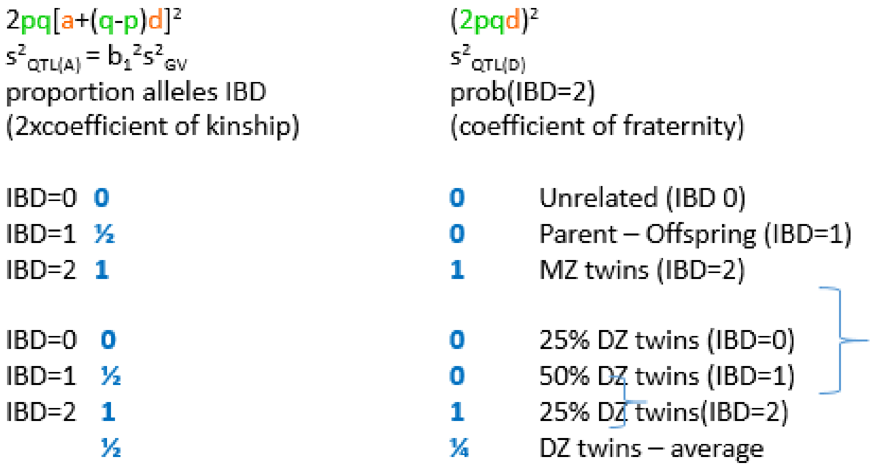
A value of D′ on its own is not very informative</td></tr>
<tr><td><b>What is the cause of linkage</b></td><td>Linkage disequilibrium decays over generations as a function of theta</td></tr>
</table>

disequilibrium in nature? And linkage equilibrium?	 © Elsevier. Nussbaum et al: Thompson and Thompson's Genetics in Medicine 7e - www.studentconsult.com Disequilibrium happens due to a mutation Equilibrium happens due to crossing over The higher the theta, the faster disequibrilium decays The smaller the theta, the less likely it is to break the loci apart If the loci are in two different chromosomes, it only takes one generation to reach equilibrium
What is r squared related to Linkage Disequilibrium? How can it be calculated?	R squared - Direct function of LD Same as correlation between two loci squared R square is 0, LE R square is 1, LD
Why should we use markers in GWAS studies instead of measuring the disease locus directly?	Direct and indirect allelic association D = Disease locus  Direct - Measure disease relevance directly Usually not possible because you don't know what it is  Indirect - Measure markers to assess trait effects on D

	Hits on a GWAS does not mean much -> you still don’t know what D is, only a range of where it can be -> posterior studies are necessary Significance of markers (red squares) - Indicate upper and lower boundaries of where the disease locus is 
How can GWAS studies be significant if the effect of each QTL is small and the alpha level is tiny?	GWAS - Measure GV and phenotype -> linear or logistic analysis Equation E[pheno|GV] = $b_0 + b_1$*GV (conditional mean or predicted value) E[pheno=1|GV] = 1 / {1-exp(-($b_0 + b_1$*GV))} (conditional probability) If effect size is small and alpha is small, you need a very large sample size to pick up values
What happens when you included marker and disease locus in a regression?	Example Red lines - Not significant Green lines - Significant S - Markers Q - Causal loci to phenotype  s1 s2 Q1 s3 0 0 b_{Q1} 0 phenotype $b_{Q1}≠0$ s1 s2 s3 b_{S1} b_{S2} b_{S3} phenotype $b_{S1}≠0, b_{S2}≠0, b_{S3}≠0$ If Q is included, markers will not explain anything (Q absorbs) If Q is not included, markers will explain something (because they are correlated with Q)

What are the three common forms of association in GWAS?	Association in GWAS: 3 common forms Direct association - Measure of SNP of interest that is causing the disease Indirect association - Markers Spurious association - You reject the null hypothesis, but the SNP is not related at all
What is a probable cause of spurious associations in GWAS?	Cause of spurious associations: Population stratification If you put two different populations together, you may get significant results

3b. Classical twin design

Why are twin studies used to study genetic effects?	Example: CHRM2 gene - Feedback and regulation of Ach release, related with higher cognitive processing Find a candidate gene first Posterior study confirmed the results
Why are GWAS meta-analyses becoming more and more common?	GWAS meta-analysis Compendium of dozens of studies - Increase sample size GWAS opens the black box - One SNP at a time There are too many QTLs for complex traits like IQ Variance of IQ = additive effects + dominant effects + error (environmental effects)
What is shared variance?	Shared variance - If two variables are subject to the same influence, they share variance attributable to that influence  GV_A b_A b_A phenotype MZ_1 phenotype MZ_2 ε_1 ε_2 Covariance - Twins share genetic variance
Why are dyzygotic twins share 25% of their genetic variance IDB?	Genetic resemblance - Not affected by recombination/crossing over MZ twins are genetic identical - IBD 2 DZ twins - IBD 0, 1 or 2 -> average of 1/4 Half siblings  $2pq[a+(q-p)d]^2$ $(2pqd)^2$ $s^2_{QTL(A)} = b_1^2 s^2_{GV}$ $s^2_{QTL(D)}$ proportion alleles IBD prob(IBD=2) (2xcoefficient of kinship) (coefficient of fraternity) IBD=0 0 0 Unrelated (IBD 0) IBD=1 ½ 0 Parent – Offspring (IBD=1) IBD=2 1 1 MZ twins (IBD=2) IBD=0 0 0 25% DZ twins (IBD=0) IBD=1 ½ 0 50% DZ twins (IBD=1) IBD=2 1 1 25% DZ twins(IBD=2) ½ ¼ DZ twins – average

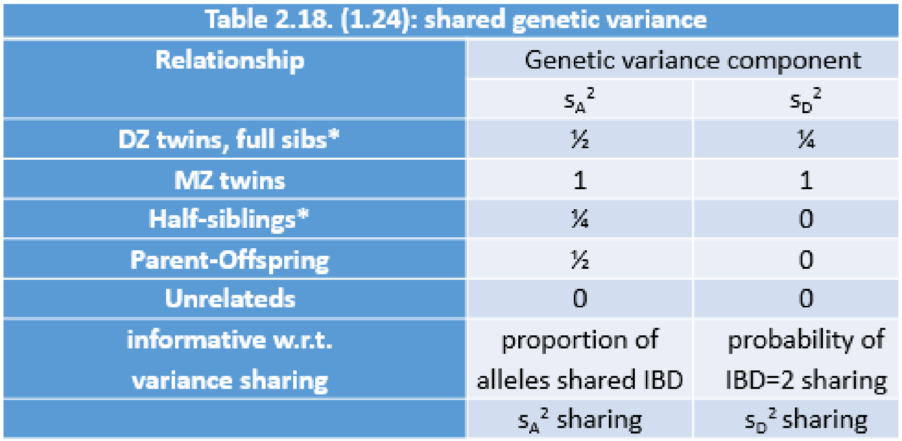


Table 2.18. (1.24): shared genetic variance

Relationship	Genetic variance component	
	s_A^2	s_D^2
DZ twins, full sibs*	½	¼
MZ twins	1	1
Half-siblings*	¼	0
Parent-Offspring	½	0
Unrelateds	0	0
informative w.r.t. variance sharing	proportion of alleles shared IBD	probability of IBD=2 sharing
	s_A^2 sharing	s_D^2 sharing

Which answer does the classical twin design try to answer?

Classical twin design

How many variance of additive and dominance effects depends on genetic relatedness

Explained variance of the genetic variant with relation to the phenotype = coefficient component + variance of the genetic variant

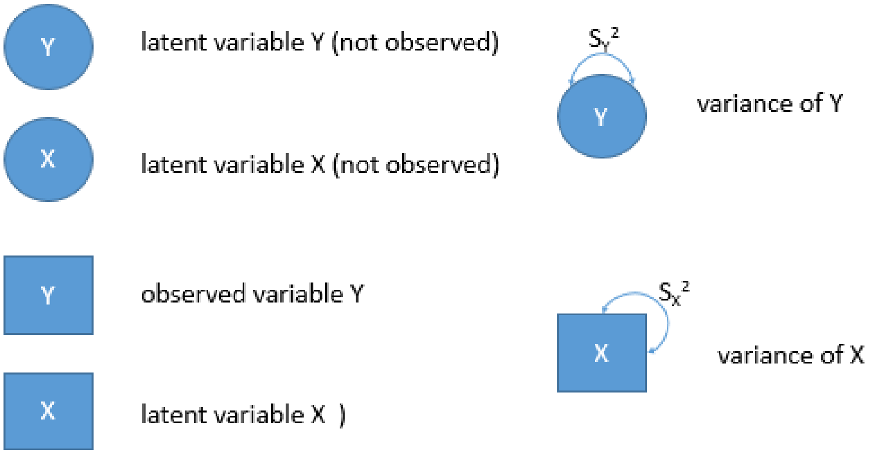
Why is pathdiagram notation useful?

Pathdiagram notation - Regression model

Square - Observed variable

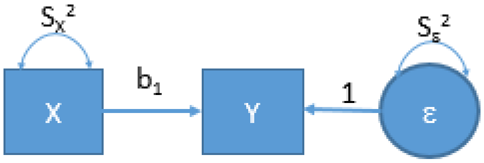
Circle - Non-observable variable

Double headed arrow - Variance of X



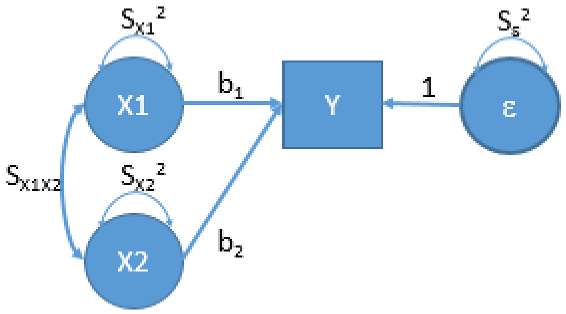
Images

Covariance



<table>
<tr>
<td></td>
<td>

Regression (direction of effect - single arrow)



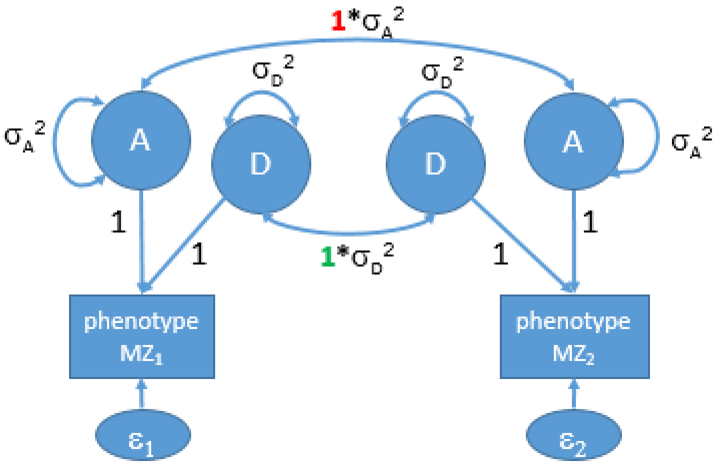
Regression with two terms - Term that represents correlation between the variables (2.b1.b2)

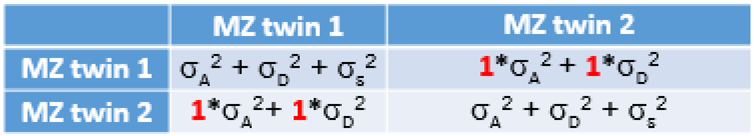
</td>
</tr>
<tr>
<td>

Draw a pathdiagram of MZ twins.

</td>
<td>

Pathdiagram of monozygotic twins





<table>
<tr><th></th><th>MZ twin 1</th><th>MZ twin 2</th></tr>
<tr><td>MZ twin 1</td><td>$\sigma_A^2 + \sigma_D^2 + \sigma_s^2$</td><td>$1*\sigma_A^2 + 1*\sigma_D^2$</td></tr>
<tr><td>MZ twin 2</td><td>$1*\sigma_A^2+ 1*\sigma_D^2$</td><td>$\sigma_A^2 + \sigma_D^2 + \sigma_s^2$</td></tr>
</table>

beh gen – Lecture 3 2018

Seconds representation is clearer
D - Hypothetical latent variable
IBD = 1

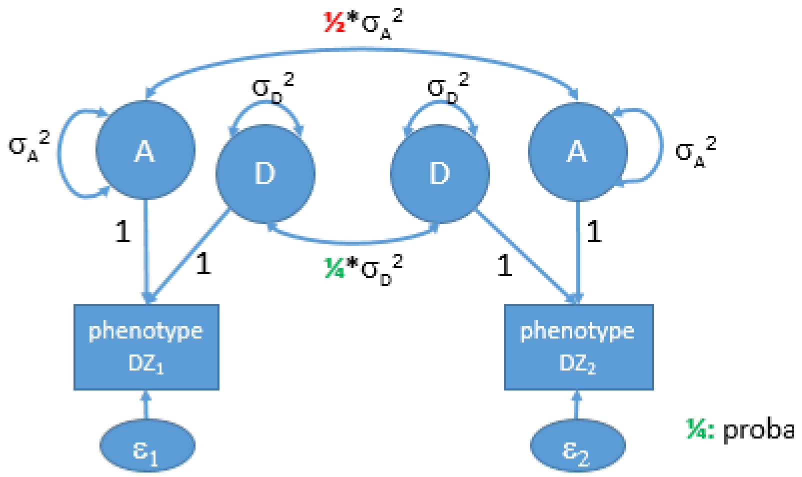
</td>
</tr>
<tr>
<td>

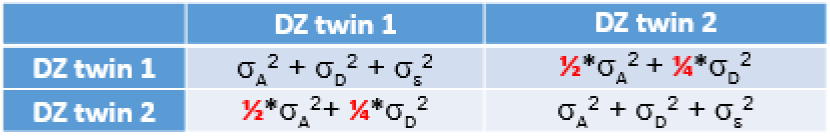
Draw a pathdiagram of dyzygotic twins.

</td>
<td>

Pathdiagram of dizygotic twins

</td>
</tr>
</table>



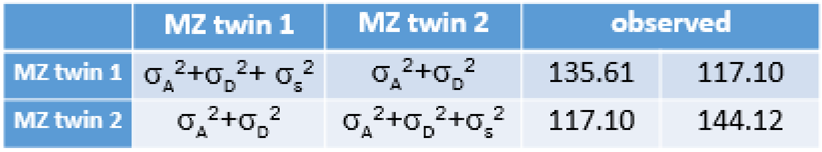


	DZ twin 1	DZ twin 2
DZ twin 1	$\sigma_A^2 + \sigma_D^2 + \sigma_\varepsilon^2$	½*σ_A^2 + ¼*σ_D^2
DZ twin 2	½*σ_A^2+ ¼*σ_D^2	$\sigma_A^2 + \sigma_D^2 + \sigma_\varepsilon^2$

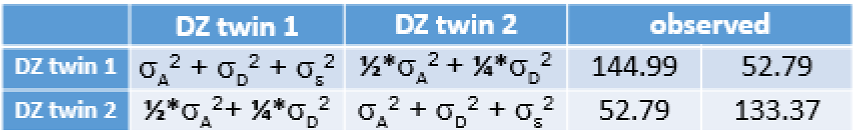
IBD = 1/2

How can you calculate the covariance matrix of twins?

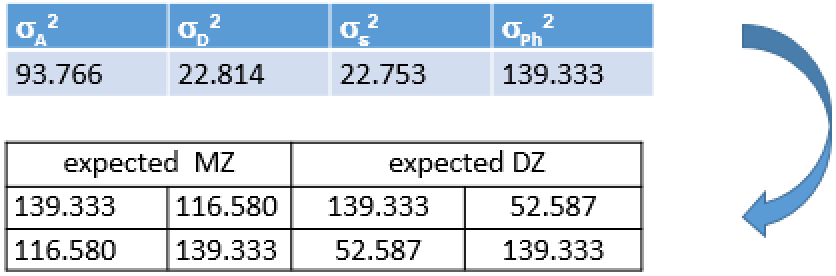
Covariance matrix of twins



	MZ twin 1	MZ twin 2	observed	
MZ twin 1	σ_A^2+σ_D^2+ σ_ε^2	σ_A^2+σ_D^2	135.61	117.10
MZ twin 2	σ_A^2+σ_D^2	σ_A^2+σ_D^2+σ_ε^2	117.10	144.12



	DZ twin 1	DZ twin 2	observed	
DZ twin 1	$\sigma_A^2 + \sigma_D^2 + \sigma_\varepsilon^2$	½*σ_A^2 + ¼*σ_D^2	144.99	52.79
DZ twin 2	½*σ_A^2+ ¼*σ_D^2	$\sigma_A^2 + \sigma_D^2 + \sigma_\varepsilon^2$	52.79	133.37



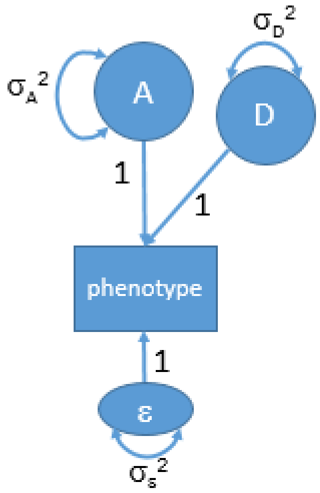
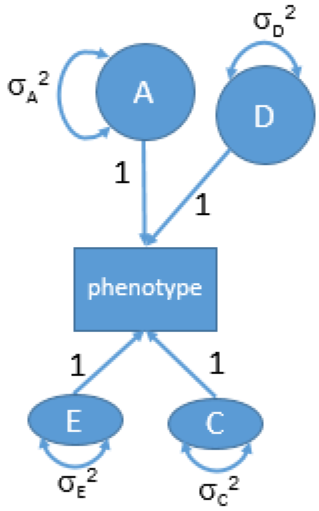
σ_A^2	σ_D^2	σ_ε^2	σ_{Ph}^2
93.766	22.814	22.753	139.333

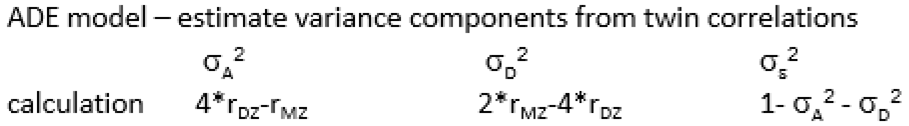
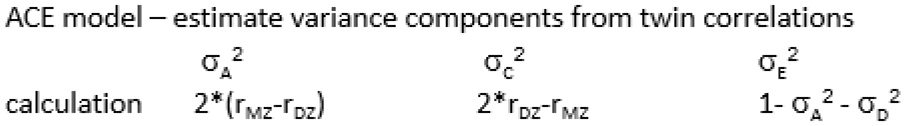
expected MZ		expected DZ	
139.333	116.580	139.333	52.587
116.580	139.333	52.587	139.333

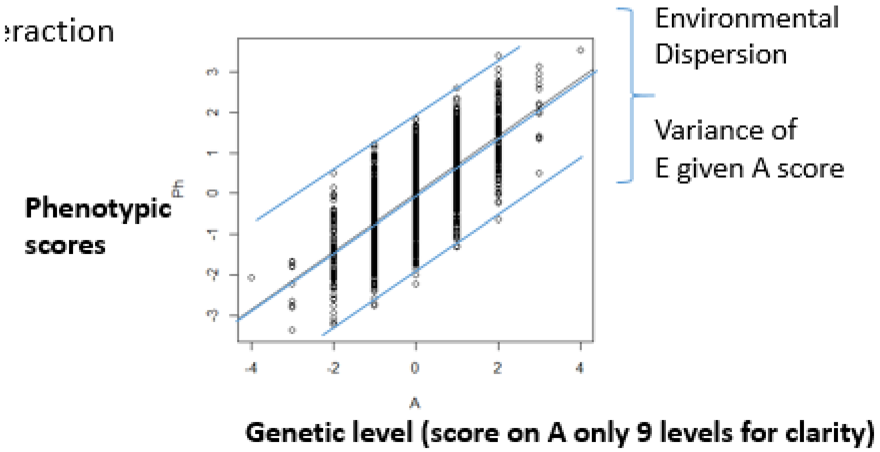
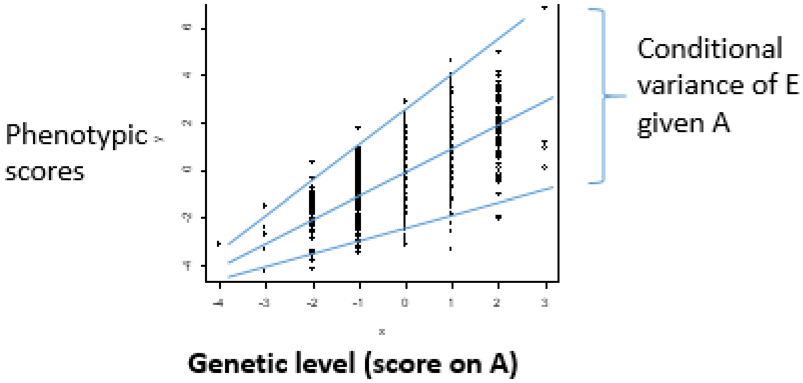
Monozygotic twin
Covariance = QTLa + QTLd

Dizygotic twin
Covariance = 1/2QTLa + 1/4QTLd

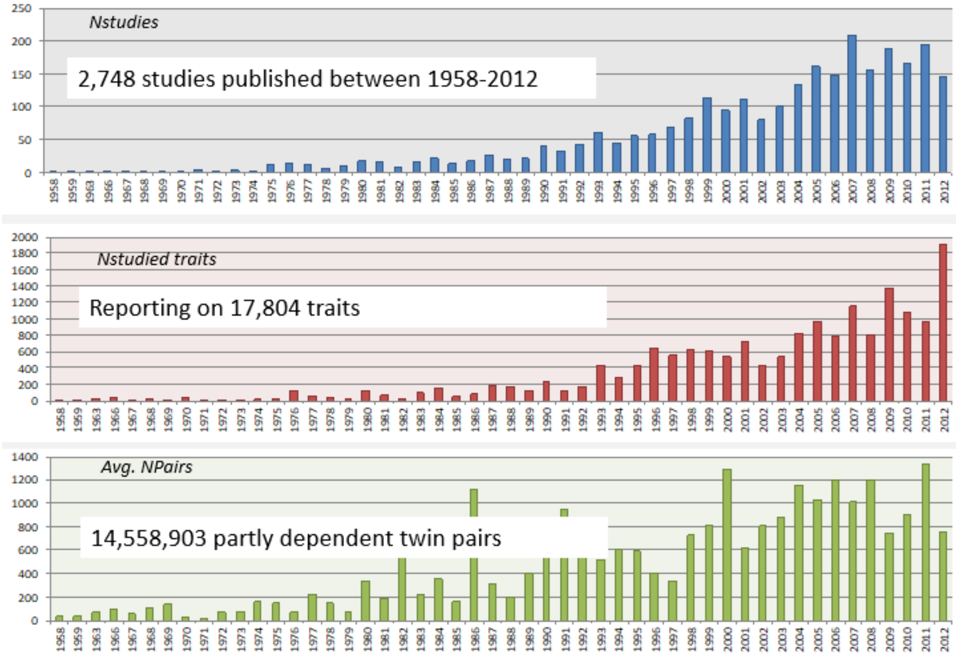
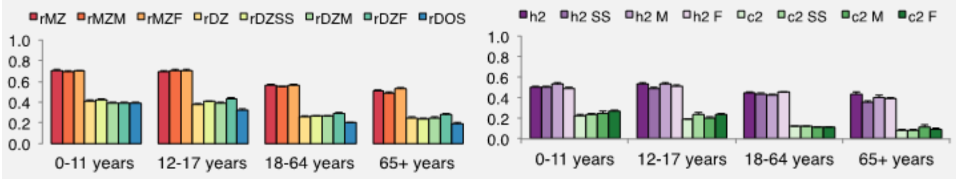
Expected model should resemble the observed effects

What does A, D, C and E stand for in a twin study model?	A and D are random variables  σ_A^2 A D σ_D^2 1 1 phenotype 1 ε σ_ε^2  σ_D^2 σ_A^2 A D 1 1 phenotype 1 1 E C σ_E^2 σ_C^2 $\sigma_{Ph}^2 = \sigma_A^2 + \sigma_D^2 + \sigma_\varepsilon^2$ $\sigma_{Ph}^2 = \sigma_A^2 + \sigma_D^2 + \sigma_C^2 + \sigma_E^2$ A - Additive effects D - Dominance effects E - Environmental variation C - Shared environmental variants E - Non-shared environmental variants Error is not included in the model
How are addivite, dominant and environmental effects calculated in a twin study?	Additive + Dominant = Genetic component (broad sense heritability) Overall effect Additive only - Narrow sense heritability σ_A^2 σ_D^2 σ_ε^2 $4*r_{DZ}-r_{MZ}$ $2*r_{MZ}-4*r_{DZ}$ $1-\sigma_A^2-\sigma_D^2$ 4*.18 - .41 2*.41-4*.18 1-.31-.10 .31 .10 .59 QTLa = 4*rdz-rmz QTLd = 2*rmz-4rdz QTLe = 1 - QTLa - QTLd (i.e. the rest)
What is the main problem with the classical twin design?	Twin design issue Distiction between genetic and environmental effect - You cannot fit a model with additive, dominant, shared environmental effects and unshared environmental effects ACDE model is not possible

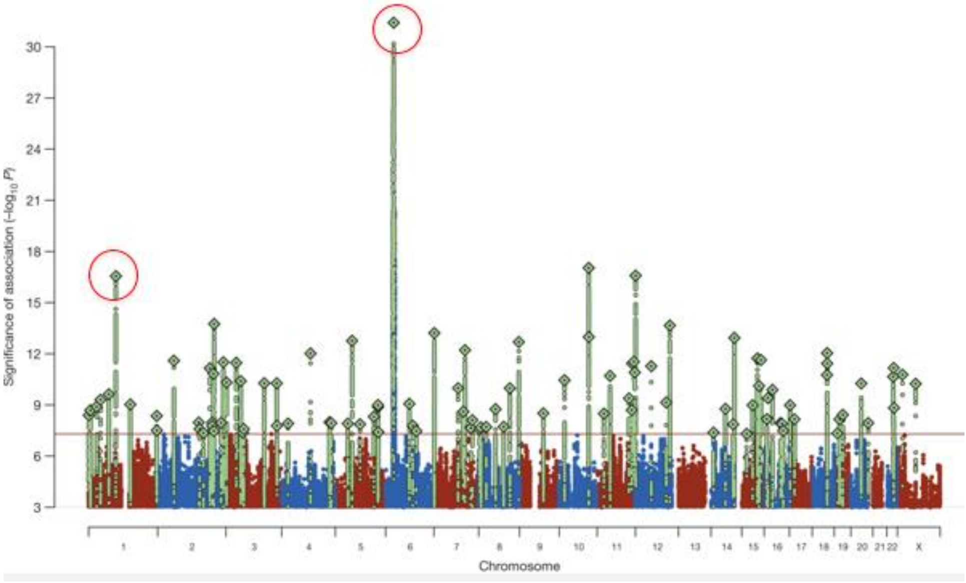
	ACE model is ok ADE model is ok E model = twins are not correlated CE model = monozygotic and dyzygotic twins are the same
What are the four types of models possible in a twin study?	 ADE model – estimate variance components from twin correlations σ_A^2 σ_D^2 σ_E^2 calculation $4*r_{DZ}-r_{MZ}$ $2*r_{MZ}-4*r_{DZ}$ $1- \sigma_A^2 - \sigma_D^2$  ACE model – estimate variance components from twin correlations σ_A^2 σ_C^2 σ_E^2 calculation $2*(r_{MZ}-r_{DZ})$ $2*r_{DZ}-r_{MZ}$ $1- \sigma_A^2 - \sigma_D^2$ If rMZ > 2*rDZ = ADE If rMZ < 2*rDZ = ACE If rMZ = 2*rDZ = AE model If rMZ = rDZ = E model When the difference is not very large -> Small dominance component
Why twin studies challenged the status quo in the 80s?	Twin studies - Highlight the importance of genetic variation In the 80's, people denied the influence of genetics for complex behaviors
What are the assumptions of the classical twin design?	Assumptions of CTD • The sample is representative of the population • Random mating - No correlation between the parents, which is often not true (high IQ children have high IQ parents) • Equal environmental assumption • No correlation between A-C or A-E - Safe assumption for personality, risky assumption for IQ (IQ correlates with high economic status, affects the offspring environment) • No Genetic-environment correlation - Younger children intelligence are more dependent on shared environmental effects, older children are more dependent on genes
What is the difference when there is AxE	No AxE correlation

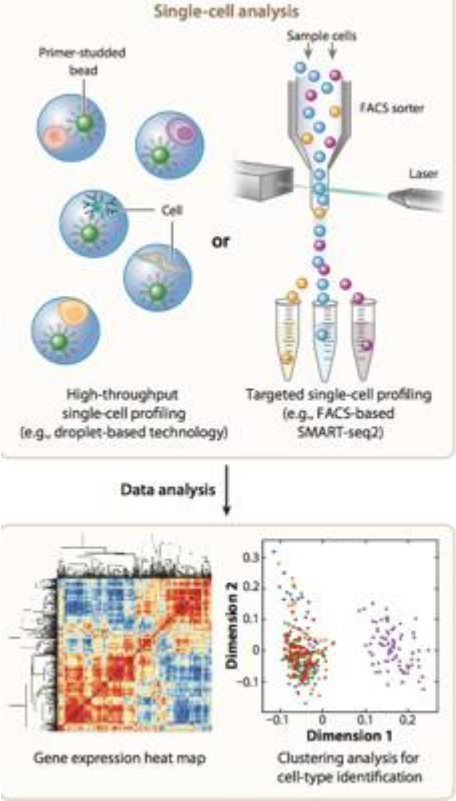
correlation and when there is not?	 AxE correlation  Depending on A score, environmental effects change
What can you do if the premises of the classical twin design are violated?	What you can do regarding violation of the classic twin design: • Extend models (include parents) • Extend assumption - Include more information • Violations of assumption are well known -> Data can be interpreted in a specific light

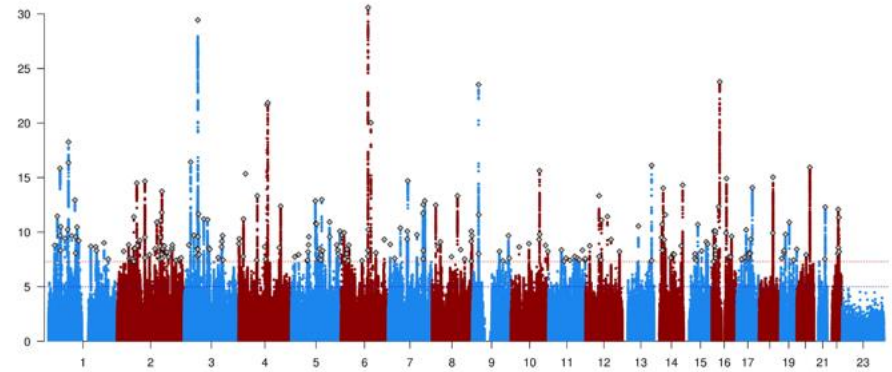
4. From heritability to functional experiments

Why is the field of human genetics developing so rapidly?	Field of human genetics is rapidly advancing in the recent past New technologies Large scale collaboration Novel disease insights
What was interesting to notice regarding the number of publications of Classic Twin Studies in the last few decades?	Meta-analysis of all twin studies Between 1900-2012 Extraction of sample size, effect size Standardize trait quantification -ICF There is still an increase of twin studies every year, even though we now can genotype individual very cheaply  2,748 studies published between 1958-2012 Reporting on 17,804 traits 14,558,903 partly dependent twin pairs Most studied traits: Weight, diseases, big-five personality traits, intelligence
How much of 'human behavior' is genetically inherited?	If you lump all traits together into 'human behavior'  Rmz = 0.636 RDZ = 0.339 Genetic inheritability = 0.49

	Shared environmental effect = 0.17
What were the main conclusion of meta-analysis of twin studies?	Main conclusion All traits are heritable to some extent Shared environmental effect is relatively small Most traits seem to have additive genetic effects Interactive website: match.ctglab.nl
What is heritability?	Heritability - Proportion of trait variance attributable to genetic variance Variation in genes underlie trait differences between individuals
How close are humans to: **a. Mice** **b. Chimpanzees** **c. Humans of the opposite gender**	If heritable - Look at the DNA 3 billion base pairs - One in every thousand base pair is different between people Human and mice - 87,5% Human and chimpanzee - 99% Males and females - 99,9% *50% in twins = Means 50% of 0.1%*
What is the difference between monogenic and	Monogenic disorders - One gene/one mutation Most genetic causes are already known Polygenic disorders - Influenced by multiple genes, each of small effect

polygenic disorders?	Genetic causes mostly unknown (correlation with environmental factors) Genetics variations (SNPs): Inside or outside genes Genetic variations can be: harmless, harmful, latent, silent
How were genetic studies made before the advent of GWAS?	Discovering genes for complex traits Candidate gene study - Before 2005 GWAS - Genotyping became cheap and accessible
What was the first GWAS hit discovered?	GWAS results First hit discovered - Macular degeneration (a few genes with large effect sizes) Most cases - Small effect sizes, large samples are needed Most genetic variance is not in the genes! Gene deserts or intronic regions
What do GWAS hits signify for researchers?	GWAS hits are hard to interpret SNPs are correlated in the same chromosome There are often hundreds of hits - it is difficult to choose a gene for a follow up study
What is a Manhattan plot?	Manhattan plot - GWAS results 
Why molecular studies can help us understand GWAS hits?	Interpreting GWAS loci Interference with DNA folding Positional information - SNP can indicate genes of interest

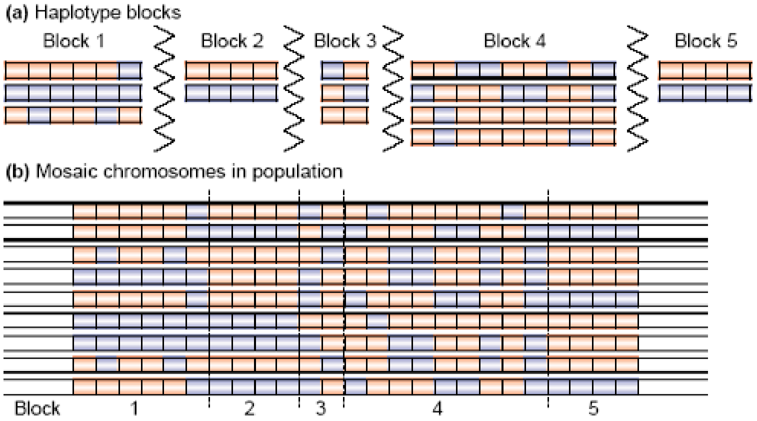
What is scRNAseq?	Novel techniques: scRNAseq - Gene expression at a cellular resolution (GWAS for specific cell types)  Single-cell analysis Primer-studded bead Sample cells FACS sorter Laser Cell or High-throughput single-cell profiling (e.g., droplet-based technology) Targeted single-cell profiling (e.g., FACS-based SMART-seq2) Data analysis Dimension 2 0.3 0.2 0.1 0 −0.1 −0.1 0 0.1 0.2 Dimension 1 Gene expression heat map Clustering analysis for cell-type identification
How to design follow up experiments from GWAS hits?	Designing follow up experiments Classical techniques (knockout, knockdown) - not viable for SNPS with small effect sizes Taking into account polygenic nature of traits: Chemo and ontogenetic techniques; induced pluripotent stem cells from patients with disease (create mini-brains in the lab)
What is the monozygotic and dizygotic twin correlation for IQ? Which model does this suggest?	Individual differences in IQ Twin correlation for IQ - 0.82 MZ and 0.45 DZ

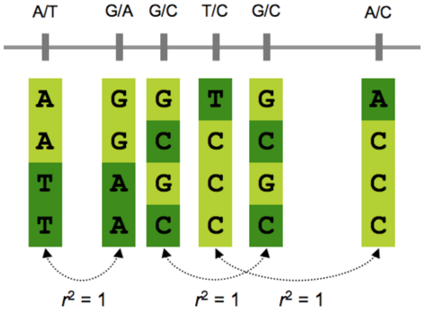
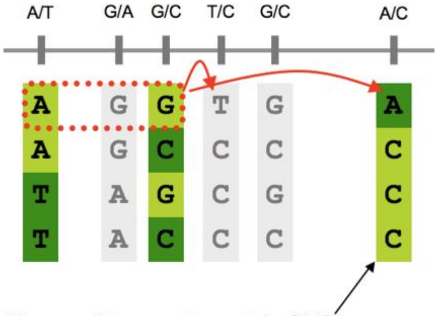
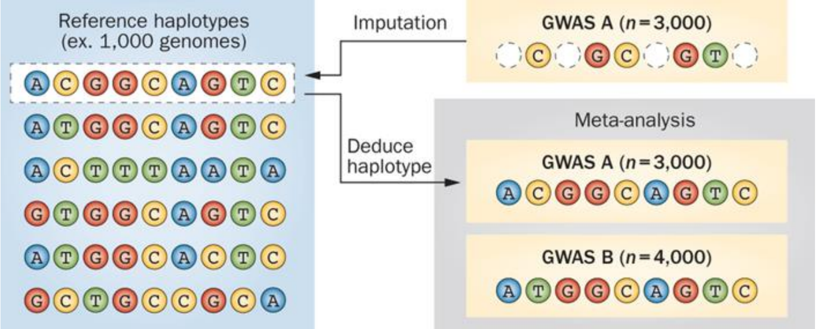


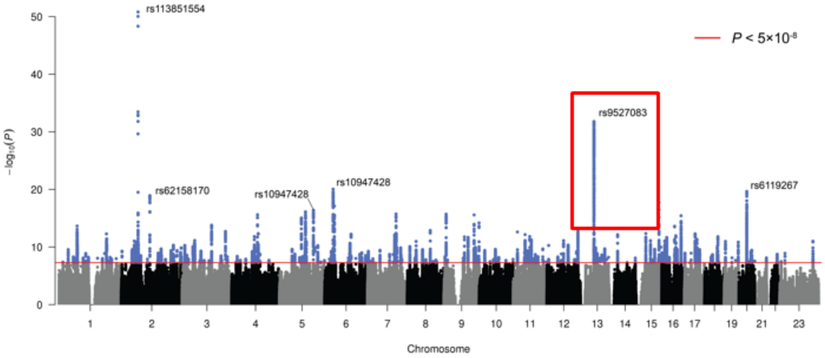
GWAS - 81 missense mutations associated with IQ
Mostly expressed in the brain
Expressed more in certain cell types - Medium spiny neurons, pyramidal CA1 neurons
Correlation with different traits

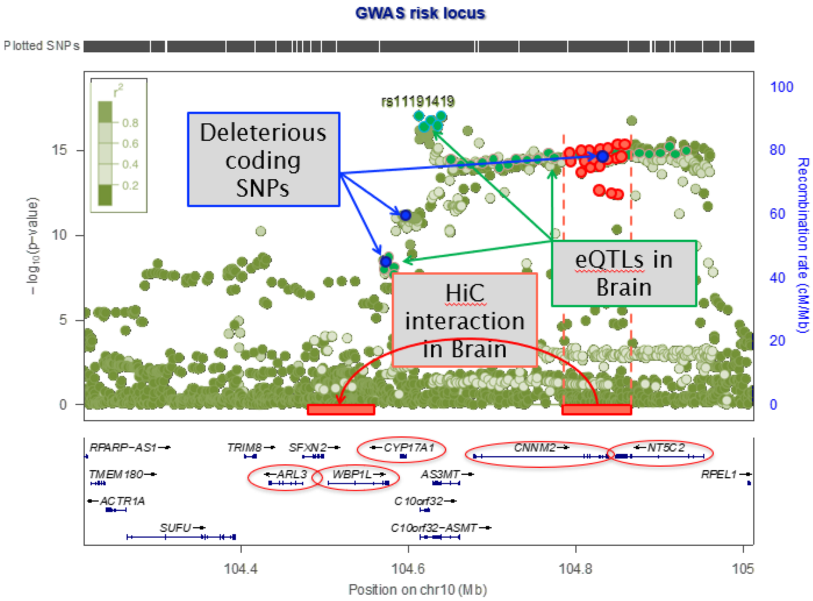
Score of GWAS cannot predict an individual's IQ score, only to distinguish individuals at a population level

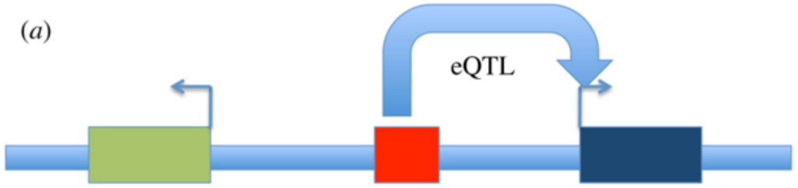
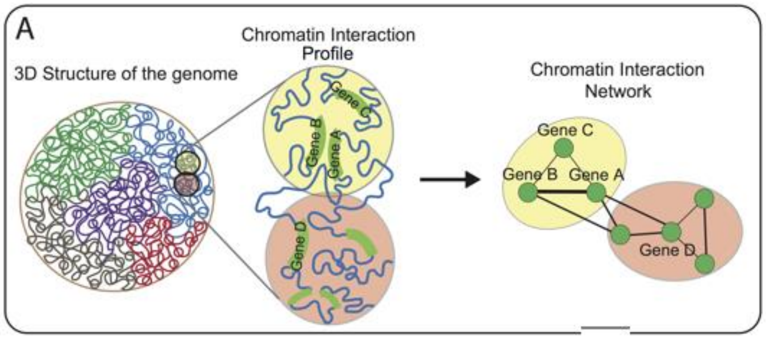
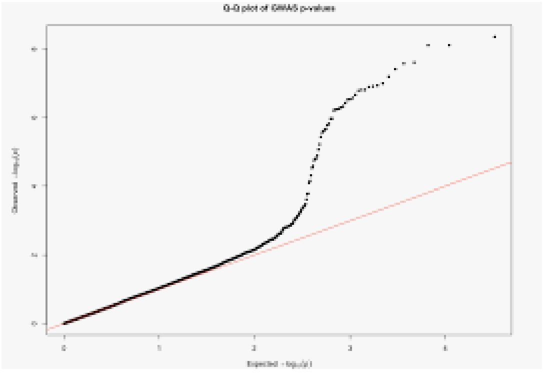
5. GWAS and interpretation of GWAS results

What is the purpose of Disease Genetics?	Disease Genetics Predict if someone will get sick Test hypothesis about relationships to other diseases or traits (comorbidity) Understand the biology of the disease so we can design better treatments and diagnostics
What is the main difference between GWAS and twin studies?	Twin and family studies do not measure the variation at all - They inform the contribution of genetics/environment
Why some GVs are correlated to one another?	What creates genetic variation Mutation - Spontaneous change in the DNA Recombination - Re-shuffles existing patterns of variation Consequences - Genetic variants are correlated because they share a history of inheritance Correlation decreases over generations due to recombination Depends on linkage disequilibrium
What did the HAPMAP project do? Why was this important?	Linkage disequilibrium map of the human genome HAPMAP projects - Genotyping of different populations with different ethnic backgrouns Limited haplotype diversity - Correlated blocks of SNPs  (a) Haplotype blocks Block 1 Block 2 Block 3 Block 4 Block 5 (b) Mosaic chromosomes in population Block 1 2 3 4 5 Only three options for a block of 6 SNPs

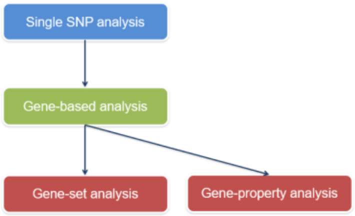
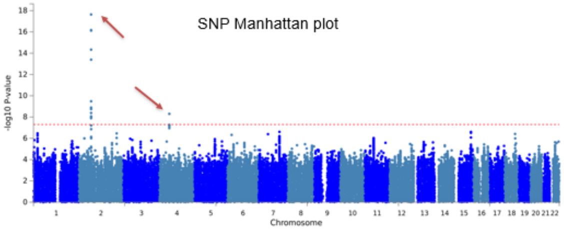
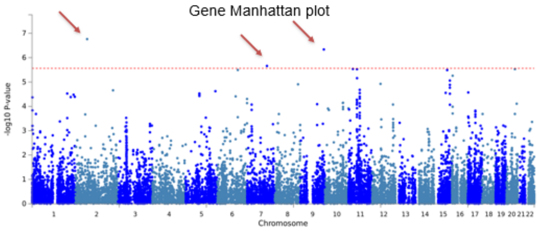
Why was the HAPMAP project important to increase genotype efficiency?	Genotyping a single SNP in a block provides all the information for all SNPs in this block - Efficient genotyping, less expensive  A/T G/A G/C T/C G/C A/C A G G T G A A G C C C C T A G C G C T A C C C C $r^2 = 1$ $r^2 = 1$ $r^2 = 1$  A/T G/A G/C T/C G/C A/C A G G T G A A G C C C C T A G C G C T A C C C C No need to genotype this SNP *Genotyping is always done in the plus strand*
What is the purpose of reimputing SNPs in a GWAS? What is the problem with this process?	Different companies genotype different SNPs - may be problematic when performing a meta-analysis  Reference haplotypes (ex. 1,000 genomes) A C G G C A G T C A T G G C A G T C A C T T T A A T A G T G G C A G T C A T G G C A C T C G C T G C C G C A Imputation GWAS A (n=3,000) C G C G T Deduce haplotype Meta-analysis GWAS A (n=3,000) A C G G C A G T C GWAS B (n=4,000) A T G G C A G T C You can reimpute the missing correlated SNPs This misses rare haplotypes or recent mutations in the population
What are the steps in GWAS?	Steps in GWAS until Manhattan plot Quality of genotyping is important - Heterozygous can be similar to homozygous Tests for genotyping error Imputation Association analysis - PLINK software Meta-analysis - Increase sample size Manhattan plot

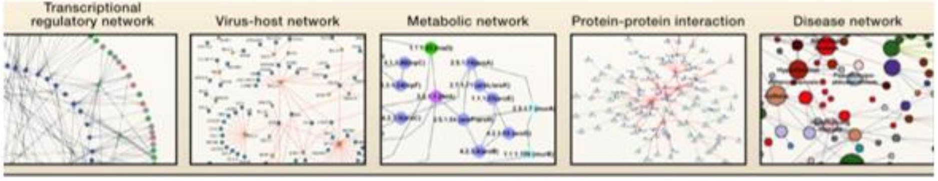
What is the common GWAS threshold?	Statistical tests for every SNP -> High number of false positives GWAS threshold is 5*10e-8 GOLD standard for association studies - Replicating association in different laboratories
What are examples of false replications and the arguments that were used to try to justify them?	Not true replications - and proposed explanations • Association to the same trait, but a different gene - Justified by Genetic heterogeneity (different populations) • Association to same trait, same gene, different SNPs - Justified by Allelic heterogeneity • Association to same trait, same gene, same SNP, opposite direction - Justified by Allelic heterogeneity/population differences • Association to different, but a different correlation phenotype - Justified by Phenotypic heterogeneity (depends on phenotype correlation!) • No association at all - Justified by "Sample size too small"
What is a true replication of a GWAS?	True replication - Same trait, same SNP, same allele, same direction of effect, different and independent population
What do single outlier dots in a GWAS might indicate?	Manhattan plot  'Line' is more reliable than single dots - Probably caused by genotyping error
How is causation proved from a GWAS analysis?	Functional studies - Where is this gene being expressed? Develop ideas for posterior causal studies Causation is proved with clinical trials/pharmacological assays
Why is it difficult to interpret GWAS hits?	Post GWAS annotation and interpretation LD complicates the interpretation of results - You do not measure causal genes ○ Some SNPs are known to have no effect ○ Some SNPs are known to have a direct effect - Change RNA structure or protein; can be located in exons or introns

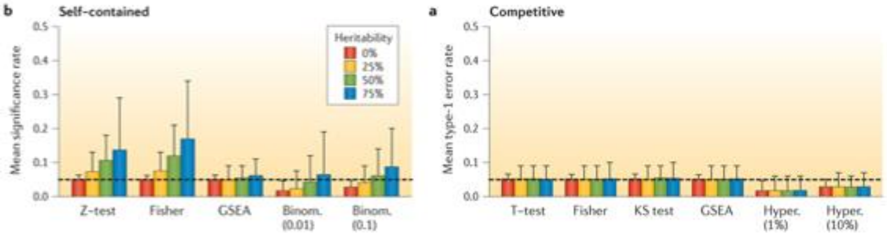
What are the functional categories of SNPs?	Functional categories of SNPs Protein coding - May alter protein structure (truncated proteins) Splicing regulation - Disrupt splicing regultation Transcriptional regulation - Disrupt gene regulation (TF binding sites, CpG islands, microRNAs) Post-translation modification - Interfer with proper posttranslation modification of proteins
If there are many GWAS hits, how can you choose SNPs for posterior studies?	How to pinpoint causal genes based on GWAS risk loci Are there functional variants in GWAS risk loci? Are there SNPs with high CADD scores or low regulomeDB scores? Are there regulatory variants or eQTLs in GWAS risk loci? Example CADD scores higher than 10 - Select important SNPs RegulomeDB score 1f HiC interaction in the Brain 
What are CADD scores?	CADD scores - Combined Annotation Dependent Depletion score Tool for scoring deleteriousness of SNPs + insertions and deletions in the human genome Higher than 10 - Predicted to be 10% most deleterious substitutions Higher than 20 - Indicates the 1% most deleterious
What are eQTLs?	Expression QTLs (eQTLs) - Increase level of expression of other genes

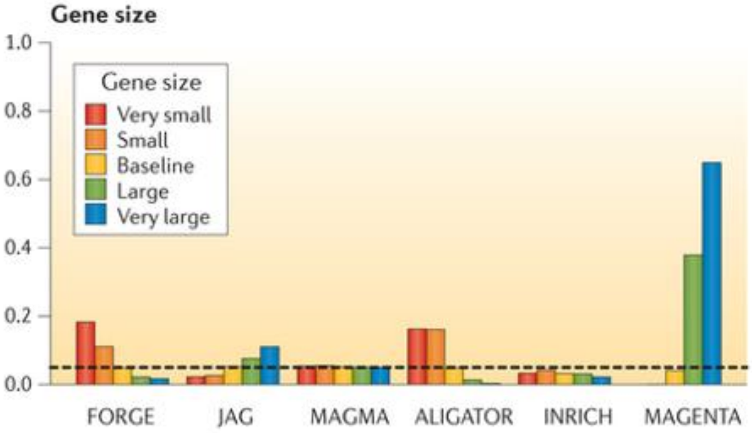
	The same regulatory regions and variant could be an eQTL for gene 2 in (a) tissue 1 and for gene 1 in (b) tissue 2, suggesting limited interrogation of tissues would be misleading for the biological signal underlying disease.  Can be present outside/farther away from the gene of interest
Why studying chromatin interaction may be useful for GWAS?	Chromatin interaction  A 3D Structure of the genome Chromatin Interaction Profile Gene C Gene B Gene A Gene D Chromatin Interaction Network Gene C Gene B Gene A Gene D DNA sequences may be distant in the linear sequence but close in the tridimensional structure
What can you study after you find a hit on GWAS?	SNP annotation implicates genes - Guide for posterior studies Explore gene functions Explore pathways of implicated genes Explore in which tissues the genes are expressed Explore which cell types are indicated
What is FUMA?	FUMA (Functional Mapping and Annotation) - Tool for GWAS annotation Q-Q plot - Red line equals no association (calculated from your alpha level)  If every SNP is associated - Suggests population stratification eQTL - Independent genes (linkage equilibrium)

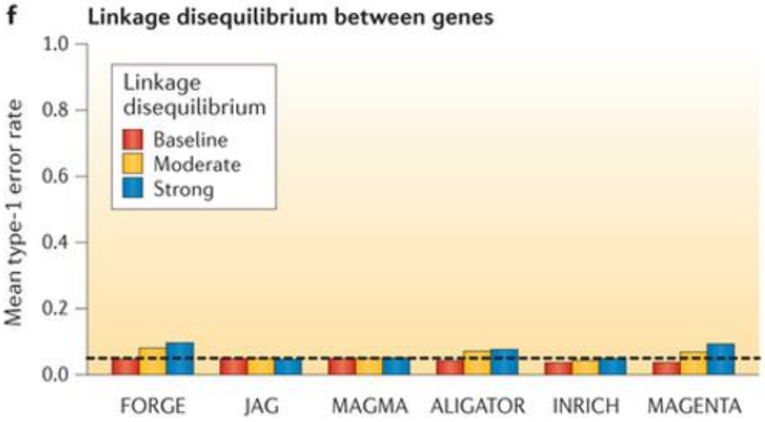
6a. Gene-set analysis

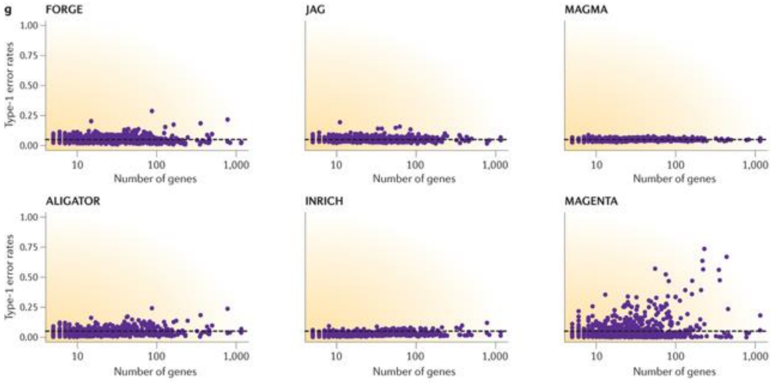
What are the options to test for functional clustering of SNPs?	Testing for functional clustering of SNP associations  Single SNP analysis Gene-based analysis Gene-set analysis Gene-property analysis Single SNP analysis Gene-based analysis - Gene as an unit of analysis Gene-set analysis (binary) - Biological pathway Gene-property analysis (continuous) - Expression levels or probability of being a member of a gene-set
What is the problem with testing for the joint association of all SNPs? How would you solve that?	Gene based analysis • Tests for joint association effects of all SNPs in a gene, taking into account LD (correlation between SNPs -> results in false positives ->multicollinearity must be avoided) (**Exam question)** No single SNP needs to reach GWAS significance -> the gene-based may still be significant
What does a single dot in a gene Manhattan plot might represent?	SNP Manhattan Plot Every dot is a single SNP  Gene Manhattan Plot Every dot is a single gene  Single dot - Implies that the individual SNPs have a really small effect, but their joint effect has a large effect

What are the pros and cons of gene-based analysis?	Gene-based analysis Pros: Reduce multiple testing, accounts for heterogeneity in gene (small effects of single SNPs), immediate gene-level interpretation (easier to come up with follow-up hypothesis) Cons: Disregards regulatory/non-genic information, still a lot of tests
What are the pros and cons of using genes as a unit of analysis?	Using gene as an unit of analysis Pros Reduce multiple testing Increases statistical power Deals with genic heterogeneity Provides immediate biological insights Cons Selecting reliable sets of genes is difficult • Different levels of information • Different quality of information
On what basis can you choose gene sets?	Choosing gene-sets can be based on:  Transcriptional regulatory network Virus-host network Metabolic network Protein-protein interaction Disease network Protein-protein interaction networks Yeast-two hybrid or immunoprecipitations - Real differentiated cells, extract real interactions! Co-expression - Protein interactions expressed at the same in the same tissue/organelle Transcription regulatory networks Biological pathway
What is the problem with public databases of gene sets?	Choosing gene sets KEGG (provides cascade of events), Gene Ontology, Ingenuity (paid), Biocarta (functional relationships between proteins involved), String Database, Human Protein Interaction Database; or you may create a manually curated by experts lists Public databases - Biased (not all genes are included), disease genes tend to be investigated more often, genes that are more investigated will have more interactions), not always reliable (interactions are often predicted, not validated; if experimentally tested, it is unknown how reliable it is)
What was the problem when	Comparing public databases and manually curated SPL list - Synapses

the SPL list (manually curated) was compared to public databases?	438/1043 of SPL genes have no KEGG pathway 388/1043 of SPL genes have no gene ontology 655 are not synaptic according to the database
On which aspects do tools for genes-set analysis differ?	Tools for gene-set analysis Differ in: • Self-contained or competitive tests • Different statistical algorithms test different alternative hypothesis • Different sensitivity for LD, number of SNPs, number of genes, background heritability
What is the difference between self-contained and competitive tests?	Self-contained - H0: Genes in gene-set are not associated with traits; problematic because polygenic traits will always be associated with a large part of the genome Competitive tests - H0: Genes in gene-set are not more strongly associated than genes **NOT** in the gene-set
What happens with type one erros when you use a competitive test? And when you use a self-contained test?	The more polygenic your trait is, the more likely you are to get a false positive when using self-contained tests (higher background heritability -> effect of the entire genome on the trait)  This doesn't occur with competitive tests Never use self-contained tests
What is the difference between minimal P-value and combined p-value?	Different statistical algorithms test different alternative hypothesis Minimal P-value - At least one SNP needs to be significant Combined P-value - It doesn't matter if SNPs are signicant
What happens with type 1 error when you	Effect of gene size for signifance value

<table>
<tr>
<td><b>increase gene size?</b></td>
<td>


MAGENTA should not be used
</td>
</tr>
<tr>
<td><b>What happens with type 1 error when you increase LD?</b></td>
<td>
LD for significance value


FORGE/ALLIGATOR should not be used
</td>
</tr>
<tr>
<td><b>What happens with type 1 error when you increase number of genes?</b></td>
<td>
Number of genes for significance value


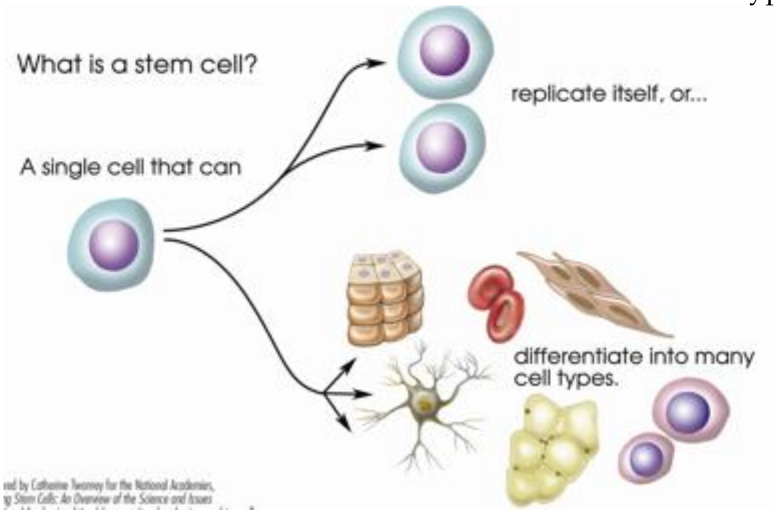
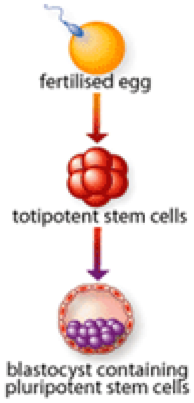
</td>
</tr>
</table>

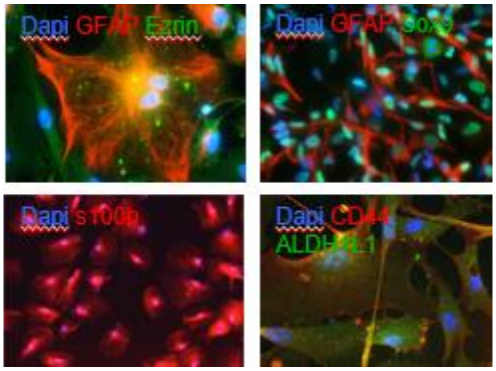
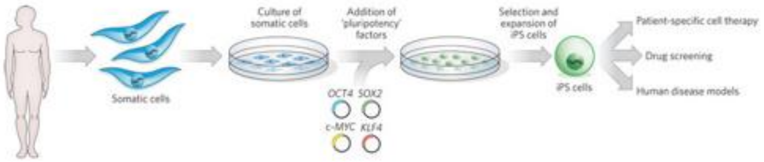
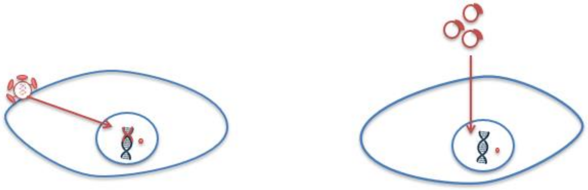
6b. MAGMA

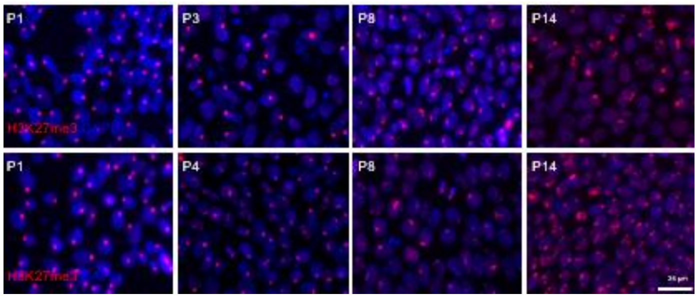
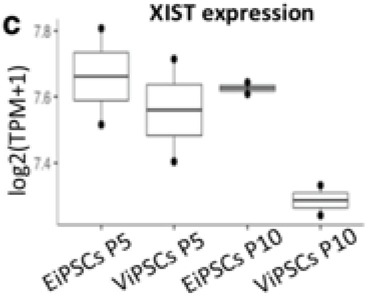
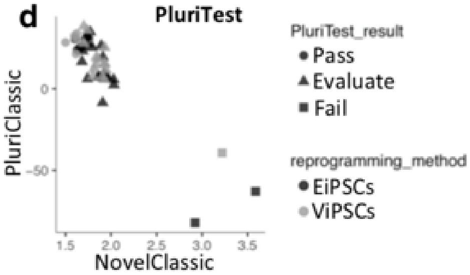
What is MAGMA?	MAGMA - Software tool for gene and gene-set analysis Input - Genotype and phenotype data; summary statistics for GWAS + gene definition and gene sets
What are the three main steps of using MAGMA?	Three main steps Annotation - Map SNPs onto genes Gene analysis - Compute association of genes with phenotype Gene-set analysis - Gene association in gene sets Extension modeling options: Advanced analysis (conditional and joint analysis)
What is the annotation stage based on?	Annotation - Based on physical location of SNPs/genes in the DNA Window around gene - Variable, mainly used for upstream/transcription start site of gene Genes can be non-protein-coding as well
What are the three models for SNP association? Which one is the most recommended?	Gene analysis Joint association of all SNPs in a gene with the phenotype Different analysis models have greater sensitivity to different genetic architectures MAGMA has different models with different sensitivies • Principal component linear regression - Requires raw genotype data • SNP-wise mean - Mean SNP association • SNP-wise Top - Strongest SNP association • SNP-wise Multi - Combines SNP-wise mean and top
Why does the gene-set analysis use a one-sided statistical test?	Gene-set analysis - Essentially a t-test Unit of analysis: Genes One-sided test - High negative association scores are not interesting Self-contained analysis (single sample t-test) - H0 - $\mu s = 0$ Competitive analysis (dual sample t-test) - H0 - $\mu s = \mu 0$
What are some stastical challenges you might face when using MAGMA?	Statistical challenges Outlier effects - Small subset of strongly associated genes driving the association Independent observations - Linkage disequilibrium Model covarience Confounding - Apparent effect may be induced by an overlap with truly associated set

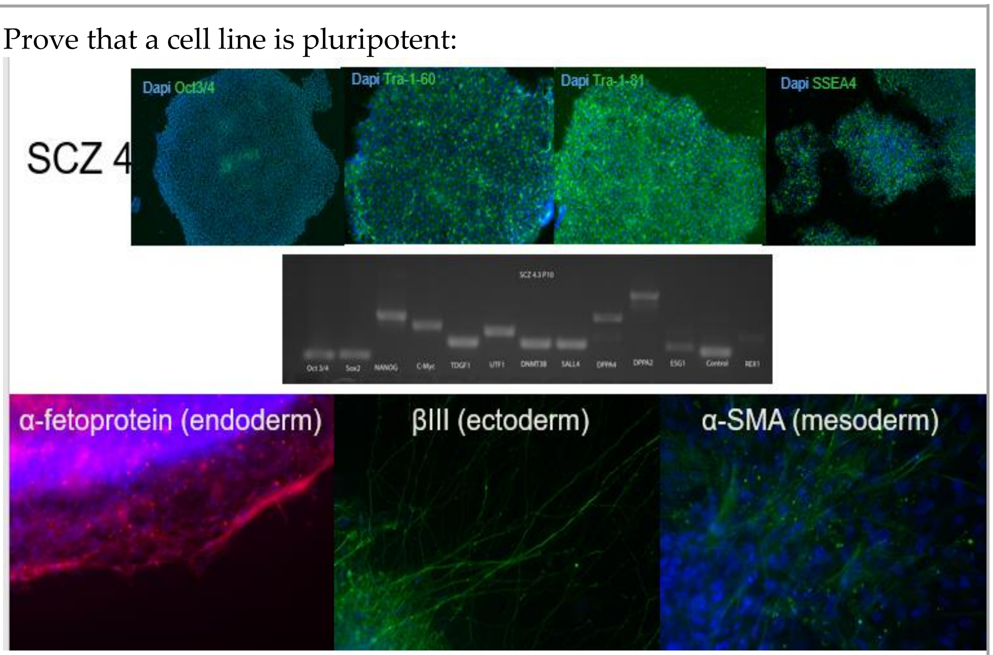
	Confounding factors can only be ruled out by experimental design
Which statistical test is used in the competitive analysis of MAGMA?	Competitive analysis Linear regression model - Two sample t-test (much more flexible) Example: Brain-specific gene expression Lower p = higher Z
How can you detect an effect when a gene has a correlated expression across tissues?	Correlated expression across tissues Solved by multiple regression - If the effect is true, the apparent effect of correlated variables will disappear • Requires true effect variable to be present • Very large number of variables (reduces power) • This does not rule out confounding effects
Why might you want to divide your gene set into different subsets?	Interaction analysis Multiple linear regression - Analyse the overlapping of pathways Example: Division of miRNA-145 gene set into four subsets: Three non-significant, one very significant -> This effect would be undetectable by regular gene-set analysis

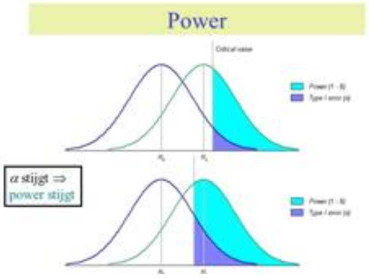
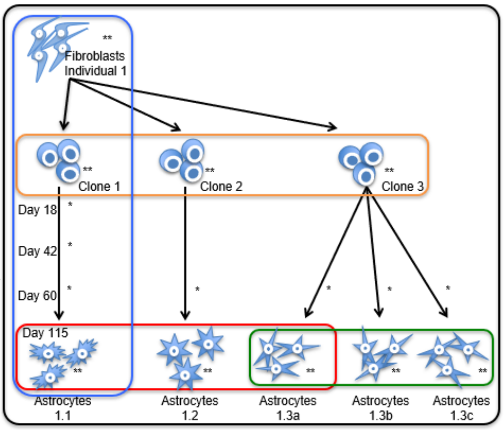
7a. Modeling genetically complex disorders using induced pluripotent stem cells

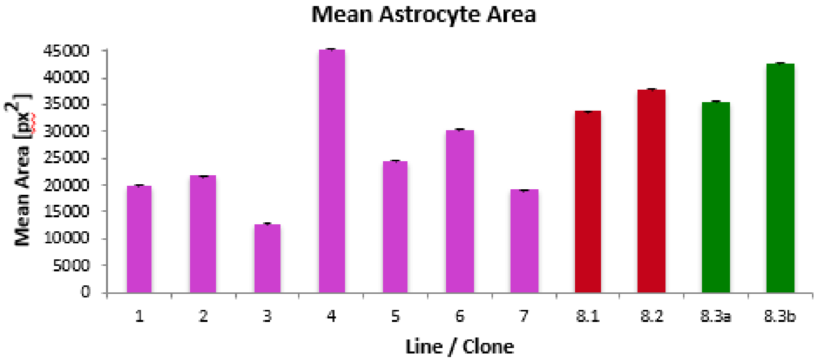
What is the definition of a stem cell?	Stem cells - Can divide or differentiate into different cell types  What is a stem cell? A single cell that can replicate itself, or... differentiate into many cell types. ed by Catherine Twomey for the National Academies, g Stem Cells: An Overview of the Science and Issues
What do trophoblasts and blastocyst form?	Trophoblast - Forms placenta Blastocyst - Forms the person  fertilised egg totipotent stem cells blastocyst containing pluripotent stem cells
What are the types of stem cell potency?	Types of potency Totipotent - Can make an entire human being (extraembryonic/placental cells + all tissues) Before 16-morula stage Pluripotent - Can make all tissues (but cannot form an individual/form placental tissue) Multipotent - Can make more than one cell type, but not all (haematopoietic stem cells)

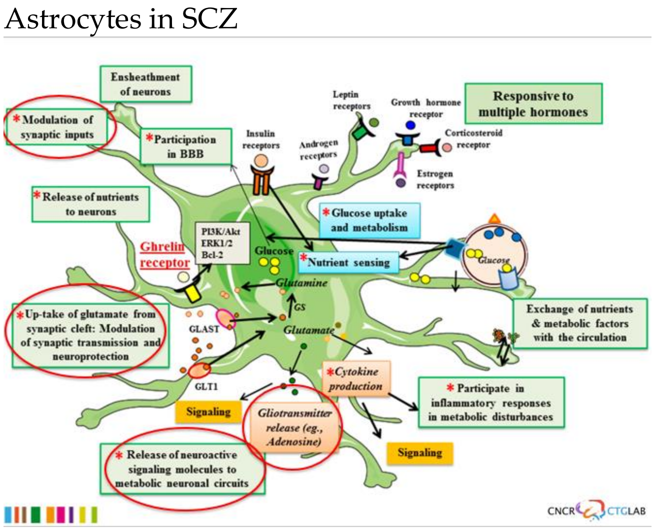
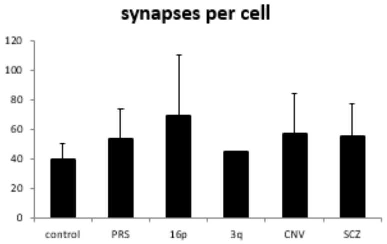
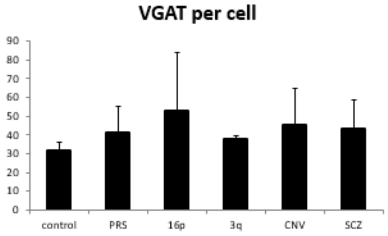
How would you know if your stem cells have differentiated into the tissue you wanted?	Model tissue in vitro Embryonic stem cells from the inner cell mass (pluripotent) Differentiation can be mimicked in vitro - Inserting the same factors in vivo How to know you have made the right tissue - Antibodies for cell markers  Dapi GFAP Ezrin Dapi GFAP Sox9 Dapi s100b Dapi CD44 ALDH1L1
What are the four Yamanaka factors? How were they discovered?	Induced pluripotent stem cells Yamanaka factors - Turn differentiated cells into pluripotent cells (OSKM)  Somatic cells Culture of somatic cells Addition of 'pluripotency' factors OCT4 SOX2 c-MYC KLF4 Selection and expansion of iPS cells iPS cells Patient-specific cell therapy Drug screening Human disease models Process of discovery: 100 factors genes expressed in PSC -> narrowed to 24 (transcription factors since they need to modify expression of other genes); leave factors off one at a time
How do Yamanaka factors work?	Yamanaka factors: Somatic genes are silenced, pluripotent genes will be switched on Chromatin remodeling Epigenetic remodeling
What are two options to deliver stem cell gene factors to a cell?	Delivery of factors  Integrating - Lentivirs Not integrating - Episomal (Plasmids + eletroporation (still reaches nucleus)) Safer - Does not disrupt native DNA

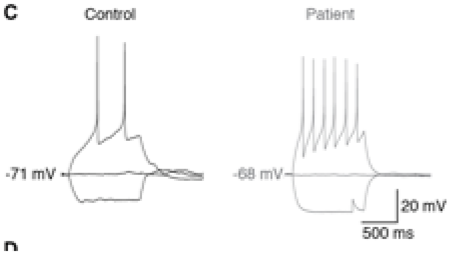
What causes X-inactivation?	Females are mosaics - Different inactivation of X chromosome XIST - RNA transcript activates PCR2 (adds methylation to DNA) H3K27 - Repressive chromatin marker of inactivation
Which method of stem cells delivery factors would be safer for ex-vivo therapy for women? Why?	 P1 P3 P8 P14 H3K27me3 P1 P4 P8 P14 H3K27me3 25 μm Inducing pluripotency removes X chromosome inactivation using Lentivirus (first row) X chromosome inactivation is maintained when using plasmid (second row)
What happens to XIST expression during lentivirus treatment?	XIST expression remains normal in episomal reprogramming, not in lentivirus reprogramming  X-linked disorders could be treated with X chromosome re-activation of lentivirus
Describe a pluripotency test. What is the X and Y axis? What you expect if a cell was converted successfully into PSC?	Pluripotency test Ratio of gene expression of PSC (Y axis) compared to differentiated cells (X axis)  There is no difference between lentivirus and episomal

How would you prove that a cell line is pluripotent?	Prove that a cell line is pluripotent:  SCZ 4 Dapi Oct3/4 Dapi Tra-1-60 Dapi Tra-1-81 Dapi SSEA4 SCZ 4.3 P10 Oct 3/4 Sox2 NANOG C-Myc TDGF1 UTF1 DNMT3B SALL4 DPPA4 DPPA2 ESG1 Control REX1 α-fetoprotein (endoderm) βIII (ectoderm) α-SMA (mesoderm) Immunostaining - Check for expression of PSC factor Immunoblot - Expression of these protein Three markers for the the three germ layers
Why should we use iPSC?	Benefits of iPSC Ethical - No embryos needed IPSCs are identical to donor - Safer transplantation, good for genetic disorders that are very complex (e.g. schizophrenia with 108 associated SNPs)
Define the main differences between CNVs and SNPs associated with schizophrenia.	Genetics of SCZ Heritability: 81% Lifetime risk: 1% CNVs (copy number variants): Rare, large effect, few patients, low penetrance SNPs: Common, very small effects (combined large effects), majority of patients
How has been SCZ been studied in the past?	Studying SCZ Post-Morten studies Animal studies: Pharmacology, transgenic mice (pick CNV or CRISPR 108 -> incredibly laborious) Picking CNV - DISC1, miR-137, NRXN1 -> Create synaptic deficit (they are all synpase proteins); creates a biased sample

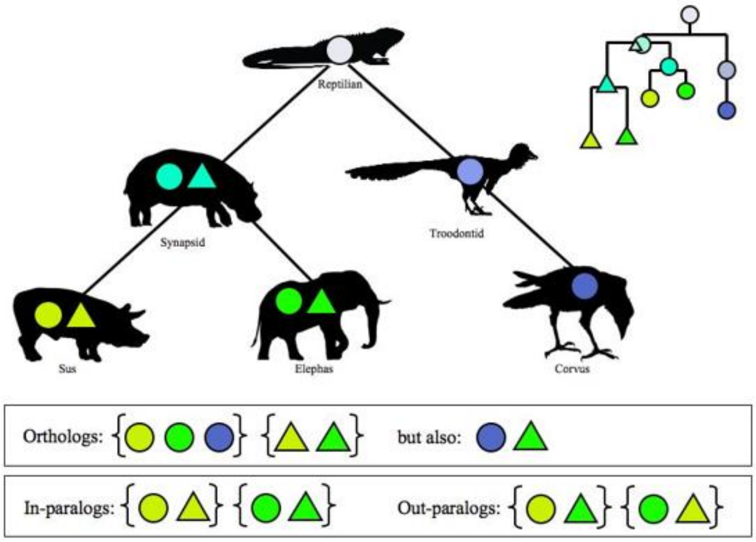
How could you improve the power of your iPSC study?	IPSC for SCZ Heterogeneous population -> decreases power Pick extreme SNPs from polygenic risk score Pull graphs apart - Controls Low-risk score and patients with high-risk score  Problems: Expensive, laborious, various protocols (variability between studies) Accumulation of mutation Defects in reprogramming Inaccuracy of protocol
Describe how you could validate your protocol to convert differentiated cells into iPSC.	Determine variability and source  Fibroblasts Individual 1 Clone 1 Clone 2 Clone 3 Day 18 Day 42 Day 60 Day 115 Astrocytes 1.1 Astrocytes 1.2 Astrocytes 1.3a Astrocytes 1.3b Astrocytes 1.3c Three individuals - 3 cell lines from each; analyse variability of cells before and after differentiation from the same individual Whole-proteome from single cells - Correlation of each cell to one another Differentiated cells are more similar to one another than to other people - Means that protocol is reliable; undifferentiated cells are not Older colonies become more alike - Hard to interpret Technical replication - Same sample multiple times

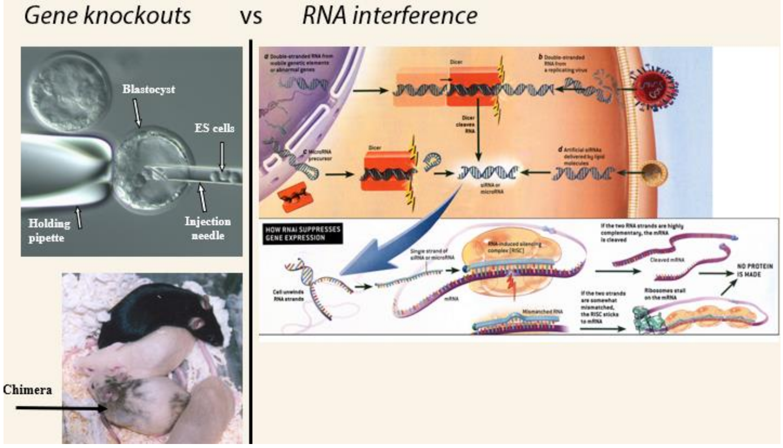
	Cellomics Astrocyte area - Variation is larger between individuals (pink) compared to clones from the same individual (red) or technical replicates (green) 
How many clones should we use per individual to optimize resources?	Conclusions 1 clone per individual and multiple people Sequence entire genome from each cell - Check for unexpected mutations (PRS, CNVs)
Why should we not use lentivirus to study rare diseases?	Episomal reprogramming is better for rare disease - Lentivirus requires integration, may disrupt DNA in unexpected ways
What are some neuronal dysfunctions of SCZ patients?	Neurons in SCZ Dysregulation of excitation-inhibition network Synaptic pruning -> brain volume decreases SMAD inhibition - Block pathways of differentiated cells Forms rosettes -> neural tube-like, makes both neurons and glial cells HSHH and Valproic acid - GABAergic cells Neuritis length is reduced in SCZ patients Dendrite length stays the same Axonal length is reduced in SCZ patients

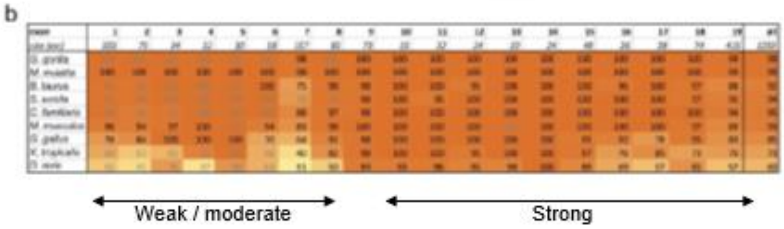
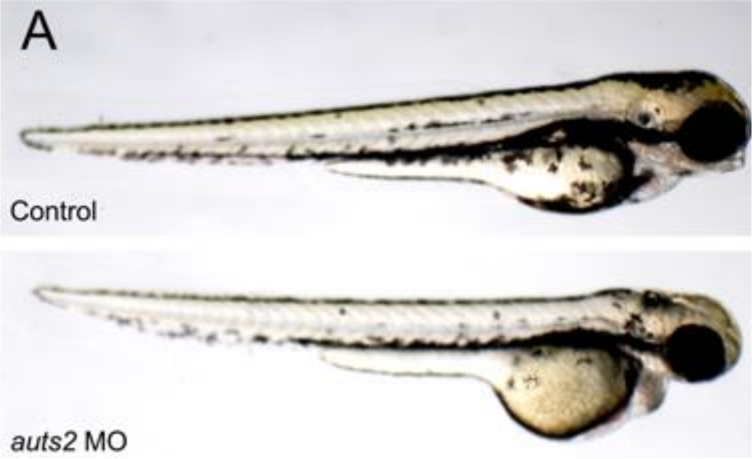
What happens with astrocytes in SCZ?	 Astrocytes in SCZ Ensheathment of neurons *Modulation of synaptic inputs *Participation in BBB Insulin receptors Androgen receptors Leptin receptors Growth hormone receptor Responsive to multiple hormones Corticosteroid receptor Estrogen receptors *Release of nutrients to neurons *Glucose uptake and metabolism Ghrelin receptor PI3K/Akt ERK1/2 Bcl-2 Glucose *Nutrient sensing Glucose Glutamine GS Glutamate *Up-take of glutamate from synaptic cleft: Modulation of synaptic transmission and neuroprotection GLAST GLT1 Exchange of nutrients & metabolic factors with the circulation *Cytokine production *Participate in inflammatory responses in metabolic disturbances Signaling Gliotransmitter release (eg., Adenosine) Signaling *Release of neuroactive signaling molecules to metabolic neuronal circuits CNCR CTGLAB 31 astrocyte specific gene sets in SCZ Coexpressed genes (neurons and astrocytes) - CACNA1C Less density of astrocytes and GFAP in SCZ patients
What happens when SCZ astrocytes were co-cultured with neurons?	Results SCZ Nestin/CD44 - Expressed in stressed cells No difference in astrocytes alone When co-culturing SCZ astrocytes with neurons - VGAT (gabaergic terminals) is lower for SCZ patients; ratio of GABAergic and glutamatergic neurons is altered   Sandwich culture - Cells not in physical contact, but astrocytes can secrete factors
What happens with oligodendrocytes in SCZ?	Oligodendrocytes in SCZ White matter abnormalities in patients Less MBP in chimeras (human iPSC in mice)
Why should we use organoids to study SCZ?	Organoids - three dimensional -> allows myelination to occur Reelin - Produced in the outer layer of the brain Ctip2 - Layer 5/6 MBP - Myelinating Oligodendrocytes

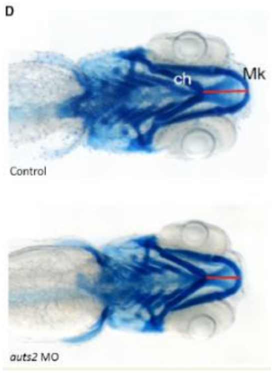
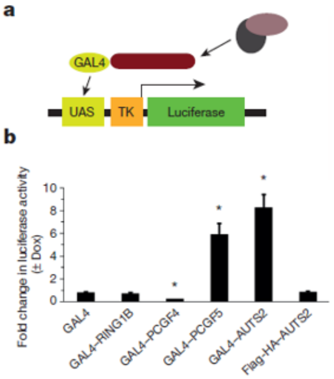
What happens to the half-width of action potentials in SCZ patients? What does that mean?	 Half-width of action potential - How long does a neuron take to go back to resting state Measure from depolarization and hyperpolarization peaks Half-width is longer for SCZ patients - It takes longer to recover

7b. Gene function analysis in Neuroscience for Mendelian Disorders

How could we use yeast to study a human disease, like Wilson's?	Wilson's disease - ATP7B Monogenic disease with a known gene - The observed allelic variant of the patient will or will not explain the disease ATP7B structure - If the mutation is on the ligand site, you may argue that the mutation is causal Wilson's disease in yeast - ccc2 gene from yeast is similar to ATPB7 Copper transport -> important for iron metabolism Reinserting ATP7B in yeast restores function in knockout yeast Tests for different patient mutation Cheap, ethical and quick (three days)
What are homologs, orthologs and paralogs?	Homologs - Genes shared from a common ancestor; divided in ortholog and paralogs  Reptilian Synapsid Troodontid Sus Elephas Corvus Orthologs: but also: In-paralogs: Out-paralogs: Orthologs - Similar structure and different function Paralogs- Genes related by duplication in a genome - have different function
What can you do with a knockout?	Knockout functions Investigate gene function Test causality of candidate disease mutation Develop and test therapies in disease model

What are two common methods to inactive genes?	Methods to inactive genes  Gene knockouts vs RNA interference Blastocyst ES cells Holding pipette Injection needle Chimera a Double-stranded RNA from mobile genetic elements or abnormal genes Dicer b Double-stranded RNA from a replicating virus Dicer cleaves RNA c MicroRNA precursor Dicer d Artificial siRNAs delivered by lipid molecules siRNA or microRNA HOW RNAi SUPPRESSES GENE EXPRESSION Cell unwinds RNA strands Single strand of siRNA or microRNA RNA-induced silencing complex (RISC) mRNA Mismatched RNA If the two RNA strands are highly complementary, the mRNA is cleaved Cleaved mRNA NO PROTEIN IS MADE If the two strands are somewhat mismatched, the RISC sticks to mRNA Ribosomes stall on the mRNA Gene knockout RNA interference - Translation will be blocked or mRNA will be degraded very quickly Very efficient in zebrafish - Double stranded modified RNA to be more stable (morpholinos)
What is the phenotype of someone with AUTS2 deletion?	AUTS2 in a syndromes form of autism AUTS2 deletion - Autistic behavior, IQ decline, short stature, microencephaly, feeding difficulties,generalized hypotonia (muscle weakness) *Heterozygous loss of gene is sufficient to cause the phenotype*
What correlates with the severity of phenotype of AUTS2-deletion patients?	All mutations were intronic - Variant of unknown significance AUTS2 deletions are observed in patients with these symptoms Heterogeneity of phenotypes - Sum of presence of scores = AUTS2 score Phenotypes are more severe in patients with more deletions in the C-terminus - Unexpected because N-terminus truncation should be more severe

What is the explanation for the fact the deletions in the C-terminus are more severe?	Evolutionary conservation of AUTS2 Amino acid identity - C-terminus is more conserved evolutionarily than N-terminus  Zebrafish may have a functional ortholog to the human protein
What method could you use to identify an alternative transcriptional splice site?	Alternative transcriptional splice site Present in the middle of exon 9 - There is two gene products inside the same gene RACE - Rapid Identification of cDNA Ends
How could we use zebrafish to study a human deletion like AUTS2?	Zebrafish  Morpholinos - Knockdown of AUTS2 causes microencephaly (establishes causality between gene and phenotype) Phenotype is rescued by the insertion of human AUTS2 cDNA into zebrafish cDNA does not have introns - It is easier to produce in silico Microencephaly is also rescued by the C-terminal short isoform (cDNA)
How can facial dysmorphism be quantified in zebrafish?	Facial dysmorphism

	 Blue - Stain for cartilage Larger distance between Ch and Mk (ceratohyal and Meckel's)
What happens to the number of neurons in AUTS2 zebrafish?	Reduced proliferation of neuronal progenitors in auts2 morphants Reduction of differentiated neurons Phosphohistone-h3 antibody - Epigenetic markers (proliferating neurons) There are less proliferating precursor cells
What is the main advantage of studying mendelian disorder, as opposed to complex traits?	Mendelian disorders - Sample size can be much smaller Causal role of mutation - Large effect size
What is the mechanistic function of AUTS2? Why is it unusual?	AUTS2 gene product function - Still unknown Possible transcription factors Enhancer regions in the middle of AUTS2 gene AUTS2 is part of Polycomb complex - Positively regulates transcription of targeted genes (while most other components of this complex reduce transcription)  AUTS2 binds to promoter regions Mouse models confirm the finding