umcg

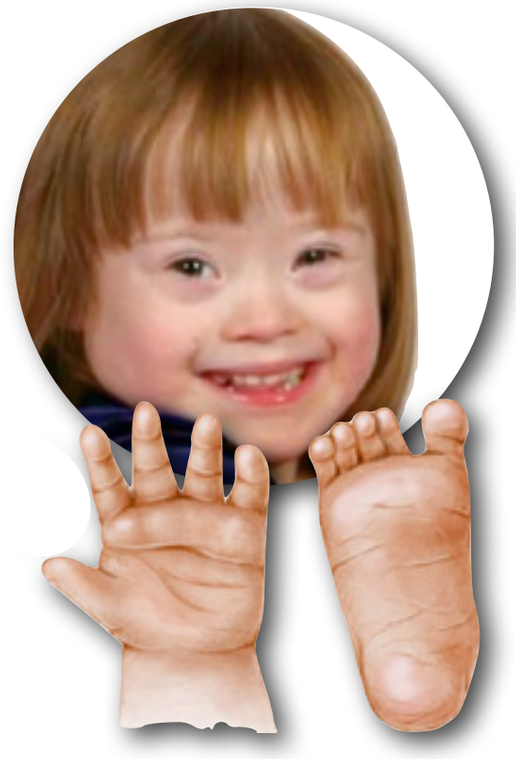
Lude Franke >

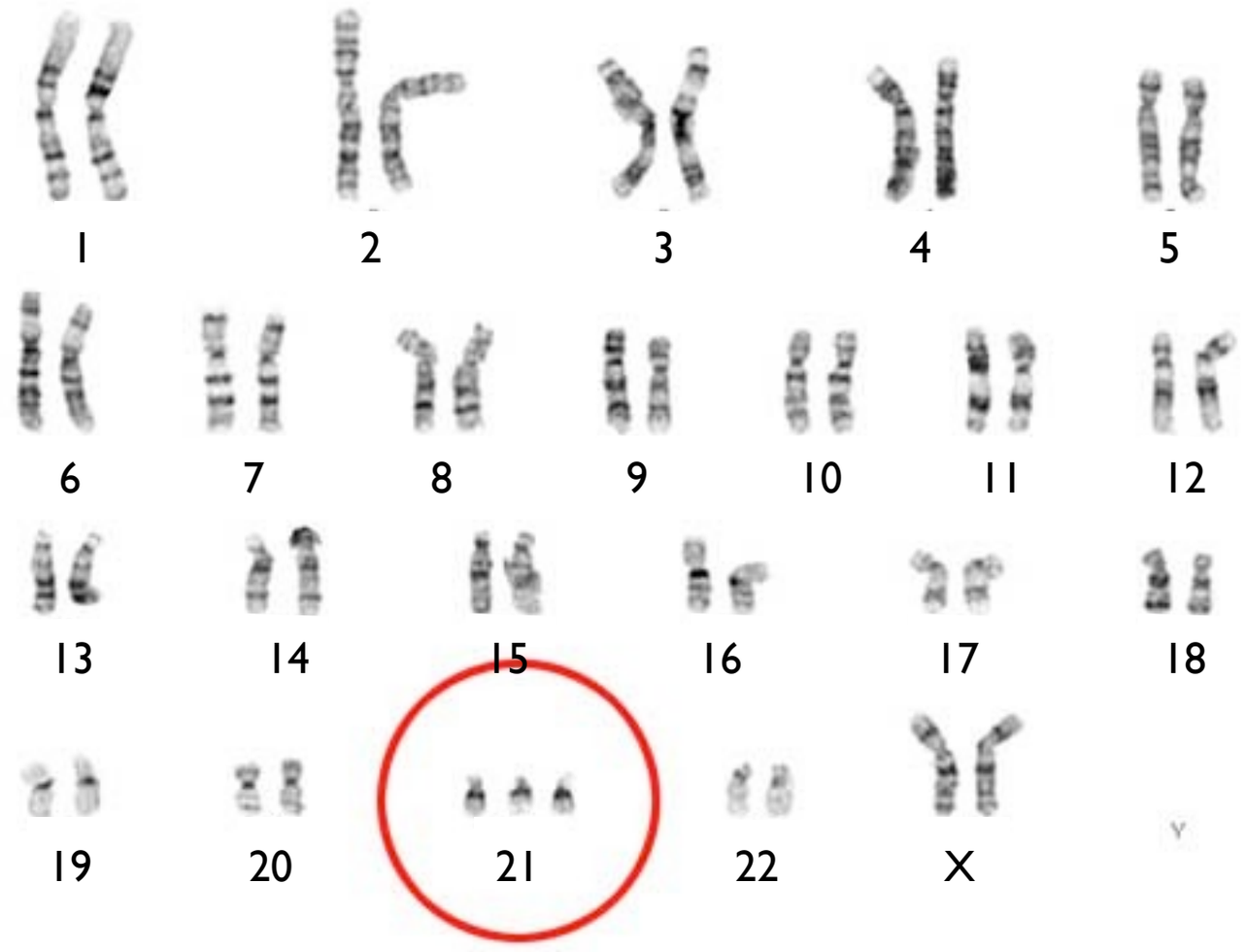
DNA microarrays: Genome-wide association studies

Department of Genetics, UMC Groningen, the Netherlands

Down's syndrome:

- First described in 1866 by John Langdon Down
  - Prevalence: 1 / 700
  - Impairment in cognition & physical growth
  - Particular set of (facial) characteristics
- Cause found in 1959: trisomy chromosome 21





Down's syndrome:

Why do Down's syndrome children get sick?

They do not miss any genetic material!

Can it be that an extra chromosome 21 results in increased gene expression of chromosome 21 genes?

Could that cause Down's syndrome?

But what about chromosome X?

- Males: one copy chromosome X
- Females: two copies chromosome X, chromosome X inactivation

Other diseases due to chromosomal aberrations:

Klinefelter syndrome	Normal chr. 1 - 22, but XXY
Turner syndrome	Chromosome X missing (X0)
Patau syndrome	Trisomy chromosome 13
Edwards syndrome	Trisomy chromosome 18

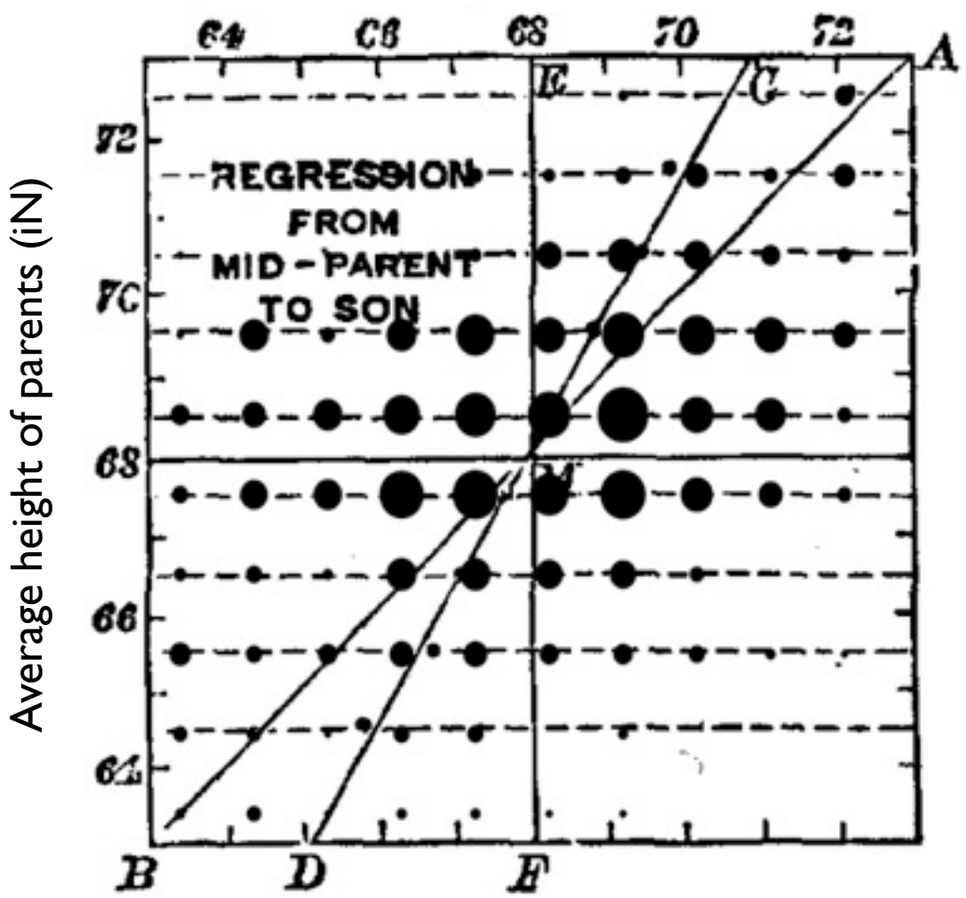
As such severe genetic mutations (here duplications or deletions of entire chromosomes) can cause disease

But what about all those other disease? Is every disease genetic? How can we determine that?

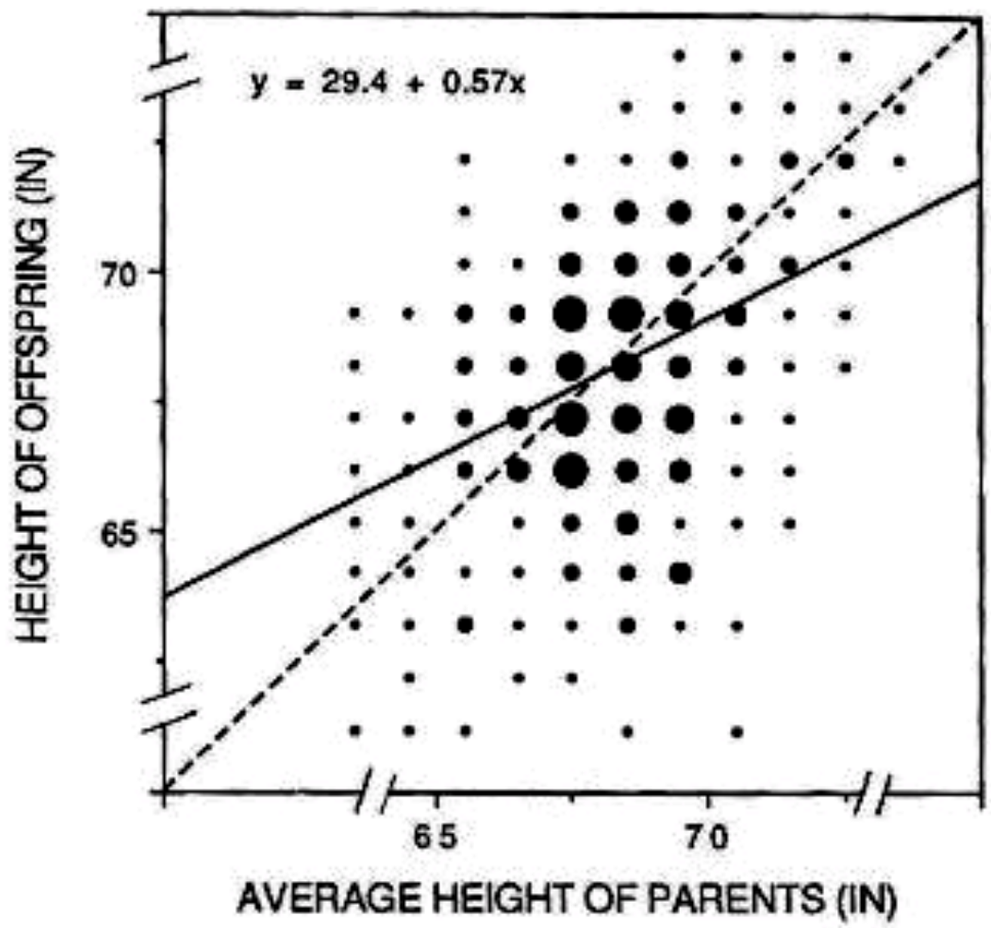
Human adult height:

Sir Francis Galton showed in 1889 a clear relationship between average height of parents and the height of offspring:

Original 1889 figure



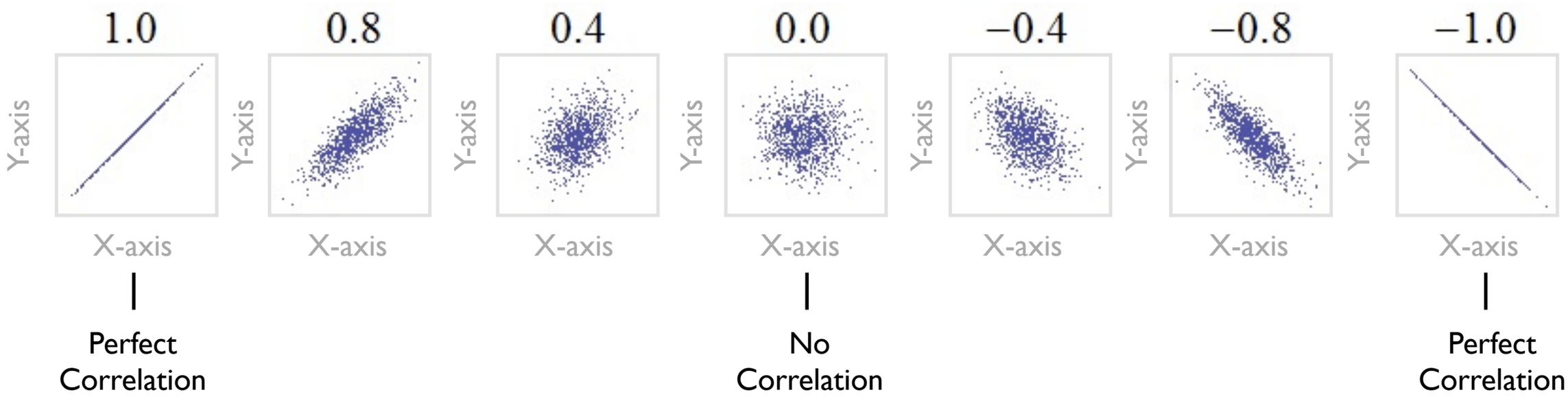
Cleaned-up figure



Human adult height:

- Strongly genetically determined, but to what extent?
- Can be calculated by a Pearson correlation coefficient
  (invented by Francis Galton and his student Karl Pearson)

Pearson correlation coefficient

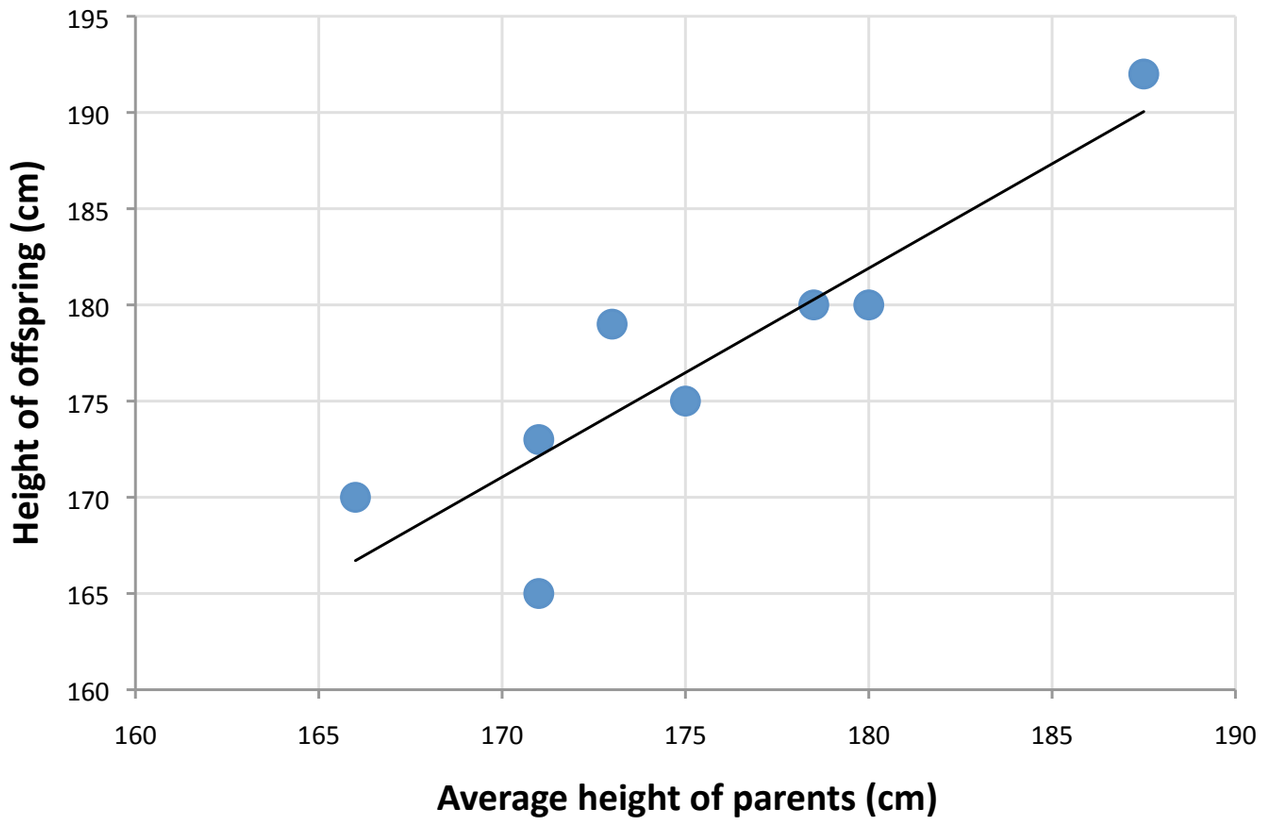


http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

Human adult height: example

Lude°Franke

Height father	Height mother	Average height parents	Height offspring
180	170	175	175
175	167	171	165
192	165	179	180
202	173	188	192
169	163	166	170
173	169	171	173
182	164	173	179
174	186	180	180



Pearson correlation coefficient: r = 0.89
Proportion explained variance: **r^2 = 0.79**

Based on this data [illegible] for 79% genetically determined

Real estimates:
78-80% in white men
75-80% in white women

Human adult height: strongly heritable

Lude°Franke

Adult height is strongly heritable, so genetic variation must play an important role.

Can we find the genetic variants that influence adult height?

Types of genetic disorders

Lude°Franke

Mendelian: one single mutation causes the disease.

The inheritance can be e.g. recessive or dominant

On / Off Switch

Complex: multiple genetic variants can each increase or decrease disease risk

So you will have a few unfortunate variants and a few fortunate variants. If you have many unfortunate variants, you will get ill.

The environment also plays an important role (smoking, exercise, diet, etcetera)

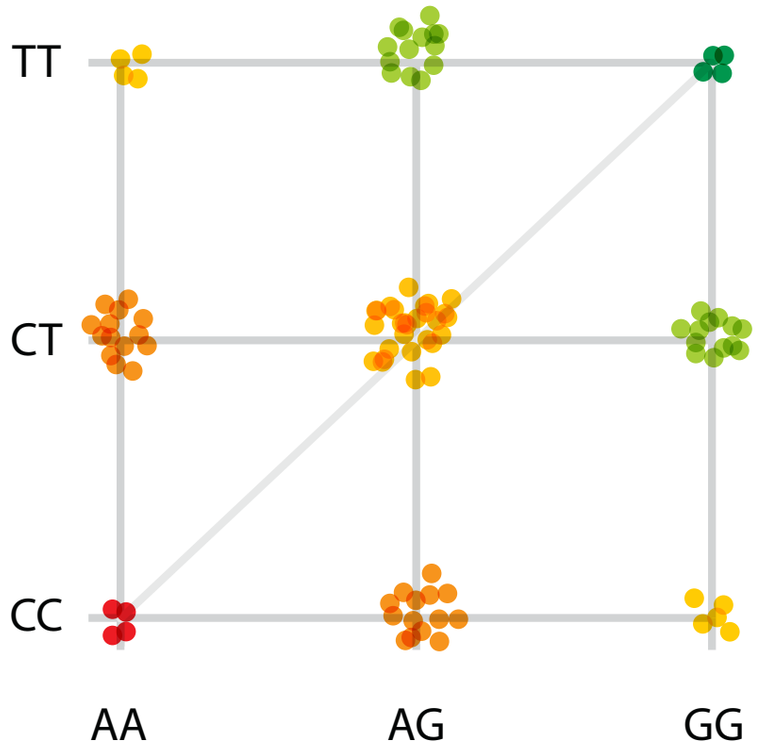
Classes of genetic variants:

		Number known in dbSNP / DGV
Single nucleotide variant	ATTGGCCTTAACCCCCGATTATCAGGAT ATTGGCCTTAACCTCCGATTATCAGGAT	>6,000,000
Insertion–deletion variant	ATTGGCCTTAACCCGATCCGATTATCAGGAT ATTGGCCTTAACCC---CCGATTATCAGGAT	>100,000
Block substitution	ATTGGCCTTAACCCCCGATTATCAGGAT ATTGGCCTTAACAGTGGATTATCAGGAT	?
Inversion variant	ATTGGCCTTAACCCCCGATTATCAGGAT ATTGGCCTTCGGGGGTTATTATCAGGAT	?
Copy number variant	ATTGGCCTTAGGCCTTAACCCCCGATTATCAGGAT ATTGGCCTTA-------ACCTCCGATTATCAGGAT	>50,000

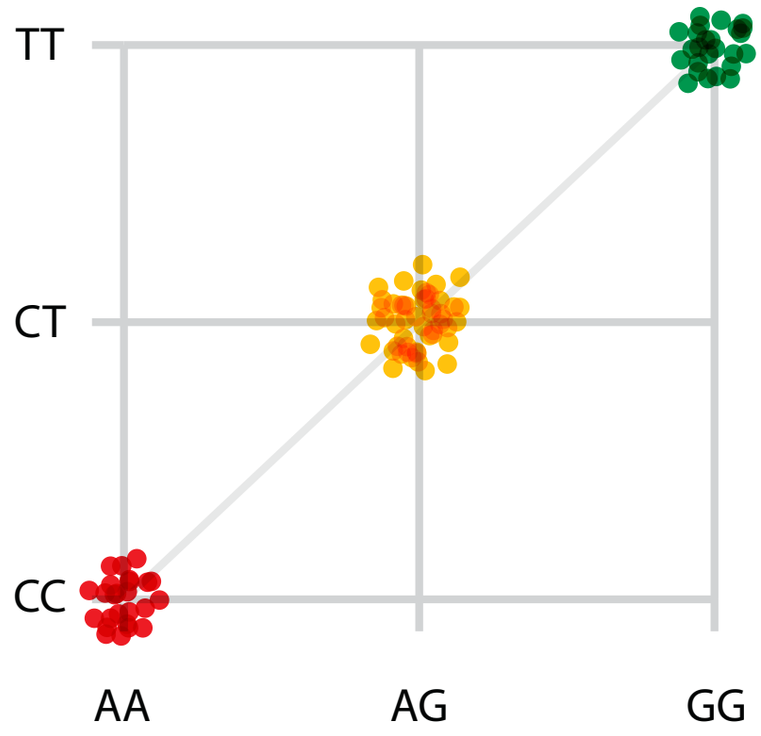
Structural variants

Do we need to investigate all these variants?

SNP X and Y uncorrelated:
linkage equilibrium



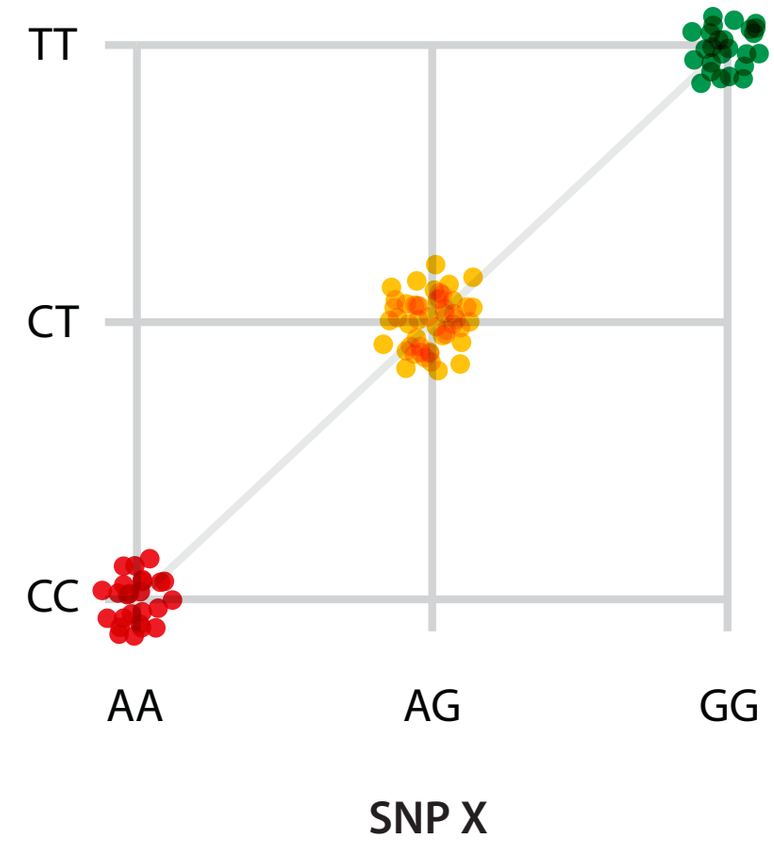
SNP X and Y correlated:
linkage disequilibrium



SNP X and Y correlated:
linkage disequilibrium

By genotyping SNP X,
genotypes of SNP Y
can be predicted

Not necessary to investigate all SNPs!



Availability of genome-wide arrays in 2006

Lude°Franke

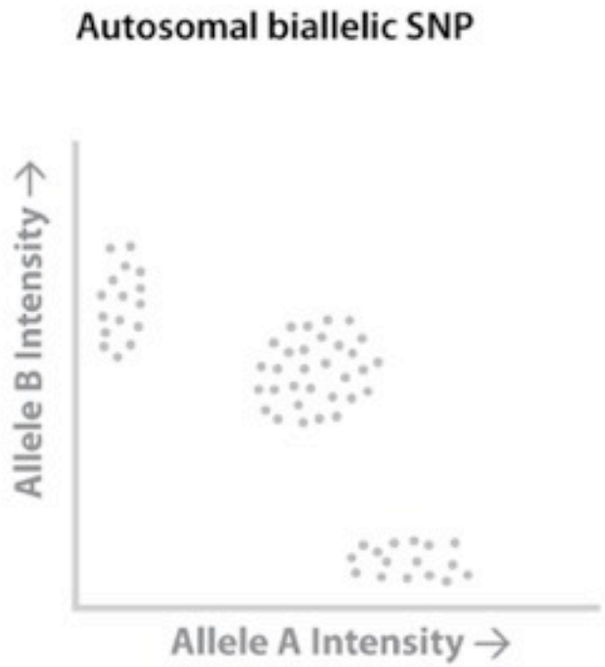
Available technique:

Oligonucleotide arrays ('DNA-chip')
(e.g. Illumina or Affymetrix)

Contain > 300,000 SNPs that have been designed to capture most of the known genetic variation well.

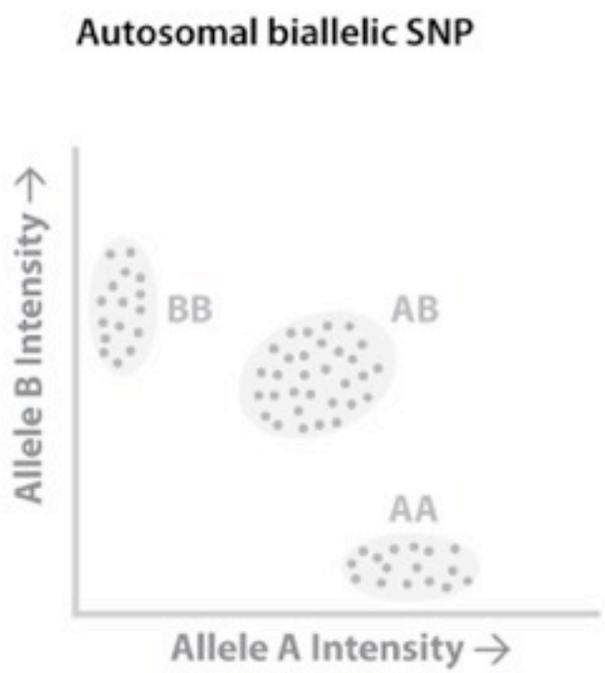
How oligonucleotide arrays call genotypes

Lude°Franke



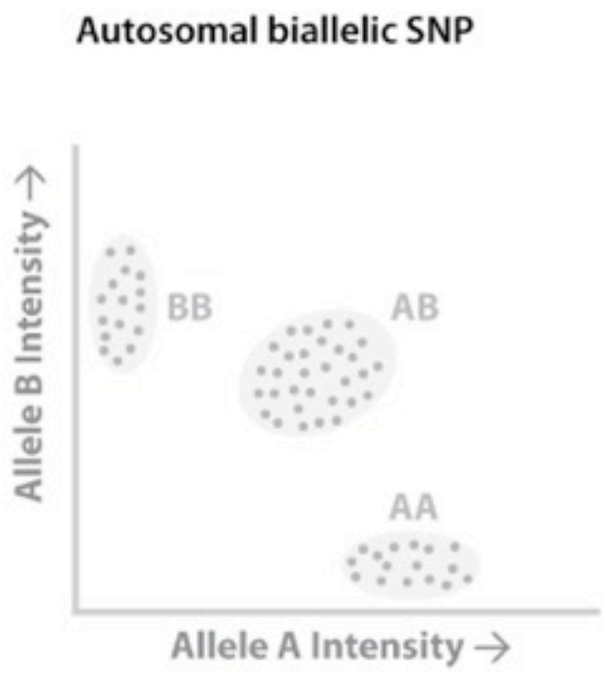
How oligonucleotide arrays call genotypes

Lude°Franke



How oligonucleotide arrays call genotypes

Lude°Franke



Genotype	AA	AB	BB
Frequency	25%	50%	25%
Allele Frequency	A: (2AA + AB) / 2 = 50%		B: (AB + 2BB) / 2 = 50%
Genotype Expected	AA: A^2 = 25%	AB: 2AB = 50%	BB: B^2 = 25%

nature | October 2010

doi:10.1038/nature09410

Hundreds of variants clustered in genomic loci and biological pathways affect human height

A full list of authors and their affiliations appears at the end of the paper.

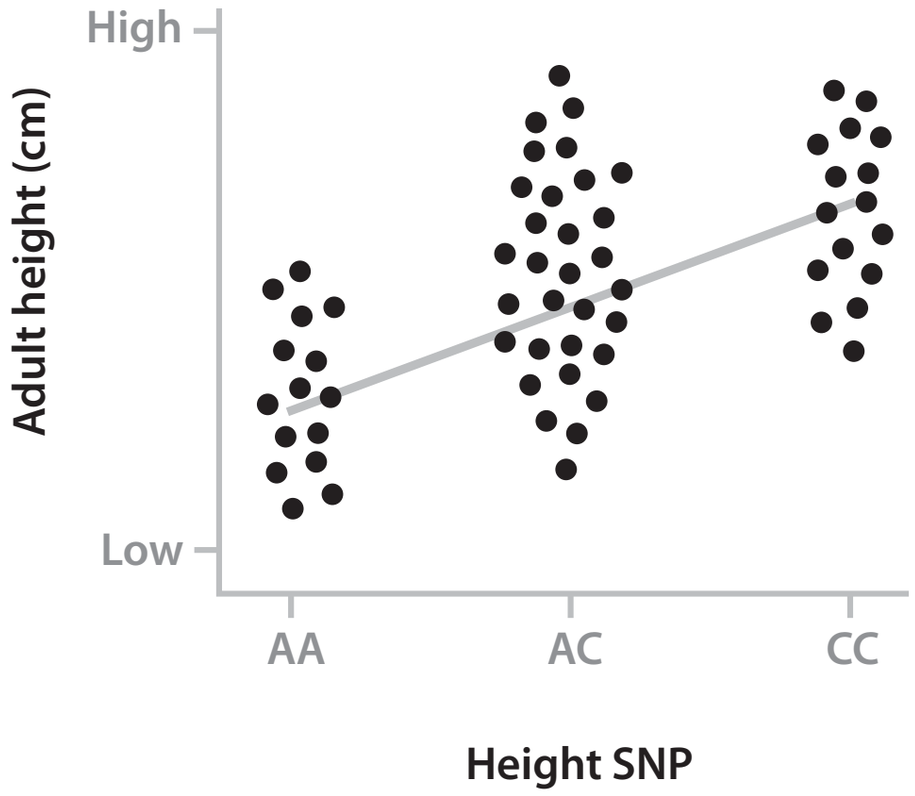
180 SNPs affect adult human height!

Analysis of 180,000 people

Lude°Franke

However, each of the 180 SNPs have a very small effect on height:

Each of the SNPs explain only between ~0.3% and ~0.5% of height variation

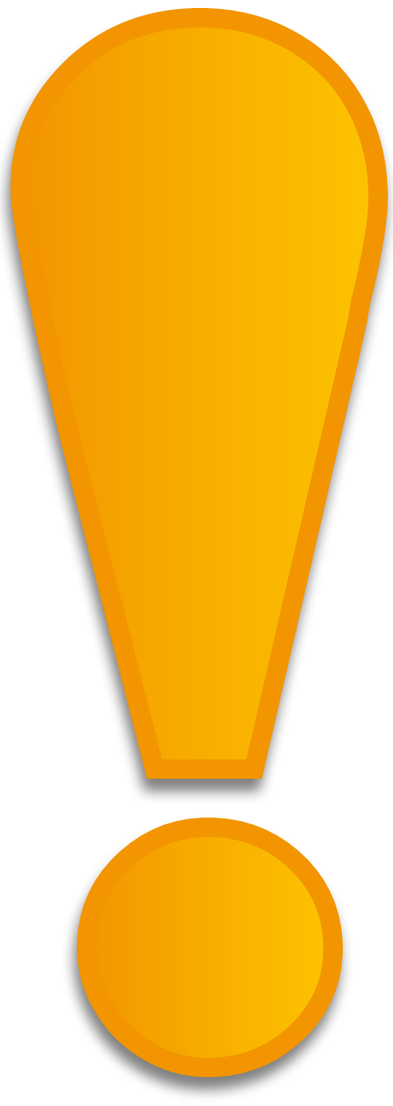


But severe mutations in the genes where these SNPs are located sometimes give dramatic effects:

SNP variant	Gene where SNP is located	Mutations in this gene cause
rs1042725	HMGA2	Pygmy mice
rs6060373	GDF5	Chondrodysplasia (abnormally short and deformed limbs); brachydactyly (short digits) DuPan syndrome; multiple synostoses syndrome.

180 variants explain hardly anything!

Lude°Franke



Adult height is strongly heritable, SNP variants in180 SNP variants have been found.

But together they explain only 10.5% of the variation in height!

What is going on? We do not understand yet!

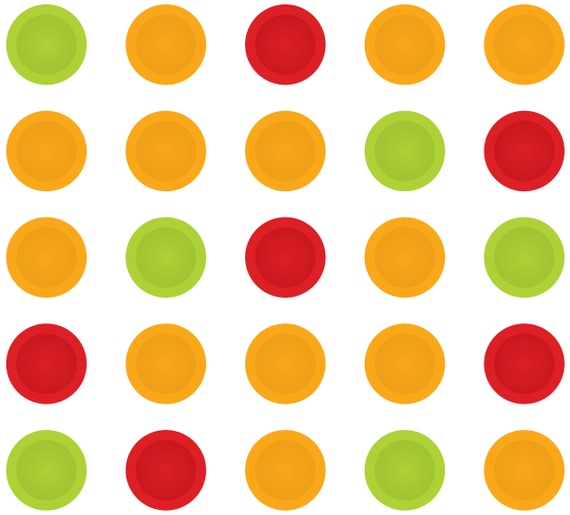
SNP Genotype

AA

AC

CC

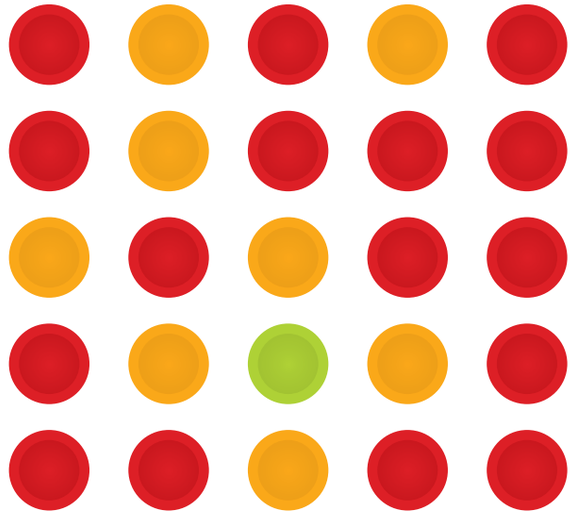
Healthy people



AA: 6 people (24%)
AC: 13 people (52%)
CC: 6 people (24%)

A allele freq: 50%
C allele freq: 50%

Sick people



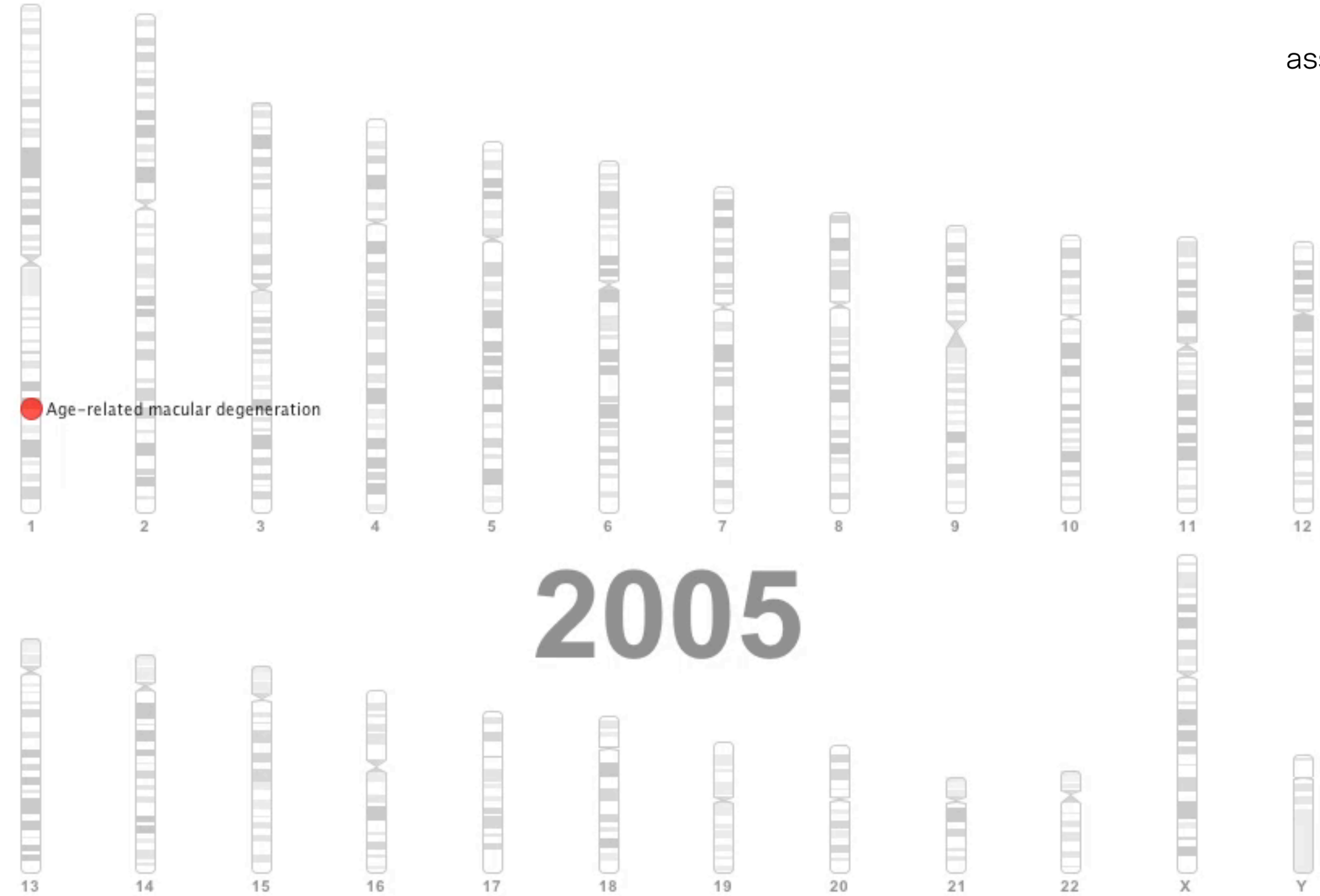
AA: 1 person (4%)
AC: 8 people (32%)
CC: 16 people (64%)

A allele freq: 20%
C allele freq: 80%

Seven years of GWAS studies

Gene atlas

6,054
disease
associations



Teri Manolio *et al*: A catalog of published genome-wide association studies

Celiac disease: common (1% prevalence) small intestinal inflammatory condition induced by dietary wheat, rye and barley.

June 2007 → March 2008

nature genetics

A genome-wide association study for celiac disease identifies risk variants in the region harboring *IL2* and *IL21*

David A van Heel[1], Lude Franke[2,17], Karen A Hunt[1,17], Rhian Gwilliam[3,17], Alexandra Zhernakova[2], Mike Inouye[3], Martin C Wapenaar[4], Martin C N M Barnardo[5], Graeme Bethel[3], Geoffrey K T Holmes[6], Con Feighery[7], Derek Jewell[8], Dermot Kelleher[7], Parveen Kumar[1], Simon Travis[9], Julian RF Walters[10], David S Sanders[11], Peter Howdle[12], Jill Swift[13], Raymond J Playford[1], William M McLaren[3], M Luisa Mearin[14,15], Chris J Mulder[16], Ross McManus[7], Ralph McGinnis[3], Lon R Cardon[8], Panos Deloukas[3] & Cisca Wijmenga[2,4]

nature genetics

Newly identified genetic risk variants for celiac disease related to the immune response

Karen A Hunt[1], Alexandra Zhernakova[2], Graham Turner[3], Graham A R Heap[1], Lude Franke[2], Marcel Bruinenberg[4], Jihane Romanos[4], Lotte C Dinesen[5], Anthony W Ryan[3], Davinder Panesar[1], Rhian Gwilliam[6], Fumihiko Takeuchi[6], William M McLaren[6], Geoffrey K T Holmes[7], Peter D Howdle[8], Julian R F Walters[9], David S Sanders[10], Raymond J Playford[1], Gosia Trynka[4], Chris J J Mulder[11], M Luisa Mearin[12,13], Wieke H M Verbeek[11], Valerie Trimble[3], Fiona M Stevens[14], Colm O'Morain[3], Nicholas P Kennedy[3], Dermot Kelleher[3], Daniel J Pennington[1], David P Strachan[15], Wendy L McArdle[16], Charles A Mein[17], Martin C Wapenaar[4], Panos Deloukas[6], Ralph McGinnis[6], Ross McManus[3,18], Cisca Wijmenga[2,4,18] & David A van Heel[1,18]

2010 nature genetics → 2011 nature genetics

Celiac disease as example

Lude°Franke

The start:
778 unrelated UK celiac patients
1,422 unrelated UK controls

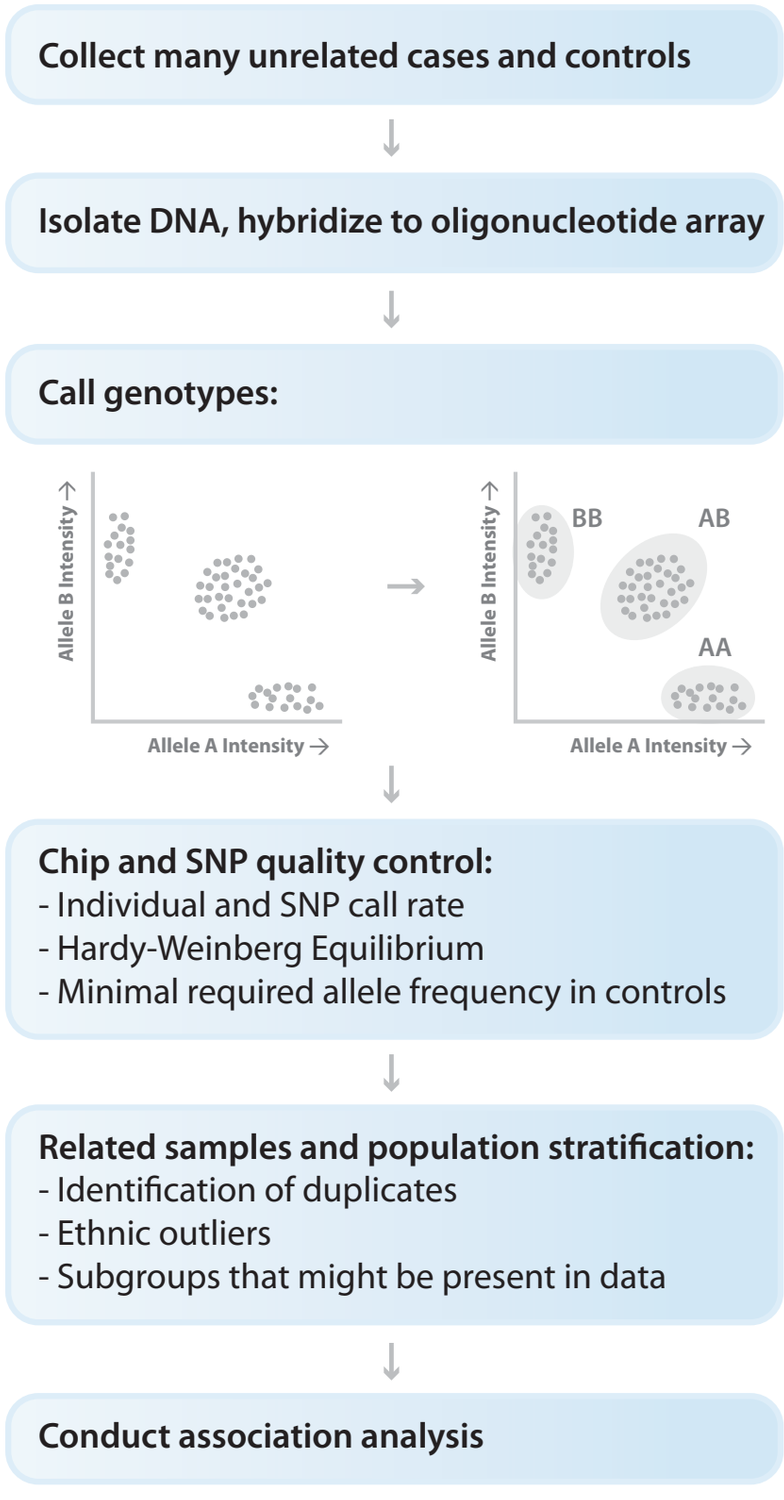
Genotyping: Illumina Hap300
Infinium oligonucleotide array

Calling of genotypes:
In-house developed calling algorithm (IllumiTyper)

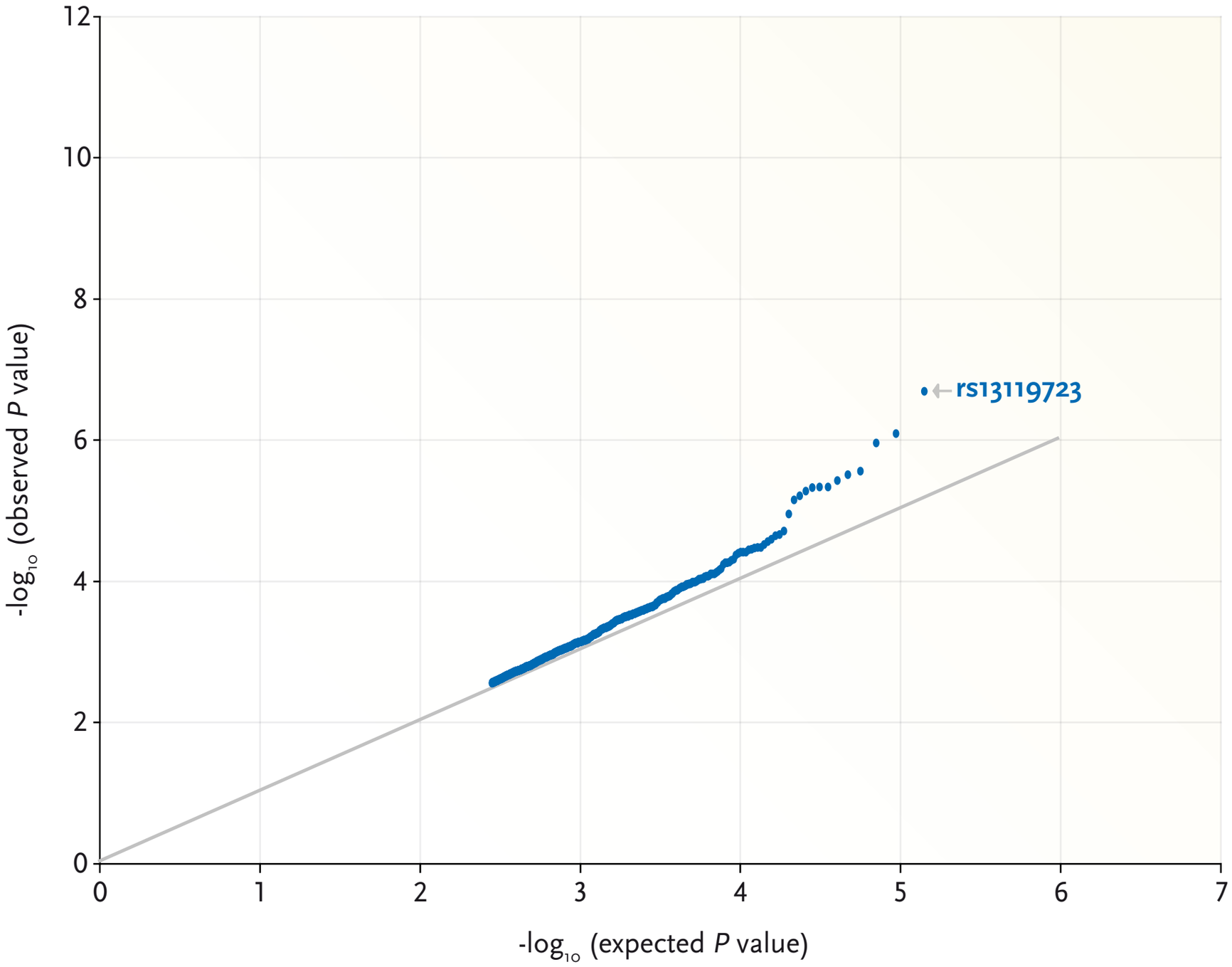
Chip and SNP quality control:
- Individual and SNP call rate > 95%
- HWE Exact P-Value > 0.001
- MAF > 0.05

Related samples:
- Duplicates removed
- Stratification assessed using Eigenstrat

Association analysis: Allele frequency
P-Value test (chi^2 test, 1 df)

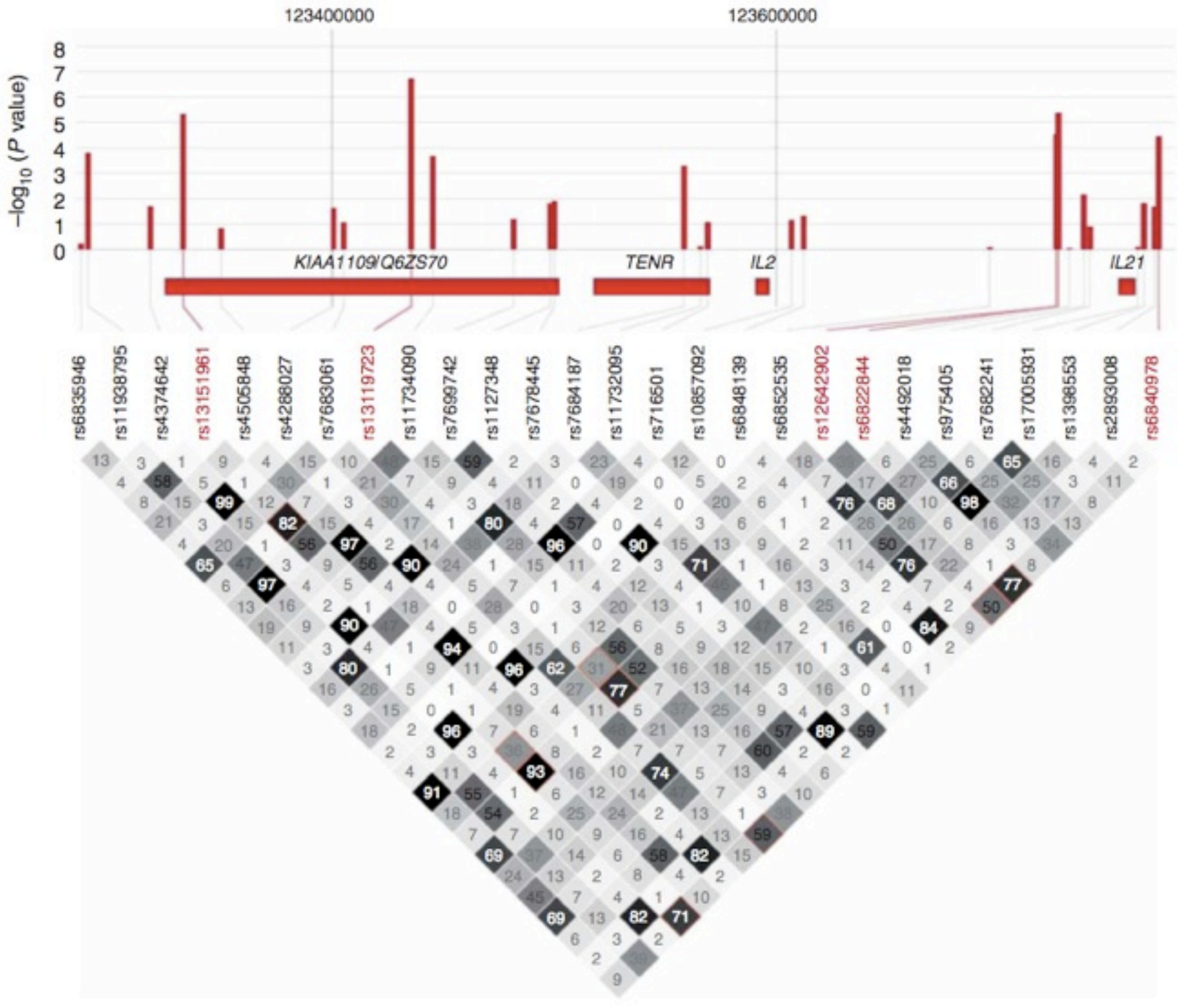


Quantile-quantile plot:



Celiac disease as example

4q27 locus:

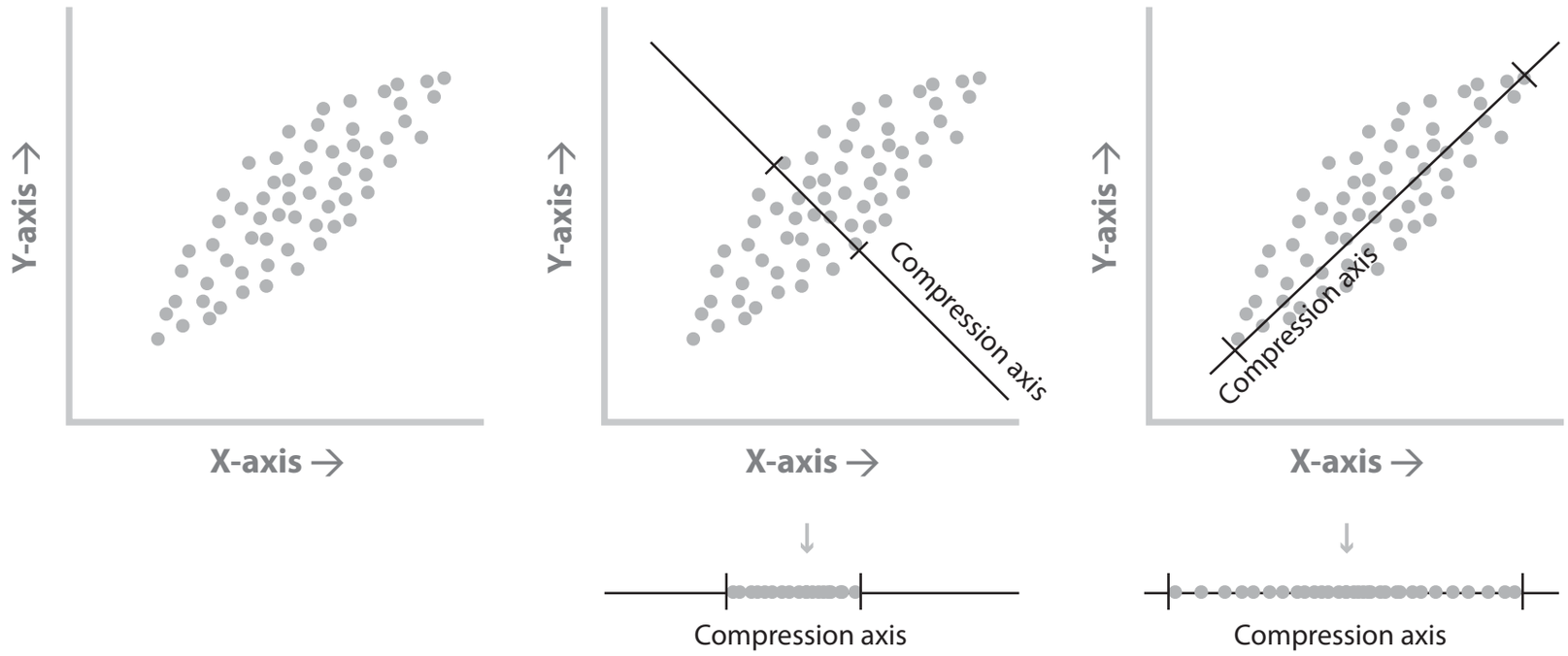


Population stratification

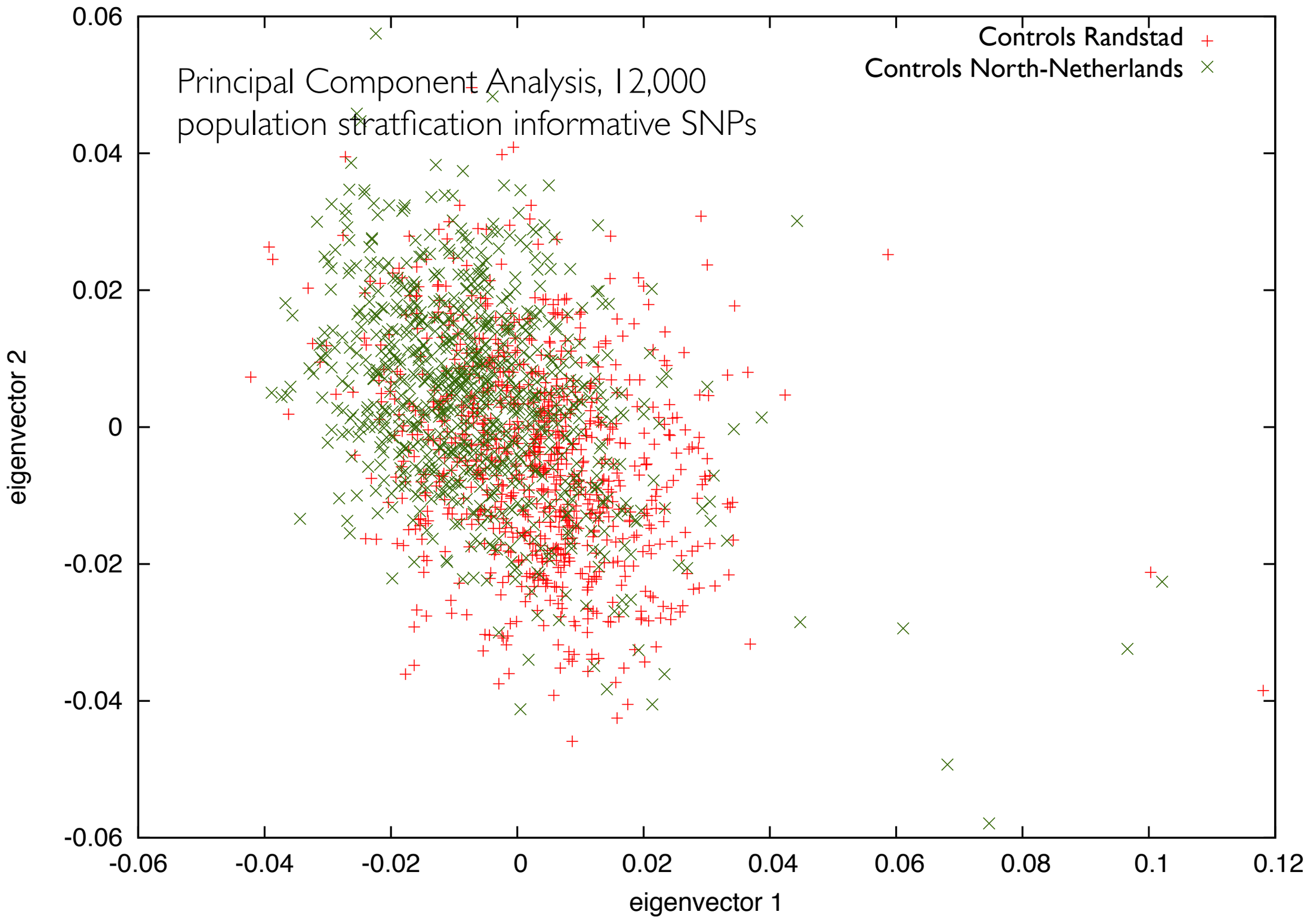
Lude°Franke

- St. Pietersberg
- Groningen, Amsterdam, Utrecht, Rotterdam
- Schiphol

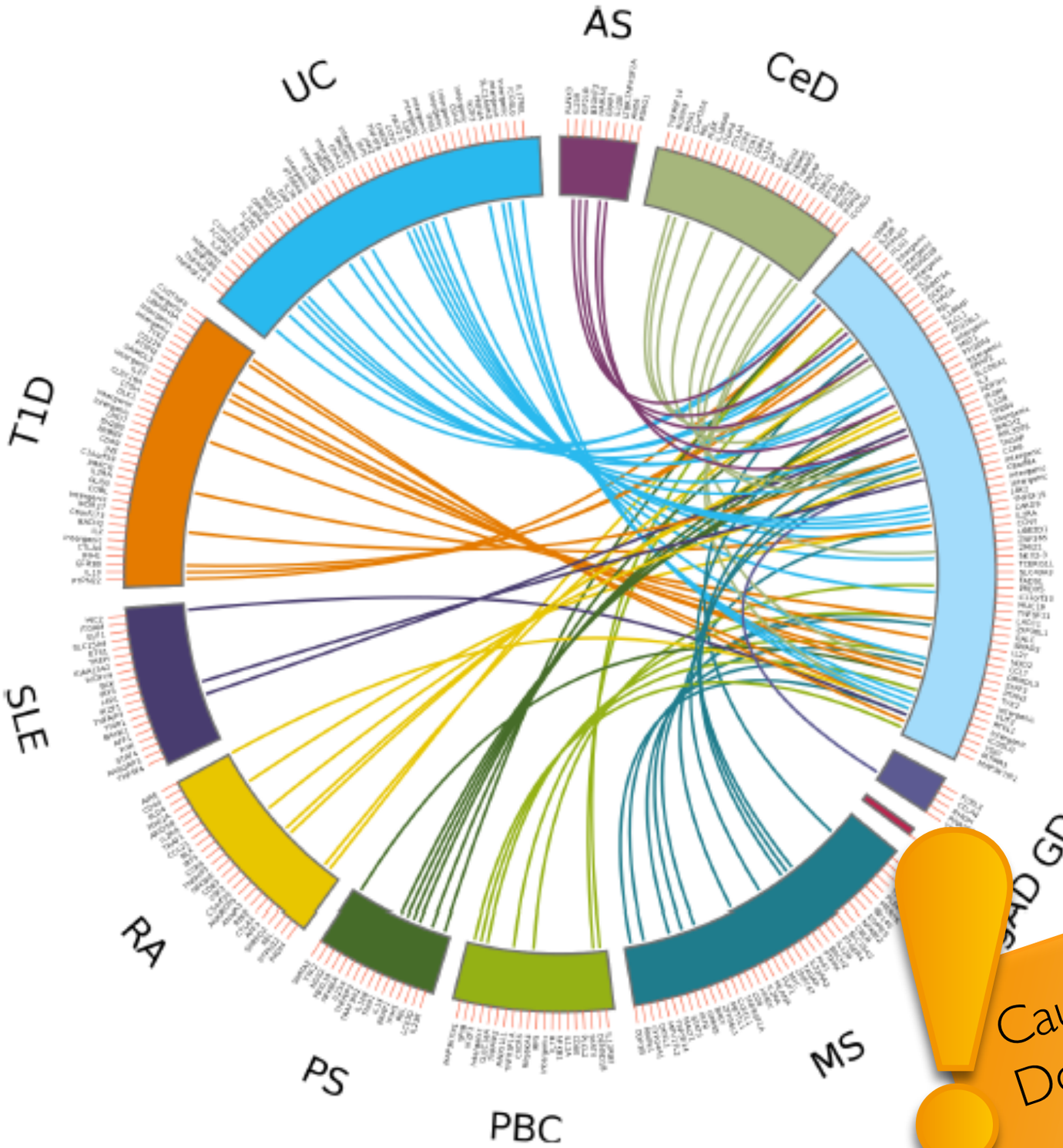
- Groningen
- Zwolle
- Utrecht
- Eindhoven
- Maastricht



Population stratification



Large overlap between immune related diseases



Crohn's Disease

Causal genes often unknown
Downstream effects mostly unclear

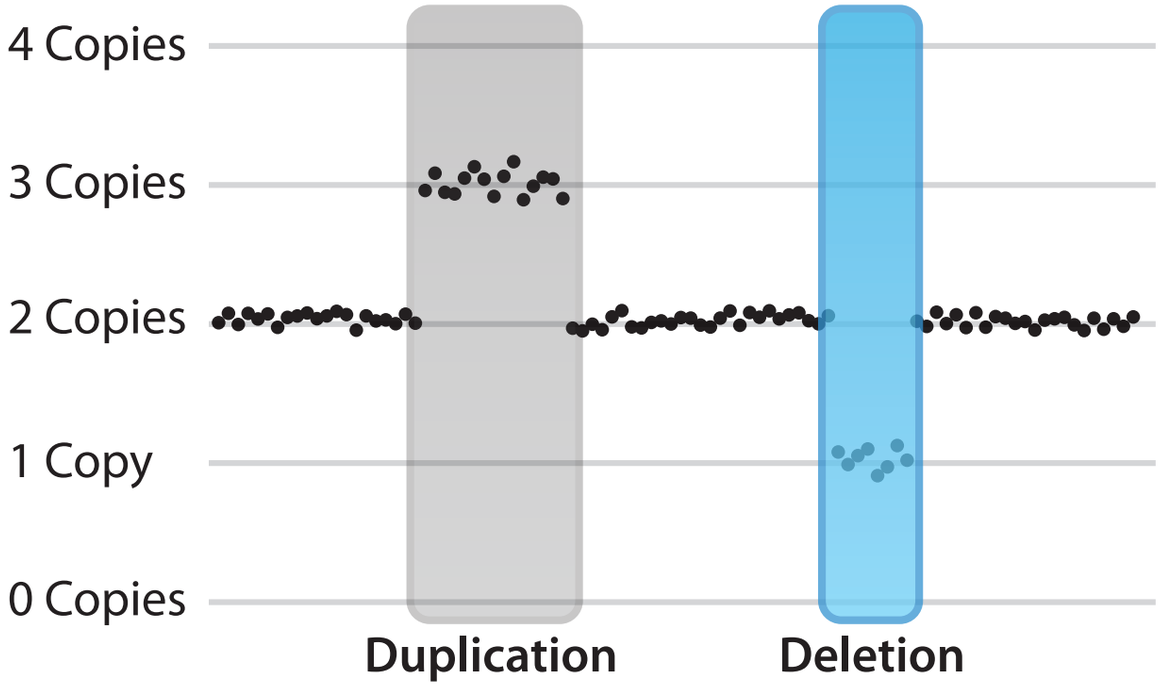
Ponce and Wijmenga, Annual Reviews of Genomics and Human Genetics, *in press*

Copy number variants

Lude°Franke

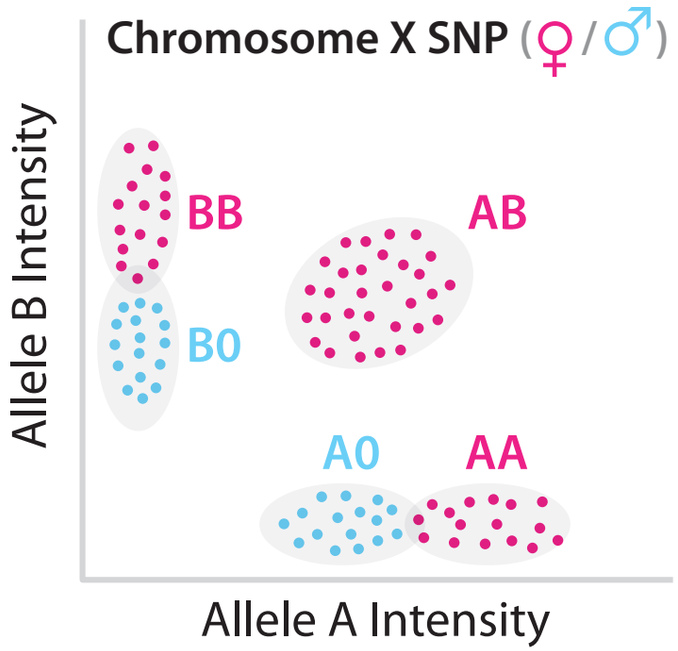
CGH / Oligonucleotide Arrays

Investigate intensity of probes



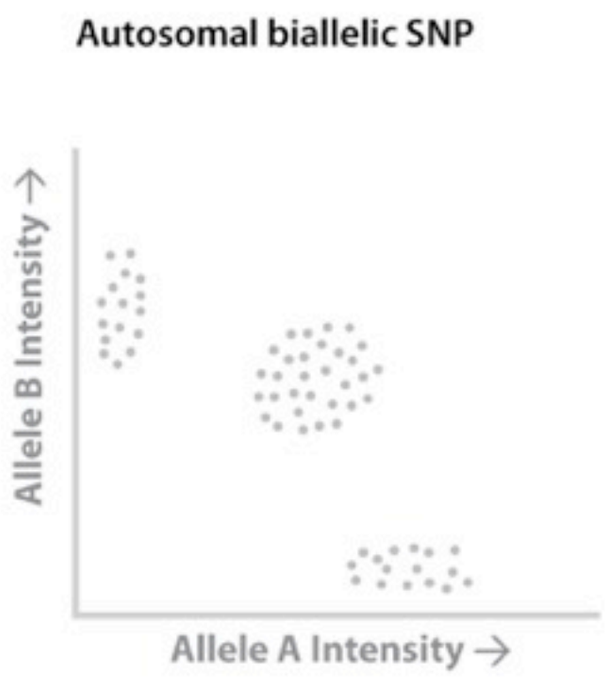
Oligonucleotide Arrays:

Take advantage of genotypes



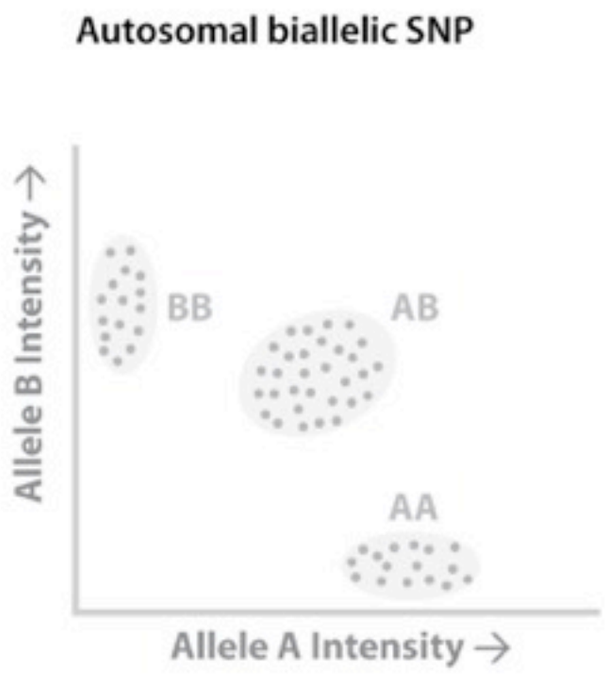
Identification of triallelic SNPs

Lude°Franke



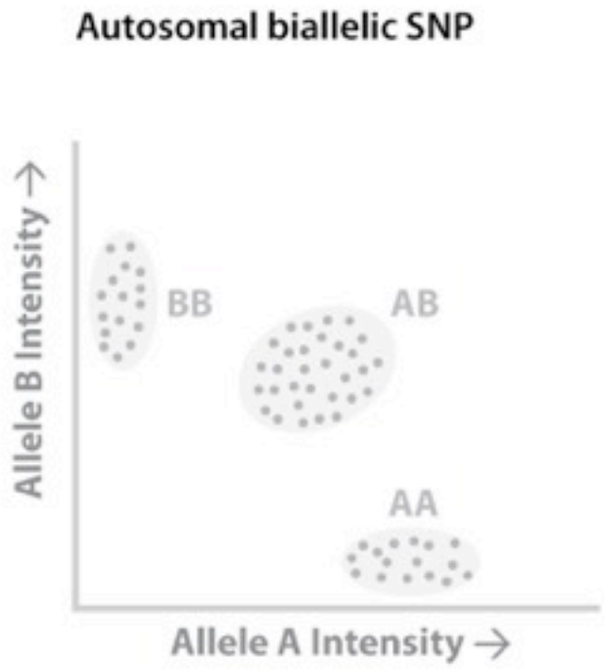
Identification of triallelic SNPs

Lude°Franke



Identification of triallelic SNPs

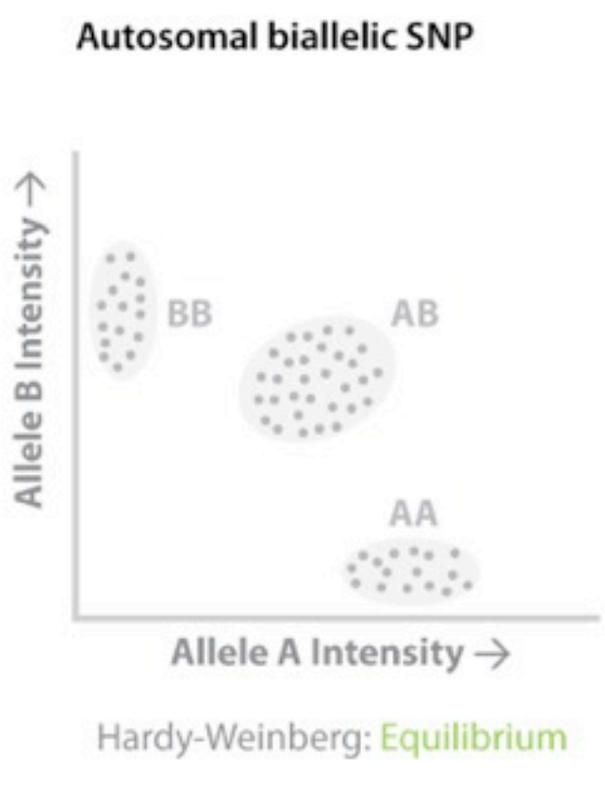
Lude°Franke



Genotype	AA	AB	BB
Frequency	25%	50%	25%
Allele	A		B
Frequency	(2AA + AB) / 2 = 50%		(AB + 2BB) / 2 = 50%
Genotype	AA	AB	BB
Expected	A^2 = 25%	2AB = 50%	B^2 = 25%

Identification of triallelic SNPs

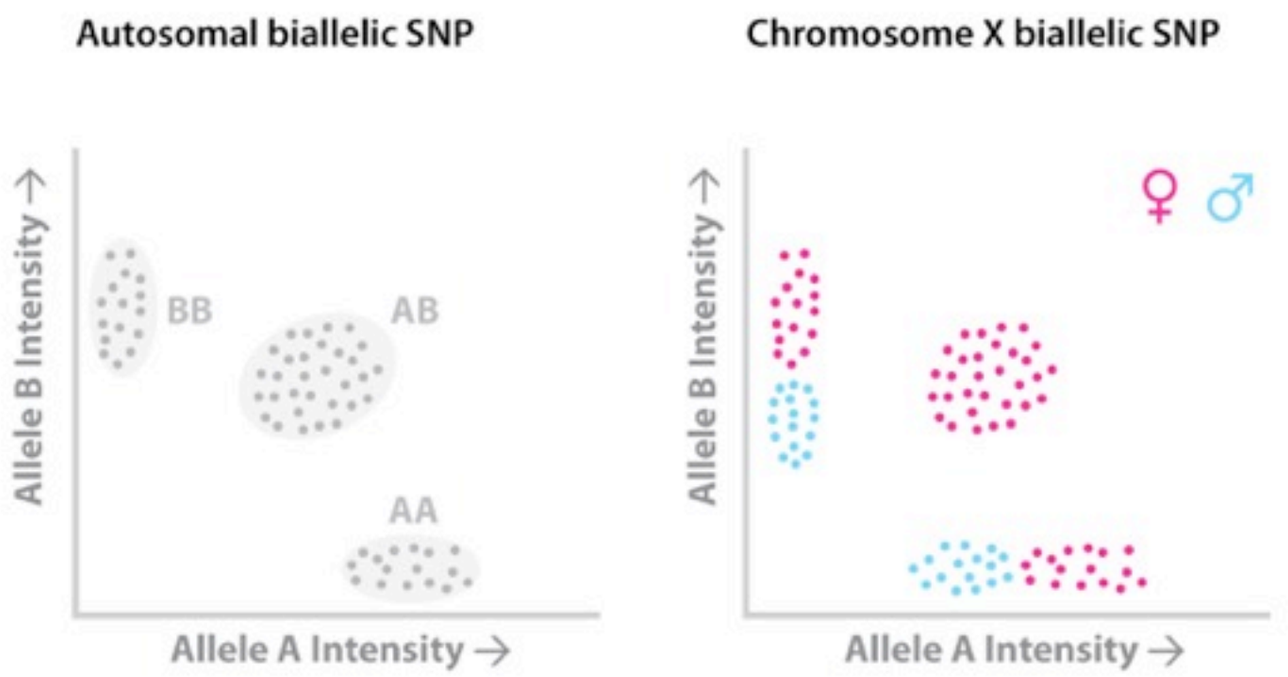
Lude°Franke



Hardy-Weinberg: Equilibrium

Identification of triallelic SNPs

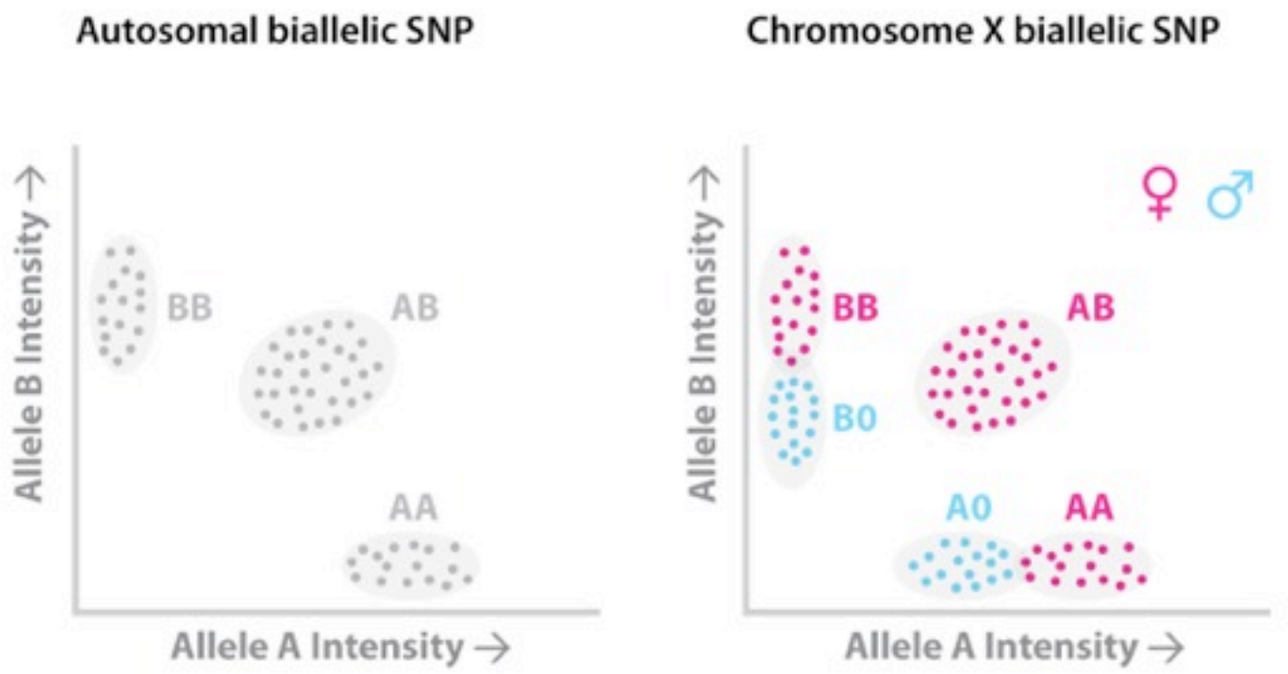
Lude°Franke



Hardy-Weinberg: Equilibrium

Identification of triallelic SNPs

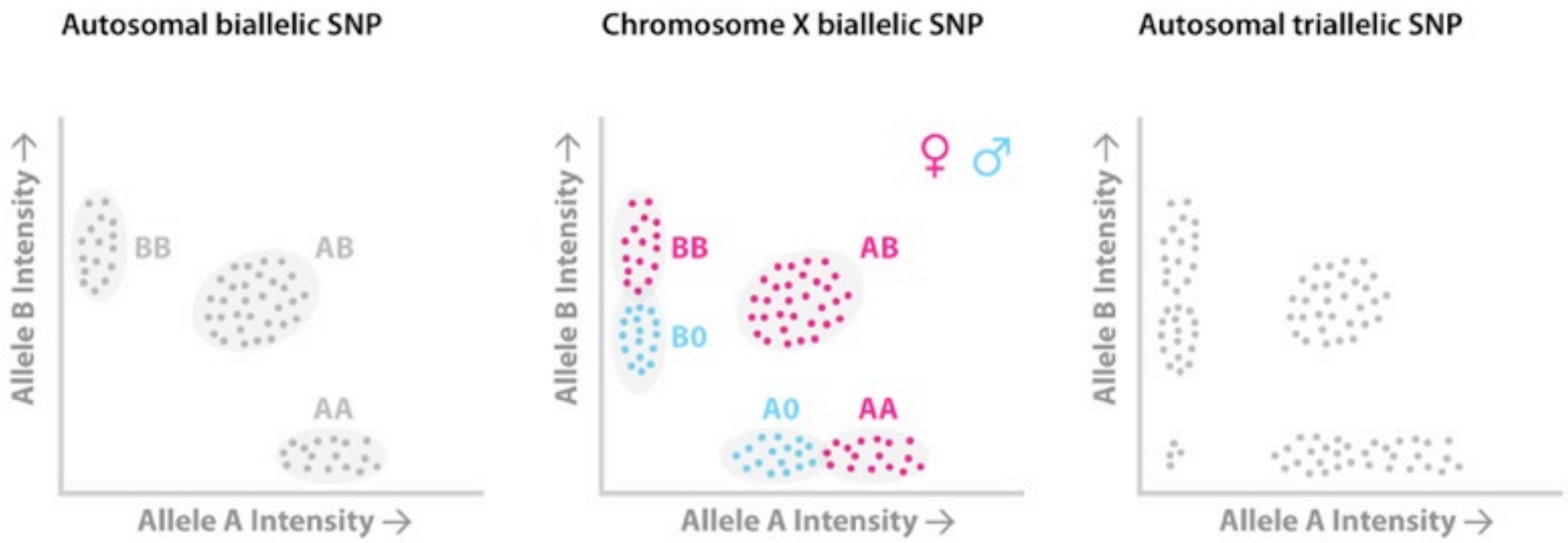
Lude°Franke



Hardy-Weinberg: Equilibrium

Identification of triallelic SNPs

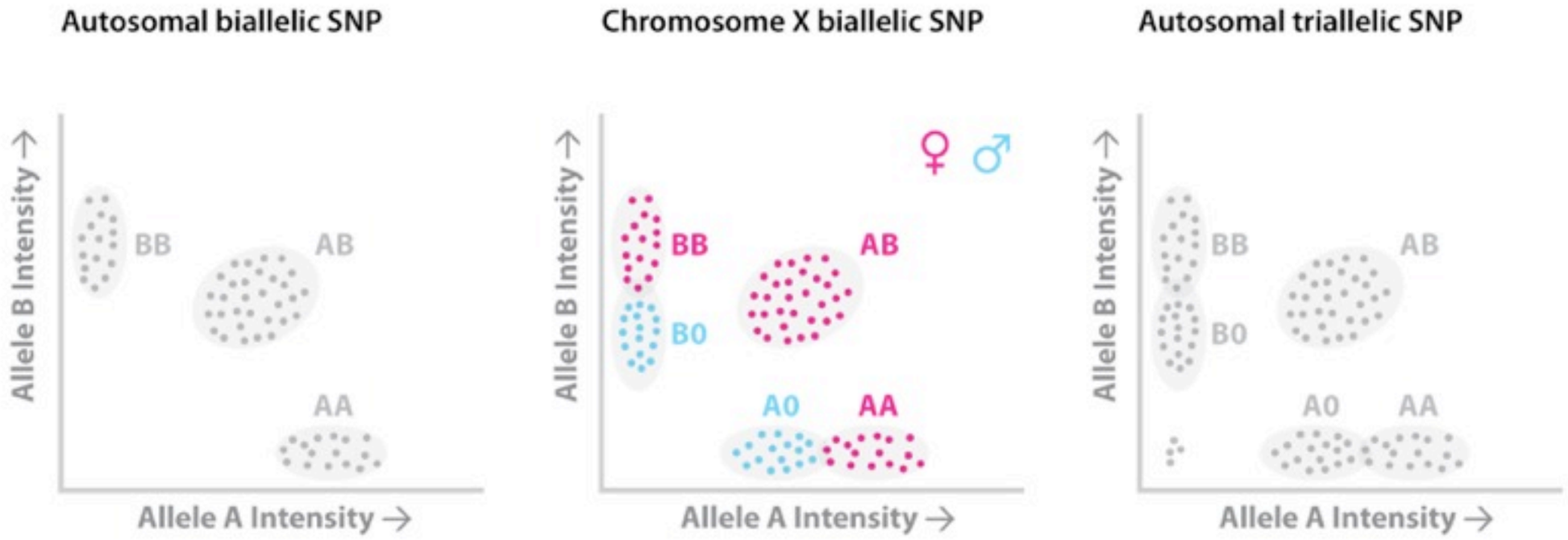
Lude°Franke



Hardy-Weinberg: Equilibrium

Identification of triallelic SNPs

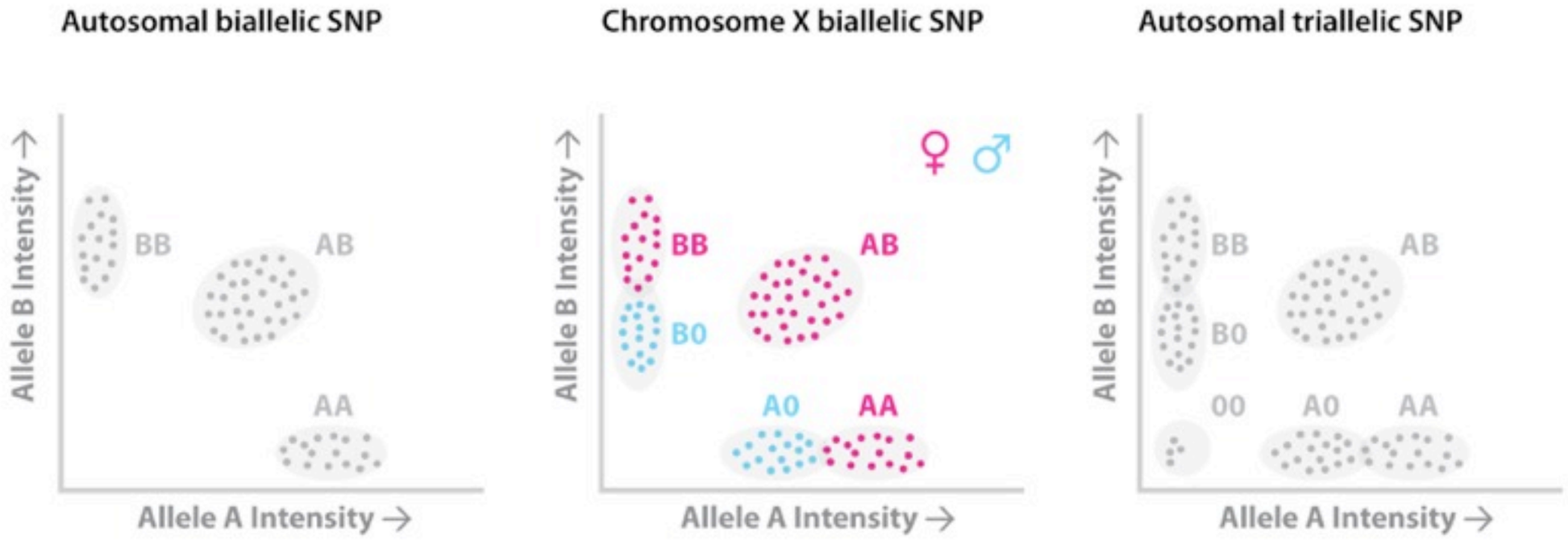
Lude°Franke



Hardy-Weinberg: Equilibrium

Identification of triallelic SNPs

Lude°Franke

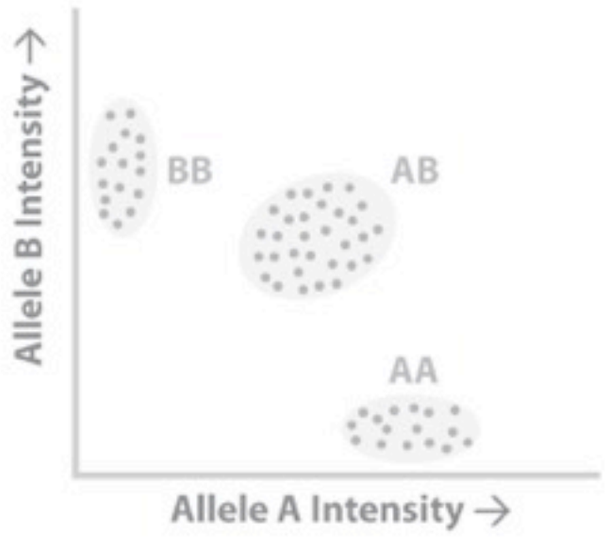


Hardy-Weinberg: Equilibrium

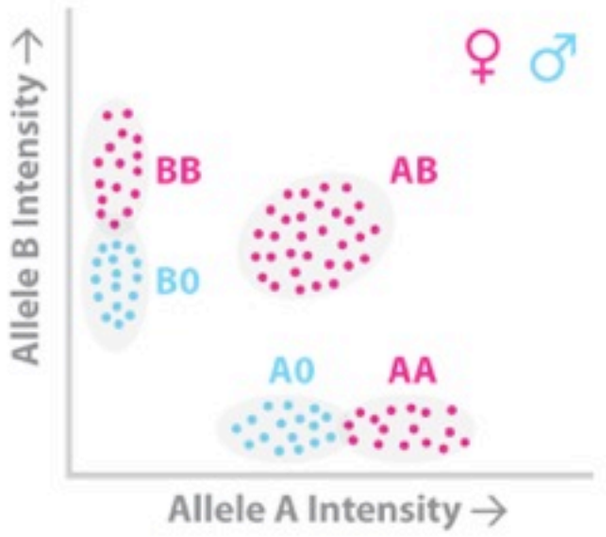
Identification of triallelic SNPs

Lude°Franke

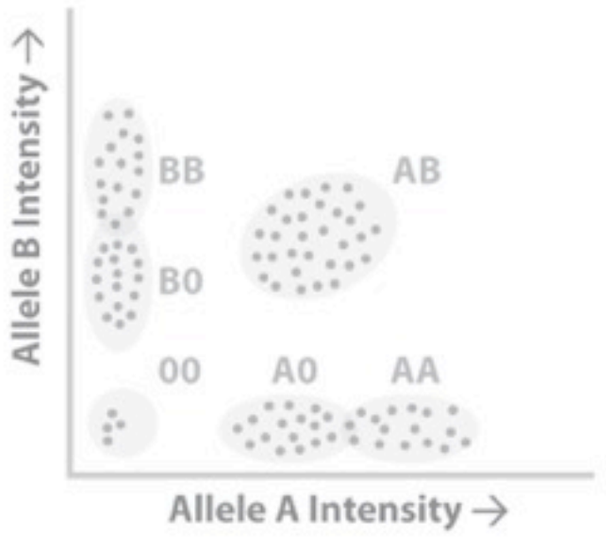
Autosomal biallelic SNP



Chromosome X biallelic SNP

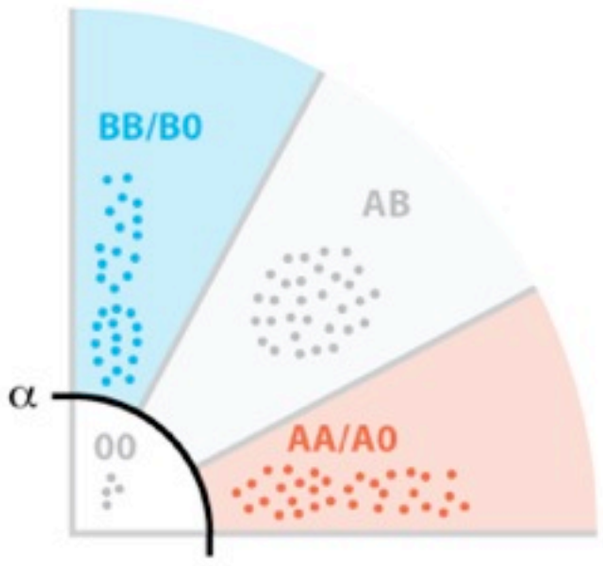


Autosomal triallelic SNP



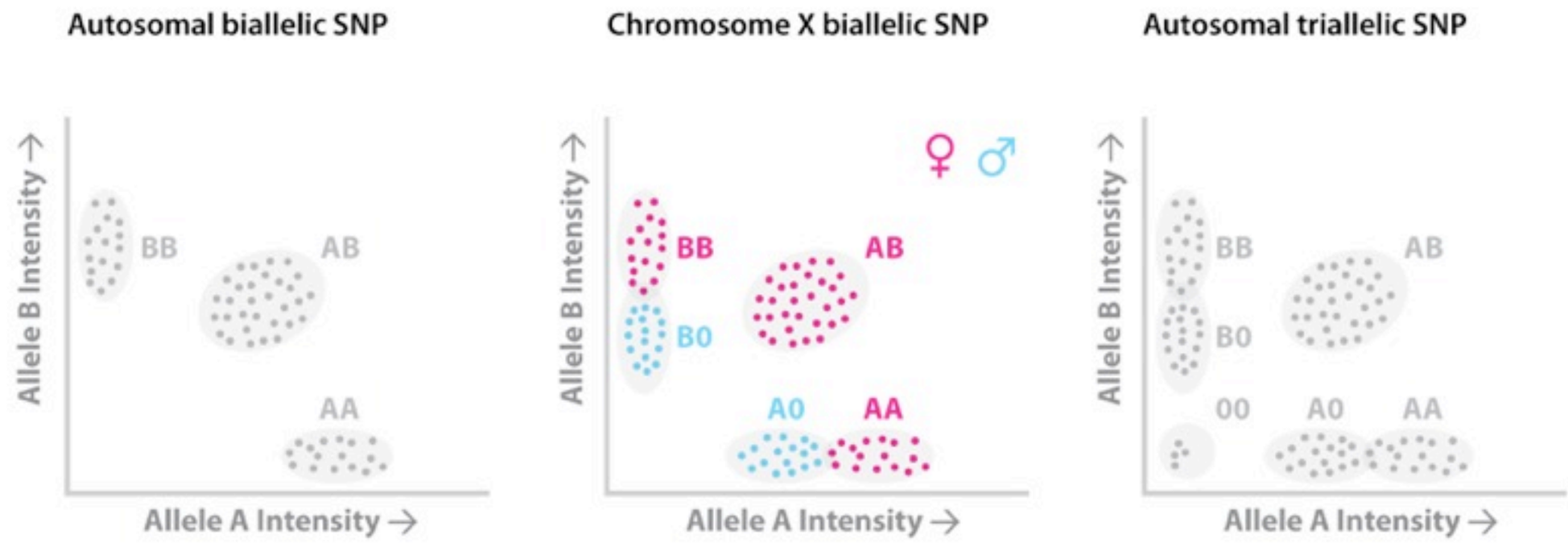
Hardy-Weinberg: Equilibrium

Genotype Assignment, Step 1:
00 Assignment (parameter α).
AA/A0, BB/B0 and AB Identification



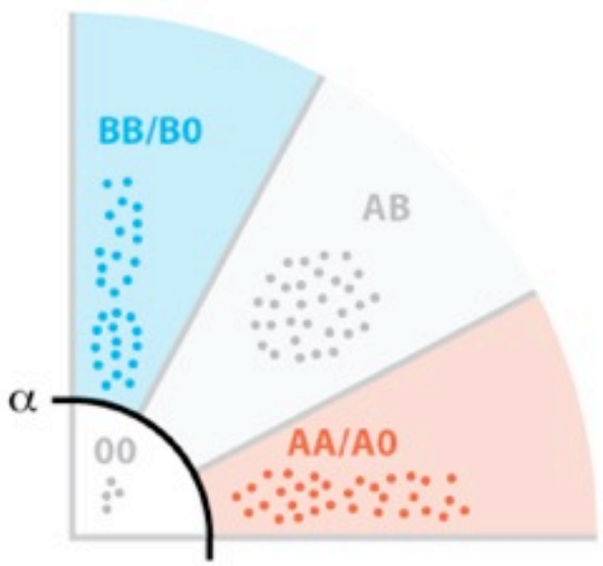
Identification of triallelic SNPs

Lude°Franke

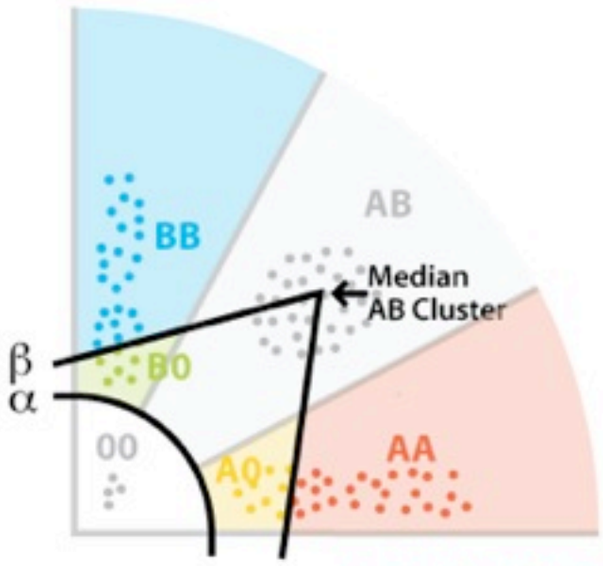


Hardy-Weinberg: Equilibrium

Genotype Assignment, Step 1:
00 Assignment (parameter α).
AA/A0, BB/B0 and AB Identification



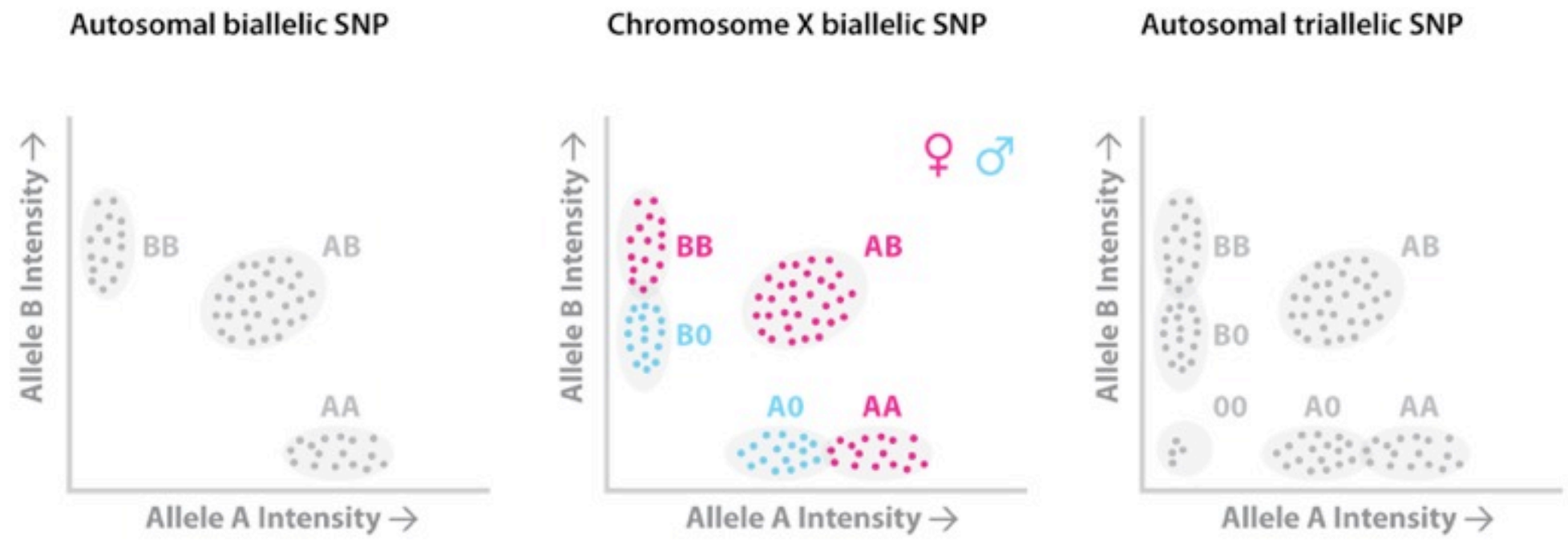
Genotype Assignment, Step 2:
Discrimination between AA and A0
and between BB and B0 (parameter β)



Hardy-Weinberg: Disequilibrium

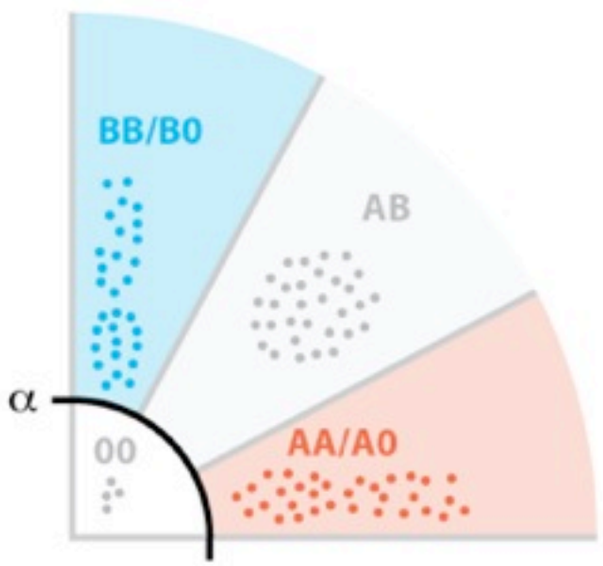
Identification of triallelic SNPs

Lude°Franke

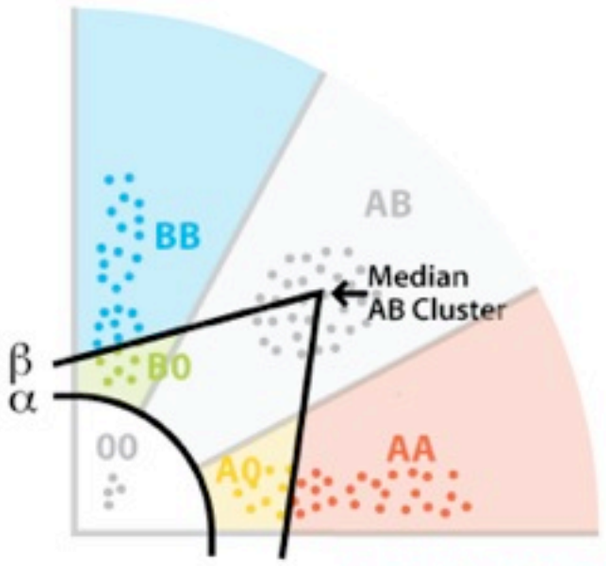


Hardy-Weinberg: Equilibrium

Genotype Assignment, Step 1:
00 Assignment (parameter α).
AA/A0, BB/B0 and AB Identification

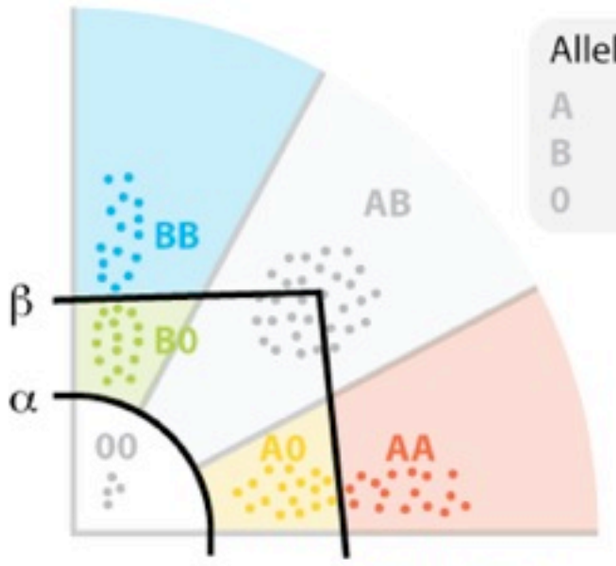


Genotype Assignment, Step 2:
Discrimination between AA and A0
and between BB and B0 (parameter β)



Hardy-Weinberg: Disequilibrium

Genotype Assignment, Step 3:
Maximum likelihood optimization of
HWE by adjusting α and β parameters



Allele	Frequency
A	40%
B	40%
0	20%

Hardy-Weinberg: Equilibrium

Linkage disequilibrium again

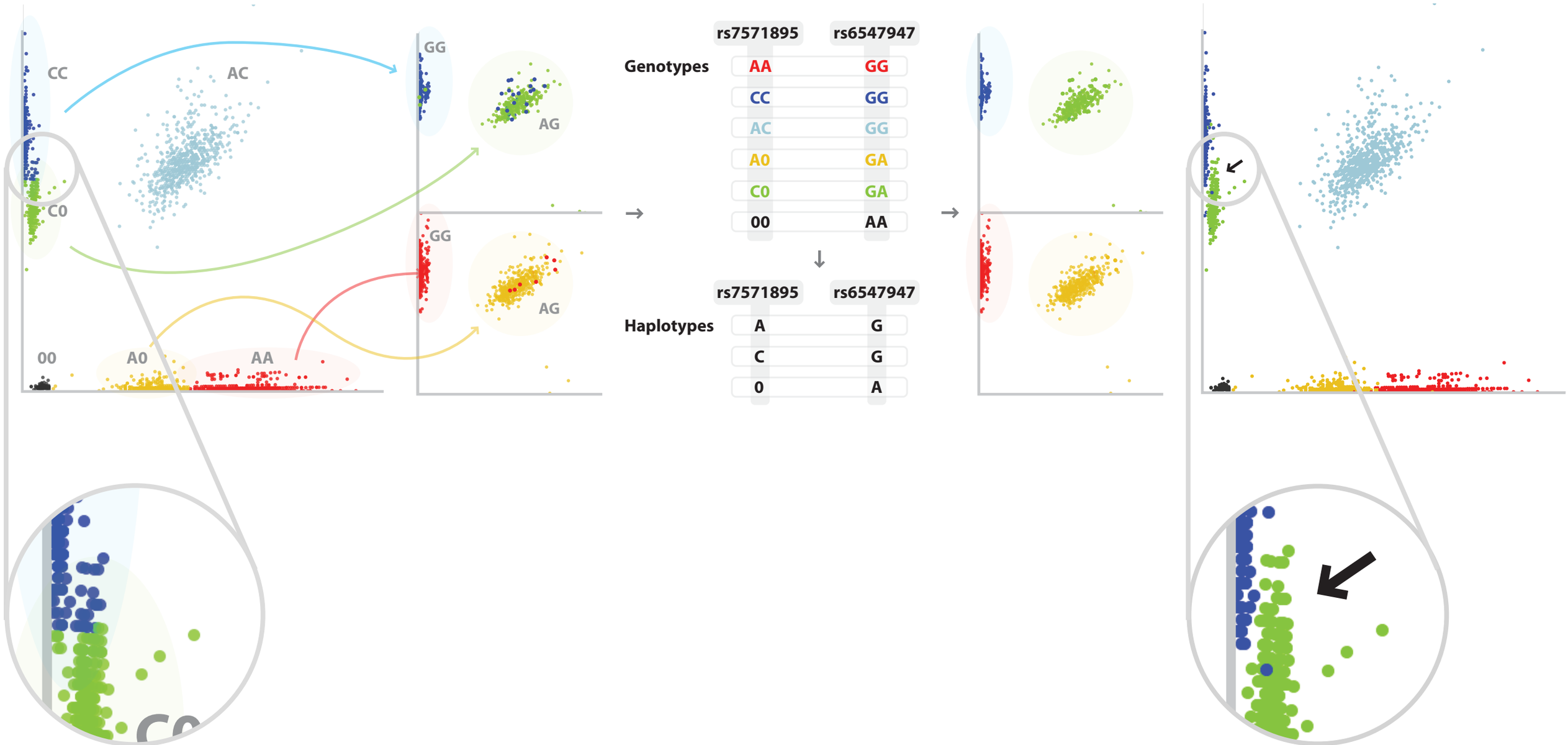
Lude°Franke

a) Initial genotype fit (rs7571895)

b) Haplotype analysis nearby SNP (rs6547947)

c) Genotype fit adjustments

d) Final genotype fit

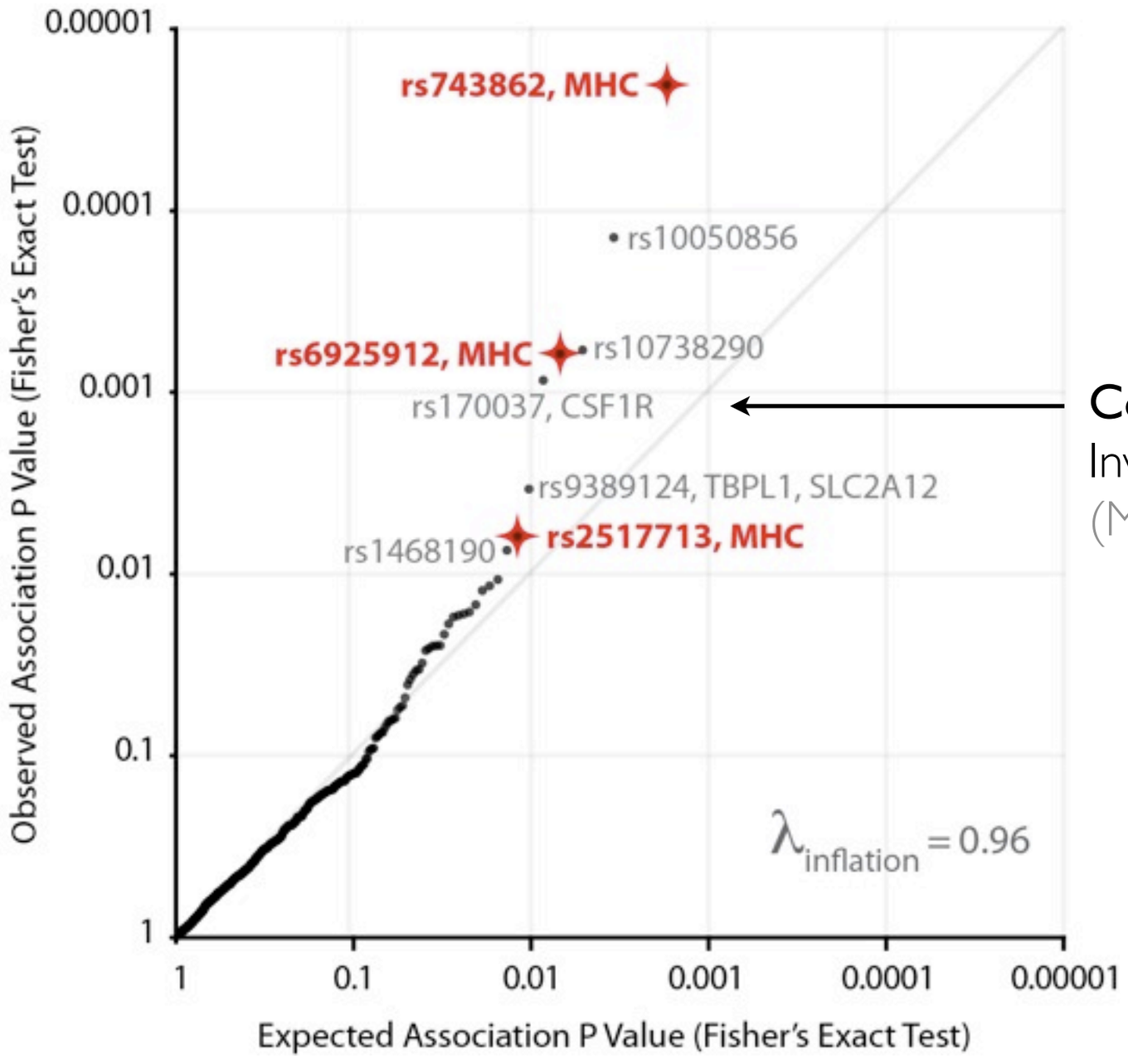


1,880 identified common human deletions

Franke *et al*, American Journal of Human Genetics, 2008

Association analysis in coeliac disease

Van Heel, Franke *et al*, A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21, Nature Genetics 2007



Colony Stimulating Factor Receptor 1:
Involved in macrophage differentiation
(Mapping in significantly linked locus at 5q)