H3ABioNet

Pan African Bioinformatics Network for H3Africa

UNIVERSITY OF THE WITWATERSRAND

SCIENTIA ET LABORE

JOHANNESBURG

Introduction to Genome-Wide Association Studies

H3ABioNet Data Management Workshop

Shaun Aron

June 2014

Outline

- Introduction to GWAS
- GWAS concepts
- GWAS design considerations
- GWAS data analysis process
- PLINK data formats

Identification of disease genes

- Identification of genes that contribute to disease risk is one of the main research areas of molecular biology
- Successful methods have been used for the identification of disease genes for several single gene disorders
- In most single gene disorders a single mutation in a gene is completely responsible for the phenotype

Identification of disease genes

- Complex common diseases such as obesity, cardiovascular disease, cardiometabolic disease, diabetes etc. usually occur as a result of variations in several different genes together with environmental factors

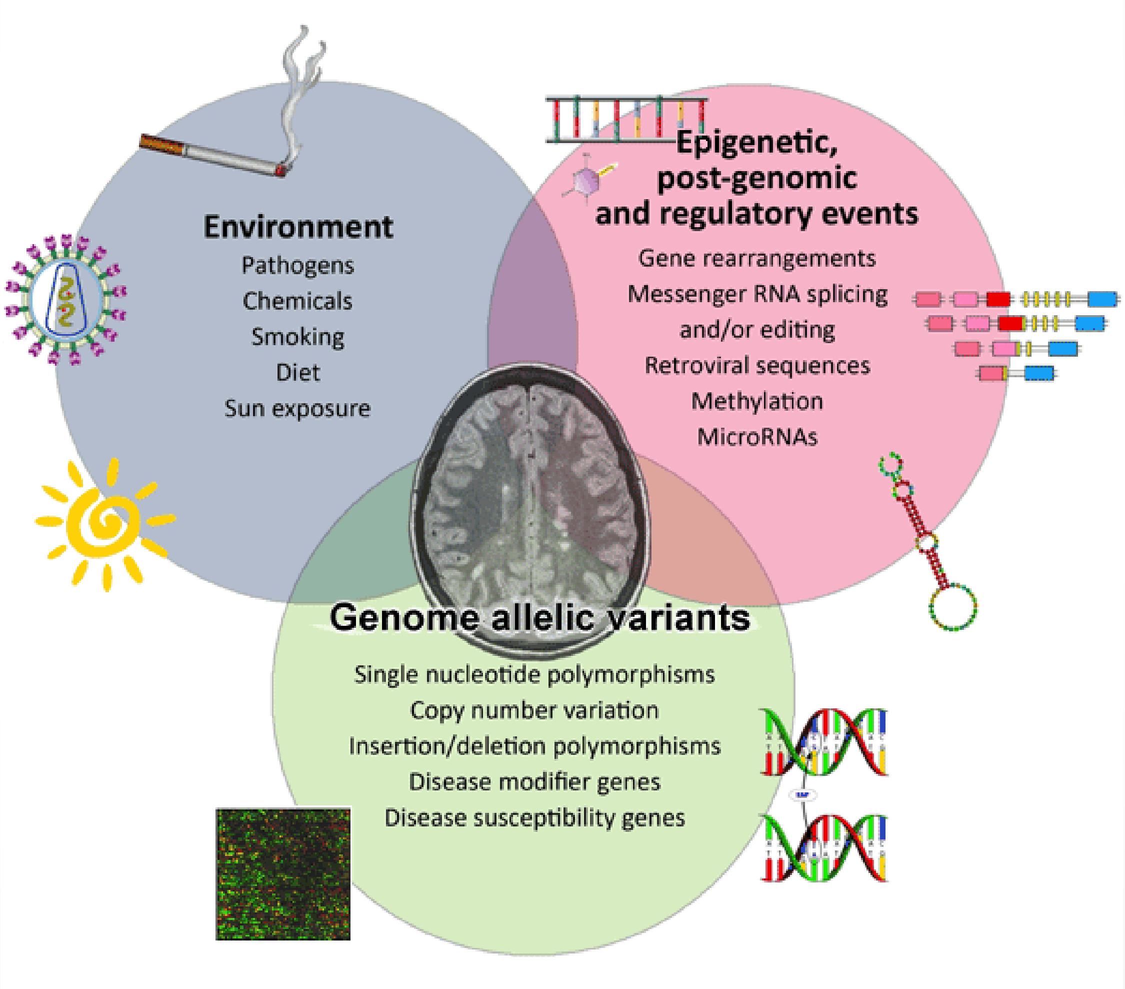
- Each variation has a small effect on the phenotype

- In most instances these are common variants

Terminology

- DNA ~ genome
  - Base pairs that make up the genome , adenosine (A), guanine (G), cytosine(C), thymine(T)
- Single nucleotide polymorphism (SNP)
  - Variations in the DNA that occur commonly in at least one population i.e. >1% e.g at a particular position in the genome there can be an A or a G in a population
- Allele
  - Each SNP has two alleles – one from each chromosome
- Mutation
  - A source of DNA variation – usually disease causal
- Point mutations
  - Changes that occur in DNA, source of SNPs and mutations

Complex diseases



Environment

Pathogens
Chemicals
Smoking
Diet
Sun exposure

Epigenetic, post-genomic and regulatory events

Gene rearrangements
Messenger RNA splicing and/or editing
Retroviral sequences
Methylation
MicroRNAs

Genome allelic variants

Single nucleotide polymorphisms
Copy number variation
Insertion/deletion polymorphisms
Disease modifier genes
Disease susceptibility genes

Classic approaches to disease gene identification

- Traditional heritability
  - Does it run in families?
- Design
  - Family, twin and adoption studies
- Molecular data
  - None
- Desired outcome
  - Gives us a clue as to whether there is a genetic component

Matt McQueen, , Colorado University, Boulder – Intro to GWAS

Classic approaches to disease gene identification

- Traditional linkage
  - Find genomic loci linked to disease
- Design
  - Family-based
- Molecular data
  - 300 – 600 repeat markers
- Desired outcome
  - Find genetic region linked to disease

Matt McQueen, , Colorado University, Boulder – Intro to GWAS

Genome-Wide Association

- GWAS
  - Find common alleles associated with disease
- Design
  - Cohort, case-control, family-based
- Molecular data
  - 500 000 – 2.5M single nucleotide polymorphisms
- Desired outcome
  - Find common genetic variation associated with disease

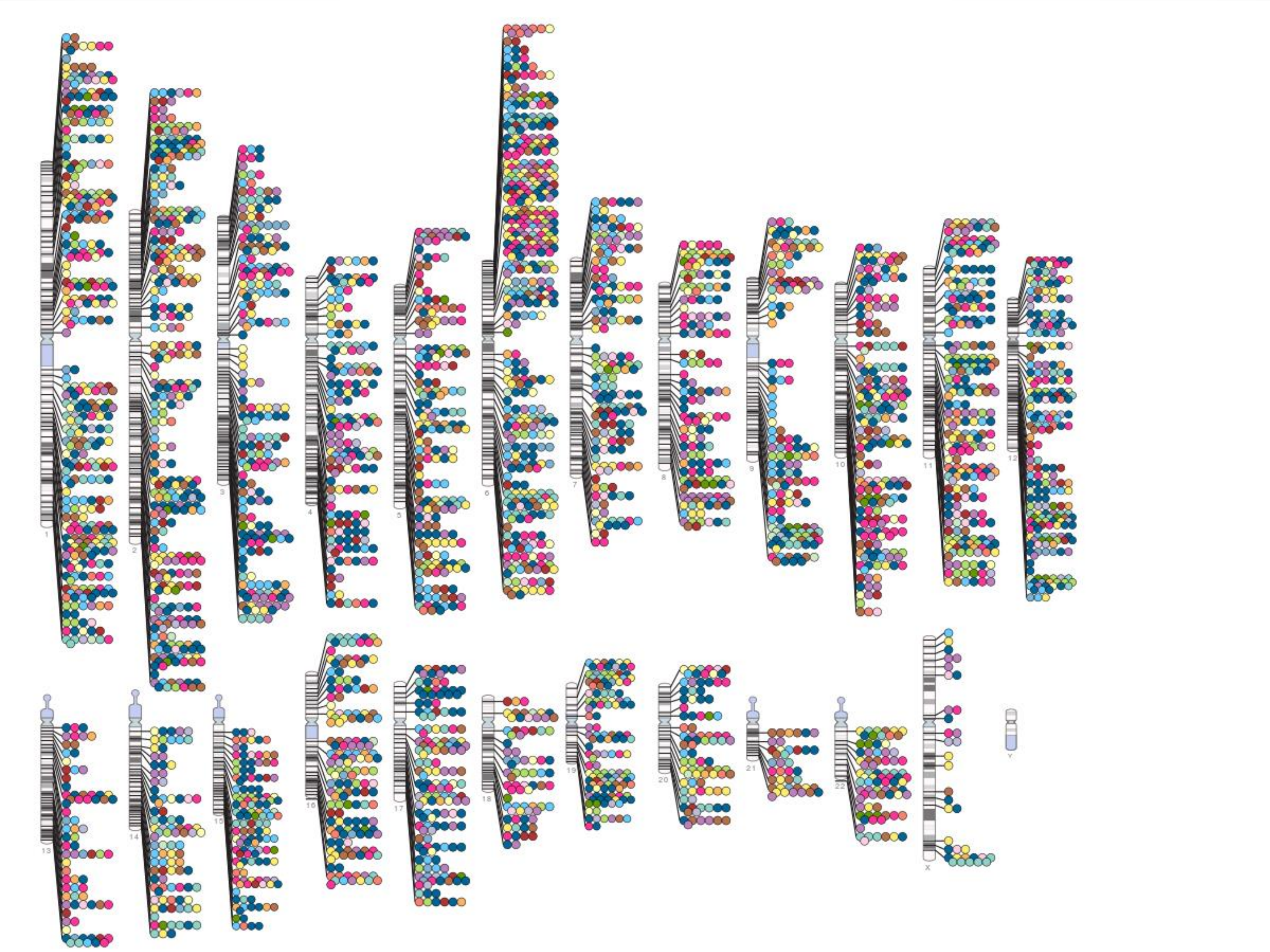
Matt McQueen, Colorado University, Boulder – Intro to GWAS

What is an association study?

- Association refers to the co-occurrence of a genetic variant with a disease trait, more frequently than can be explained by chance

- Candidate gene approach was the first type of association study
  - Select a gene linked to a phenotype based on biological function
  - Sequence that gene in affected and unaffected samples
  - Identify mutations that associate more with affected samples than unaffected samples

GWAS

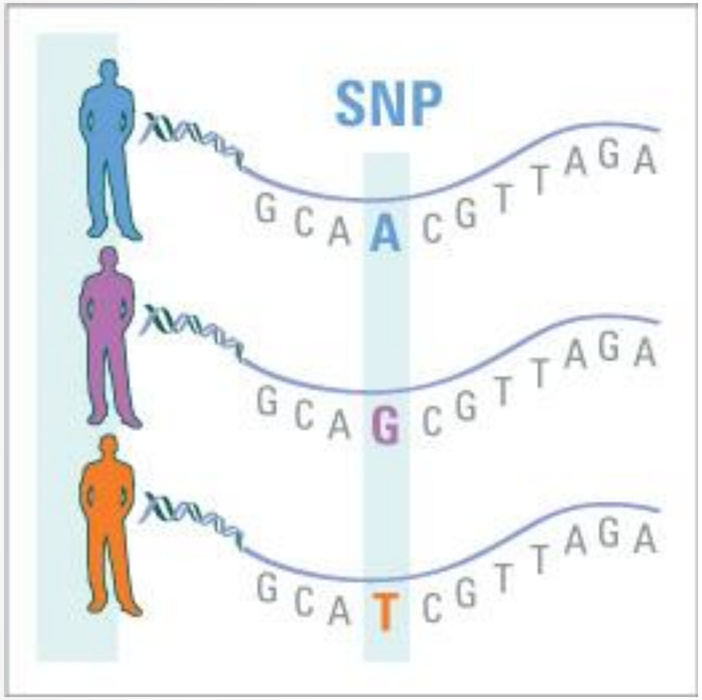


Genome-Wide Association Study Concepts

Why SNPs?

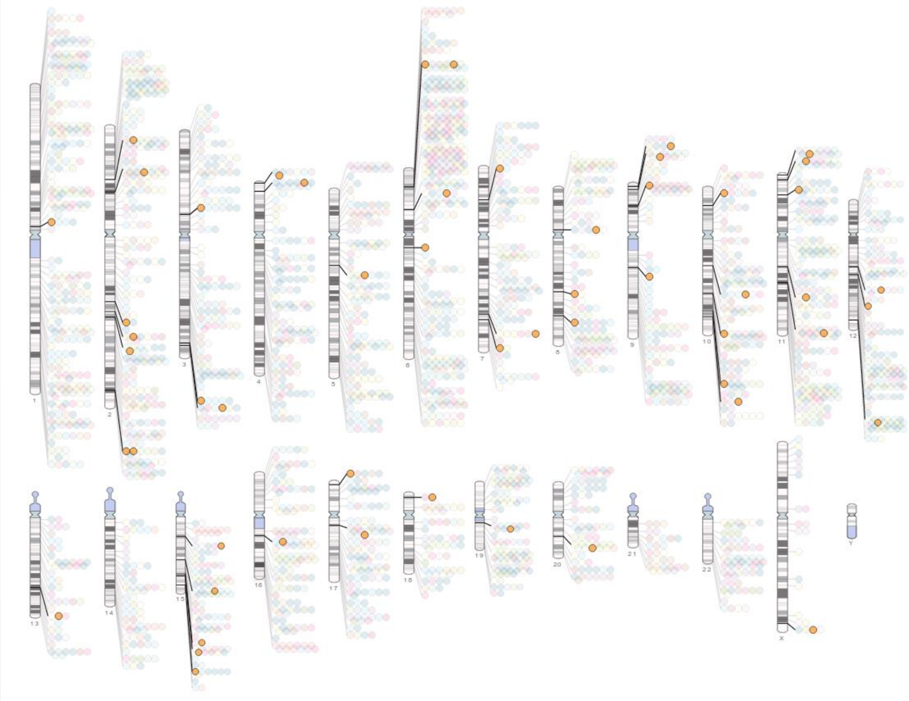
- Several types of variants in the human genome
  - SNPs
  - Insertions/deletions
  - Chromosomal mutations
  - Copy number variants
- SNPs are the most common type of genetic variation
  - In the latest version of dbSNP (141) there are 62 387 983 SNPs identified in the human genome
- SNPs occur randomly throughout the genome
- Can be accurately measured as they have previously been identified and characterised

SNPs



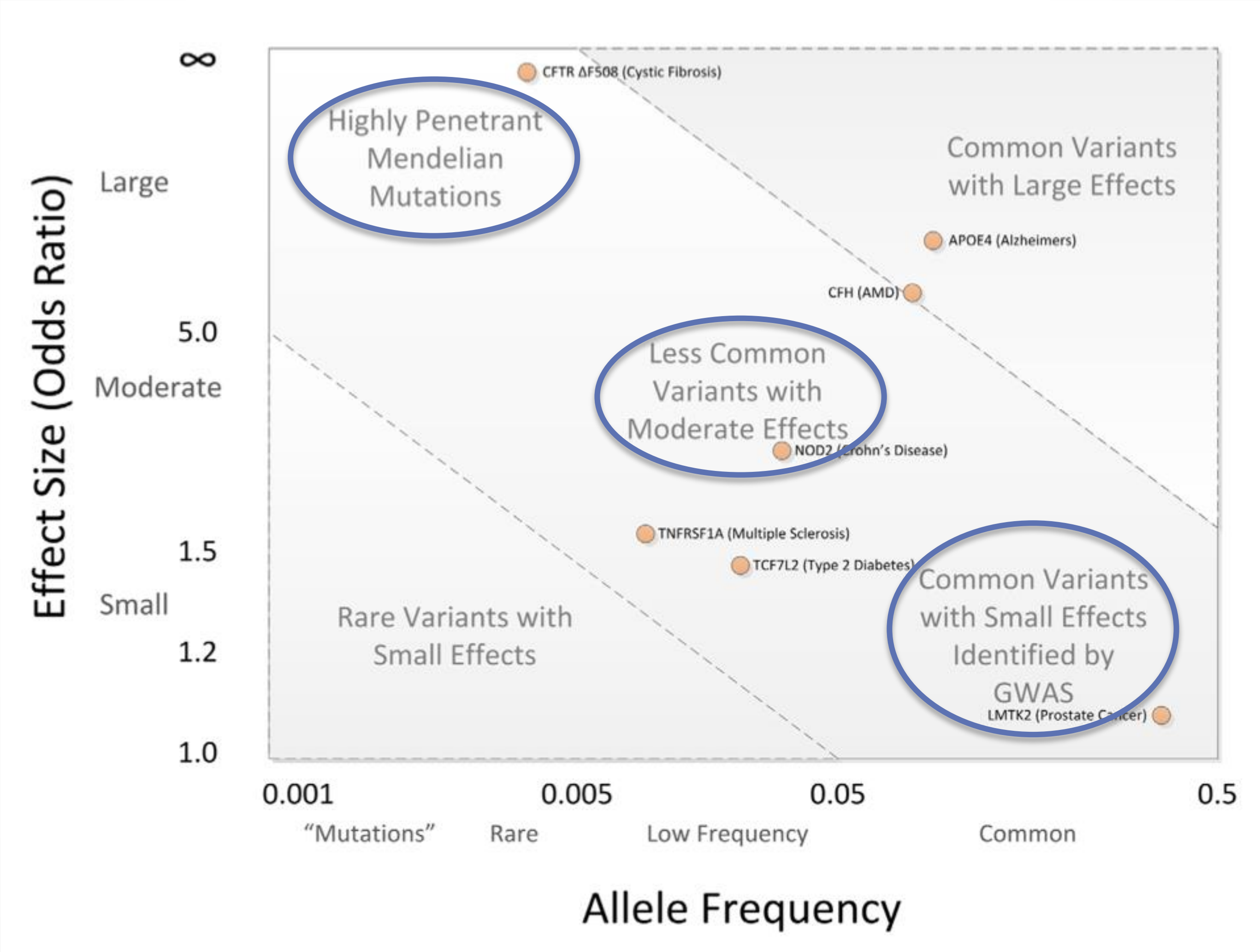
- A SNP is a position on the genome where there is a significant variation in at least one human population
- In complex diseases the co-occurrence of SNPs in various regions can contribute to a phenotype

Common disease – common variant hypothesis



- Common diseases are likely to be due to common genetic variations
- Each variation will have a small contribution to the phenotype
- Variations in multiple regions of the genome in combination with environmental factors affect susceptibility to the phenotype

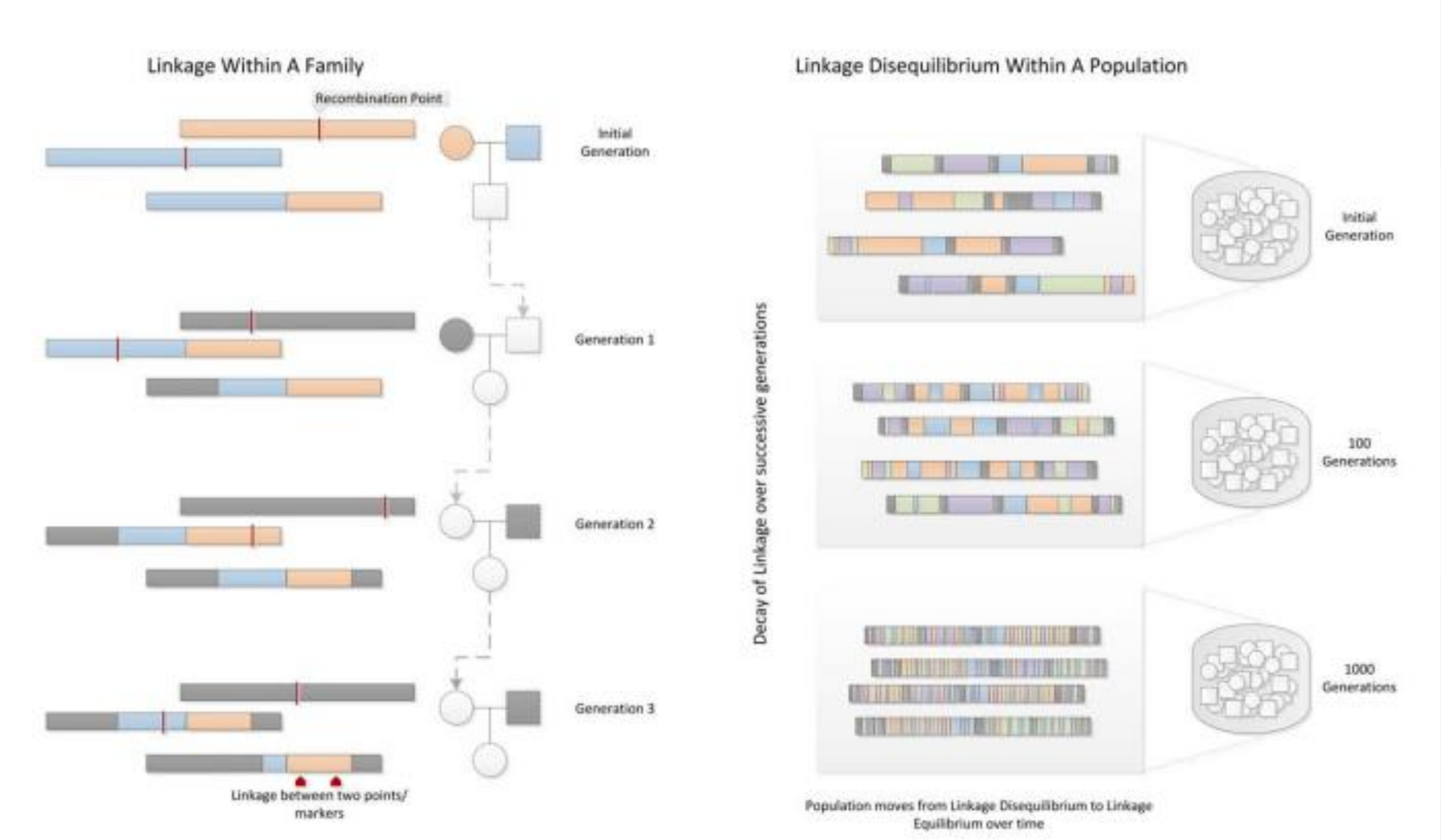
Disease allele effects



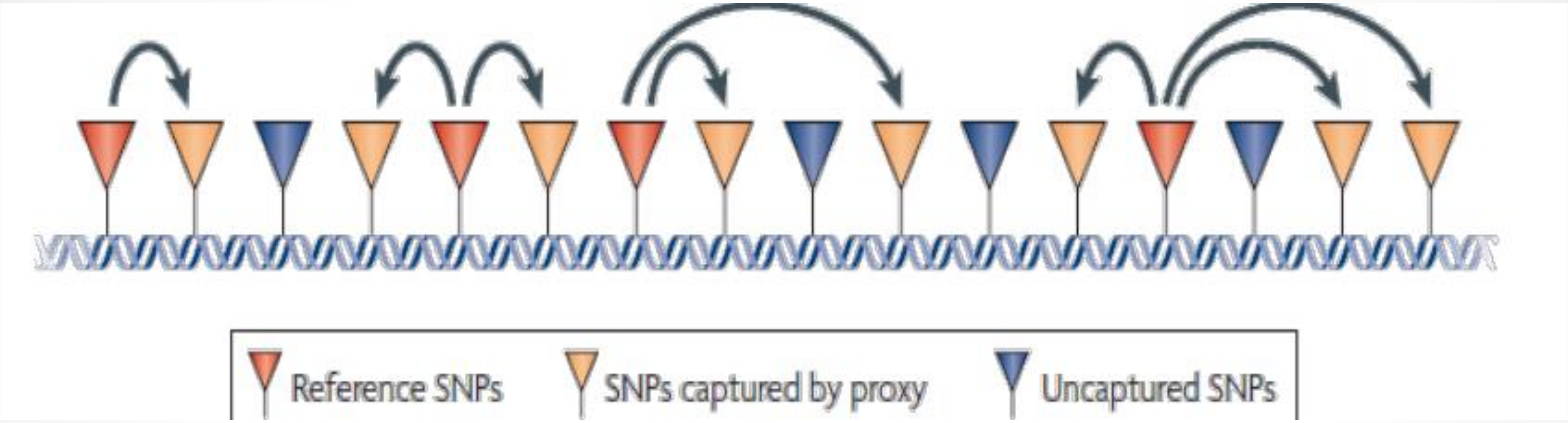
So what SNPs do we use?

- 60 million human SNPs
  - Current technologies allow for the interrogation of up to 2.5 million SNPs at a time
- Need a good method to select the most informative SNPs
- All SNPs in the human genome are not independent
- Concept of linkage disequilibrium (LD) used to select SNPs

LD and selecting tagSNPs


Linkage Within A Family
Recombination Point
Initial Generation
Generation 1
Generation 2
Generation 3
Linkage between two points/ markers
Linkage Disequilibrium Within A Population
Decay of Linkage over successive generations
Initial Generation
100 Generations
1000 Generations
Population moves from Linkage Disequilibrium to Linkage Equilibrium over time

tagSNPs and indirect associations

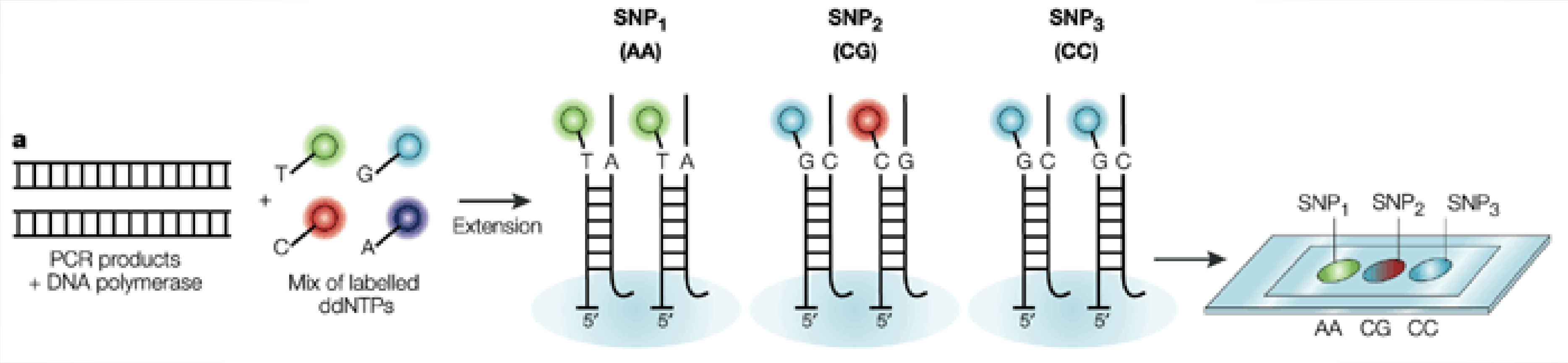


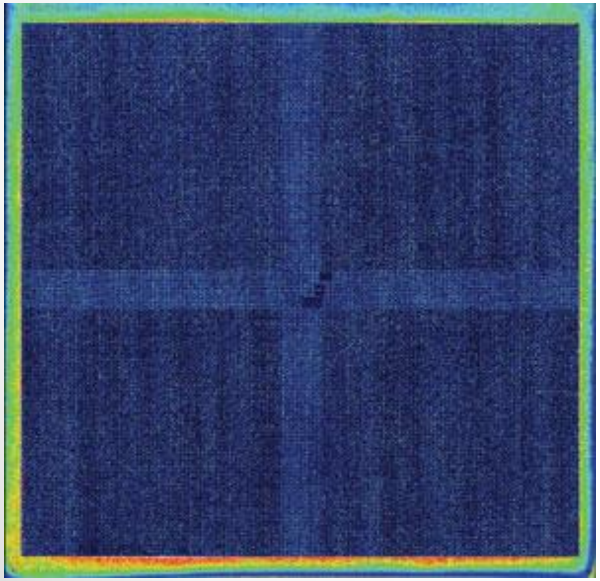
- HapMap project catalogued SNPs that occur in different populations
- Determined which regions of the genome are in LD
- Based on this we are able to select tagSNPs that are act as a proxy for other SNPs in LD
- Minimal set of SNPs required to represent most SNPs in the human genome

Important considerations

- LD varies between populations
  - HapMap data is based on a set of defined populations
  - African populations are represented by samples from
    - Ibadan, Nigeria
  - 1000 genomes project has generated further data from African populations based on sequencing data
    - Ibadan, Nigeria
    - Luhya, Kenya
    - African Americans (admixed)
- Important to note that SNP microarray chips are based on tagSNPs based on LD in European populations

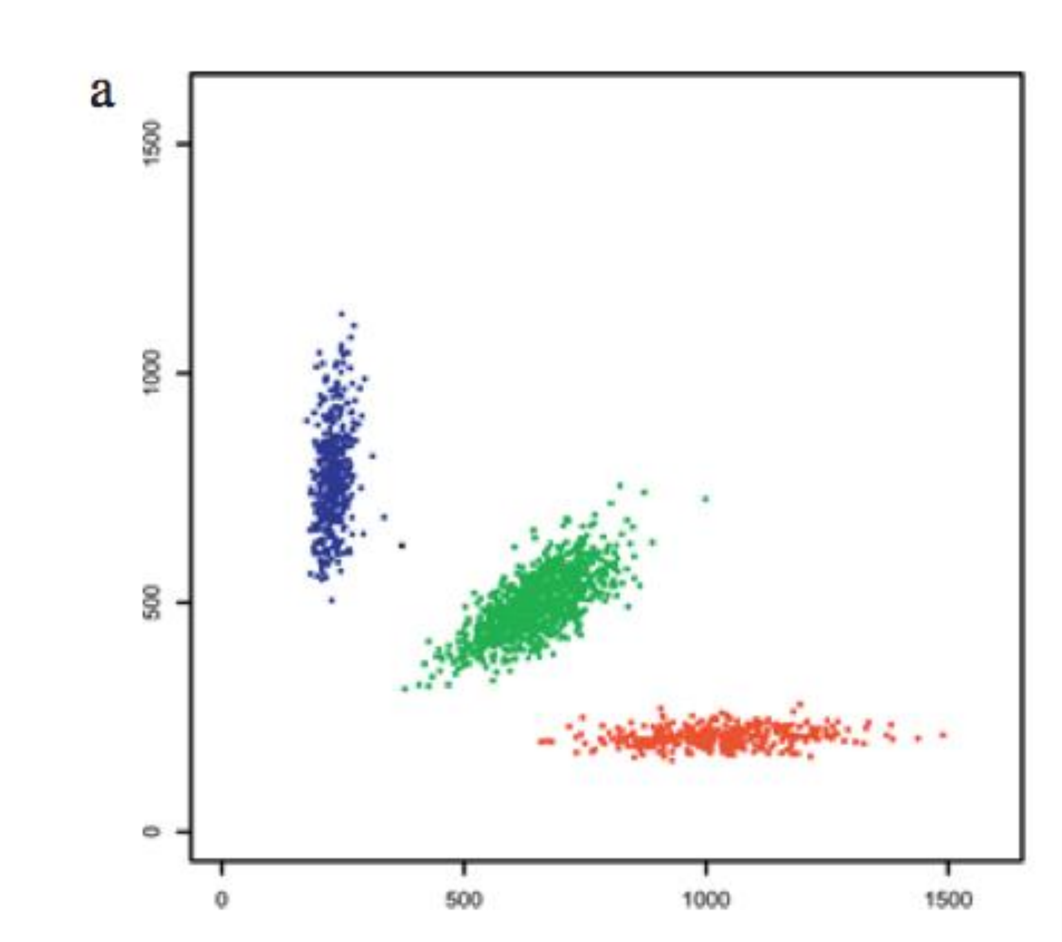
GWAS technology



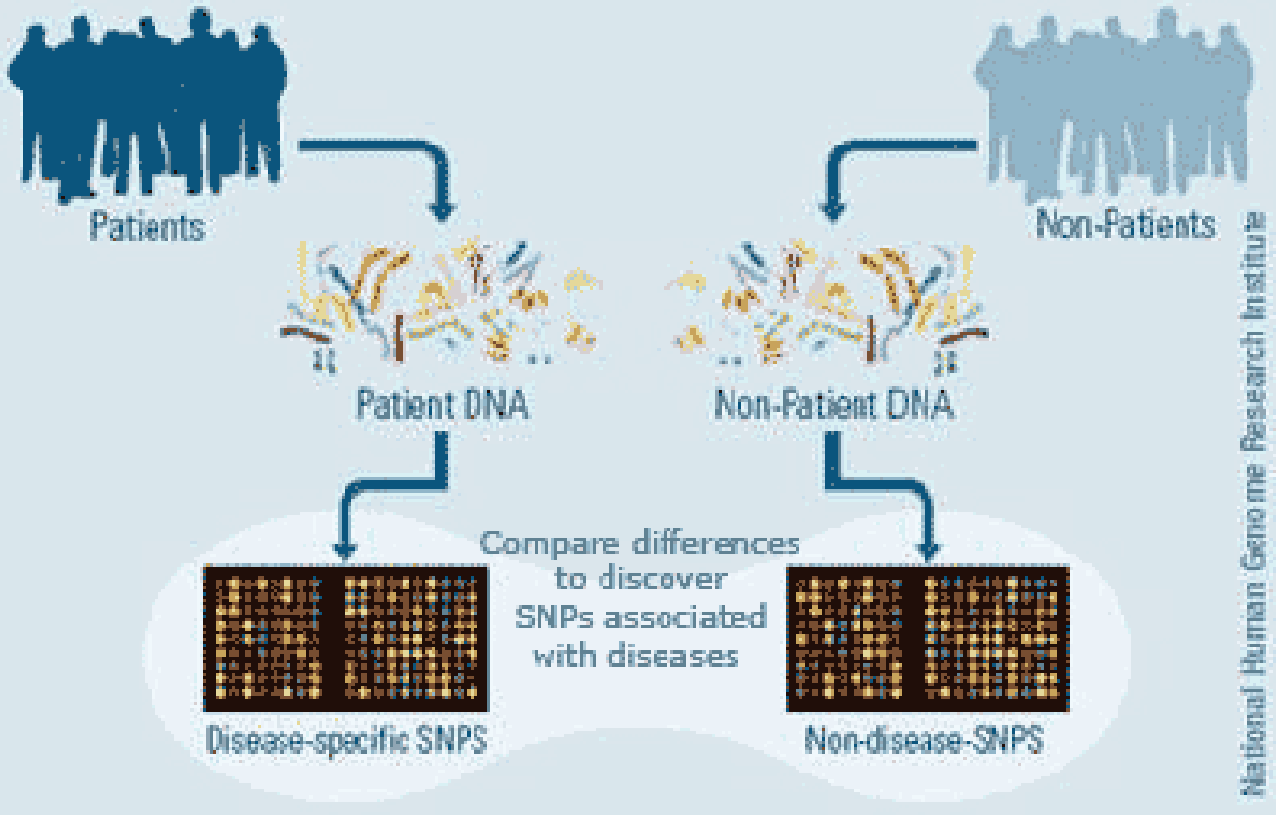


affymetrix
Biology for a better world

GWAS data



GWAS process


Patients
Non-Patients
Patient DNA
Non-Patient DNA
Compare differences to discover SNPs associated with diseases
Disease-specific SNPS
Non-disease-SNPS
National Human Genome Research Institute

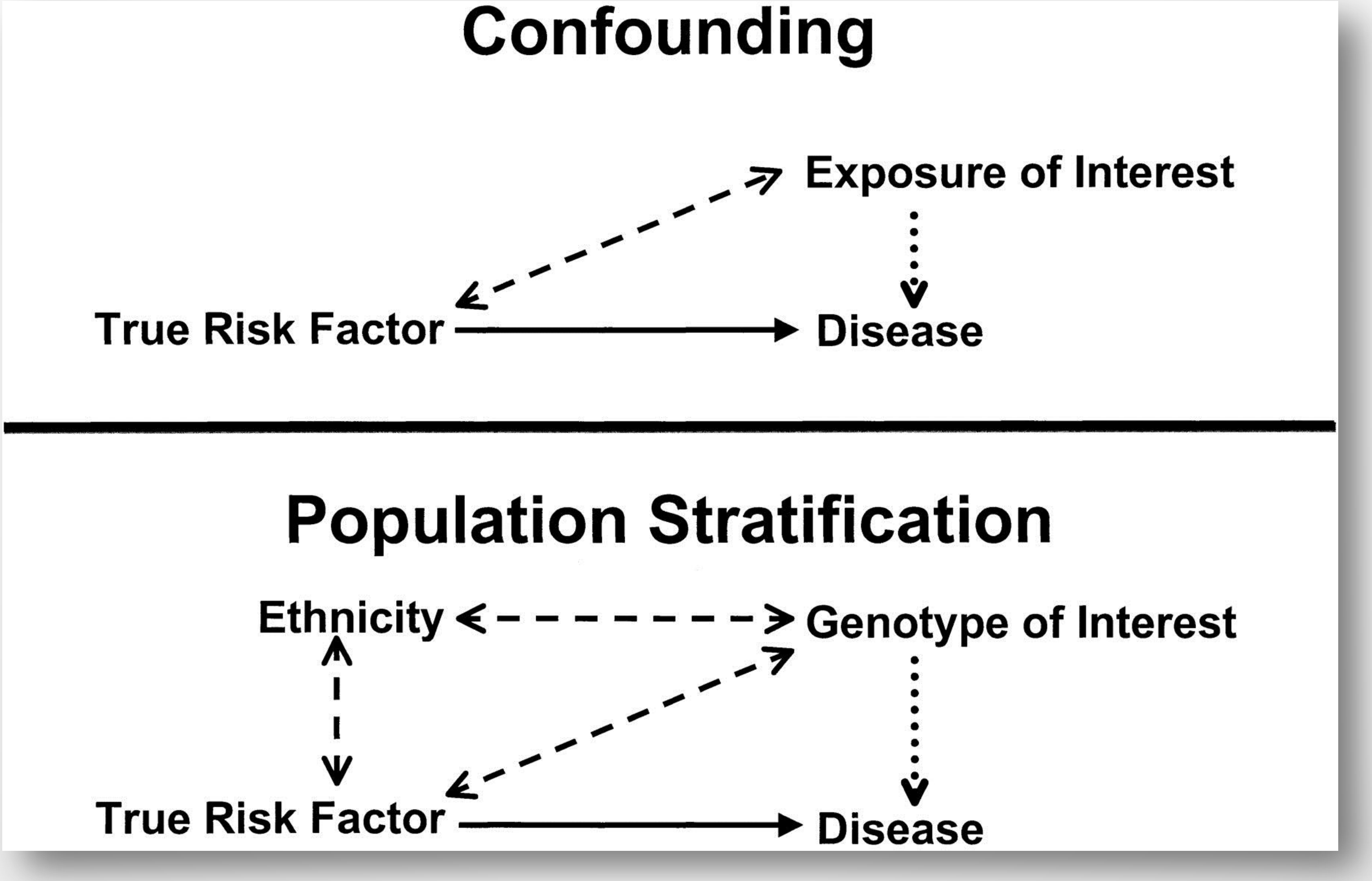
Design considerations for a GWAS

Shaun Aron - H3ABioNet Data Management Workshop June 2014

Designing a GWAS

- Study design
  - Case versus control or quantitative
  - Quantitative traits more widely used in GWAS
- Phenotype criteria and measurement
  - Standardised method for taking phenotype measurements
- Sample population homogeneity
  - Sample should be collected from individuals with similar genetic ancestry

Population stratification

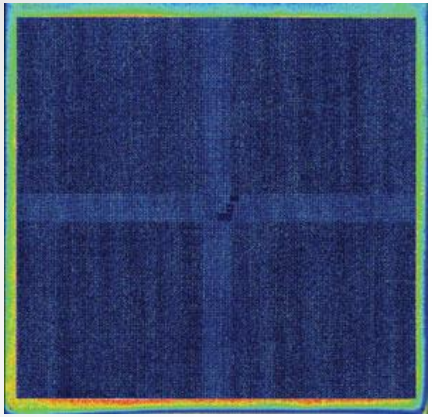


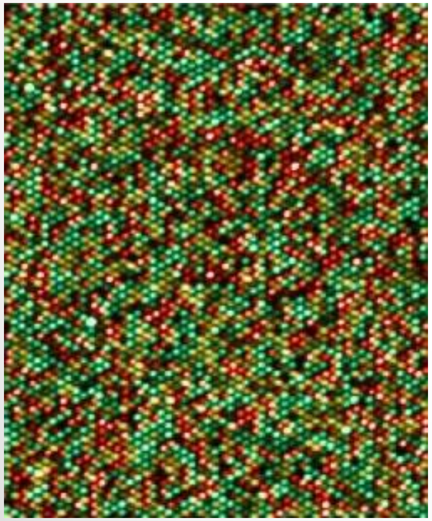
GWAS data analysis workflow

GWAS data

- Most service providers will provide you with either raw CEL. files or genotype call files
- The raw CEL files contain the raw intensity values from the microarray chip
- The genotype file will contain the actual alleles called for each SNP
- CEL. files are ~700Mb per sample (dependent on array)
- Various methods available for calling genotypes from raw cell files

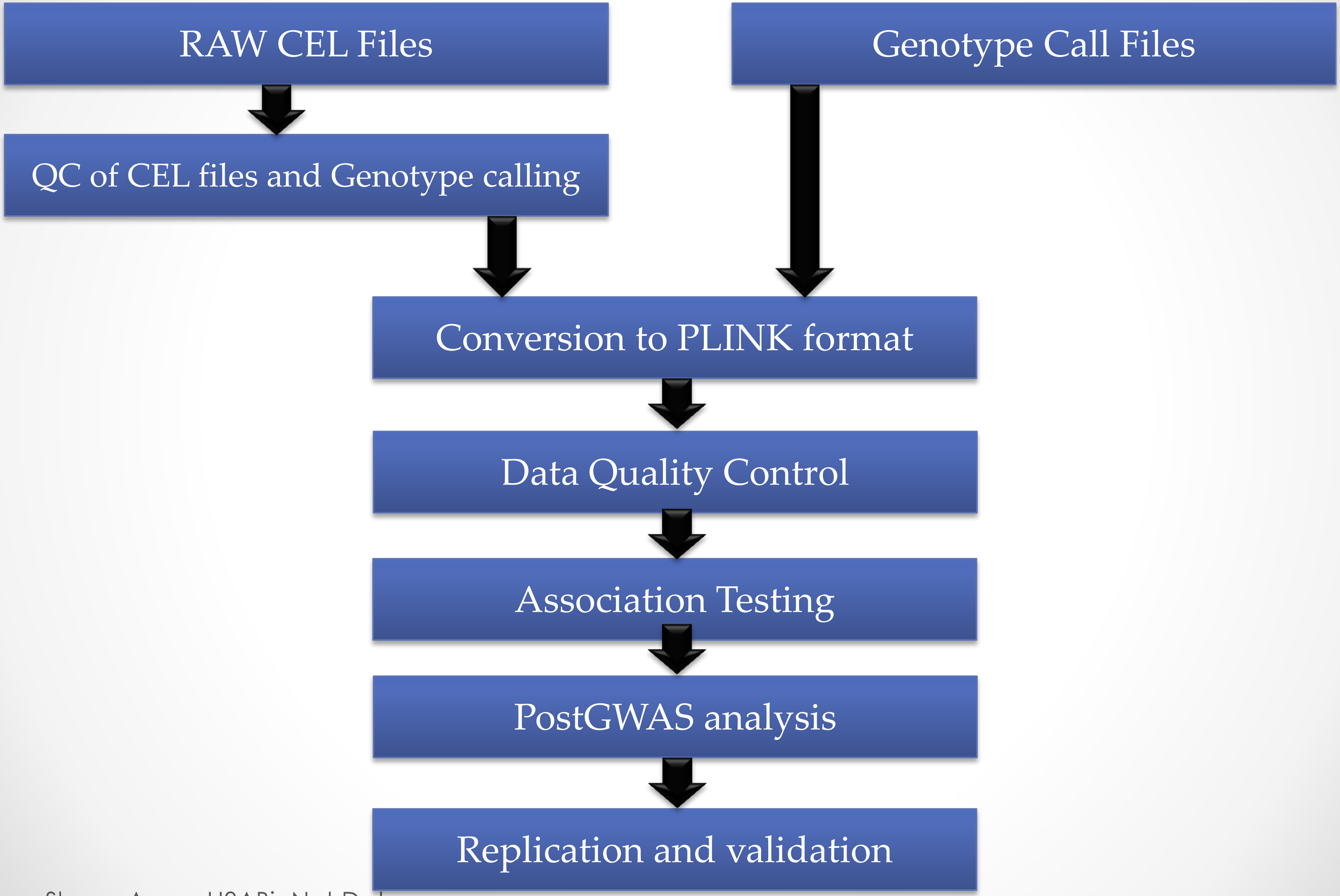
Genotype calling





Algorithm	Insitute	Reference
Birdseed	Affymetrix/Broad	Korn et.al 2008 Nat Gen 40:1253-1260
BRLMM	Affymetrix	Cawley et al. 2006
CHIAMO	WTCCC	WTCCC 2007 Nature 447:661 -78
CRLMM	John Hopkins University	Carvalho et al. 2007 Biostatistics 8:485-99
GEL	University of Chicago	Nicolae et al. Bioinformatics 22:1942-7
JAPL	Wellcome Trust, Cambridge	Plagnol et al. 2007 PLoS Genetics 3:e74
SNiPer-HD	Texas A&M University	Hua et al. 2007 Bioinformatics 23:57-63

GWAS analysis workflow



RAW CEL Files

Genotype Call Files

QC of CEL files and Genotype calling

Conversion to PLINK format

Data Quality Control

Association Testing

PostGWAS analysis

Replication and validation

PLINK data formats

PLINK data formats

- PLINK is a commonly used tool for manipulating and analysing GWAS data (Purcell, 2007)

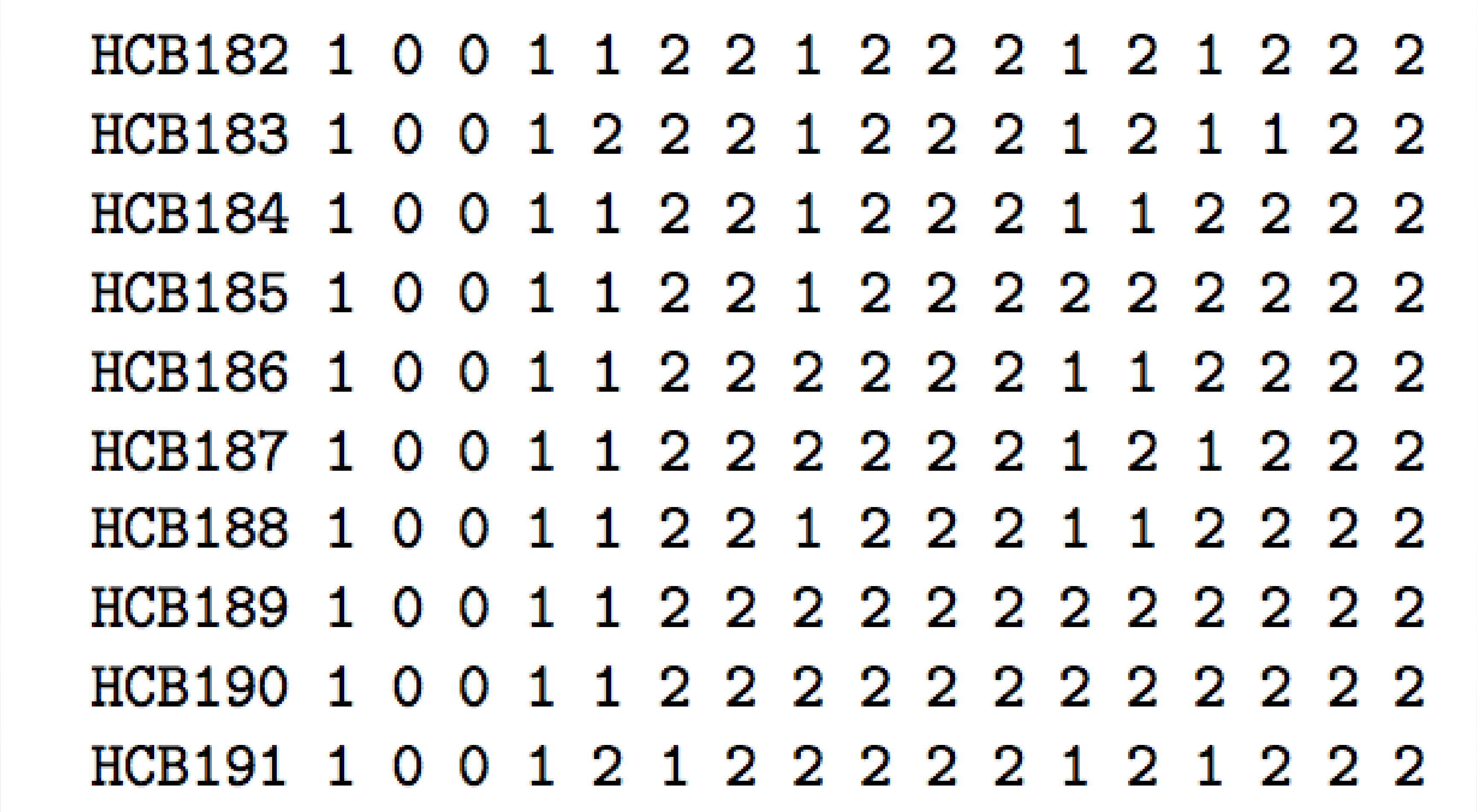
- PLINK has multiple data formats for GWAS data

- PED format
  - PED files contain individual information
  - MAP file contains SNP information

PLINK data formats

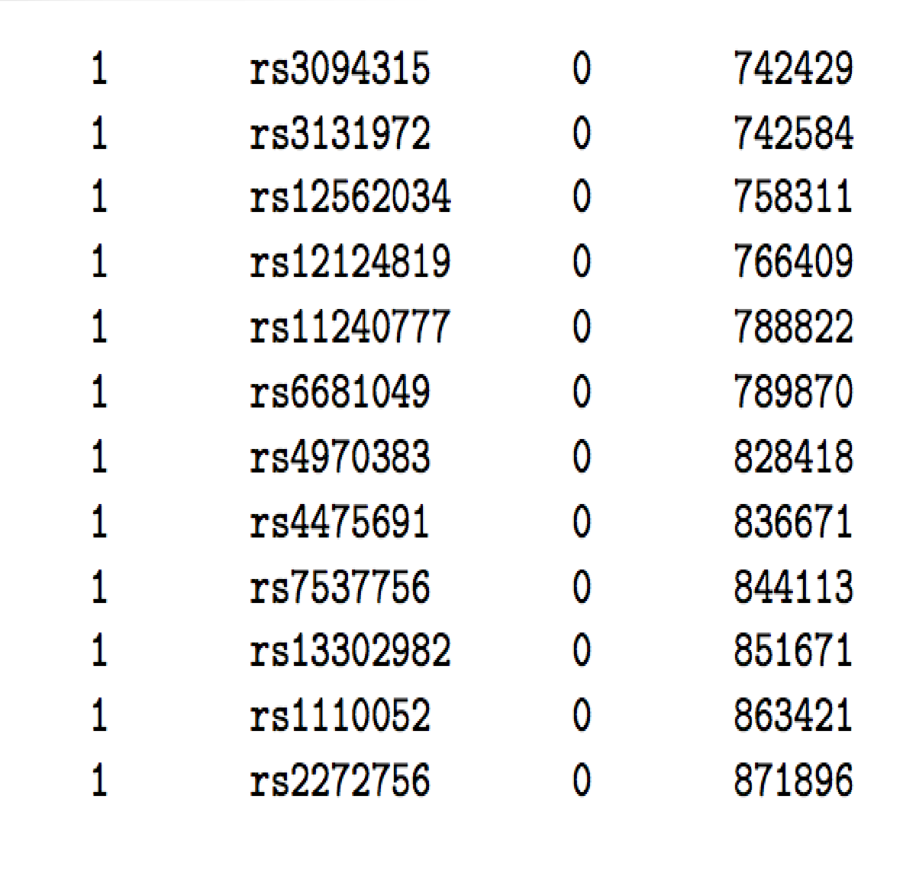
- PED file
  - One row per individual, Defined set of columns:
    - Family ID
    - Individual ID
    - Paternal ID
    - Maternal ID
    - Sex (1=male, 2=female, other=unknown)
    - Phenotype (missing = -9, control=1. case=2, or QT values)
    - Pair of columns per SNP – Different encoding formats

PED File



```
HCB182 1 0 0 1 1 2 2 1 2 2 2 1 2 1 2 2 2
HCB183 1 0 0 1 2 2 2 1 2 2 2 1 2 1 1 2 2
HCB184 1 0 0 1 1 2 2 1 2 2 2 1 1 2 2 2 2
HCB185 1 0 0 1 1 2 2 1 2 2 2 2 2 2 2 2 2
HCB186 1 0 0 1 1 2 2 2 2 2 2 1 1 2 2 2 2
HCB187 1 0 0 1 1 2 2 2 2 2 2 1 2 1 2 2 2
HCB188 1 0 0 1 1 2 2 1 2 2 2 1 1 2 2 2 2
HCB189 1 0 0 1 1 2 2 2 2 2 2 2 2 2 2 2 2
HCB190 1 0 0 1 1 2 2 2 2 2 2 2 2 2 2 2 2
HCB191 1 0 0 1 2 1 2 2 2 2 2 1 2 1 2 2 2
```

MAP file



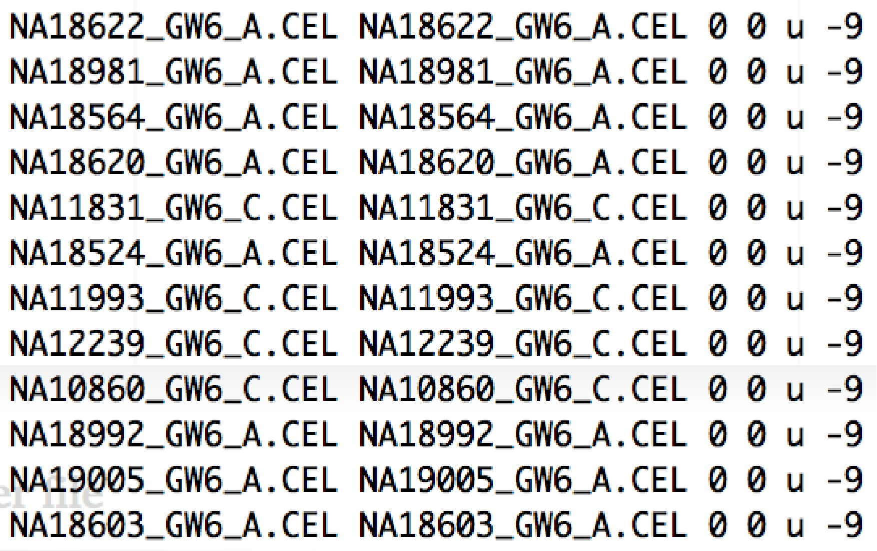
```
1      rs3094315     0     742429
1      rs3131972     0     742584
1      rs12562034    0     758311
1      rs12124819    0     766409
1      rs11240777    0     788822
1      rs6681049     0     789870
1      rs4970383     0     828418
1      rs4475691     0     836671
1      rs7537756     0     844113
1      rs13302982    0     851671
1      rs1110052     0     863421
1      rs2272756     0     871896
```

- One row per SNP, set of defined columns
  - Chromosome number 1..26 (X, Y, XY, MT)
  - SNP ID (dbSNP)
  - Genetic distance (Morgans)
  - Base pair position

Binary PED format

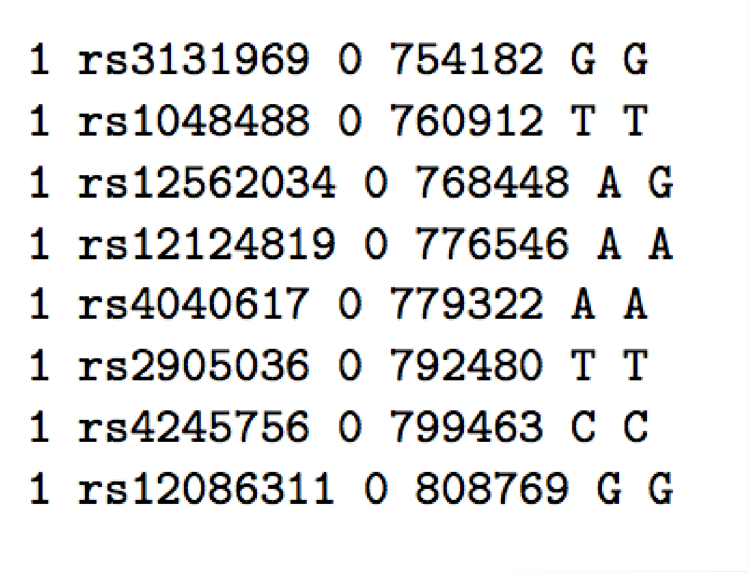
- Faster, more efficiently accessible and compact format
  - FAM file
    - One row per individual – identification information – first 6 columns of PED file). Human readable
  - BIM file
    - One row per SNP. MAP file PLUS the two alleles for that SNP. Human readable
  - BED file
    - One row per individual – genotype information (rest of the columns of the PED file). Not human readable

FAM file



```
NA18622_GW6_A.CEL NA18622_GW6_A.CEL 0 0 u -9
NA18981_GW6_A.CEL NA18981_GW6_A.CEL 0 0 u -9
NA18564_GW6_A.CEL NA18564_GW6_A.CEL 0 0 u -9
NA18620_GW6_A.CEL NA18620_GW6_A.CEL 0 0 u -9
NA11831_GW6_C.CEL NA11831_GW6_C.CEL 0 0 u -9
NA18524_GW6_A.CEL NA18524_GW6_A.CEL 0 0 u -9
NA11993_GW6_C.CEL NA11993_GW6_C.CEL 0 0 u -9
NA12239_GW6_C.CEL NA12239_GW6_C.CEL 0 0 u -9
NA10860_GW6_C.CEL NA10860_GW6_C.CEL 0 0 u -9
NA18992_GW6_A.CEL NA18992_GW6_A.CEL 0 0 u -9
NA19005_GW6_A.CEL NA19005_GW6_A.CEL 0 0 u -9
NA18603_GW6_A.CEL NA18603_GW6_A.CEL 0 0 u -9
```

BIM file



```
1 rs3131969 0 754182 G G
1 rs1048488 0 760912 T T
1 rs12562034 0 768448 A G
1 rs12124819 0 776546 A A
1 rs4040617 0 779322 A A
1 rs2905036 0 792480 T T
1 rs4245756 0 799463 C C
1 rs12086311 0 808769 G G
```