Introduction to Genetics, Genomics and Genome-wide Association Studies

jodie.lord@kcl.ac.uk

@JodieLord5

Learning Objectives

- Understand the basics of genetics of “complex” disorders.
- Understand the importance of the human genome project in:
  - enabling a better understanding of the core structure of the human genome.
  - enabling the introduction of genome-wide approaches to understanding links between genes and disease / genes and behavior.
  - paving the way for the rise of other “omics” fields within biology.
- Develop a working knowledge of core genome-wide methodology and understand the importance of genome-wide association studies.
- Understand the importance of genomics as a key strata within the wider “omics” field.

Learning Objectives

- Understand the basics of genetics of “complex” disorders.

- Understand the importance of the human genome project in:
  - enabling a better understanding of the core structure of the human genome.
  - enabling the introduction of genome-wide approaches to understanding links between genes and disease / genes and behavior.
  - paving the way for the rise of other “omics” fields within biology.

- Develop a working knowledge of core genome-wide methodology and understand the importance of genome-wide association studies.

- Understand the importance of genomics as a key strata within the wider “omics” field.

BACKGROUND: ”Classic” genetic disorders

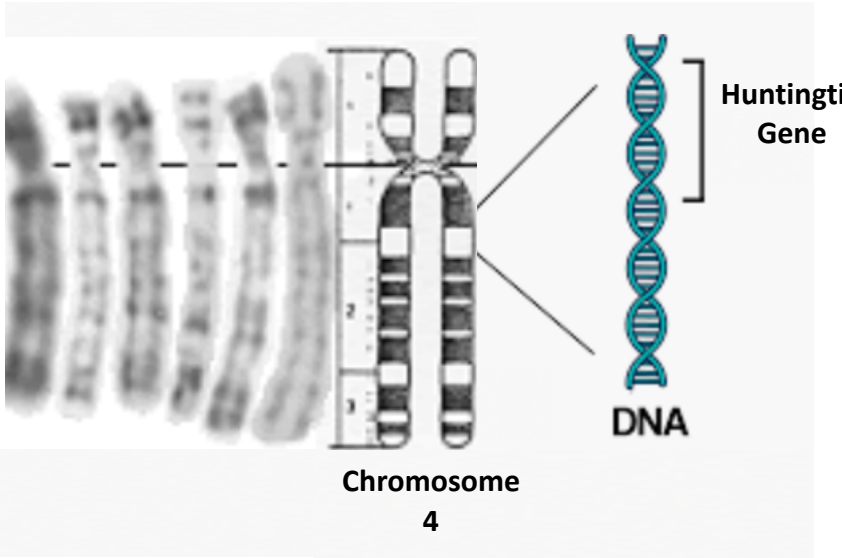
- “Mendelian” → single gene = causative and wholly predictive of future disease onset.

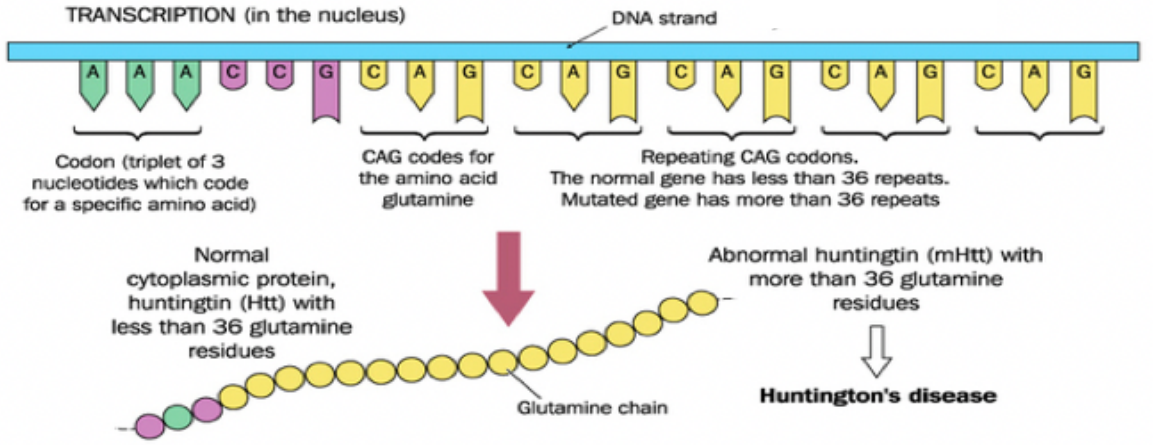
- Typically a large genetic mutation identified within a coding region of a gene → impacts amino acid sequences processed → impacts protein structures made up from the amino acids → large consequences for biological processes within the body.

- Example = Huntington’s Disease

BACKGROUND: ”Classic” genetic disorders

- Example = Huntington’s Disease:





https://ghr.nlm.nih.gov/condition/huntington-disease

- Huntingtin (HTT) gene provides genetic instructions for making huntingtin protein.
- Mutation in this gene = extra CAG sequences repeated in the gene.
- CAG = the codon for amino acid glutamine – so an excess of glutamine amino acids= made by the gene
- These then chain together when making forming the huningtin protein – impacting the structure of this protein.
- Protein structure = VERY important as different structures act as the “keys” to open specific “locks” in biological processes..

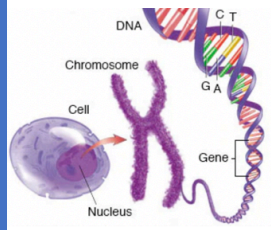
BACKGROUND: "Classic" genetic disorders

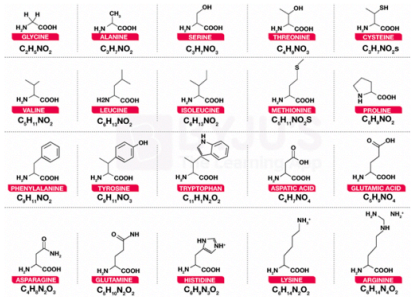
REMEMBER THIS SLIDE?…

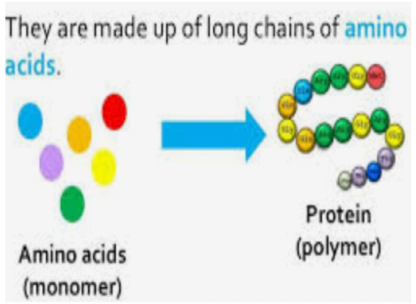
What is Omics?

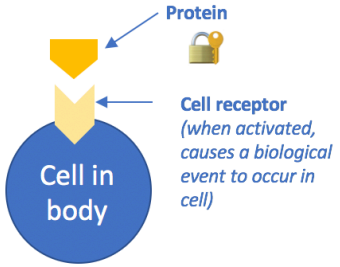
- A (very) quick biology refresher….

GENES **MAKE** AMINO ACIDS **FORM** PROTEINS **ENABLE** BIOLOGICAL PROCESSES







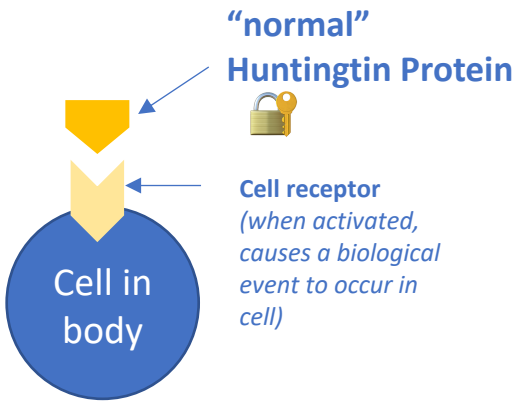


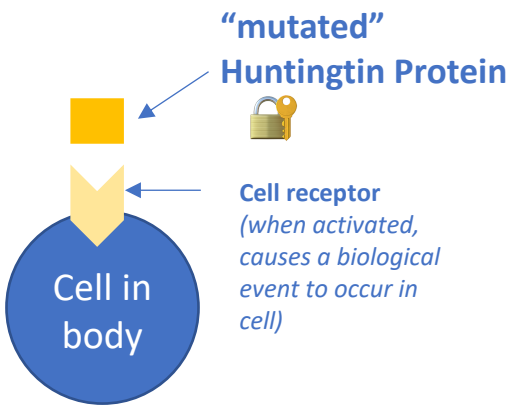
- Protein structure = VERY important as different structures act as the "keys" to open specific "locks" in biological processes..

BACKGROUND: ”Classic” genetic disorders

- Example = Huntington’s Disease:

- Huntingtin (HTT) gene provides genetic instructions for making huntingtin protein.
- Mutation in this gene = extra CAG sequences repeated in the gene.
- CAG = the codon for amino acid glutamine – so an excess of glutamine amino acids= made by the gene
- These then chain together when making forming the huningtin protein – impacting the structure of this protein.
- Protein structure = VERY important as different structures act as the “keys” to open specific “locks” in biological processes..

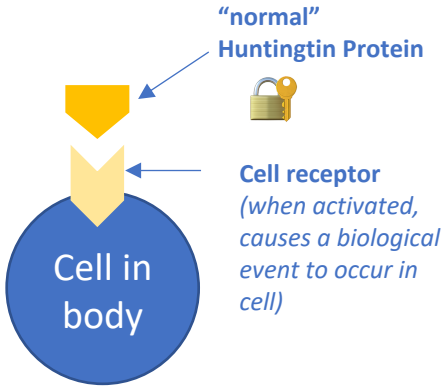


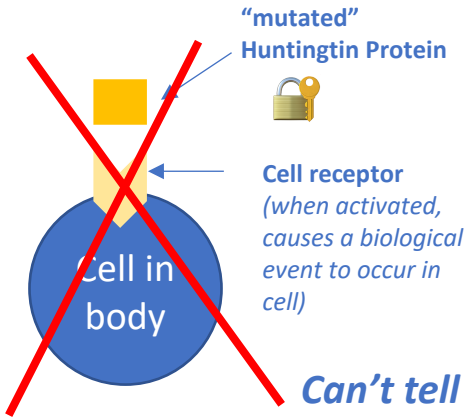


BACKGROUND: ”Classic” genetic disorders

- Example = Huntington’s Disease:

- Huntingtin (HTT) gene provides genetic instructions for making huntingtin protein.
- Mutation in this gene = extra CAG sequences repeated in the gene.
- CAG = the codon for amino acid glutamine – so an excess of glutamine amino acids= made by the gene
- These then chain together when making forming the huningtin protein – impacting the structure of this protein.
- Protein structure = VERY important as different structures act as the “keys” to open specific “locks” in biological processes..

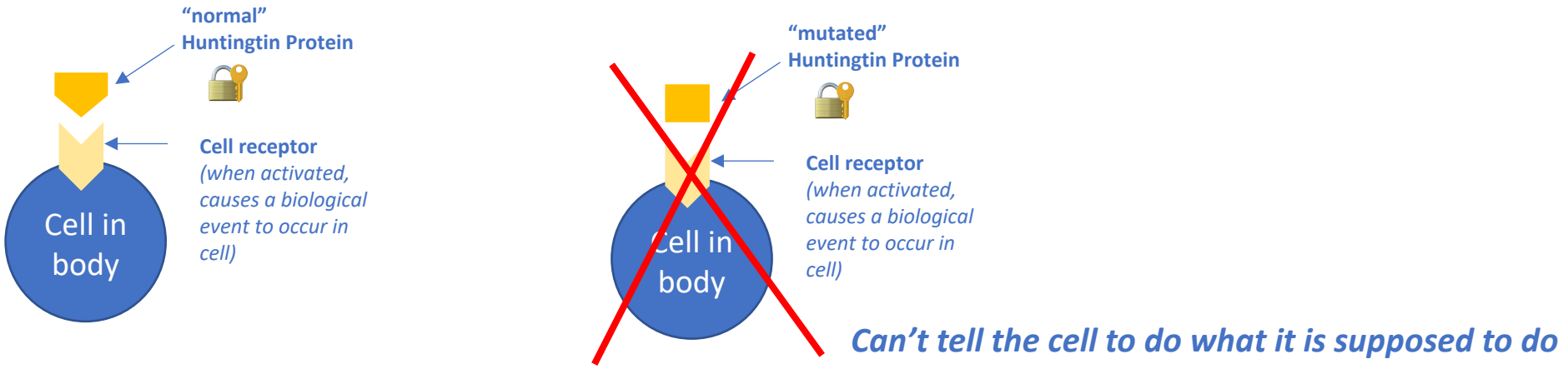




Can’t tell the cell to do what it is supposed to do

BACKGROUND: ”Classic” genetic disorders

- Example = Huntington’s Disease:



- Huntingtin protein = thought to provide important signaling instructions for nerve cells (like neurons).
- Mutated = abnormally long (due to extra glutamine amino acids in structure) → signaling messages not appropriately transmitted → incorrect / toxic accumulation within cells → abnormal functioning → eventual cell death → Huntington’s.

BACKGROUND: ”Classic” genetic disorders

- Example = Huntington’s Disease:

CAG mutations within the HTT gene = FULLY DETERMANISTIC → found with mutation – will eventually get Huntington’s…

“normal”

…hat it is supposed to do

…portant signaling instructions for nerve cells (like neurons).

…nally long (due to extra glutamine amino acids in structure) → signaling messages not …propriately transmitted → incorrect / toxic accumulation within cells → abnormal functioning → eventual cell death → Huntington’s.

BACKGROUND: ”Classic” genetic disorders

- Example = Huntington’s Disease:

These kind of “classic” genetic disorders are known as **MENDELIAN DISORDERS** – as they follow Mendel’s laws of inheritance patterns.

“normal”

…hat it is supposed to do

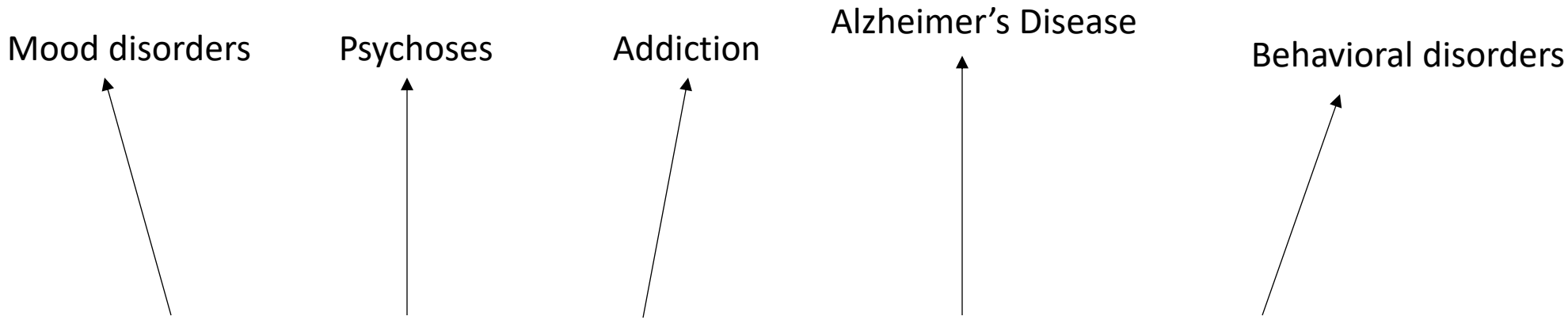
…portant signaling instructions for nerve cells (like neurons).

…nally long (due to extra glutamine amino acids in structure) → signaling messages not …propriately transmitted → incorrect / toxic accumulation within cells → abnormal functioning → eventual cell death → Huntington’s.

What is a “Complex” Disorder

- **Unlike** Mendelian disorders - no single gene or genetic mutation is wholly predictive of disorder onset.
- Risk is reflected by an interplay of various genetic and environmental factors.
- Each individual “risk factor” = very small contributory effect towards disorder.
- Risk factors exist on a normally distributed continuum – many “non-ill” individuals in wider population will have many of these risk factors also
- Risk factors = thought to be predominantly “additive”.
- When certain threshold of “liability” met → disorder observed.
- Disorders are common within population (comparably to classic genetic disorders)

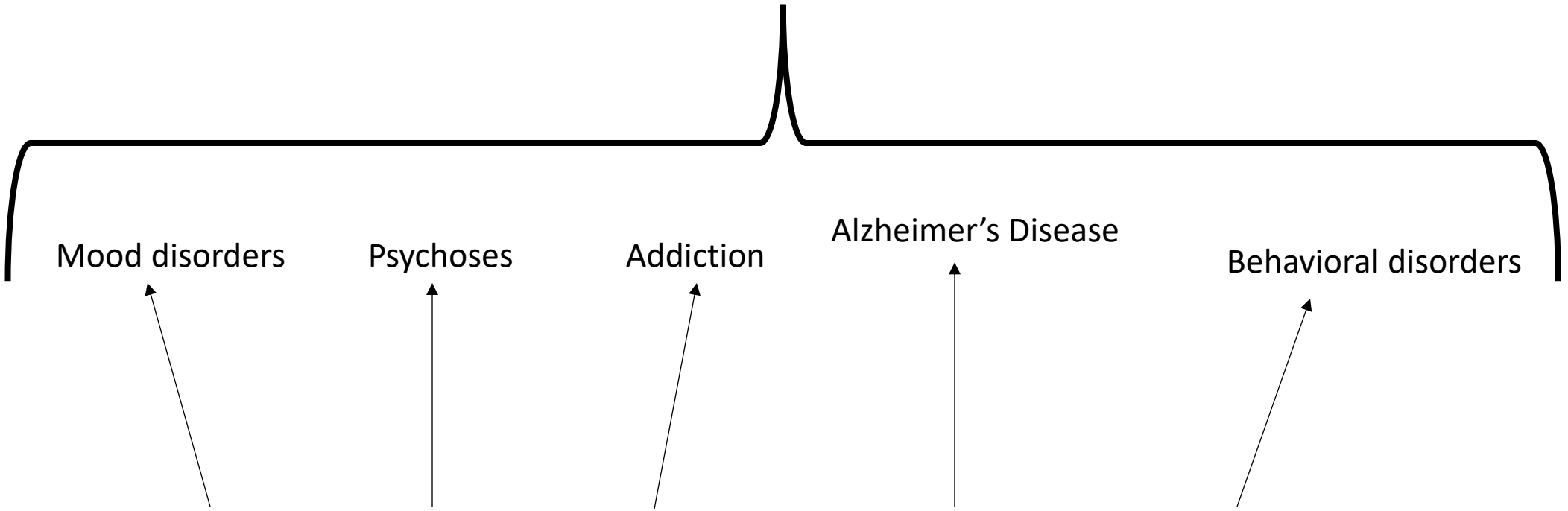
What is a “Complex” Disorder



- Disorders are common within population (comparably to classic genetic disorders)

What is a “Complex” Disorder

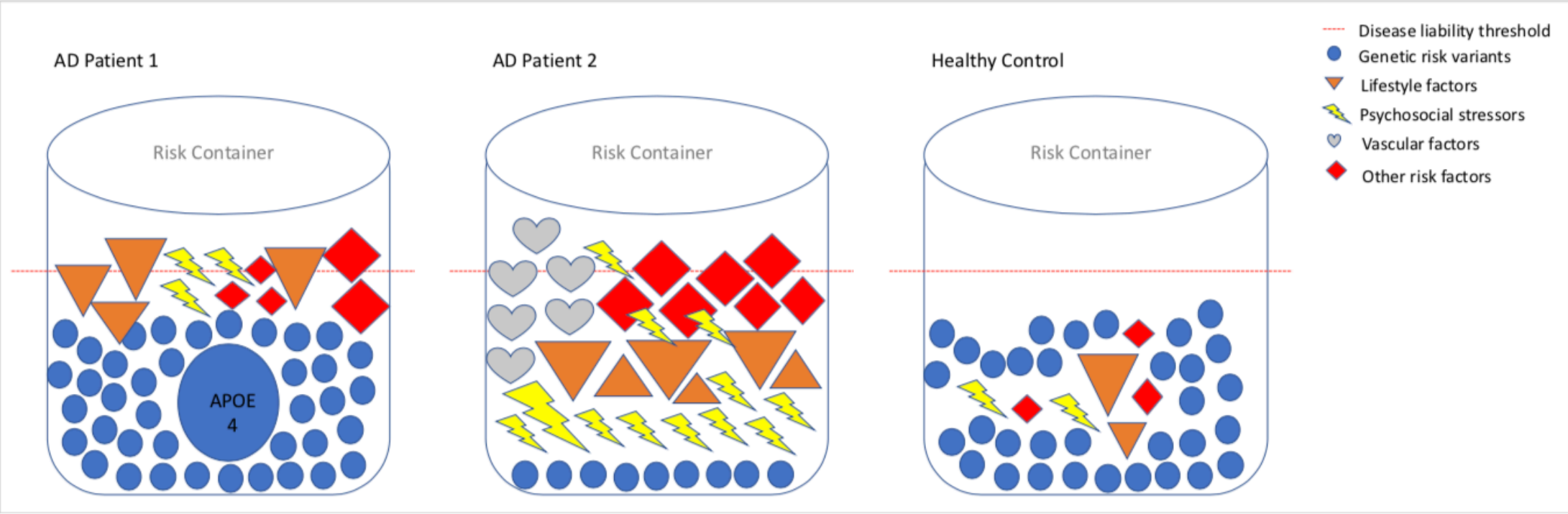
And pretty much every common behavioural / personality trait!



- Disorders are common within population (comparably to classic genetic disorders)

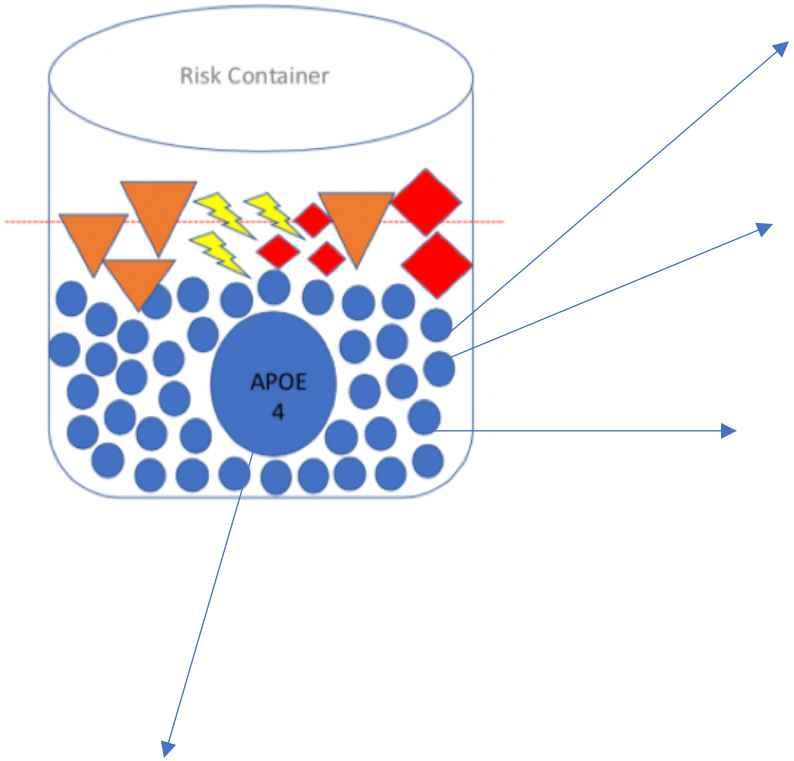
What is a “Complex” Disorder

- Liability threshold assumption.
- Example = hypothetical Alzheimer’s Disease risk:



What is a “Complex” Disorder

- **Unlike** Mendelian disorders - no single gene or genetic mutation is wholly predictive of disorder onset.

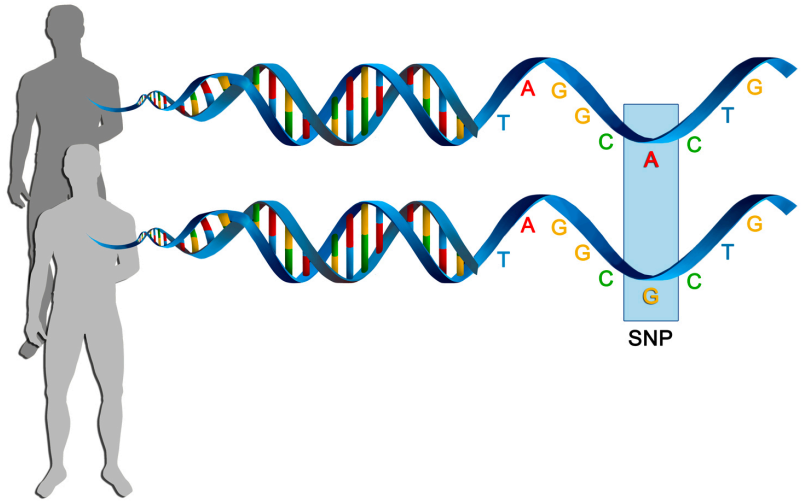


- Unlike classic disorders – mutations don’t tend to be 1) large or 2) within specific genes.
- Instead, genetic risk factors usually refer to variation within single basepairs in a particular location of the genome (e.g. an “A” becomes a “T”).
- This may or may not impact amino acid sequences – depending on whether the point of variation is within a coding or non-coding region of the genome.

- Some of these genetic risk factors have a bigger effect than others depending on their functional links (e.g. APOE4 = a particularly large genetic risk factor for Alzheimer’s). But none fully determine outcome.

What is a “Complex” Disorder

- **Unlike** Mendelian disorders - no single gene or genetic mutation is wholly predictive of disorder onset.



- Unlike classic disorders – mutations don’t tend to be 1) large or 2) within specific genes.
- Instead, **genetic risk factors usually refer to variation within single basepairs in a particular location of the genome (e.g. an “A” becomes a “T”).**

↓

- There single points of variation are referred to as: “**S**ingle **N**ucleotide **P**olymorphisms” **(SNPs)**

  - ✓ *Single because they are single points of variation.*
  - ✓ *Nucleotide because they refer to variation of one of the 4 nucleotides which make up our DNA – T,A,C,or G).*
  - ✓ *Polymorphism because there are **many** of these single points of variation scattered across our genome.*

What is a “Complex” Disorder

- **Unlike** Mendelian disorders - no single gene or genetic mut
  onset.

Because the genetic risk factors associated with complex disorders:
1. exist in both the clinical and non-clinical population, and
2. Have individually very small effects on clinical outcomes,
discovery of robust genetic association have been notoriously difficult
→ Classic genetic studies (single gene analyses) didn’t cut it.

of variation are referred to as:
Nucleotide Polymorphisms” **(SNPs)**

- ✓ *Single because they are single points of variation.*
- ✓ *Nucleotide because they refer to variation of one of the 4 nucleotides which make up our DNA – T,A,C,or G).*
- ✓ *Polymorphism because there are **many** of these single points of variation scattered across our genome.*

Learning Objectives

- Understand the basics of genetics of “complex” disorders.

- Understand the importance of the human genome project in:
  - enabling us to better understand the core structure of the human genome.
  - enabling the introduction of genome-wide approaches to understanding links between genes and disease / genes and behavior.
  - paving the way for the rise of other “omics” fields within biology.

- Develop a working knowledge of core genome-wide methodology and understand the importance of genome-wide association studies.

- Understand the importance of genomics as a key strata within the wider “omics” field.

The Human Genome Project

- 1990 – 2003

- Aims:
  - To fully sequence and map out the human genome.
  - To provide an accessible database of this information which can be used as a reference.
  - To better understand the human genome and the role in plays in health and disease.

OUR CODE OF LIFE… 😎

The Human Genome Project

- 1990 – 2003

- Aims:
  - To fully sequence and map out the human genome.
  - To provide an accessible database of this information which can be used as a reference.
  - To better understand the human genome and the role in plays in health and disease.

The Human Genome Project

- Demonstrated previous assumptions about genetic architecture of human traits = grossly underestimated.
- Mapped out 3,000,000,000 nucleotide base pairs.
- Provided a map of all major genes existing within the genome.
- Demonstrated >50% of genome = repetitive sequence.
- Mostly consists of non-coding regions!
- Made the data open access to allow use for future research.

The Human Genome Project

Paved the way for subsequent genome-wide understanding…

- **International HapMap Consortium** → using insights provided by the human genome project were able to produce a map of SNPs commonly inherited together (in so called "linkage disequilibrium").
  - Provides a way to predict information for many SNPs using information obtained from only one within the same correlated block.

THESE TWO INITIATIVES HAVE BEEN PIVOTAL IN ALLOWING US TO MOVE AWAY FROM HYPOTHESIS DRIVEN, SINGLE GENE ANALYSES (which were unsuccessful for complex disorders) AND BROUGHT TO LIGHT THE KNOWLEDGE NEEDED TO MOVE FORWARD AT THE GENOME-WIDE LEVEL.

Learning Objectives

- Understand the basics of genetics of “complex” disorders.
- Understand the importance of the human genome project in:
  - enabling a better understanding of the core structure of the human genome.
  - enabling the introduction of genome-wide approaches to understanding links between genes and disease / genes and behavior.
  - paving the way for the rise of other “omics” fields within biology.
- Develop a working knowledge of core genome-wide methodology and understand the importance of genome-wide association studies.
- Understand the importance of genomics as a key strata within the wider “omics” field.

The Rise of Genome Wide Association Studies

- GWAS utilize knowledge across the entirety of the genome.

- Generated using high density micro-arrays containing probes capable of assaying hundreds of thousands to millions of SNPs in a single array

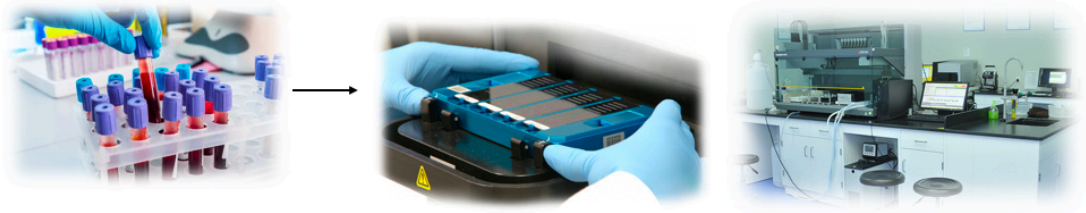
The Rise of Genome Wide Association Studies

- GWAS utilize knowledge across the entirety of the genome.

- Generated using high density micro-arrays containing probes capable of assaying hundreds of thousands to millions of SNPs in a single array

REMEMBER THIS?…

What is Omics?



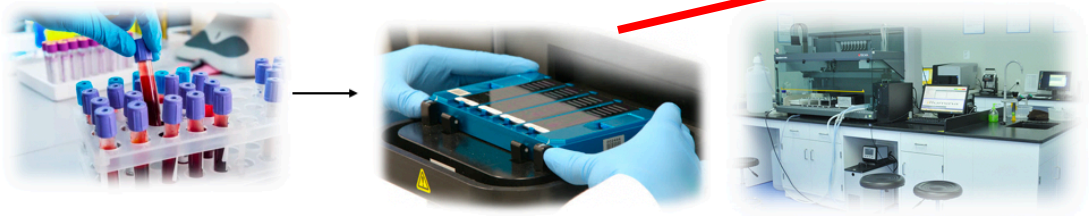
The Rise of Genome Wide Association Studies

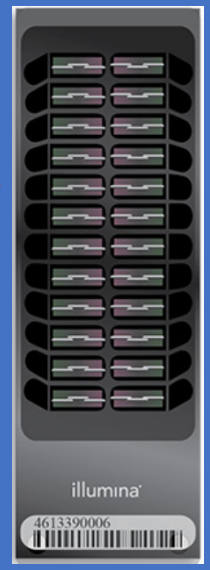
- GWAS utilize knowledge across the entirety of the genome.

- Generated using high density micro-arrays containing probes capable of assaying hundreds of thousands to millions of SNPs in a single array

REMEMBER THIS?…

What is Omics?



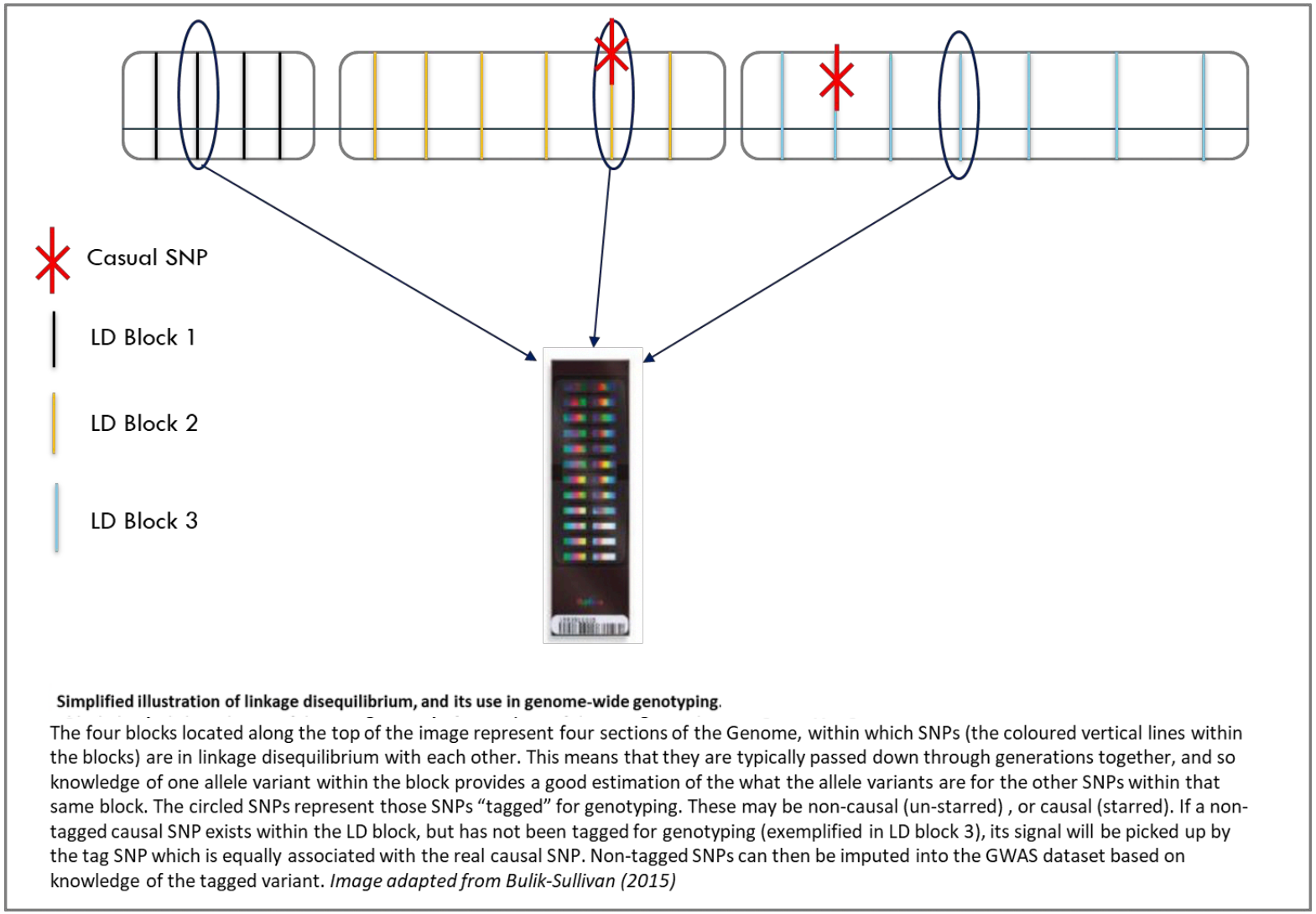


SNP genotyping array

The Rise of Genome Wide Association Studies

- GWAS utilize knowledge across the entirety of the genome.
- Generated using high density micro-arrays containing probes capable of assaying hundreds of thousands to millions of SNPs in a single array.
- Knowledge of LD structure from the HapMap consortium allows for strategic selection of variants which “tag” ungenotyped SNPs
  - These = then imputed into the data afterwards, using knowledge from their correlated genotyped marker.
  - This allows data for millions of SNPs through direct genotyping of just a few hundred-thousand.

Genome Wide Association Studies



Simplified illustration of linkage disequilibrium, and its use in genome-wide genotyping.

The four blocks located along the top of the image represent four sections of the Genome, within which SNPs (the coloured vertical lines within the blocks) are in linkage disequilibrium with each other. This means that they are typically passed down through generations together, and so knowledge of one allele variant within the block provides a good estimation of the what the allele variants are for the other SNPs within that same block. The circled SNPs represent those SNPs "tagged" for genotyping. These may be non-causal (un-starred) , or causal (starred). If a non-tagged causal SNP exists within the LD block, but has not been tagged for genotyping (exemplified in LD block 3), its signal will be picked up by the tag SNP which is equally associated with the real causal SNP. Non-tagged SNPs can then be imputed into the GWAS dataset based on knowledge of the tagged variant. *Image adapted from Bulik-Sullivan (2015)*

Genome Wide Association Studies

- When investigating disorders → typically get blood samples from large numbers of individuals with the disorders and large numbers of people without the disorder.
- Run each of their blood samples through the micro-arrays.
- Upload the genotype (their SNP information at the different locations across the genome – measured by the arrays) and phenotype (affection, sex, age etc) information onto a high performance computing space.
- Perform pairwise association tests for each one of the SNPs measured against the phenotype (disorder) of interest → typically logistic regression.

Genome Wide Association Studies

- Because you perform pairwise associations between each SNP for a single sample → have to perform statistical corrections to correct for multiple-testing.

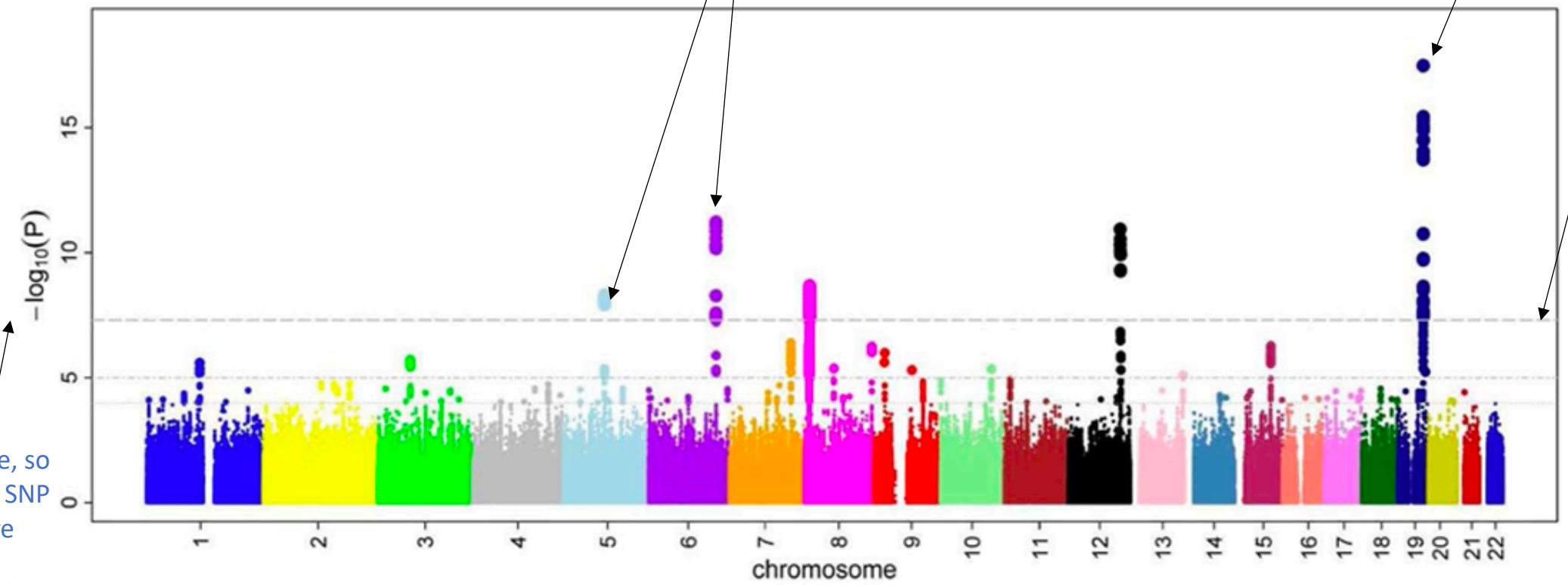
- This is typically a corrected threshold of $5*10^{-08}$
  - Conventional significance threshold = 0.05 (allowing for a 5% chance of a false association being identified).
  - As you perform more and more tests, the chances of you detecting an association simply by chance increases.
  - To correct for this increased likelihood of identifying an incorrect association, we divide the conventional significance threshold by the number of tests performed.
  - As there are typically ~1million SNP-phenotype associations tested within standard GWAS – we divide 0.05 by 1million, which = $5*10^{-08}$.
  - The consequence of this = that in order to claim a "significant association", the effect size of the association between the SNP and phenotype needs to be very large (which we know, from previous knowledge it is unlikely to be) or sample sizes need to be HUGE, so that we have enough power to see an effect if it is there.

Genome Wide Association Studies

- Typical output:

Different colours = SNPs on different chromosomes

Each point = a SNP



Line denoting ”significance” – anything about = considered to be sig associated with the disorder. Anything below = cannot say with confidence there is an association

Minus log scale, so the higher the SNP point, the more significant its association with the disorder being measured.

Genome Wide Association Studies

- Because you perform pairwise associations between each SNP for a single sample → have to perform statistical corrections to correct for multiple-testing.

- This is typically a corrected threshold of $5*10^{-08}$
  - Conventional significance threshold = 0.05 (allowing for a 5% chance of a false association being identified).
  - As you perform more and more tests, the chances of you detecting an association simply by chance increases.
  - To correct for this increased likelihood of identifying an incorrect association, we divide the conventional significance threshold by the number of tests performed.
  - As there are typically ~1million SNP-phenotype associations tested within standard GWAS – we divide 0.05 by 1million, which = $5*10^{-08}$.
  - The consequence of this = that in order to claim a "significant association", the effect size of the association between the SNP and phenotype needs to be very large (which we know, from previous knowledge it is unlikely to be) or sample sizes need to be HUGE, so that we have enough power to see an effect if it is there.

Genome Wide Association Studies

- Because you perform pairwise associations between
sample → have to perform statistical corre
testing.

- This is t

= One of the key challenges for GWAS – whilst we see many SNP-disorder associations, many fail to reach the stringent level of significance demanded. This is improving as sample sizes grow, but there is still a high degree of so called "missing heritability", where the genetics explained by GWAS does not align with heritability estimated by traditional family studies.

association being identified).

g an association simply by chance increases.

g an incorrect association, we divide the conventional significance

ied.

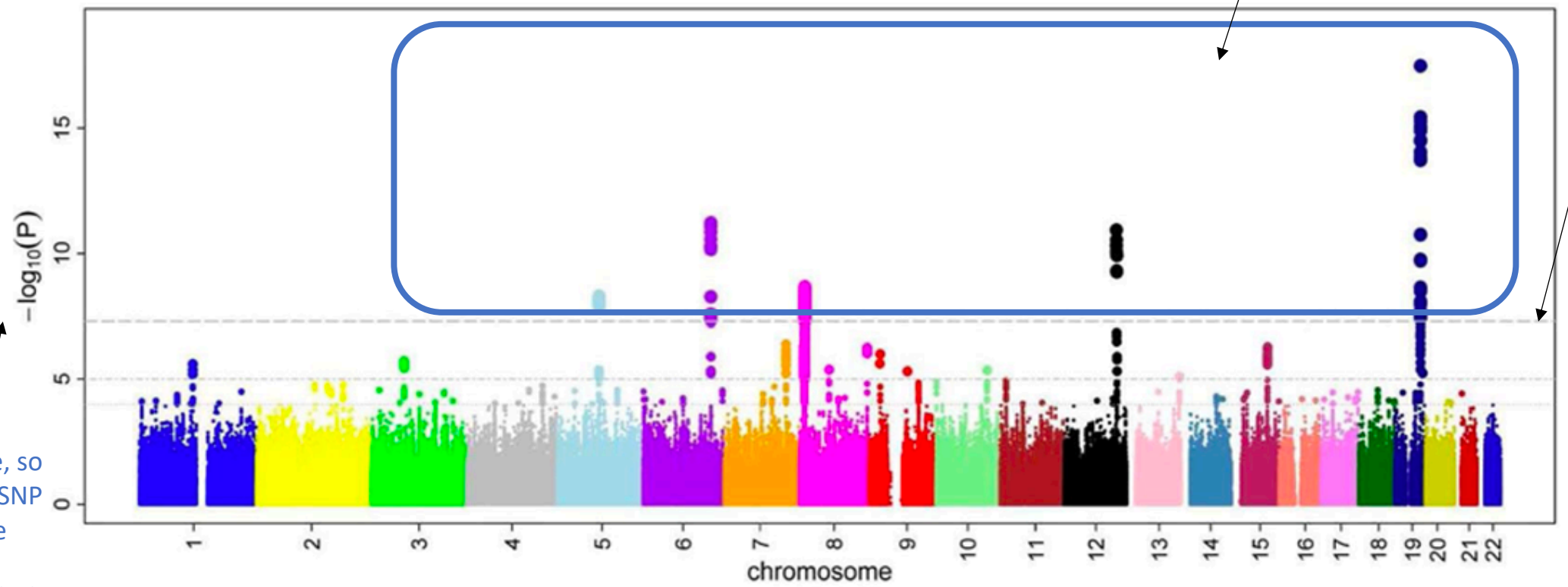
ion SNP-phenotype associations tested within standard GWAS – we divide 0.05 by 1million, which

consequence of this = that in order to claim a "significant association", the effect size of the association between the SNP and phenotype needs to be very large (which we know, from previous knowledge it is unlikely to be) or sample sizes need to be HUGE, so that we have enough power to see an effect if it is there.

Genome Wide Association Studies

- Typical output:

Only a few SNPs making the cut – which, when effects are added together, explain only a small proportion of genetic liability

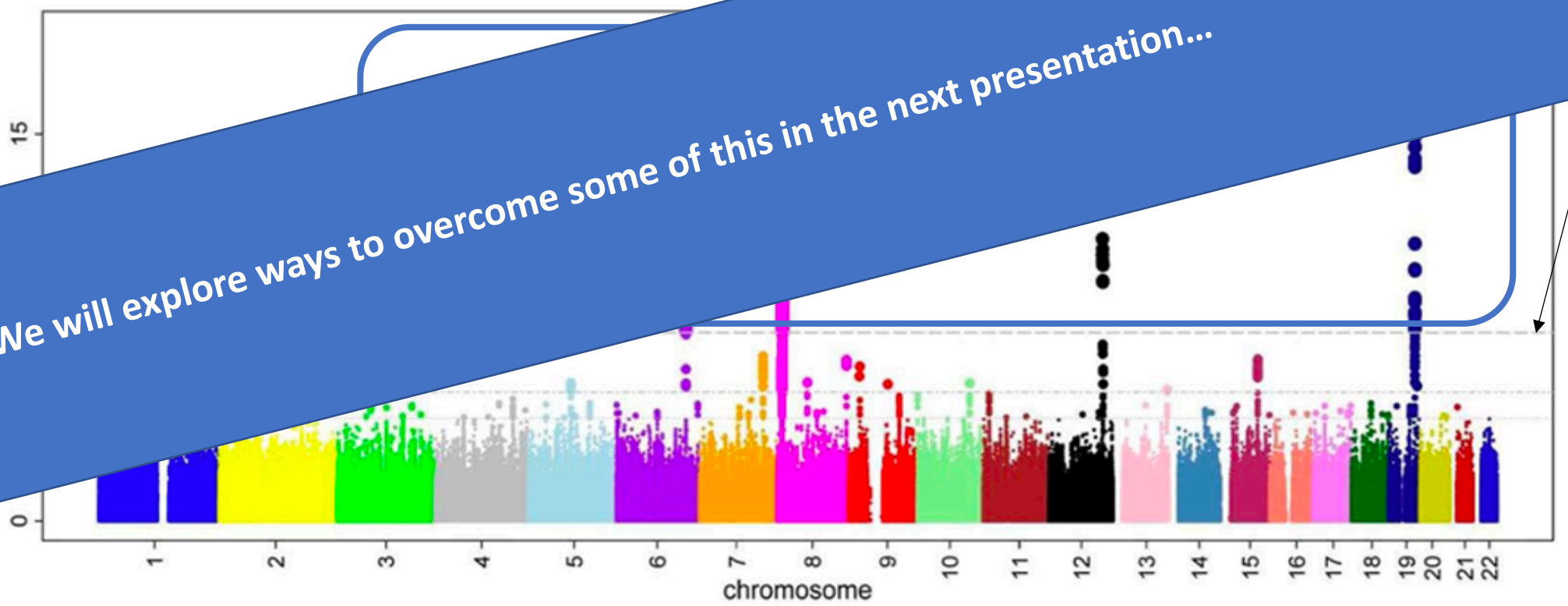


Line denoting ”significance” – anything about = considered to be sig associated with the disorder. Anything below = cannot say with confidence there is an association

Minus log scale, so the higher the SNP point, the more significant its association with the disorder being measured.

Genome Wide Association Studies

- Typical output:



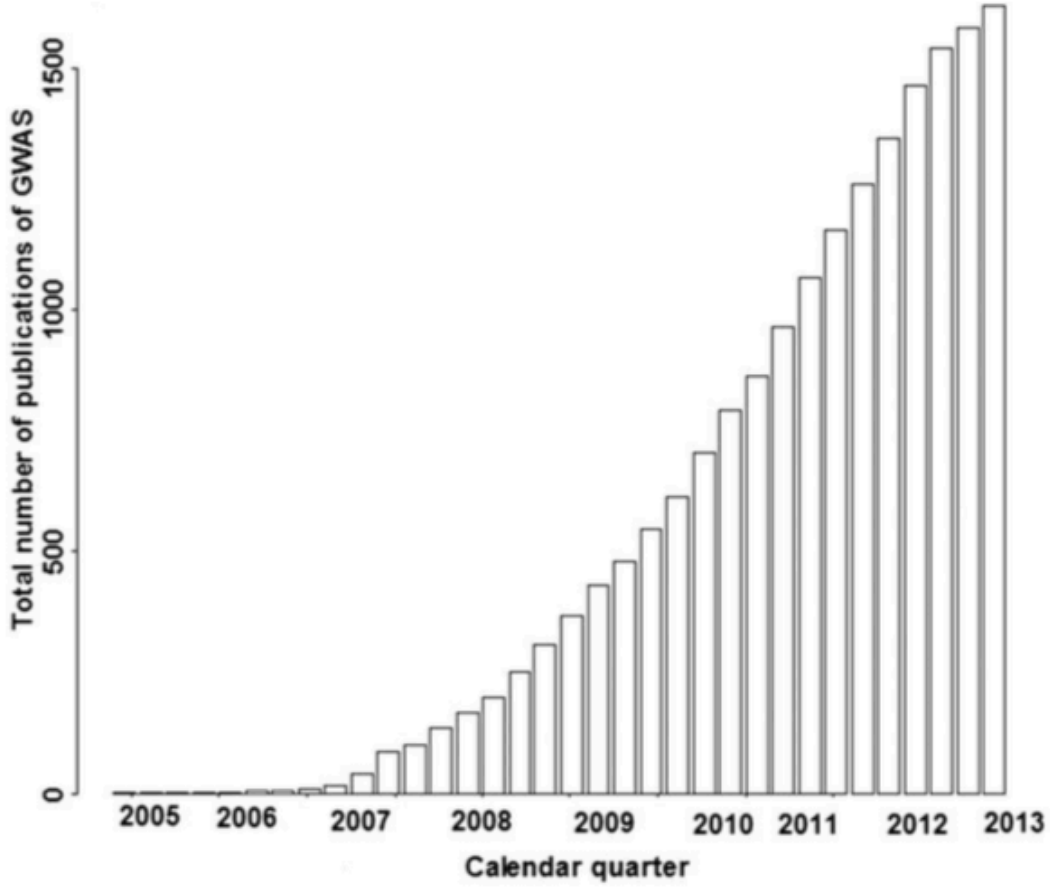
Only a few SNPs making the cut – v…
when effects are added t…
only a small…

…ing
…cance
…hing
…t =
considered to be sig associated with the disorder. Anything below = cannot say with confidence there is an association

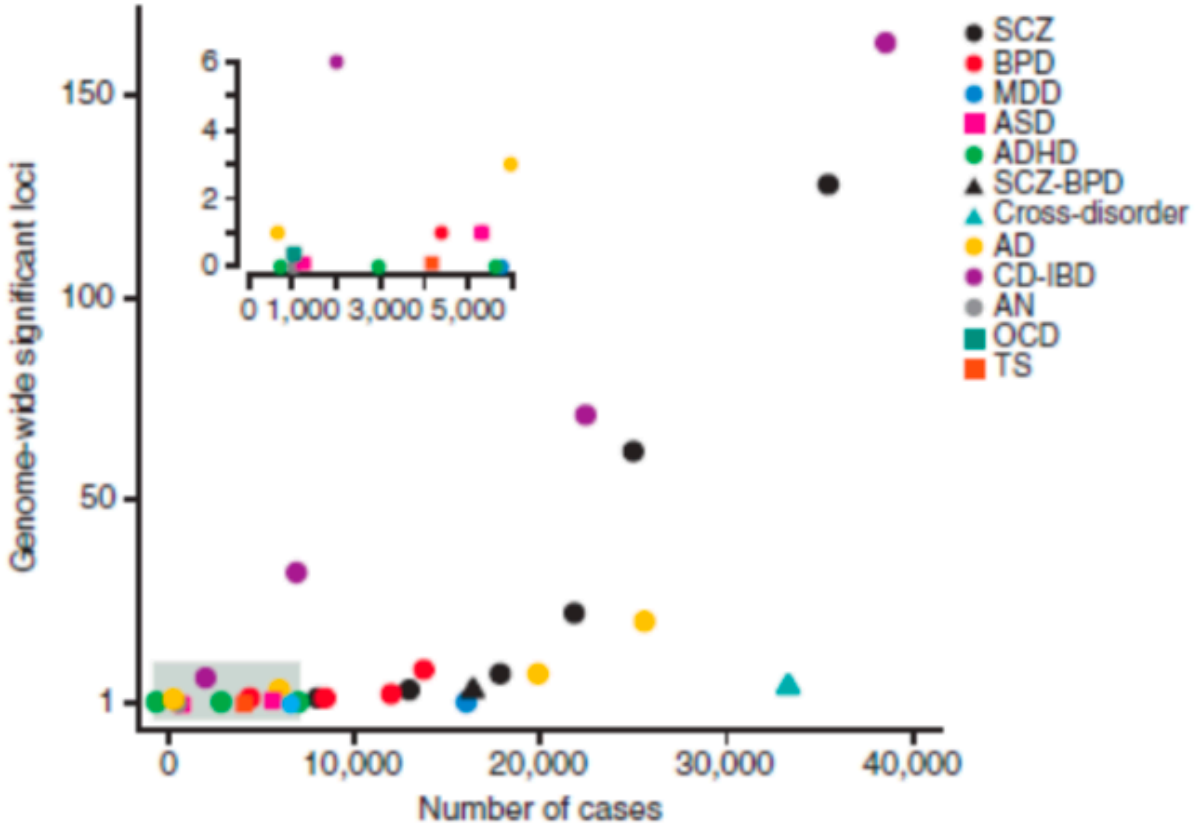
Minus lo…
the highe… SNP
point, the more significant its association with the disorder being measured.

We will explore ways to overcome some of this in the next presentation…

Genome Wide Association Studies



Genome Wide Association Studies



- As studies and sample sizes increase, we are seeing an increase in the number of genetic associations uncovered at genome-wide significance between SNPs and disorders

Learning Objectives

- Understand the basics of genetics of “complex” disorders.
- Understand the importance of the human genome project in:
  - enabling us to better understand the core structure of the human genome.
  - enabling the introduction of genome-wide approaches to understanding links between genes and disease / genes and behavior.
  - paving the way for the rise of other “omics” fields within biology.
- Develop a working knowledge of core genome-wide methodology and understand the importance of genome-wide association studies.
- Understand the importance of genomics as a key strata within the wider “omics” field.

The importance of Genomics

- Genomics paved the way for other omic technologies which followed.
- Many of the key insights we have been able to uncover across the other strata have been possible because we have been able to utilize genetic information as an “anchor” – due to the fact genes (unlike the other omic layers) remain static across the life-course.
- Our genetics lay the foundation for every other process which comes thereafter – utilizing information from our genes is therefore fundamental if we wish to integrate information across other layers.
- Genomic information allows us to link together data across different strata which may be from completely separate individuals – making multi-omic approaches more feasible.

Takeaways…

- Most disorders have a complex/multi-factorial genetic architecture.
  →Many variants, small effect, none singularly causative.

- Classic, hypothesis driven, single gene approaches often not appropriate.

- The human genome project allowed for the advent of genome-wide association studies → look at variants across the genome → hypothesis free → including non-coding regions.

- GWAS have been revolutionary in our understanding of disorder genetic architecture but require very large sample sizes → much missing heritability still exists.

- The wider “omics” field has relied on insights gained from genomics prior.