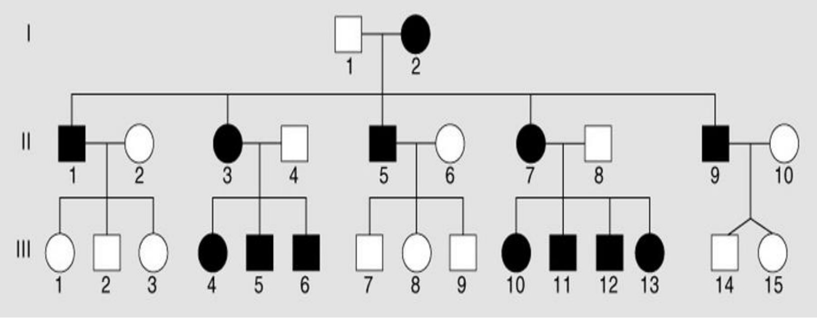
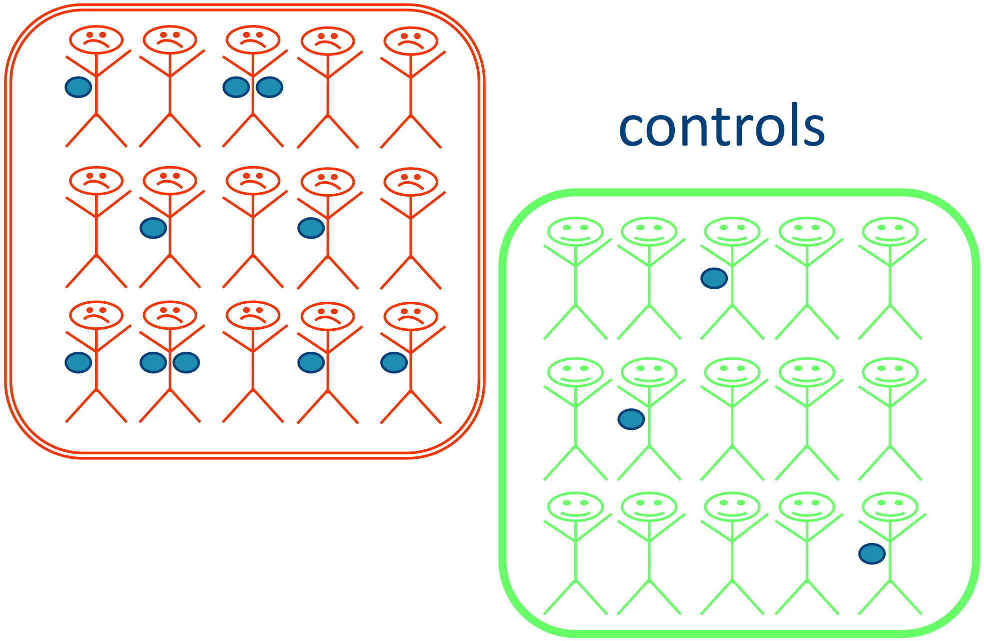
Linkage versus association

- Linkage studies



- Association studies

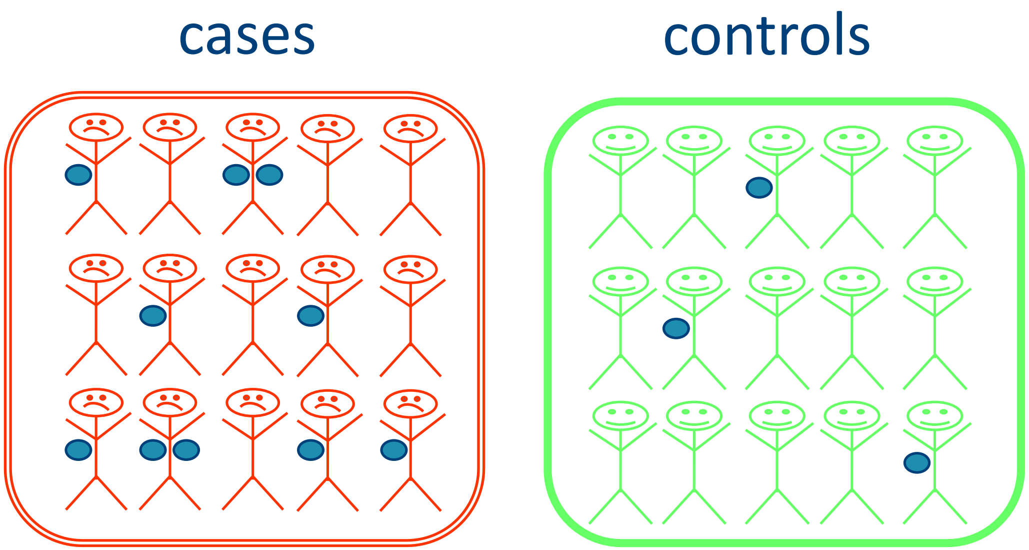
cases



controls

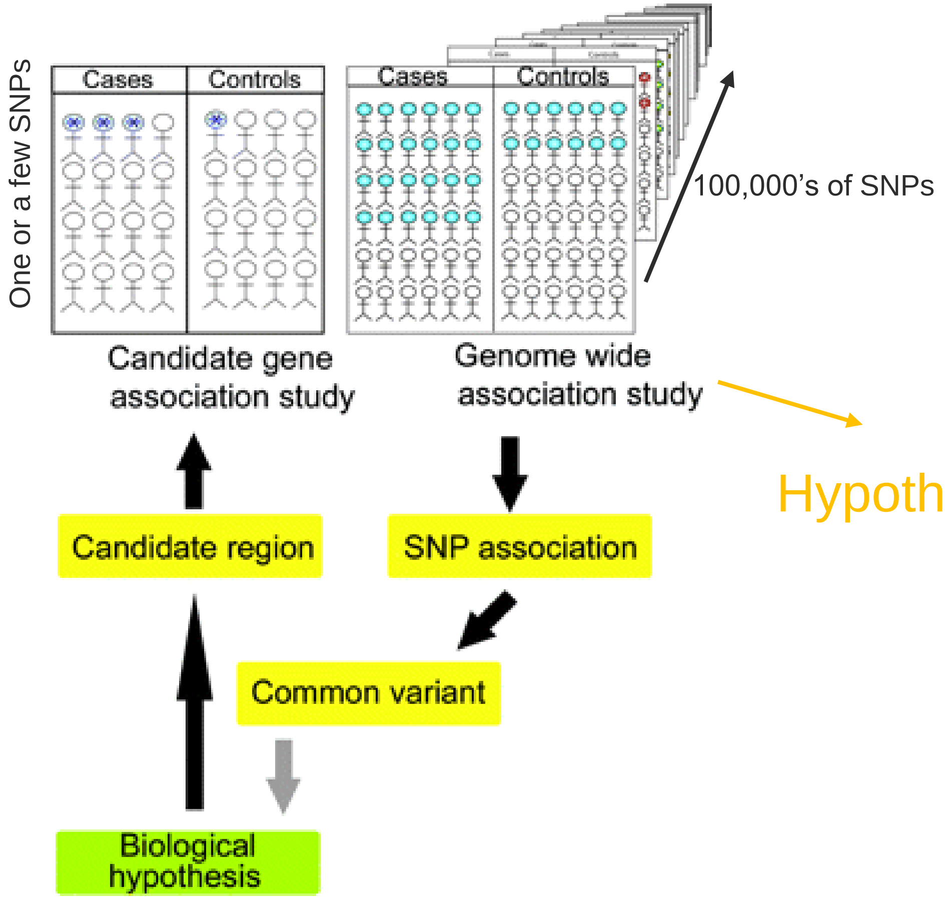
Case-control association study

- Statistical significant differences between frequency of variants in patients compared with control individuals



KU LEUVEN

Candidate-gene vs genome-wide



One or a few SNPs

Cases

Controls

100,000's of SNPs

Candidate gene association study

Genome wide association study

Candidate region

SNP association

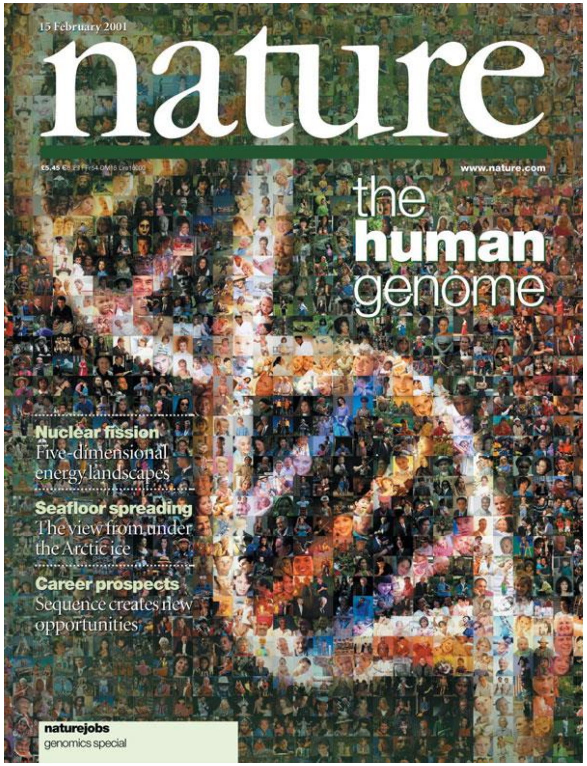
Common variant

Biological hypothesis

Hypothesis-free!

Candidate-gene vs genome-wide

Scientific and technological breakthroughs



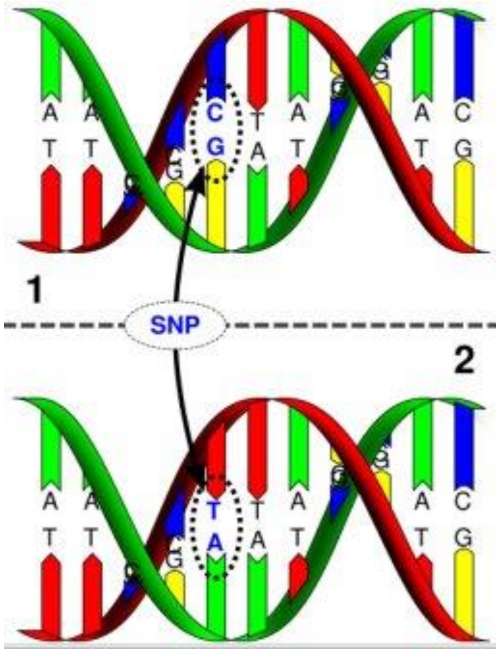
Key concepts to understand GWAS

Key concepts I – SNP = single nucleotide polymorphism

Key concepts I – SNP = single nucleotide polymorphism

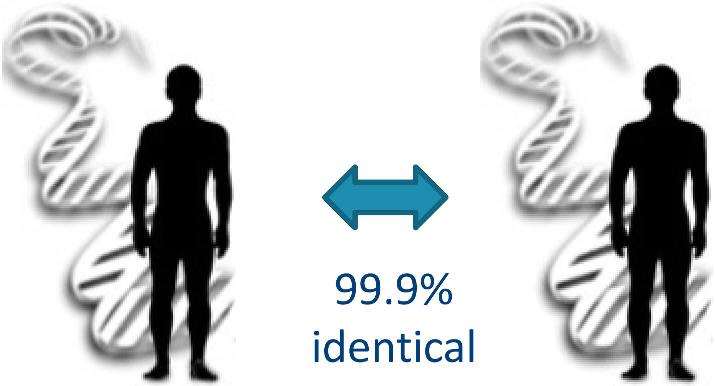
- a DNA sequence variation occuring when a single nucleotide (A,T,C or G) in the genome differs between members of a species, or between paired chromosomes in an individual **at a particular locus**

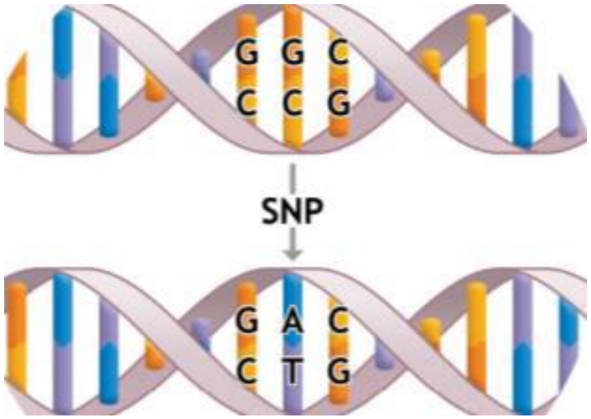
- E.g.



Potential genotypes for a person at this locus: CC/CT/TT

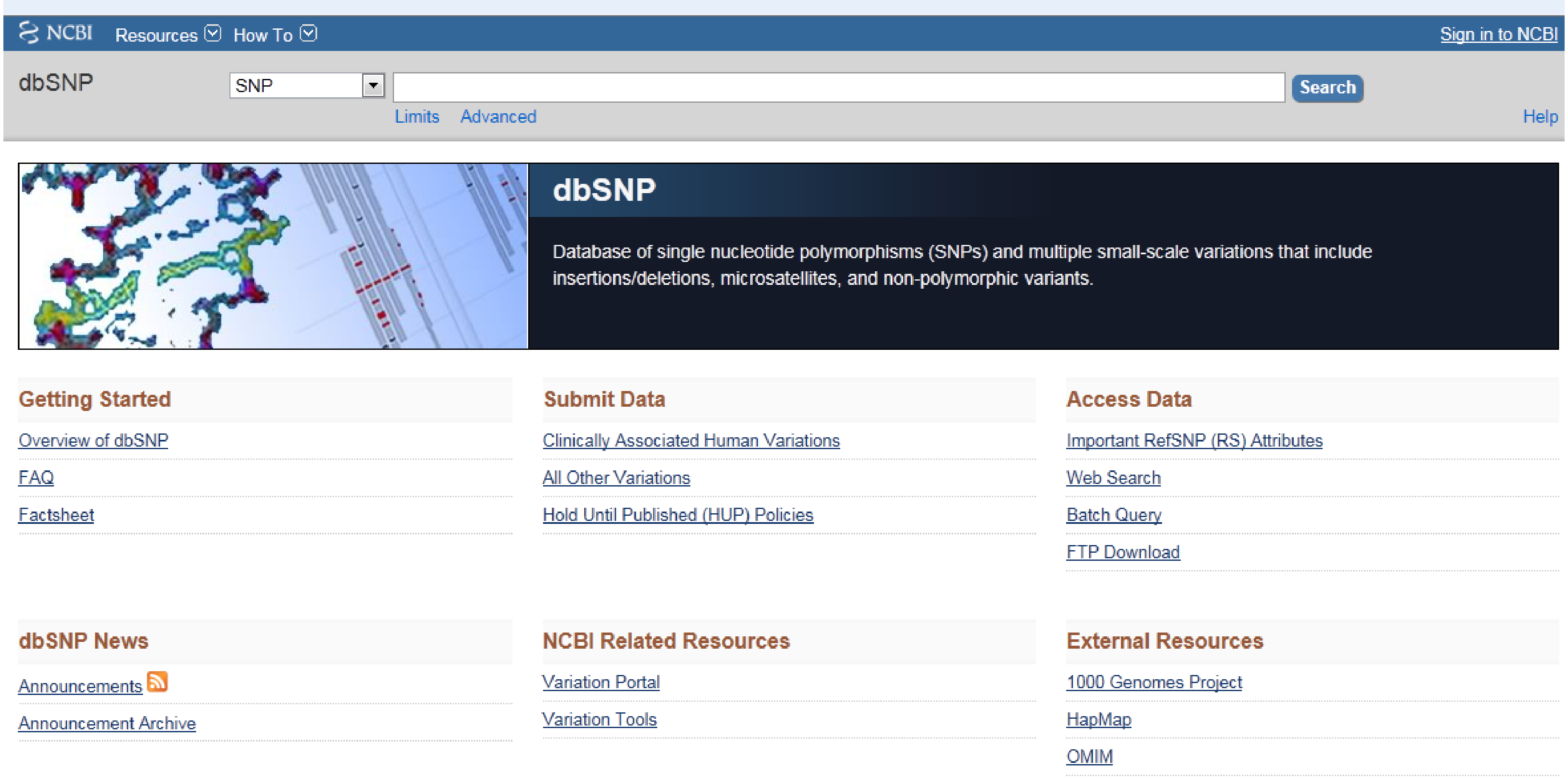
Key concepts I – Single Nucleotide Polymorphisms (SNPs)





- Phenotypically normal individuals
- SNP based genome variation about 0.1%
- Believed to explain majority of
  - our phenotypic variation
  - inherited Mendelian and complex disorders

Each SNP has a unique rs id



dbSNP

Database of single nucleotide polymorphisms (SNPs) and multiple small-scale variations that include insertions/deletions, microsatellites, and non-polymorphic variants.

Getting Started

- Overview of dbSNP
- FAQ
- Factsheet

Submit Data

- Clinically Associated Human Variations
- All Other Variations
- Hold Until Published (HUP) Policies

Access Data

- Important RefSNP (RS) Attributes
- Web Search
- Batch Query
- FTP Download

dbSNP News

- Announcements
- Announcement Archive

NCBI Related Resources

- Variation Portal
- Variation Tools

External Resources

- 1000 Genomes Project
- HapMap
- OMIM

Key concepts II – MAF = minor allele frequency

- The frequency of the SNP's less frequent allele in a given population
  - As the total allele frequency is 1 (100%), a MAF must always be less than 0.5 (50%), otherwise it would be a major allele
    - E.g. if we genotype a variant (A/G) in 1000 people
      - 550 are (A,A), 400 are (A,G) and 50 are (G,G)
      - There are 2000 alleles in total
      - The G allele is less common, accounting for 500 alleles
      - Therefore, the MAF is 500/2000 = 0.25 or 25%

- **Rare variants**: MAF < 0.5% (76% of all variants)
- **Low-frequency variants**: MAF 0.5-5% (14% of all variants)
- **Common variants**: MAF > 5% (10% of all variants)
  - Estimated to be > 10 million common variants in human
  - Focus of GWAS usually on common variants (SNPs)

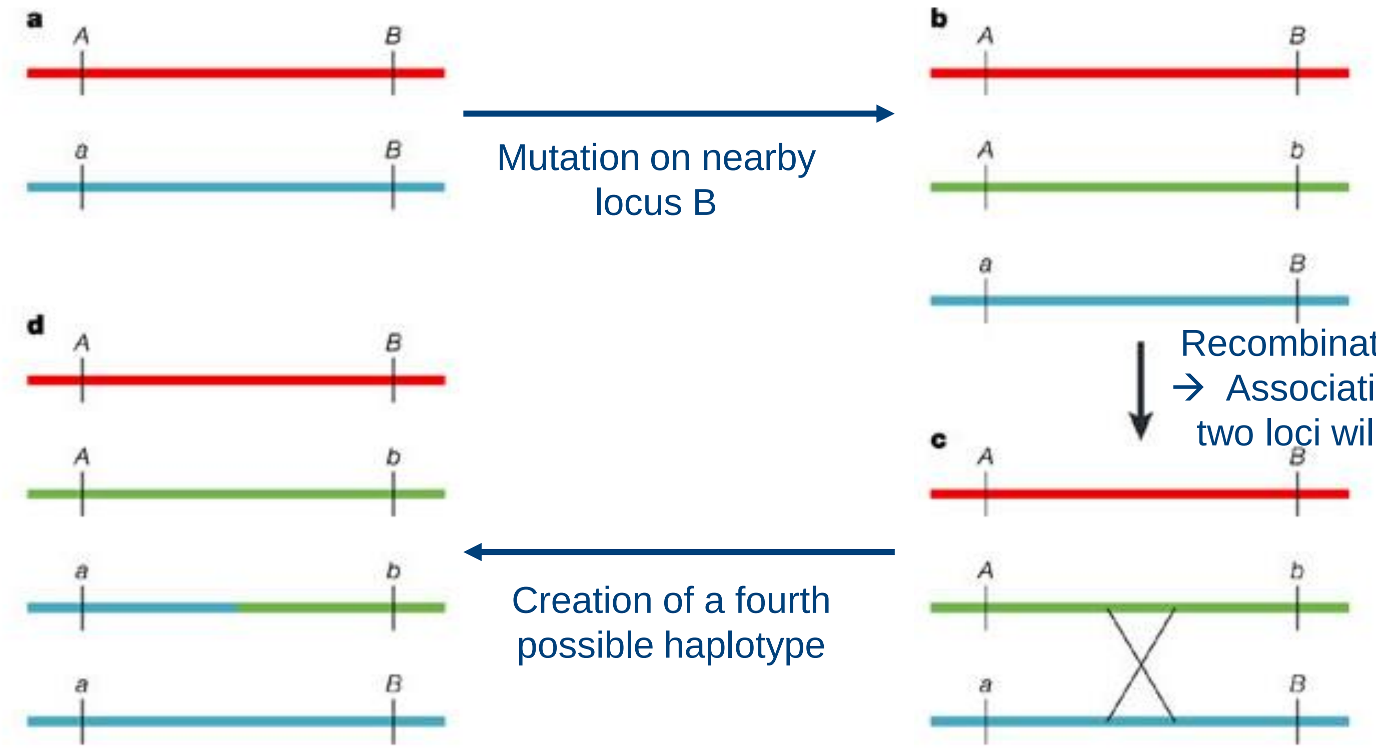
Key concepts III – Hardy-Weinberg equilibrium

- HWE = Hardy-Weinberg equilibrium
  - Both **allele and genotype frequencies remain constant** in a population unless specific disturbing influences are introduced
    - mutation, migration, non-random mating etc.
    - selection
  - Under HWE, genotype frequencies can be estimated from allele frequencies
    - Assume alleles A1 with P(A1)=p and A2 with P(A2)=q
      - P(A1A1)=p2
      - P(A1A2)=2pq
      - P(A2A2)=q2

	A1	A2
A1	A1A1 (p*p=p^2)	A1A2 (p*q)
A2	A2A1 (q*p)	A2A2 (q*q=q^2)

  - X^2-test to test for deviances from the expected

Key concepts IV – Linkage disequilibrium



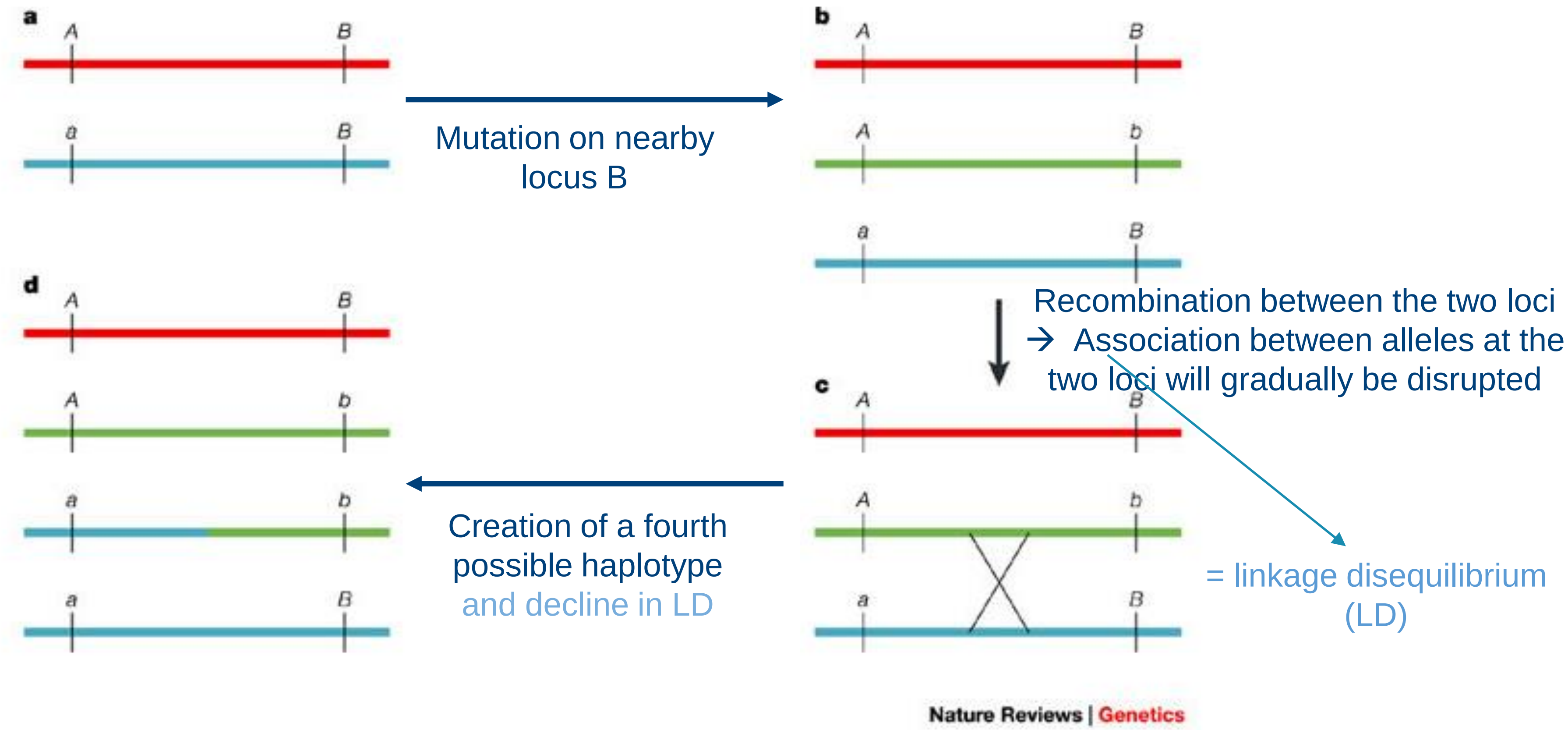
Mutation on nearby locus B

Recombination between the two loci
→ Association between alleles at the two loci will gradually be disrupted

Creation of a fourth possible haplotype

Nature Reviews | Genetics

Key concepts IV – Linkage disequilibrium



Mutation on nearby locus B

Recombination between the two loci → Association between alleles at the two loci will gradually be disrupted

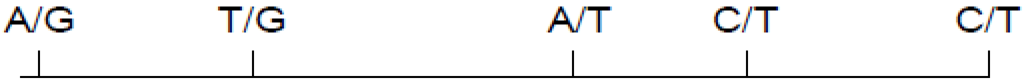
= linkage disequilibrium (LD)

Creation of a fourth possible haplotype and decline in LD

Nature Reviews | Genetics

Key concepts IV – Linkage disequilibrium

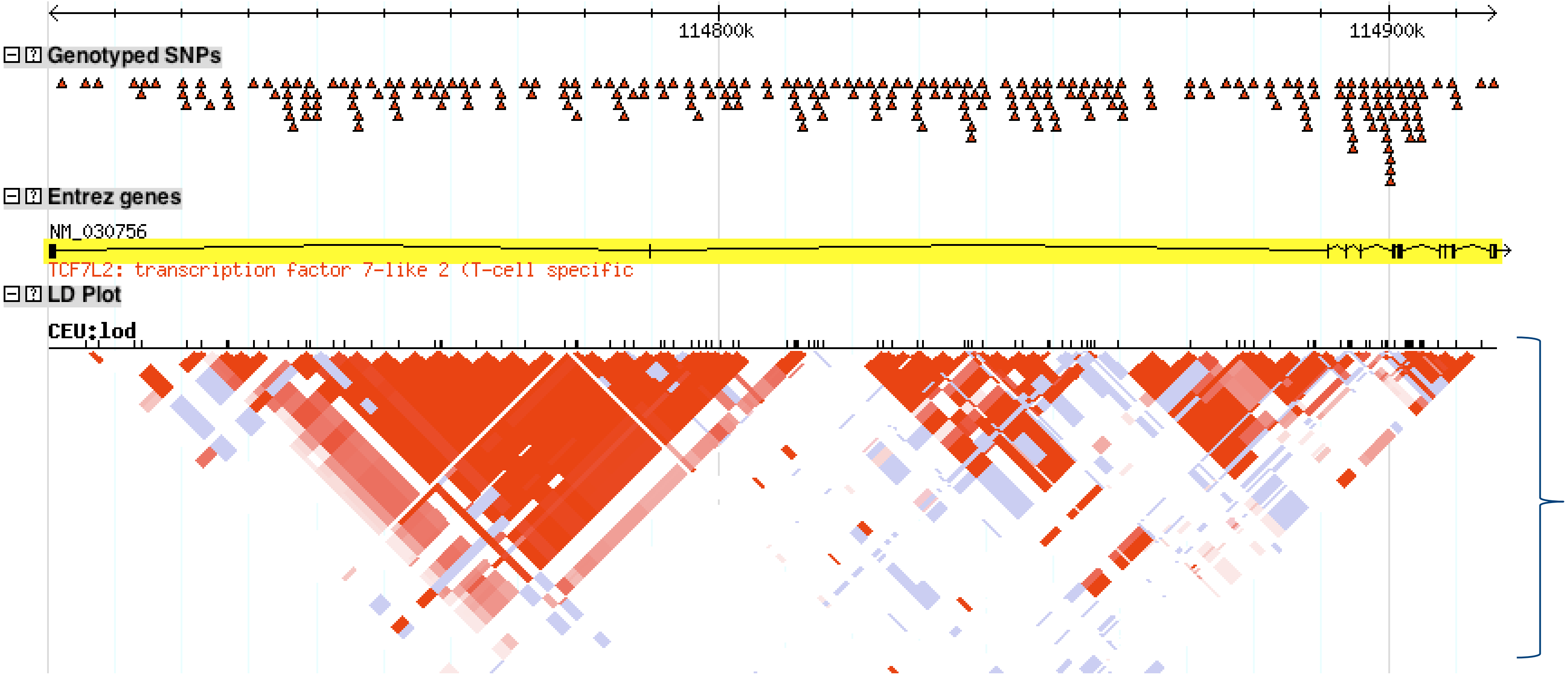
- Specific region of the genome
- 5 SNPs with two frequent alleles each



- Theoretically there are 2^5 different combinations (haplotypes)
- Practically there will only be a few
  - Eg if A, then always ATACT
  - Eg if G, then almost always GGACC

➢ These alleles are in LD

Patterns of variation



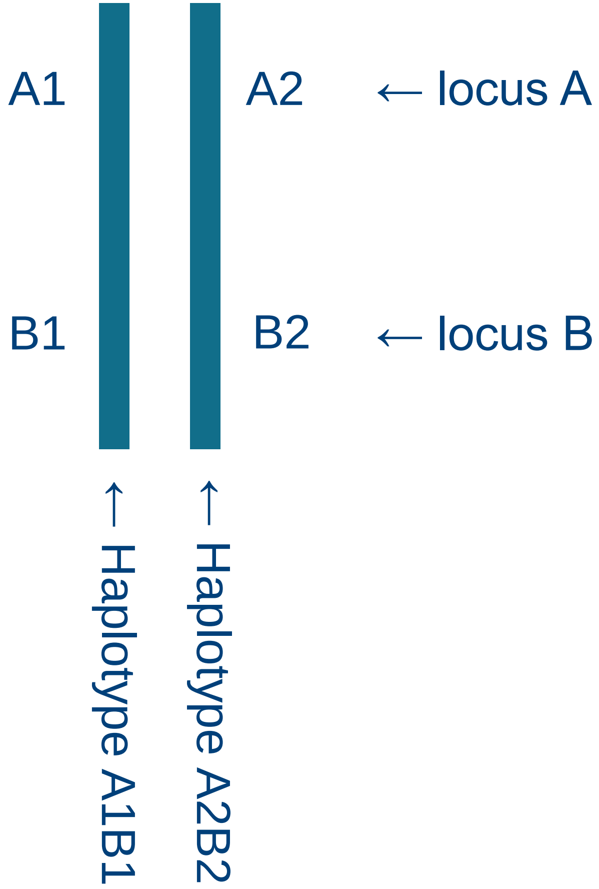
LD pattern

Key concepts IV – Linkage disequilibrium

Locus	Allele	Observed frequency
A	A1	p1
A	A2	p2
B	B1	q1
B	B2	q2

Haplotype	Expected frequency[1]
A1B1	p1 * q1
A1B2	p1 * q2
A2B1	p2 * q1
A2B2	p2 * q2

[1] under linkage equilibrium



Key concepts IV – Linkage disequilibrium

Locus	Allele	Observed frequency
A	A1	p1
A	A2	p2
B	B1	q1
B	B2	q2

Haplotype	Expected frequency[1]	Observed frequency
A1B1	p1 * q1	*x*11
A1B2	p1 * q2	*x*12
A2B1	p2 * q1	*x*21
A2B2	p2 * q2	*x*22

[1] under linkage equilibrium

Key concepts IV – Linkage disequilibrium

Locus	Allele	Observed frequency
A	A1	p1
A	A2	p2
B	B1	q1
B	B2	q2

When 2
alleles occur
on the same
haplotype
more often
than expected

Haplotype	Expected frequency[1]	Observed frequency	A2 and B2 in positive LD
A1B1	p1 * q1	*x*11	*x*11 > p1 * q1
A1B2	p1 * q2	*x*12	*x*12 < p2 * q2
A2B1	p2 * q1	*x*21	*x*21 < p2 * q1
A2B2	p2 * q2	*x*22	*x*22 > p2 * q2

[1] under linkage equilibrium

Key concepts IV – Linkage disequilibrium

Locus	Allele	Observed frequency
A	A1	p1
A	A2	p2
B	B1	q1
B	B2	q2

When 2 alleles occur on the same haplotype more often than expected

When 2 alleles occur on the same haplotype less often than expected

Haplotype	Expected frequency[1]	Observed frequency	A2 and B2 in positive LD	A2 and B2 in negative LD
A1B1	p1 * q1	*x*11	*x*11 > p1 * q1	*x*11 < p1 * q1
A1B2	p1 * q2	*x*12	*x*12 < p2 * q2	*x*12 > p2 * q2
A2B1	p2 * q1	*x*21	*x*21 < p2 * q1	*x*21 > p2 * q1
A2B2	p2 * q2	*x*22	*x*22 > p2 * q2	*x*22 < p2 * q2

[1] under linkage equilibrium

Key concepts IV – Linkage disequilibrium (LD)

- The **non-random association** of alleles at two or more loci such that they are inherited together more frequently than expected by chance
- Observed across the entire genome, not only nearby coding regions or genes causing disease
- Extent of LD varies greatly depending on the region of the genome
- When LD is strong, need fewer SNPs to capture variation in a region (‘tag SNPs’)
- Measures of LD: D, D’, and r^2

Measures of LD

- D
  - = difference between observed and expected haplotype frequencies
  - = $x_{11} - (p_1 * q_1)$

Haplotype	Expected frequency[1]	Observed frequency
A1B1	p1 * q1	*x*11 = p1 * q1 + D
A1B2	p1 * q2	*x*12 = p1 * q2 - D
A2B1	p2 * q1	*x*21 = p2 * q1 - D
A2B2	p2 * q2	*x*22 = p2 * q2 + D

  - Hard to interpret…
    - Can be negative (with arbitrary sign, depending on which one is set as A1, B1 or A2, B2)
    - Range depends on allele frequencies, and is sensitive to allele frequencies at extreme values of 0 to 1 → hard to compare markers

Measures of LD

- D’
  - = normalized D
  - = divide D by its theoretical maximum for the observed allele frequencies (ie absolute maximal possible value of D)
  - = D/D_{max}

D	D’
D > 0	D′ = (x11 − p1 * q1) / min($p1*q2$, $p2*q1$)
D < 0	D′ = (x11 − p1 * q1) / min($p1*q1$, $p2*q2$)
D = 0	D′ = 0

  - Ranges between -1 to +1
    - ±1 implies at least one of the observed haplotypes was not observed

More on D’

- Pluses:
    - D’ = 1 or D’ = -1 means no evidence for recombination between the markers
    - If allele frequencies are similar, high D’ means the markers are good surrogates for each other
- Minuses:
    - D’ estimates inflated in small samples
    - More likely to take extreme values when allele frequencies are small
    - D’ estimates inflated when one allele is rare

Measures of LD

- r^2
  - equivalent to Pearson correlation coefficient
  - $= D^2/(p1 * p2 * q1 * q2)$

  - Ranges between 0 to +1
    - 1 when the two markers provide identical information
    - 0 when they are in perfect equilibrium

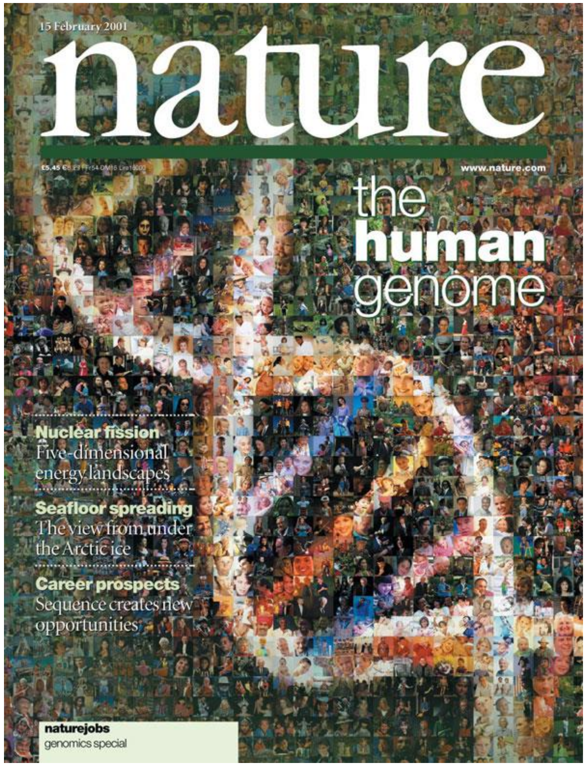
  - The measure preferred by population geneticists
    - Most commonly used in genetic association studies for human complex traits (common variants!)
    - Although it will drop with low allele frequencies (potential false negatives)

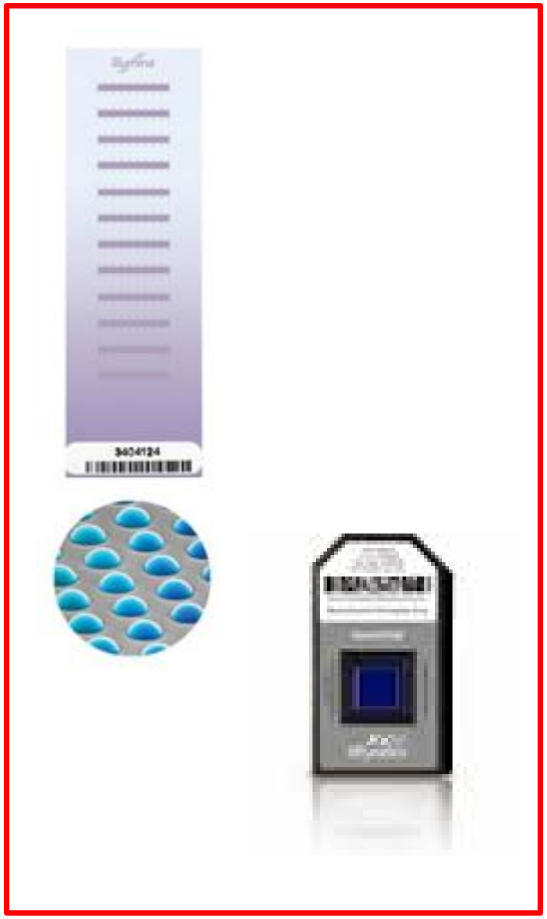
➢ When deciding which measure to use, allele frequencies are often key

Linkage disequilibrium – need to knows

- Linkage vs linkage disequiliborum
  - Linkage = between two loci located close to each other
  - Linkage disequilibrium = between two alleles of linked loci
- SNPs that are physically far away from each other are usually not well correlating because of recombination
  - In general, LD between two SNPs decreases with physical distance
- The ‘age’ of a SNP also defines its correlation with neighboring SNPs
  - Small chance on recombination between two neighboring SNPs, but when time long enough recombination possible
- Recombination hotspots also define the correlation between neighboring SNPs

Scientific and technological breakthroughs

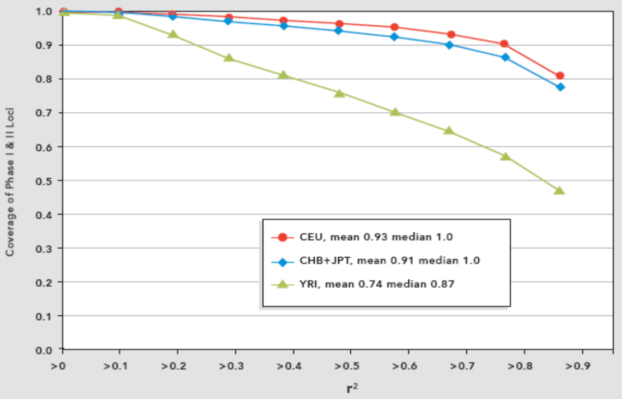




‘High throughput’ genotyping platform

- Commercially available tag sets
- Determination of alleles of different SNPs using SNP microarrays (DNA chips)
- Platform for 100,000’s of SNPs – each with 2 probes (one for each allele)
- **Unbiased** survey of the human genome

FIGURE 2: HUMANHAP550 GENOMIC COVERAGE BY POPULATION



The HumanHap550 Genotyping Beadchip content covers the majority of HapMap Phase I+II common variation.

Commercial SNP arrays

- Affymetrix
  - Affymetrix GeneChip Human Mapping 500K
  - Axiom™ Biobank Plus Genotyping Array
  - Affymetrix SNP 6.0 (900k)

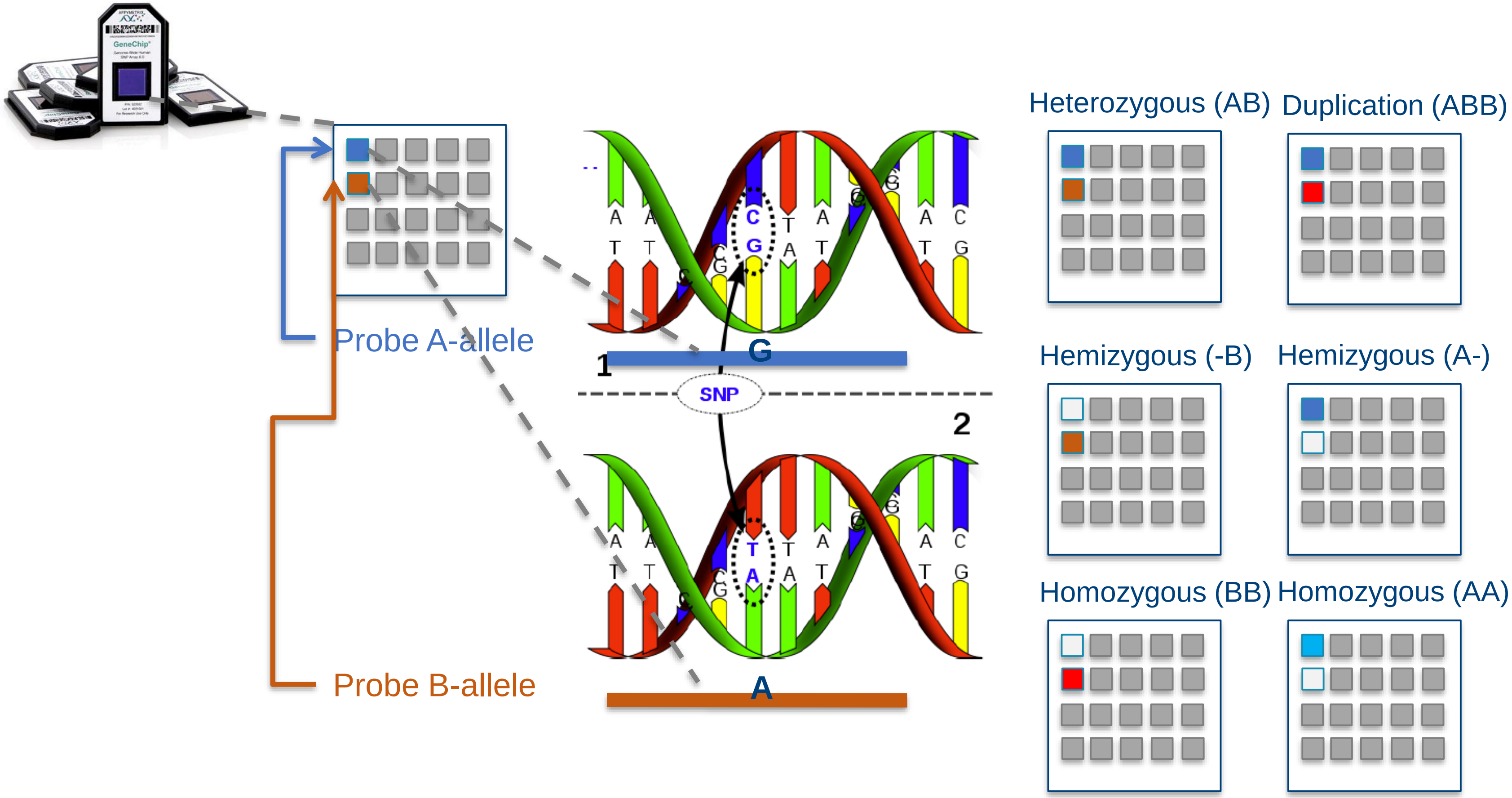
- Illumina
  - Illumina Human Hap 300, 550, 650, 1M, 2.5M, 5M
  - Illumina CardioMetaboChip, ImmunoChip
  - Illumina ExomeChip, CoreExomeChip
  - Illumina Global Screening Array (GSA)

Illumina SNP chip

- Developed for different species (animal and plant):
  - Humans (>4,3 million SNP/array)
  - Livestock (>750 000 SNP/array)
  - Horse (> 50 000 SNP/array)
  - Sheep (> 50 000 SNP/array)
  - Pig (> 60 000 SNP/array)
  - Chicken (> 50 000 SNP/array)
  - Dog (> 50 000 SNP/array)
  - Goat (> 50 000 SNP/array)
  - Maize (> 50 000 SNP/array)
  - Brassica (> 50 000 SNP/array)
  - Potato (> 5 000 SNP/array)
  - Tomato (> 5 000 SNP/array)
  - …

KU LEUVEN

SNP chip: Principle



Probe A-allele

Probe B-allele

Heterozygous (AB)

Duplication (ABB)

Hemizygous (-B)

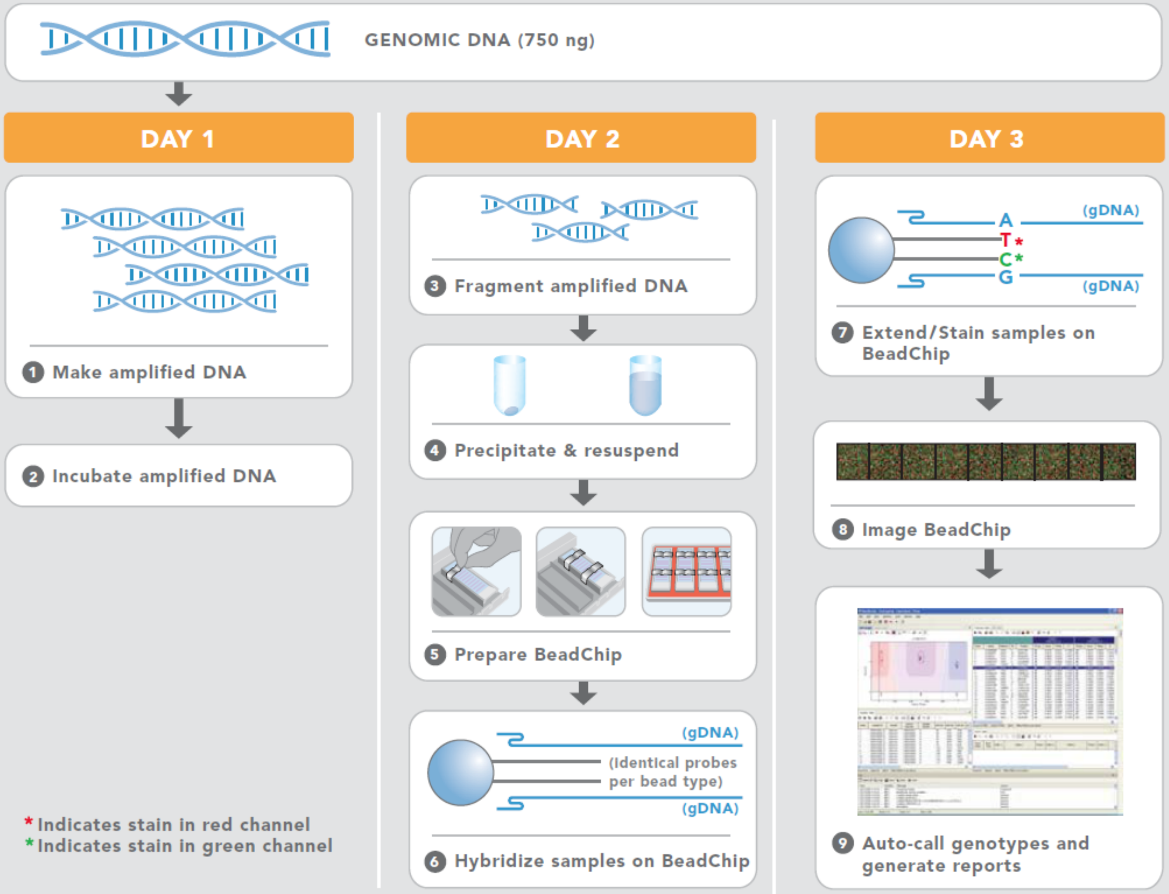
Hemizygous (A-)

Homozygous (BB)

Homozygous (AA)

Illumina Workflow

FIGURE 1: INFINIUM II ASSAY PROTOCOL



GENOMIC DNA (750 ng)

DAY 1

1. Make amplified DNA
2. Incubate amplified DNA

DAY 2

3. Fragment amplified DNA
4. Precipitate & resuspend
5. Prepare BeadChip
6. Hybridize samples on BeadChip

DAY 3

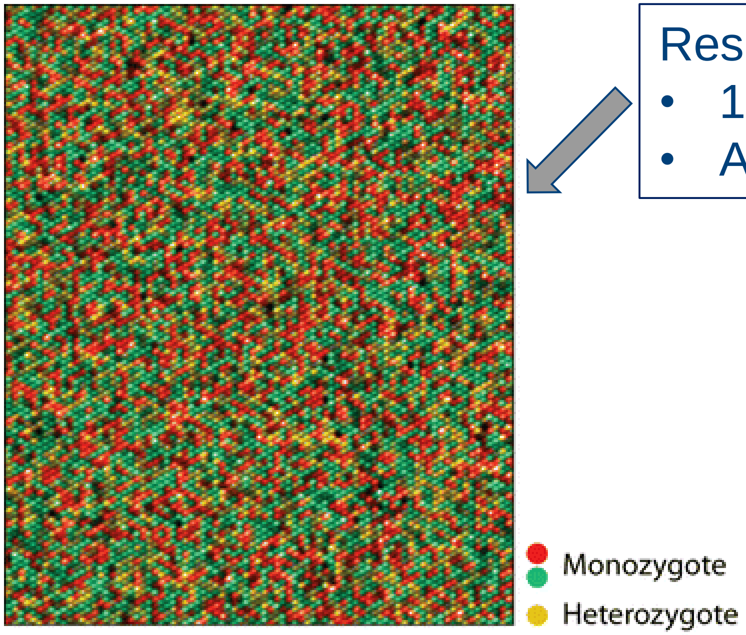
7. Extend/Stain samples on BeadChip
8. Image BeadChip
9. Auto-call genotypes and generate reports

* Indicates stain in red channel
* Indicates stain in green channel

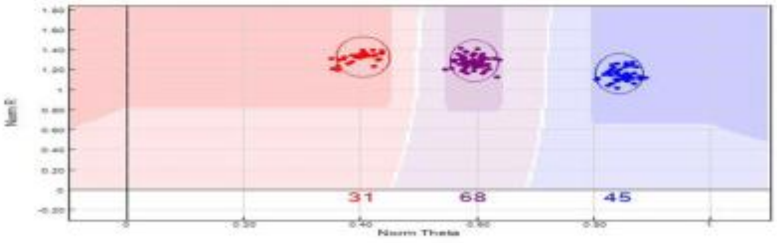
Illumina SNP chip

Results for

- 1 individual
- All SNP genotyped



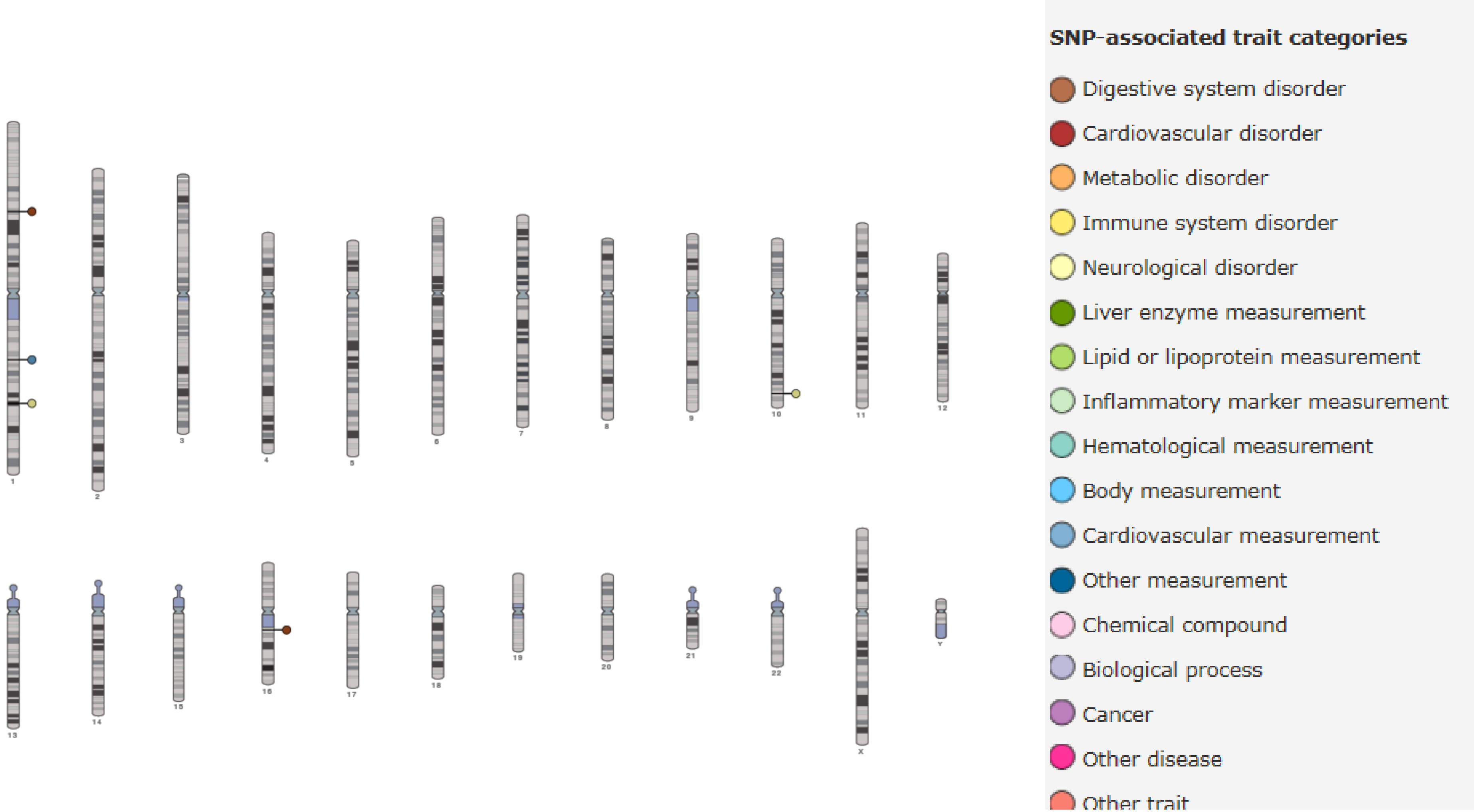
Results for all individuals genotyped are visualised in plots per SNP



Results for all individuals genotyped are translated into genotypes

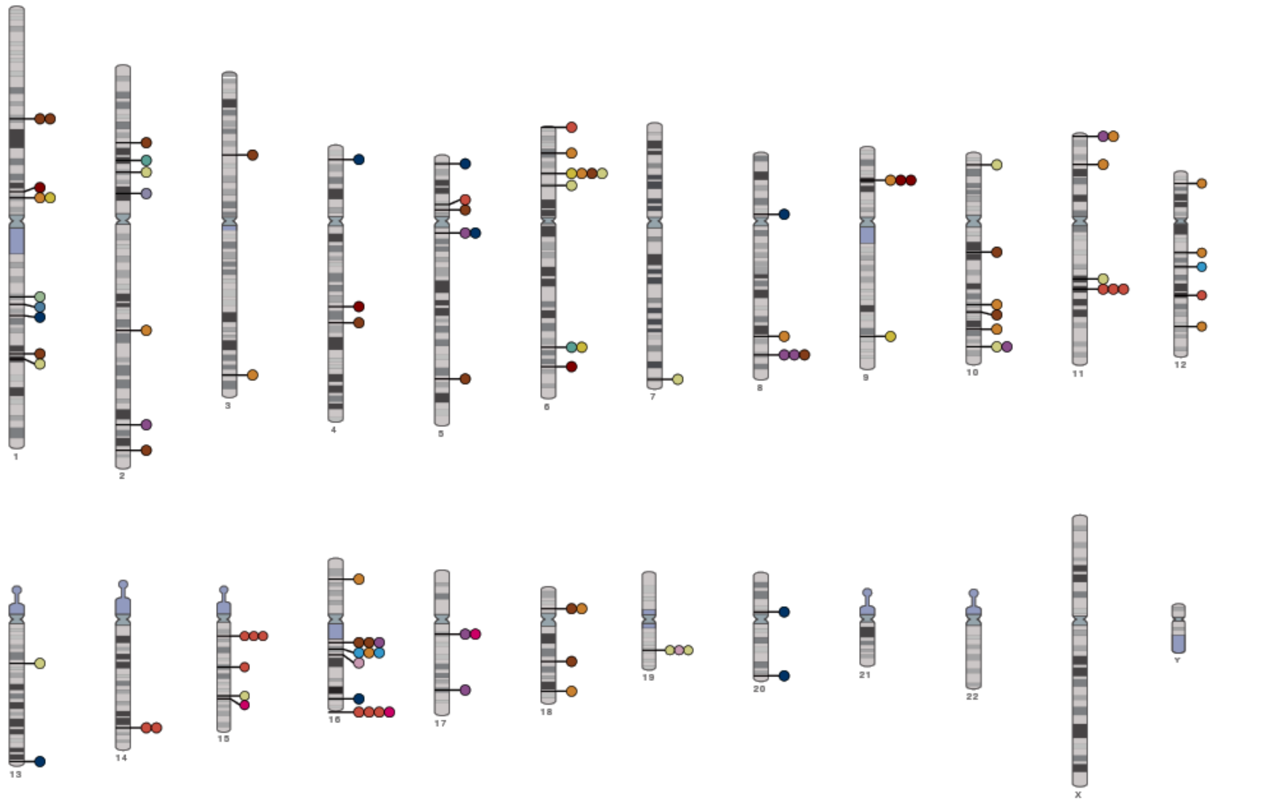
- Ind1 SNP1 AG
- Ind2 SNP1 GG
- …

Genome wide association studies Dec 2006


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
SNP-associated trait categories
● Digestive system disorder
● Cardiovascular disorder
● Metabolic disorder
● Immune system disorder
● Neurological disorder
● Liver enzyme measurement
● Lipid or lipoprotein measurement
● Inflammatory marker measurement
● Hematological measurement
● Body measurement
● Cardiovascular measurement
● Other measurement
● Chemical compound
● Biological process
● Cancer
● Other disease
● Other trait

GWAS Catalog

Genome wide association studies Dec 2007

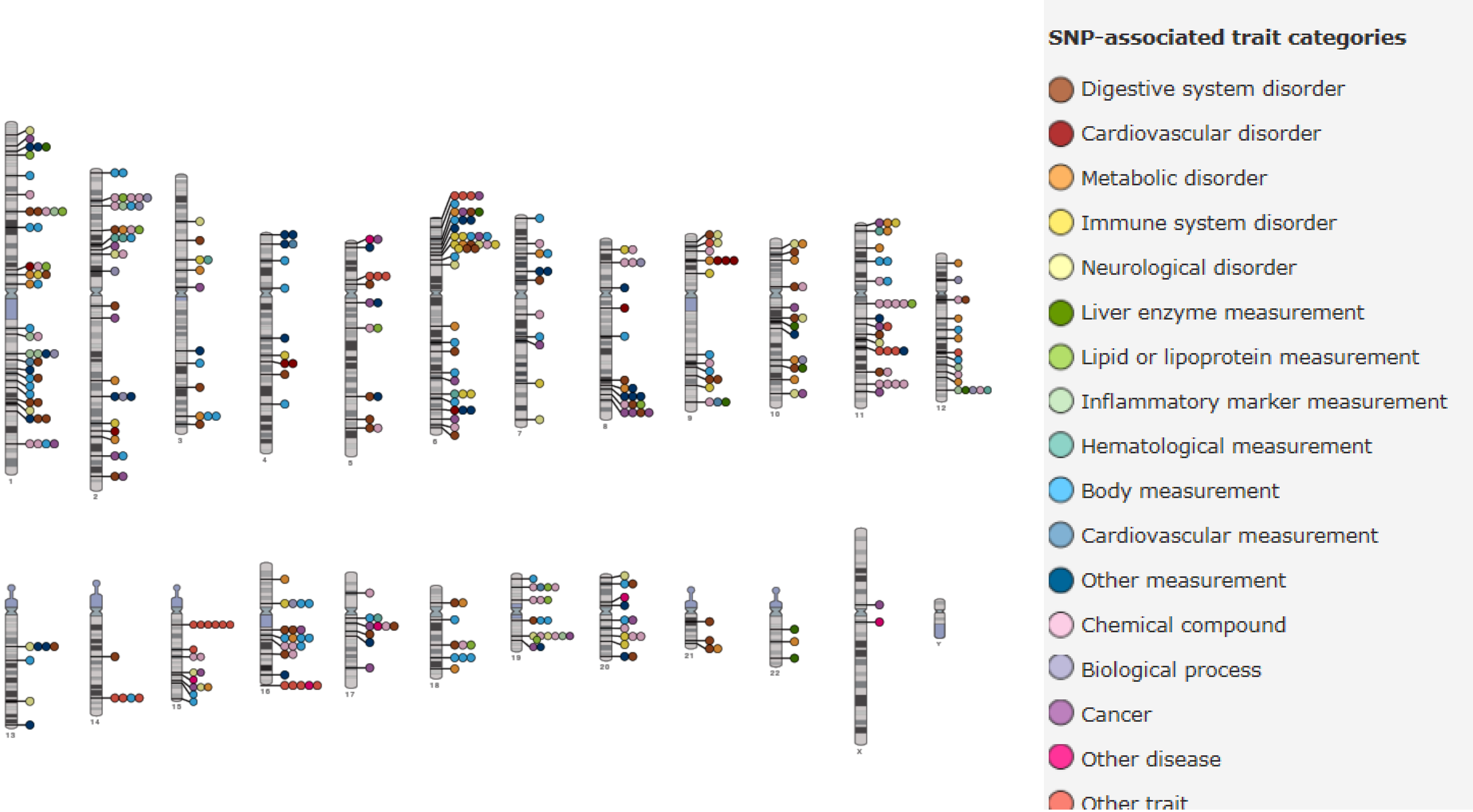


SNP-associated trait categories

- ● Digestive system disorder
- ● Cardiovascular disorder
- ● Metabolic disorder
- ● Immune system disorder
- ● Neurological disorder
- ● Liver enzyme measurement
- ● Lipid or lipoprotein measurement
- ● Inflammatory marker measurement
- ● Hematological measurement
- ● Body measurement
- ● Cardiovascular measurement
- ● Other measurement
- ● Chemical compound
- ● Biological process
- ● Cancer
- ● Other disease
- ● Other trait

GWAS Catalog

Genome wide association studies Dec 2008

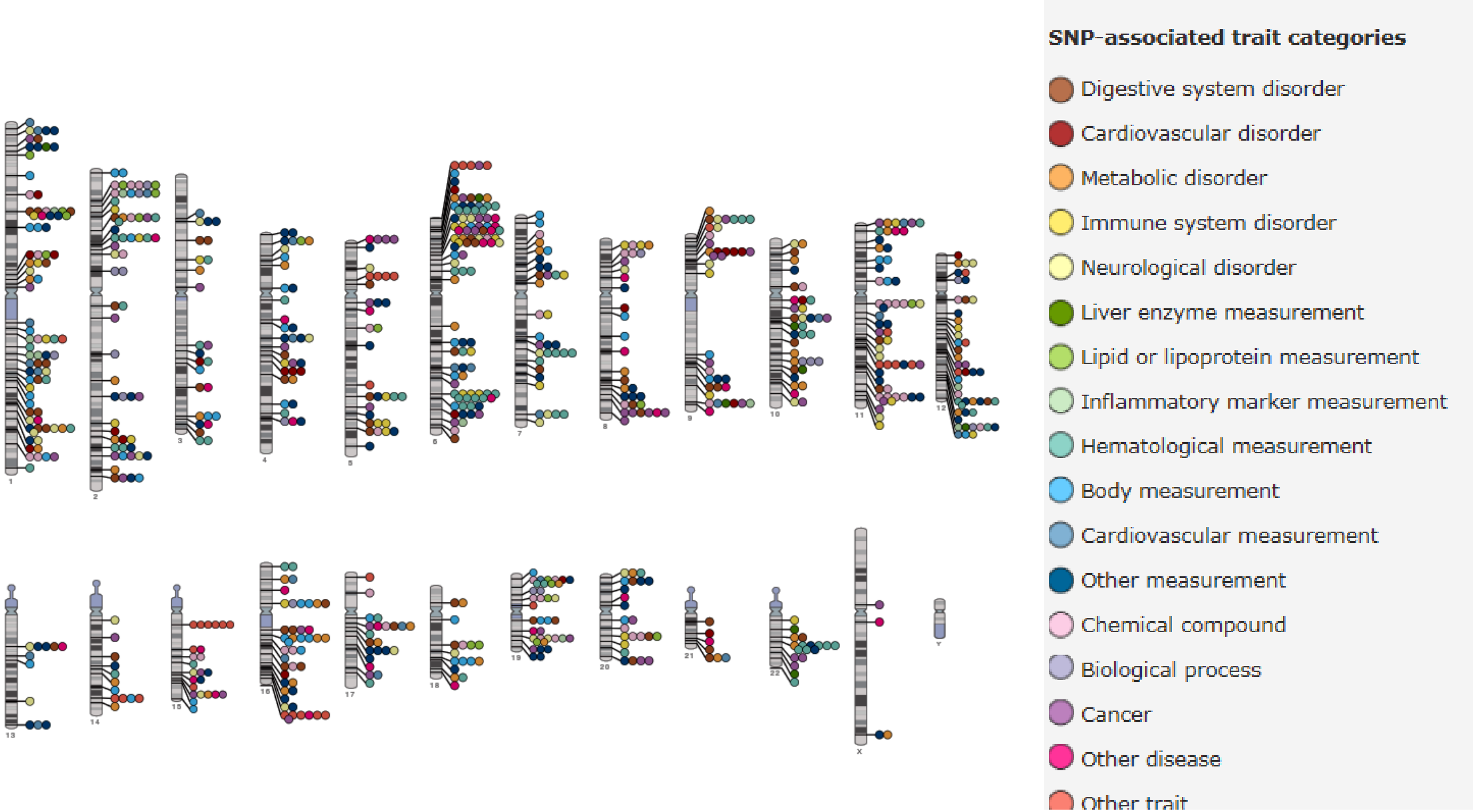


SNP-associated trait categories

- ● Digestive system disorder
- ● Cardiovascular disorder
- ● Metabolic disorder
- ● Immune system disorder
- ● Neurological disorder
- ● Liver enzyme measurement
- ● Lipid or lipoprotein measurement
- ● Inflammatory marker measurement
- ● Hematological measurement
- ● Body measurement
- ● Cardiovascular measurement
- ● Other measurement
- ● Chemical compound
- ● Biological process
- ● Cancer
- ● Other disease
- ● Other trait

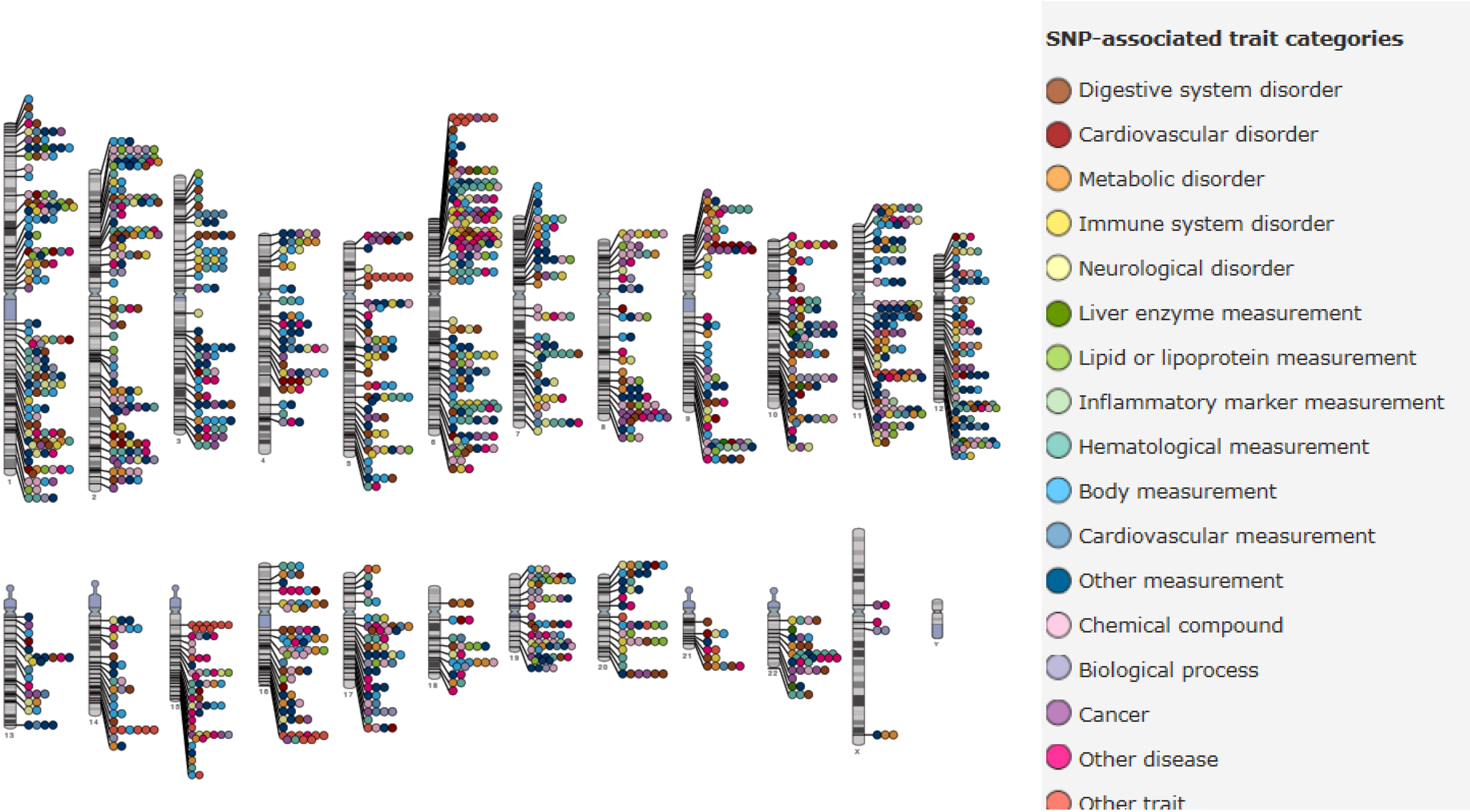
GWAS Catalog

Genome wide association studies Dec 2009



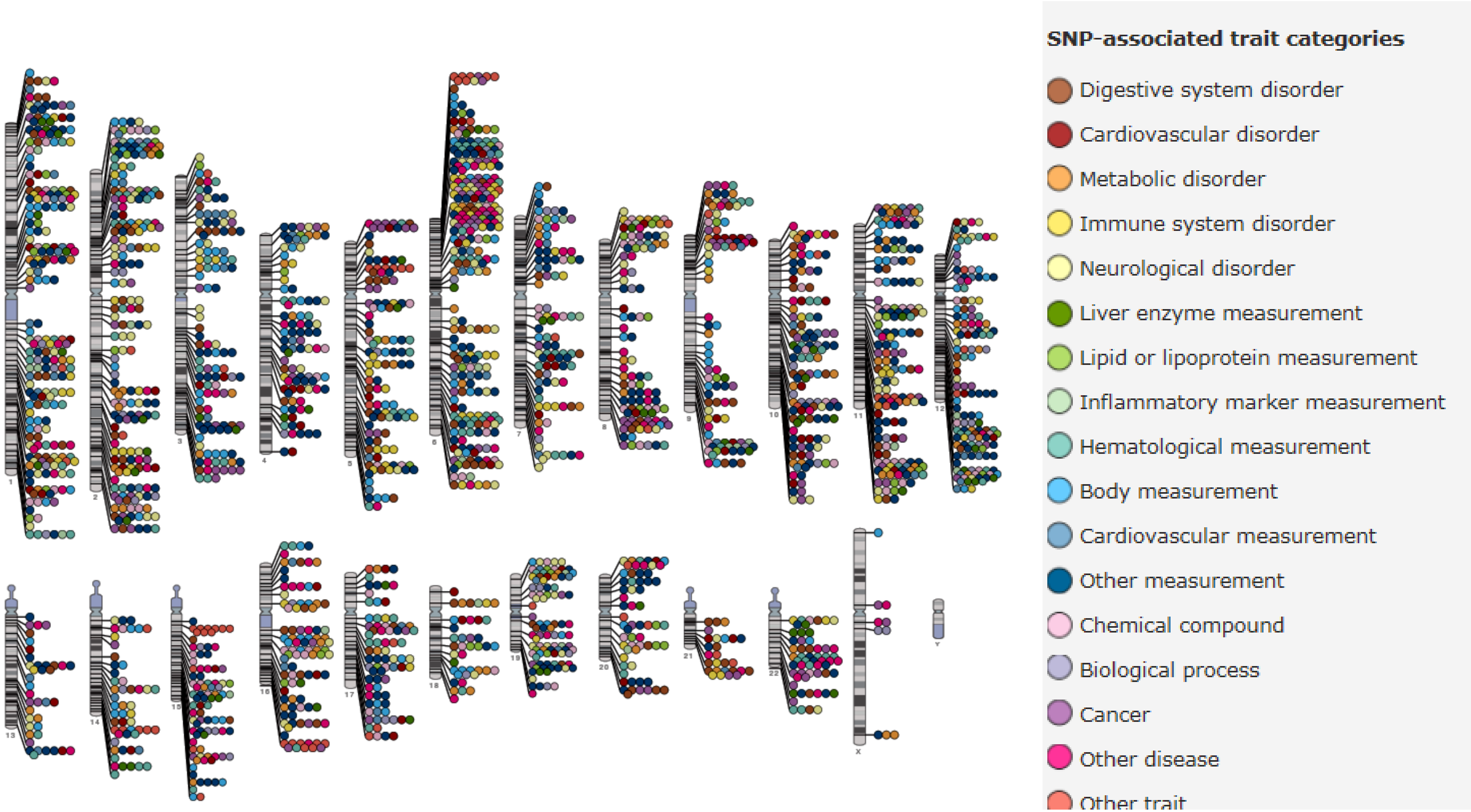
GWAS Catalog

Genome wide association studies Dec 2010



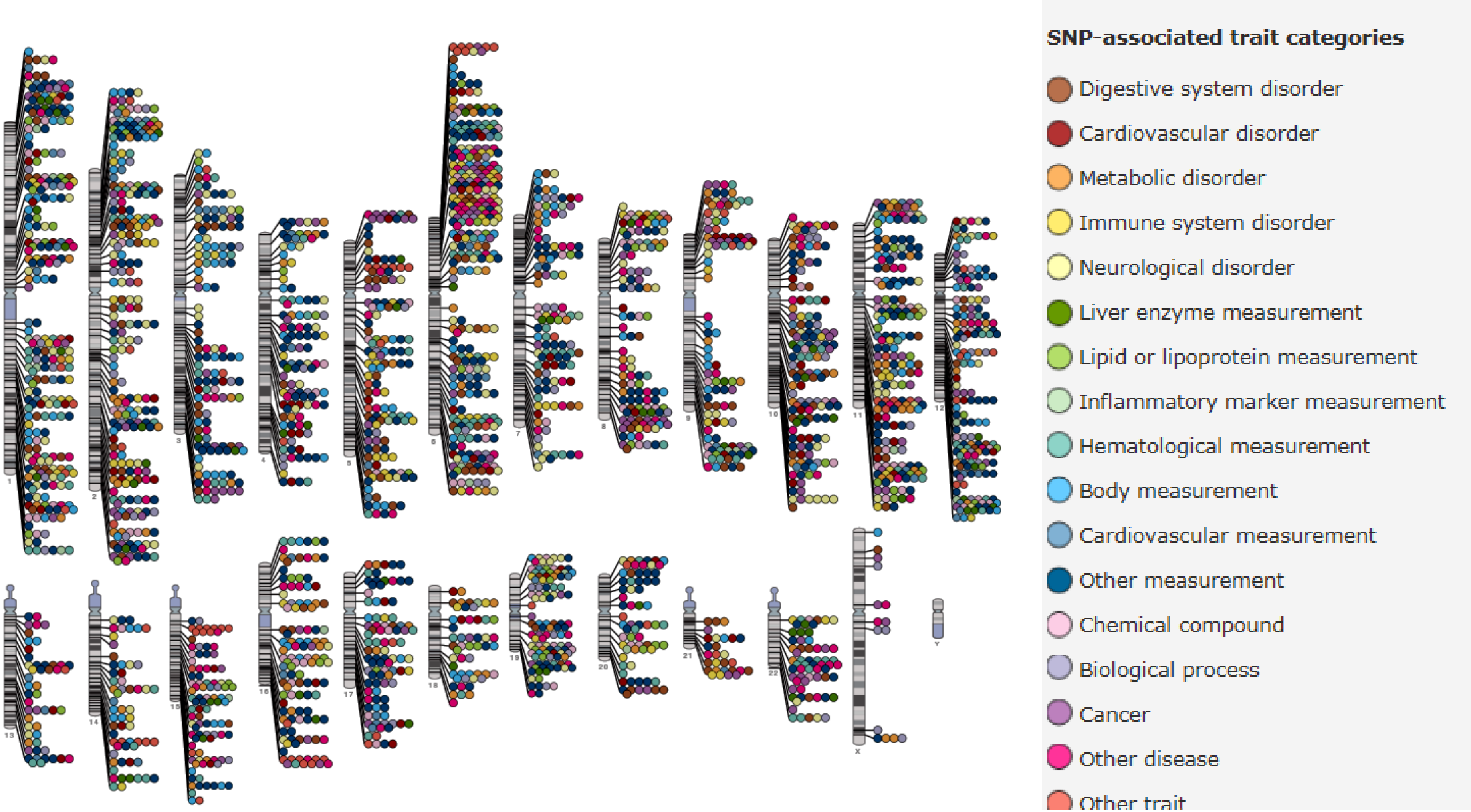
GWAS Catalog

Genome wide association studies Dec 2011



GWAS Catalog

Genome wide association studies Dec 2012



SNP-associated trait categories

- ● Digestive system disorder
- ● Cardiovascular disorder
- ● Metabolic disorder
- ● Immune system disorder
- ● Neurological disorder
- ● Liver enzyme measurement
- ● Lipid or lipoprotein measurement
- ● Inflammatory marker measurement
- ● Hematological measurement
- ● Body measurement
- ● Cardiovascular measurement
- ● Other measurement
- ● Chemical compound
- ● Biological process
- ● Cancer
- ● Other disease
- ● Other trait

GWAS Catalog

Genome wide association studies Nov 2020