Variation – the issues

What kind of variation is there ?

- diseases / harmless
- detection

Association

types of variation

- genome structure
- repeat variation
- single nucleotide polymorphisms (SNP)

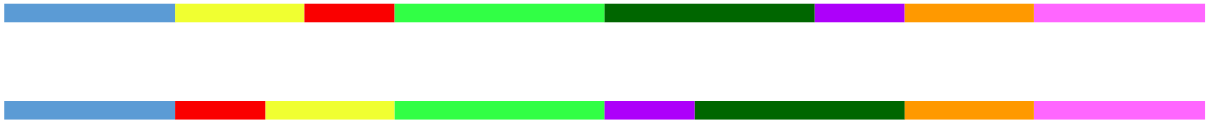
A few words on each

- remember – they are not usually associated with diseases
  - variation is common – but we are mostly healthy
  - really deadly problems are rare

Genome structure variation

Something like

- me
- you



- deletions, tandem duplications, inversions, mobile element insertions

- healthy people – 10^7 bases affected

How hard to detect ?

- might show up as hard to align if you are using a reference genome

Copy number variation

Can I convince you that this is common ?

- DNA testing (very short repeats)
- Similar copies of enzymes
  - how did α- and β-haemoglobin (and more) arise ?

Can you detect this ?

- remember problems with sequencing repeats
- can see it if you look for it (paired ends, electrophoresis…)

Common – often harmless

Single Nucleotide Polymorphisms, SNP, SNV

Mostly harmless – otherwise you would be sick

How many sites ?

- > 10^6 in a human genome

How many mutations do you have ?

- best estimates from sequencing 20-200
- estimates via mutation rates
  - $1\text{-}2\times10^{-8}$ mutations per site per generation
  - genome has 6×10^9 bases $\approx$60 – 120 mutations

What is in the SNP databases ?

- Your mutations ? probably not
- variance found in the population

SNP collections

How are they found ?

- sequencing and genome assembly

- missing bits ? probably just missing from assembly
- mismatch either
  - real variation or
  - read error

First problem – variation looks similar to an error

- There are databases of SNPs
- DNA chips with common SNPs – easy to buy

Diseases and Associations

Assumptions

- Most people are healthy
- diseases are removed by selection
- the average genome is a healthy reference

Too simple

two minutes on gene fixation

- some mutation is harmless (blue eyes to green eyes)
- it happens once – will everybody end up with green eyes ?

For a neutral mutation we might end up with green eyes

$p_{fix} = \frac{1}{2N_{pop}}$ (bit bigger than 0)

but it takes a long time and is unlikely

Will the mutation disappear ?

$p_{disappear} = 1 - p_{fix} = 1 - \frac{1}{2N_{pop}}$ which is a bit less than 1

- harmless mutations do occasionally become fixed

slightly bad mutations

Neutral mutation might become fixed

What if the mutation is very slightly bad ?

$p_{fix} = \frac{1}{2N_{pop}}$ changes, but not much

- if a mutation is slightly bad, it may also become fixed

Slightly bad mutations (very small affect on reproduction)

- colour blindness

bit worse

- albinism, phenylketonuria, haemophilia

Hamburg is full of non-optimal mutations

- is this relevant to heart disease or diabetes ?

Flow of mutations

Neutral mutations usually disappear (after a few generations)

- statistical effect

Neutral and deleterious mutations constantly reappear

- 20 – 200 per person per generation
- flow of variation (new created, old is purged)

What do you see in SNP collections ?

- SNPs that occur in many people

Are they functionally important ? 80-90 % no function

- fact will come back later

Mendel / non-Mendel inheritance

Mendelian inheritance is not often relevant

Consider conditions

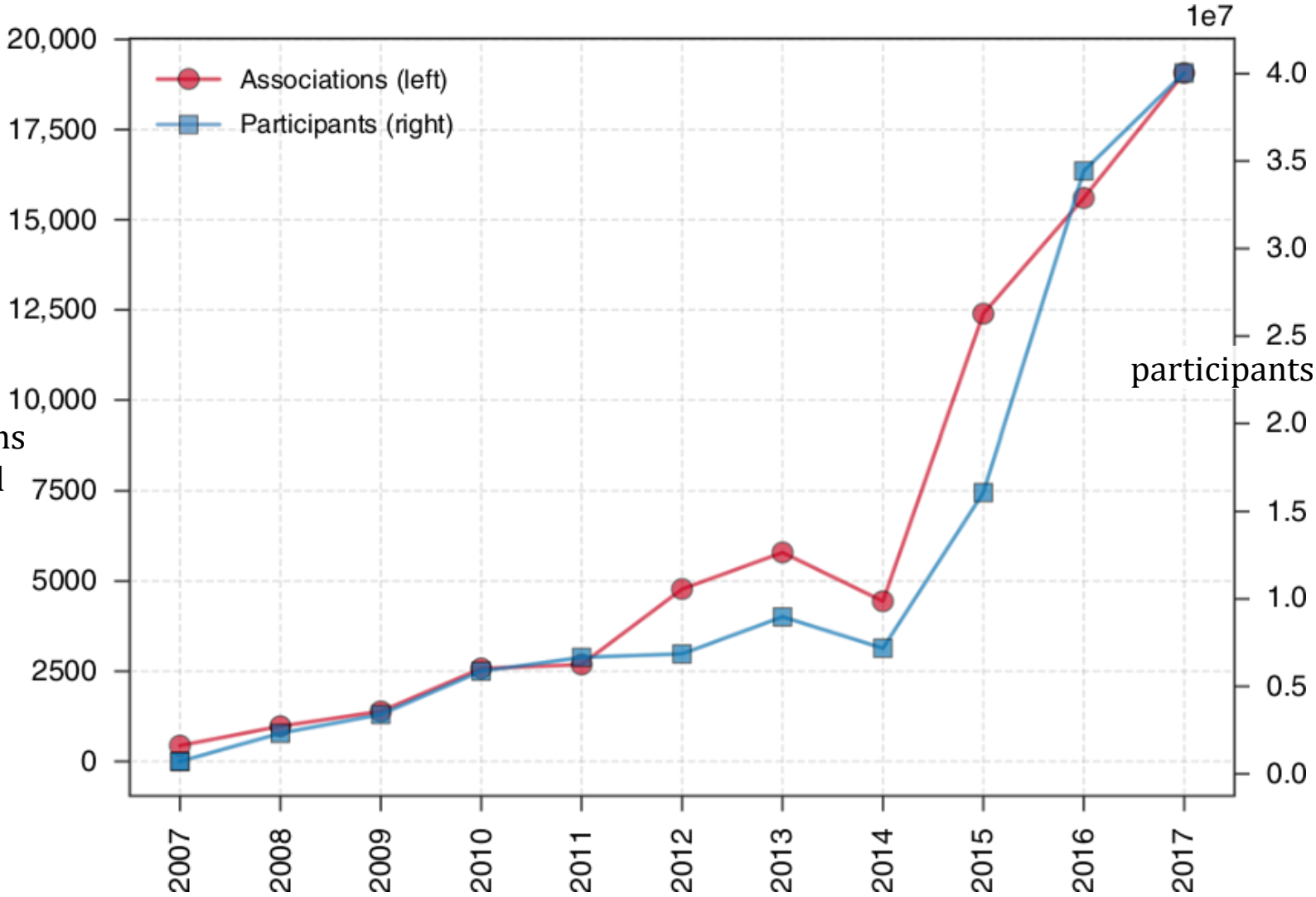
- with family component
- slightly deleterious – very small impact on reproduction
- possibly big sociological impact or happiness cost
  - heart disease, schizophrenia, diabetes, obesity, bipolar disorder, later onset multiple sclerosis, …

Environment and genetics

- alcoholism ? heart disease (I eat what my parents ate) ?

Big genetic component – but not simple Mendelian

Genome wide association data



associations discovered

participants

Mills, M.C., Rahal, C. Commun. Biol. 2, 9 (2019)

Genome Wide Association

Is an allele at a position correlated with disease incidence ?

Naïve approach

- take many people (10^4)
- sequence genomes
- compare every genome with every other
- find sites that differ
- is an allele present in sick people more than by chance ?

Not possible

- Cannot do good sequencing on so many people
- computationally gigantic
- practical approach..

Practical association studies

Cannot do whole genomes

Work with SNP data

- databases with SNPs (10^7)
- more problematic – do not work with all known SNPs – use
  - those the investigators like
  - those in some kit

The choice of candidate variations is not unbiased

Measurements and types

Collect lots of

- samples – genetic information
- phenotypes
- look at data

Two kinds of analysis

1. discrete / binary or categorical
   - had heart attack before 45, blue eyes
2. continuous (quantitative trait)
   - height, weight, pulse rate, blood pressure

Different statistical methods – conceptually the same

- my examples – just a binary trait (sick / healthy)

Looking at results

One site example

- one allele `A` or `a`, but we have two copies of the gene
  - `aa`, `aA`, `Aa`, `AA`
- phenotype ? Say healthy / disease

Example question

- is `aa` associated with disease ?

- invent some data and look at it 3 ways

Data

1000 people

- 200 have `aa` - 20 % / 0.2
- 800 are something else `aA`, `Aa`, `Aa`
- `aa` get sick more often – 40/200 = 20 % (ratio = $\frac{40}{160} = \frac{1}{4}$)
- `aA`, `Aa`, `Aa` get sick 80 / 800 = 10 % (ratio = $\frac{80}{720} = \frac{1}{9}$)
- average sickness is 120 or 12 %

	healthy	sick	total
`aa`	160	40	200
`aA`, `Aa`, `AA`	720	80	800
total	880	120	1000

1. odds ratio

	healthy	sick	total
aa	160	40	200
aA, Aa, AA	720	80	800
total	880	120	1000

aA, Aa, AA group 10 % sick (ratio $\frac{1}{9}$)

aa group 20 % sick (ratio $\frac{1}{4}$)

The ratio ? $\frac{1/4}{1/9} = 2.25$

Looks like aa = more than twice the chance of being sick

- does not account for sample size
  - need a better statistic

2. estimate p-value

	healthy	sick	total
aa	160	40	200
aA, Aa, AA	720	80	800
total	880	120	1000

- compare n_{obs} with expected n_{exp}
- for aa group, $n_{exp} = 0.12 \times 200 = 24$

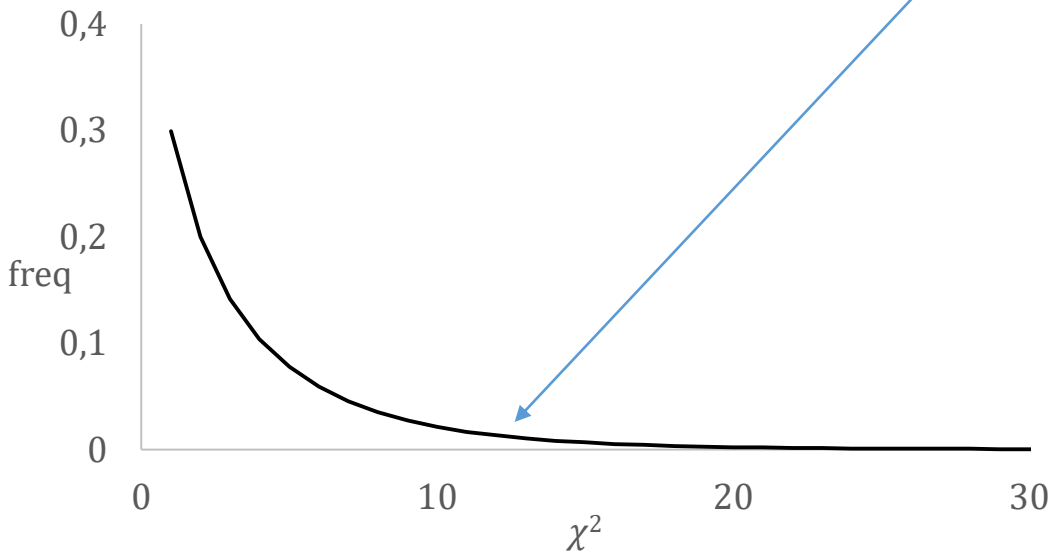
	aa	aA, Aa, aA	
n_{obs}	40	80	
n_{expect}	24	96	
diff	16	16	
$\frac{(n_{obs} - n_{expect})^2}{n_{expect}}$	10.7	2.7	$\chi^2 = 13.3$

2. estimate *p*-value

	aa	aA, Aa, aA	
n_{obs}	40	80	
n_{expect}	24	96	
diff	16	16	
$\frac{(n_{obs} - n_{expect})^2}{n_{expect}}$	10.7	2.7	$\chi^2 = 13.3$

$$\frac{(n_{obs} - n_{expect})^2}{n_{expect}}$$

properties..



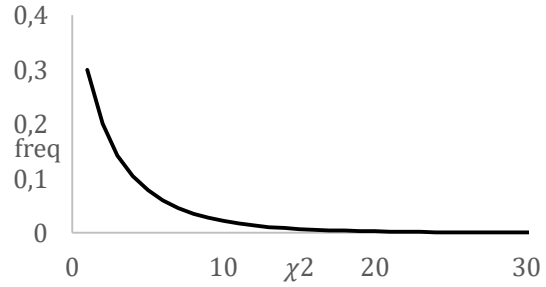
2. estimate p-value (χ^2)

$\frac{(n_{obs}-n_{expect})^2}{n_{expect}}$ is part of the χ^2 formula

Captures two properties

- how unusual is something $n_{obs} - n_{expect}$
- how much data do I have ?

Can be converted to a p-value via distribution



not for klausur

- summed over categories
- uses known distribution of χ^2
  - depends on how many categories (degrees of freedom)

3. brute force - permutation

	healthy	sick	total
aa	160	40	200
aA, Aa, AA	720	80	800
total	880	120	1000

original data: 110/1000 are sick (0.11)

- repeat 10000 times
  - repeat 200 times
    - x is random number [0:1]
    - if $x < 0.11$ $n_{sick\_\mathrm{aa}} = n_{sick\_\mathrm{aa}} + 1$
  - repeat 800 times
    - x is random number [0:1]
    - if $x < 0.11$ $n_{sick\_\mathrm{other}} = n_{sick\_\mathrm{other}} + 1$

From 10000 experiments, how often is $n_{sick\mathrm{aa}} > 40$?

summarise looking at data

Odds ratios

- intuitive
- often seen in papers – often as log $\left(\frac{ratio_{group\,1}}{ratio_{group\,2}}\right)$
- does not give an idea of significance

Significance statistic p-value

- most seen in papers
- correct statistics a bit more complicated
- no idea of size of effect

permutation tests

- easy to apply to more complicated questions (`aa` vs `aA`, ..)
- computational time high
- very general – you can use it if you do not like statistics

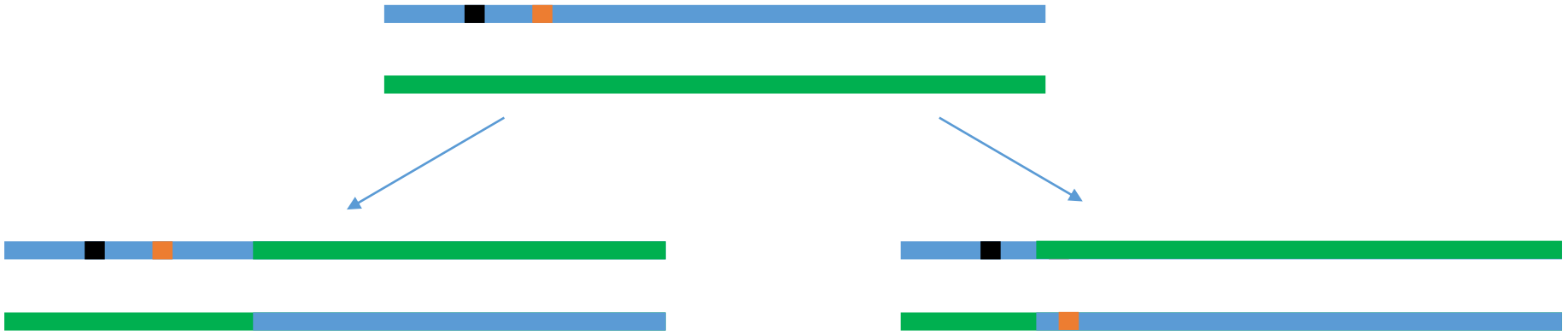
Simple examples

Vorsicht - my examples

- `aa` vs `aA`, `Aa`, `AA`, maybe you are interested in `aa`, `aA` vs `AA`
- χ^2 becomes more complicated with more variables
- can be improved upon

Genetic Linkage

You are not usually looking for an exact gene



linkage preserved

linkage broken

What is the probability of breaking linkage ?

$p_{breakage} \propto d$

If distance d is very small, ■ ■
a variant in one is probably in the other

You find ■ is associated with disease

- you accept that the relevant gene is somewhere nearby

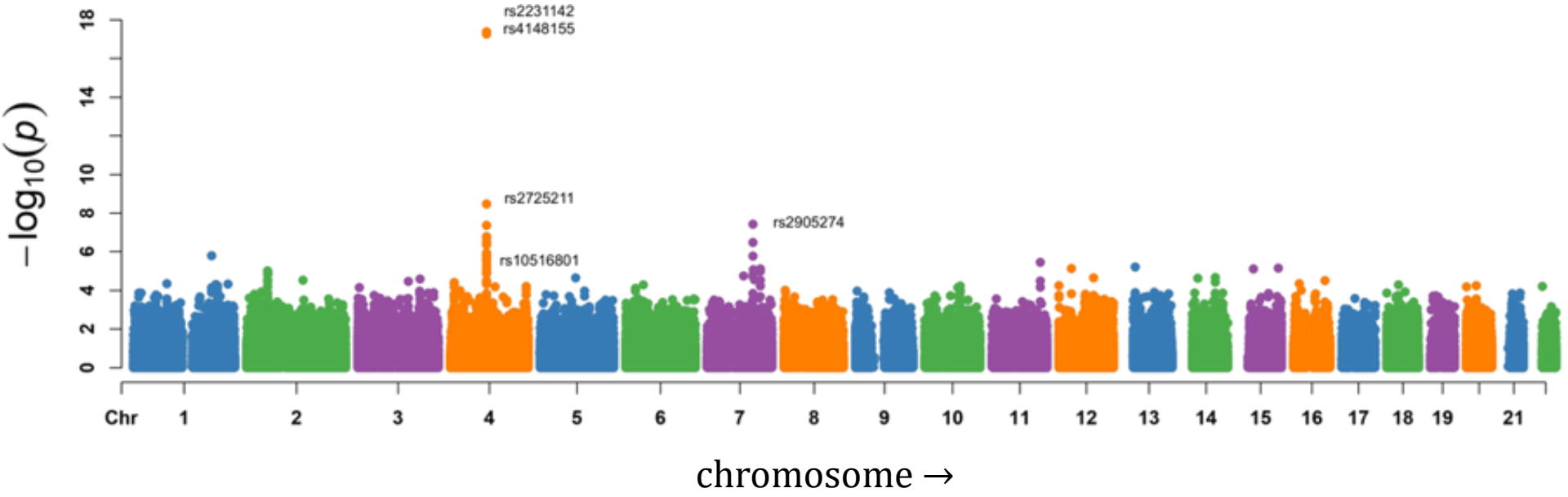
Goal – do not worry about finding an exact gene

- good result – many nearby sites correlated with disease

Consequence for site / SNP selection

- lots of sites, reasonably distributed – example …

Looking for gout (gicht) example



Lee, M-T., Hsu, .. Lee, C-C. Sci. Rep 9, 4981 (2019)

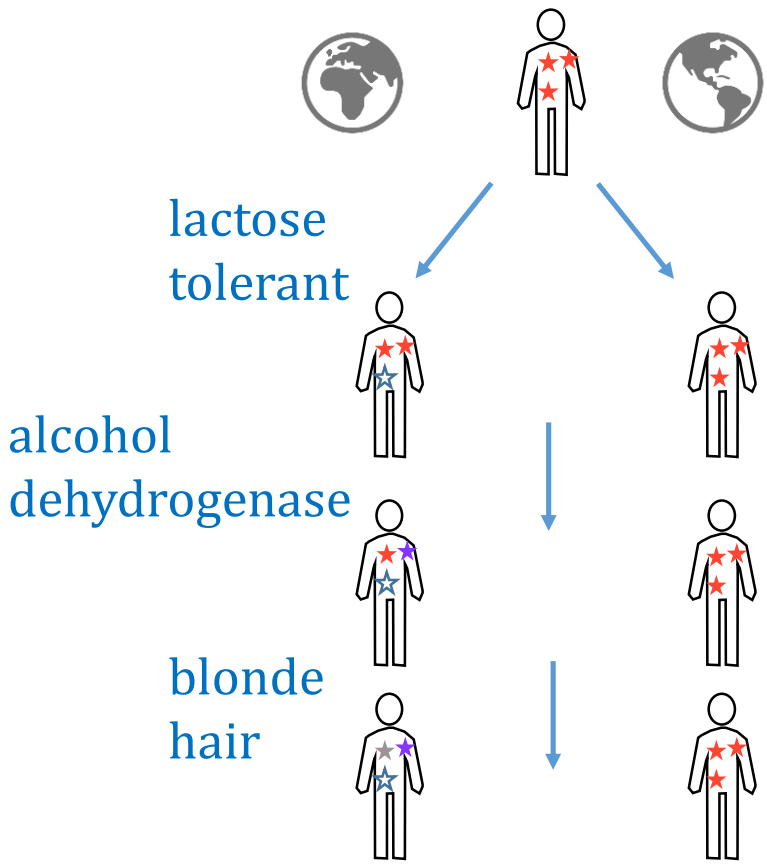
geographical linkage

Linkage may not be due to genetic organisation

- study lactose tolerance
- you find a correlation with a gene for blonde hair

Non-genetic linkage ?

- geography, sociology, religion



p-values problems

Sounds easy. Can be tricky (p probability)

- probability of something being seen by chance ?
- $p = 10^{-3}$ sounds good, $p = 0.3$ sounds insignificant

Clinical studies ?

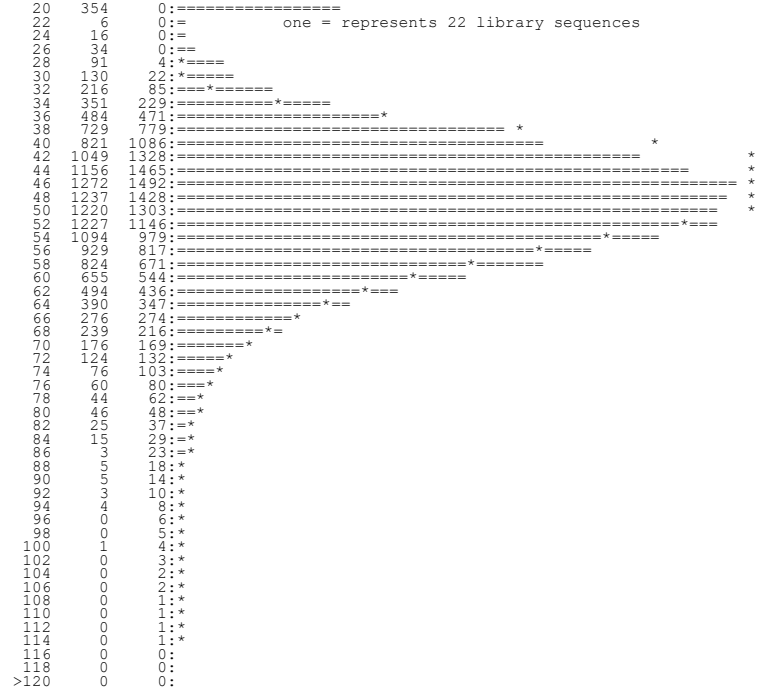
- $p = 0.05$ sacred value, significant / insignificant

More formal

- what is the chance of seeing a particular observation if the null hypothesis is true ?
- example from sequence analysis

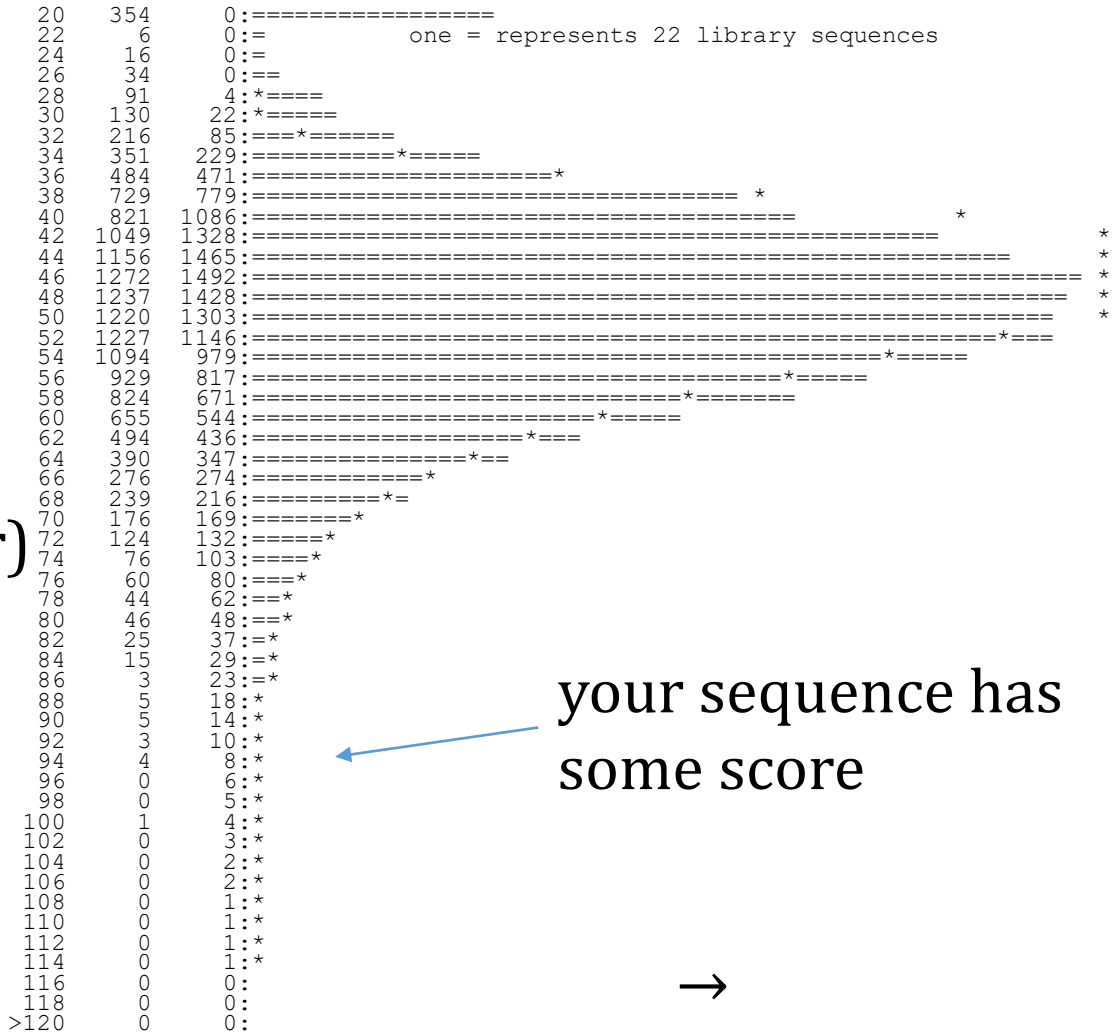
p-value example (blast)

- blast says a sequence match has $p = 10^{-9}$
- where does it come from ? what does it mean ? what is the null hypothesis ?
- Align query to databank, save scores and make histogram



the null hypothesis ?

- your protein has similarity by chance



scores
(↓ are better)

your sequence has some score

→

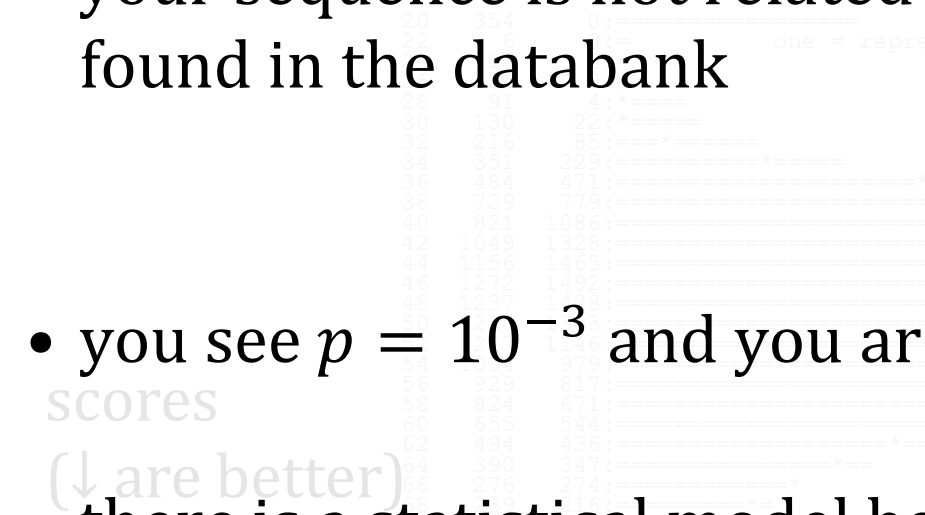
number of times seen in databank

The null hypothesis

- your sequence is not related to the one found in the databank

- you see $p = 10^{-3}$ and you are happy

- there is a statistical model behind this (distribution of scores)



Another example

p-value in clinical trial

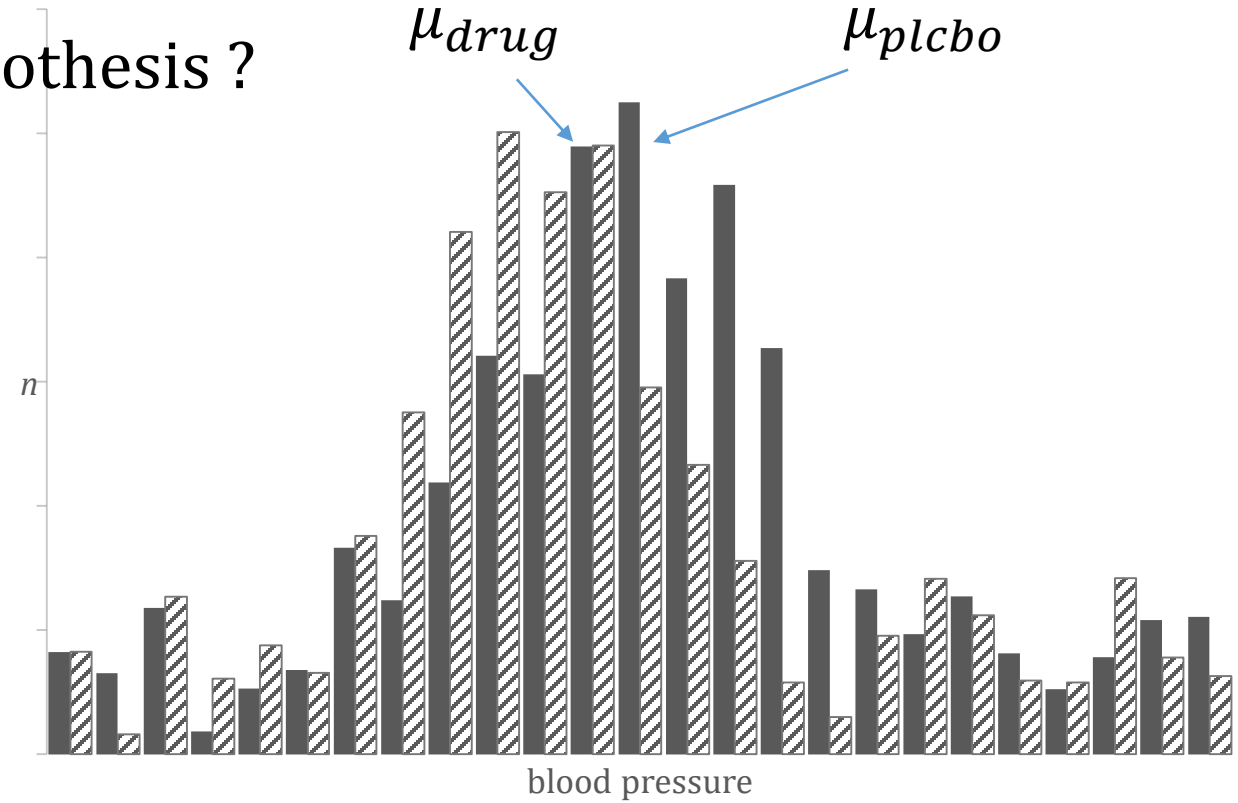
Take n subjects… $\frac{n}{2}$ get new drug, $\frac{n}{2}$ get placebo

- my statistical model considers μ (mean) … μ_{drug} and μ_{plcbo}
  - μ is blood pressure, memory recall, sugar in blood…

Has my drug helped ?

- what is the null hypothesis ?

$$\mu_{drug} = \mu_{plcbo}$$



Back to genomes..

use a t test

multiple test p-values

You like significance of $p < 0.05$

- you test 20 drugs and declare them helpful (significant)

- One of your treatments is of no value

Genome analogy

- you find many sites correlated with a disease $p = 10^{-3}$
- how many tests did you really do ?
- looked at 500 000 SNPs, you are 500 000 times more likely to find a significant correlation

What do people often do ?

p-value in association studies

Say $p < 0.05$ is significant (for one test)

- divide by a 10^6 (pretend you have done a million tests)
- typical literature requirement is $p < 0.05 \times 10^{-6}$

$$p < 5 \times 10^{-8}$$

Obviously arbitrary

p-value and size of effect

Question you asked

- is this allele / SNP correlated with disease ?

Effect size

- Take lots of people (good study)
- measure heart disease rates
- allele A – group has 0.1% disease
- allele B – group has 0.15 % disease incidence
- is this a 50 % increase ? OR
- increase from small to a bit less small ?

p values alone do not tell you if something is important

p-value reliability

Back to clinical trial

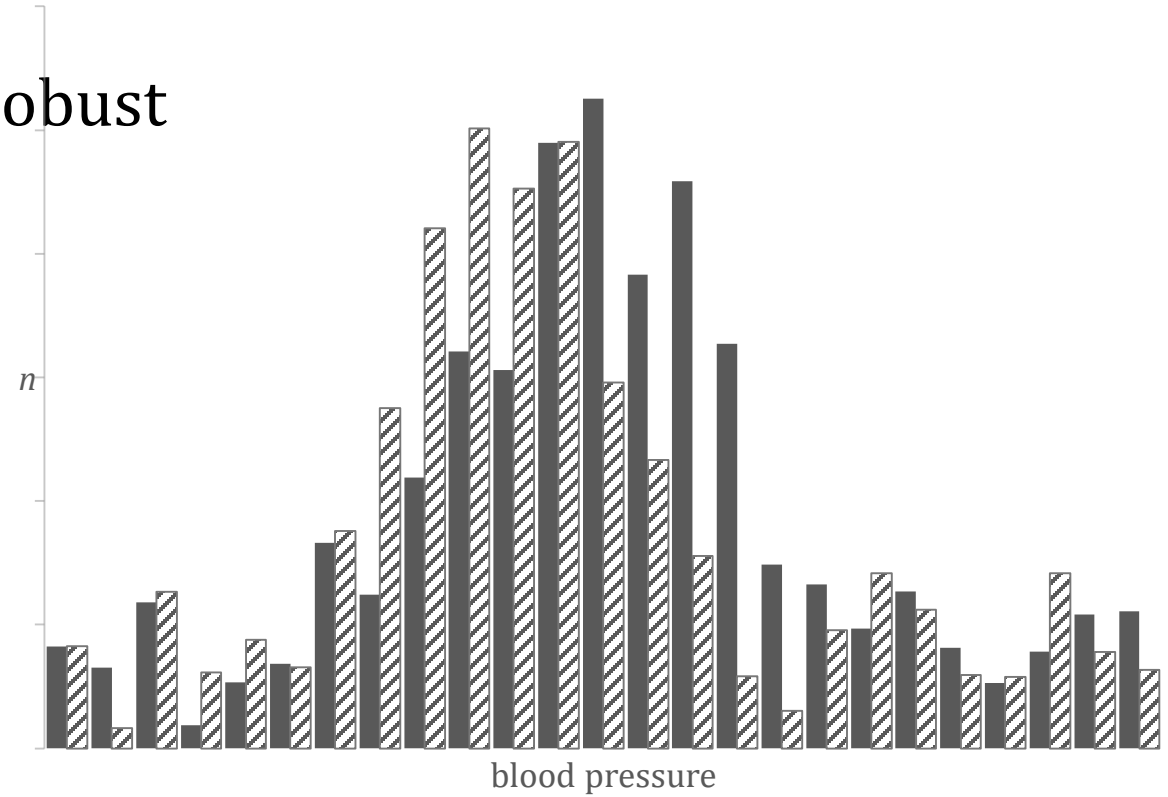
- you calculate $p = 0.05$

Change allocation of patients to groups

$p = 0.4$ or $p = 0.6$

- p values may not be robust

Association studies are sensitive to noise



multi loci statistics

First … one locus (Mendelian)

- either `A` or `a` alleles
- say `a` is present in 0.3 % of population
  - `aa` is present in 0.003 × 0.003 = 0.1 % of population
  - usually causes detectable health problem
  - in Hamburg 2×10^3 cases (common)

You do a study

- every `aa` is associated with disease – not too bad

Now consider if the disease condition requires two sites..

multi locus version

two genes each two variants **aa**/**aA**/**AA**, **bb**/**bB**/**BB**

a	18 %
A	82 %
b	18 %
B	82 %

- you have one of **aabb**, **aAbb**, **Aabb** …
- only **aabb** leads to disease
  - how often do you see this ?
    $0.18 \times 0.18 \times 0.18 \times 0.18 = 0.1\%$
    exactly the same as 0.1 in previous example
- how often do you see the combinations ? …

aa bb	0.18^4	0.1 %	sick
aA bb	$0.18 \times 0.82 \times 0.18 \times 0.18$ (× 2)	1.0 %	healthy
AA bb	$0.82 \times 0.82 \times 0.18 \times 0.18$	2.2 %	healthy
aa bB	…	1.0 %	healthy
aA bB	$0.18 \times 0.82 \times 0.18 \times 0.82$ (×4)	8.7 %	healthy
AA bB	…	19.8 %	healthy
aa BB	…	2.2 %	healthy
aA BB	$0.18 \times 0.82 \times 0.82 \times 0.82$	19.8 %	healthy
AA BB	0.82^4	45.2 %	healthy

aa	**bb**	0.18^4	0.1 %	sick
aA	**bb**	$0.18 \times 0.82 \times 0.18 \times 0.18$ (× 2)	1.0 %	healthy
AA	**bb**	$0.82 \times 0.82 \times 0.18 \times 0.18$	2.2 %	healthy
aa	**bB**	...	1.0 %	healthy
aA	**bB**	$0.18 \times 0.82 \times 0.18 \times 0.82$ (×4)	8.7 %	healthy
AA	**bB**	...	19.8 %	healthy
aa	**BB**	...	2.2 %	healthy
aA	**BB**	$0.18 \times 0.82 \times 0.82 \times 0.82$	19.8 %	healthy
AA	**BB**	0.82^4	45.2 %	healthy

- 32.8 % have an **a**, but they are nearly all healthy
- 3.3 % have an **aa**, but even they are nearly all healthy
- really only see sickness if you look for **aabb**

  but if you have 50 000 genes where would you look ?

two locus version is bad

You will only see the disease if you compare

`aabb` vs `aabB`, `aaBB`, `aAbb`, …

At the start, have no idea which pairs are bad

- imagine DNA chip lets you look at 50000 sites
- you do not know which sites are correlated

What if I have a 3-locus disease ?

Consequence

- multi-locus effects are hard to find
- to be found – a locus by itself should have a detectable effect

summarise problems with numbers

A paper says they found 20 new genes associated with schizophrenia

1. If you look at enough markers, you will find some with small p-values
   Use a corrected p threshold
2. One must look at size of effect measured
3. p-values are rarely stable
4. multi loci phenotypes will be computationally hard to detect

problems – biases and confusion

Obvious biases

- data from rich countries
  - ¾ are from US, UK or Iceland
  - > 90% white

Calculation of expected values

- everything depends on "how different from random ?"
- gene A/a pair - you find sick/healthy

What is broken ?

expected values

Hamburgers have

- Labskaus-ase gene `L`
- 90 % have defective Knödelase `k`

Bavarians have

- 90 % `l`, 90 % `K`

Average has 50/50 `L`/`l`, `K`/`k`

You do a study of tolerance of Knödel-Toleranz

- find strong association with Labskau and Knödel genes

What is wrong ? Two ways to look at it

1. geographical linkage
2. use wrong n_{expect} in $\frac{(n_{obs}-n_{expect})^2}{n_{expect}}$

expected values

What should you do ?

- Amongst Hamburgers (regardless of Knödeltoleranz)
  - get `l`/`L` and `k`/`K` ratios – use these to calculate $n_{expected}$

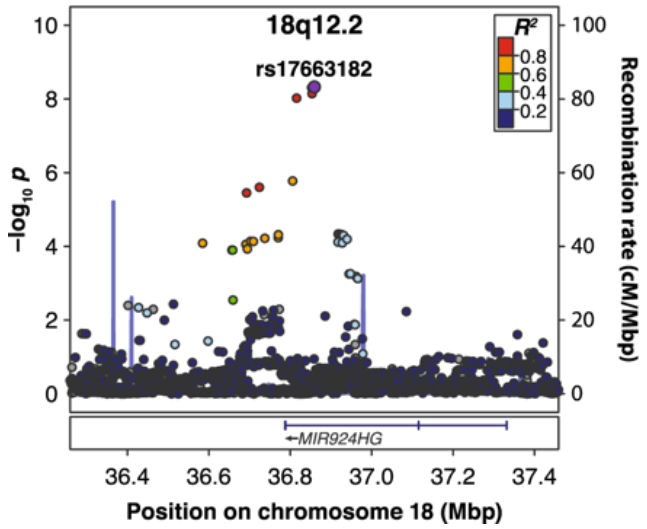
Common term for uneven background distribution

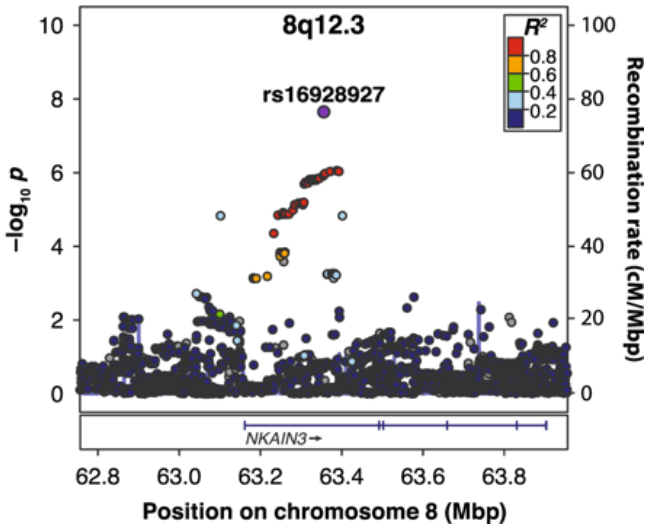
- stratification

Problem any time you have different ethnic groups in one study

phenotype problems

- is bipolar disorder easy to diagnose ?
- is there one kind of multiple sclerosis ?
- can you quantify dyslexia ? or perhaps people who admit they cannot spell





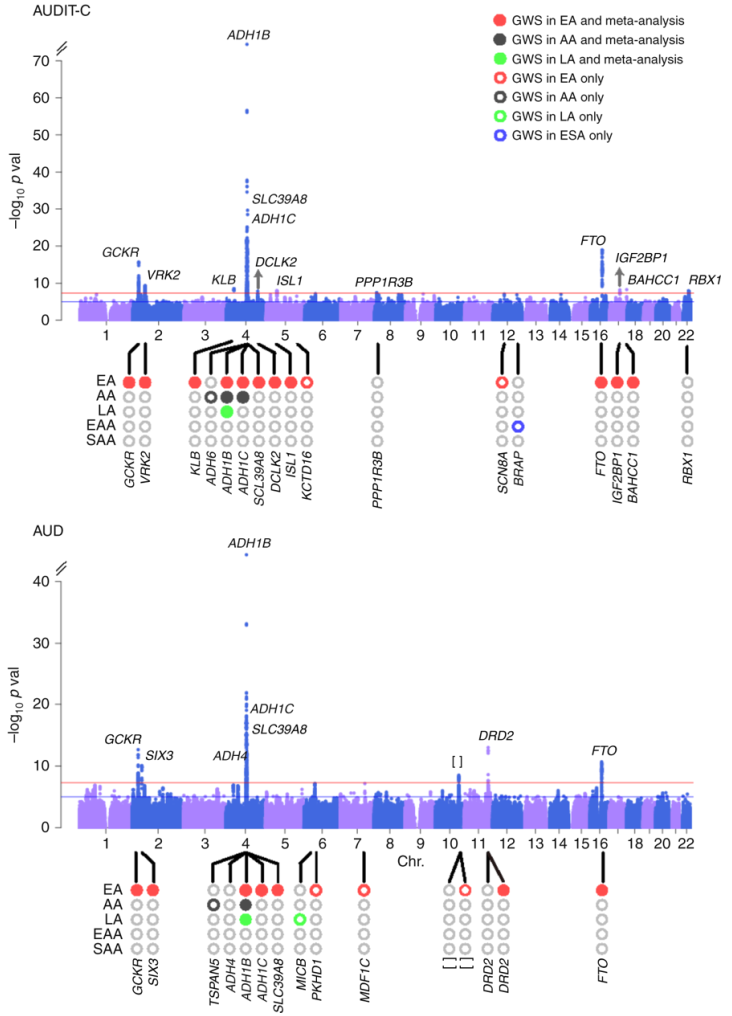
Dialluisi, A., Andlauer, F.M., …Schulte-Körne, G, Trans Psych, 9, 77 (2019)

Environment

Strong family element in

- alcoholism
- weight, …

- genetic or sociological ?



Kranzler, H.R., Zhou, H…. Gelernter, J. Nat. Commun., 10, 1499 (2019)

Summarise problems

- p-values, robustness
- stratification - linkage due to sociology, geography
- phenotypes
  - difficult to measure
  - genetic mixture

Forensics

Ask grandmother about bioinformatics

- genome sequences

Ask what DNA is for ?

- forensics

Why

- Tatort
- disaster identification
- paternity
- …

DNA

We are ≈ 99.7 % identical (bases) … about 3×10^{-3} different

- $3\times10^{-3}\times 3\times10^{9} \approx 10^{7}$ different sites
- lots of room for identification

Philosophy

- not too expensive
- work on partially degraded DNA
- statistics can be assessed
- clear, discrete results – no smeared electrophoresis
  - no difficult / expert interpretation

Early approaches

Sequencing methods ? Problems

- slow
- expensive
- what would you look at ?
  - SNPs – typically two / three variables at a site (`aa`, `aA`, `AA`)

Restriction enzyme patterns

Individuals are different, but

- bands are not easily quantified

Alleles – not good

Lots of single nucleotide polymorphisms

- you have AA**A**A I have AA**C**A – just two variants
- 10 sites gives 1 in 2^{10}=1024 (not enough)
- 20 sites gives 1 in 10^6 (not really enough)

Real statistics worse - Imagine 80 % population is A, 20% a

- 0.8^{10} = 0.11 so more than 10 % of people in one class

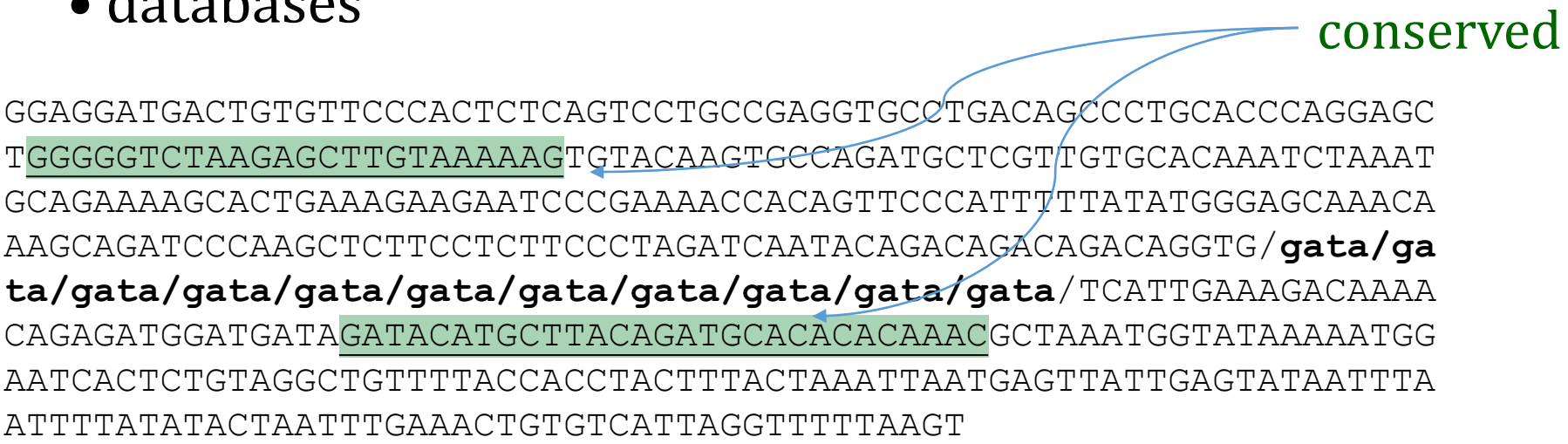
Technical problems – how would you check quickly ?

- sequencing of regions ?
- microarrays ? Probably will not work for single bases

Short tandem repeats (STR)

Properties – details soon

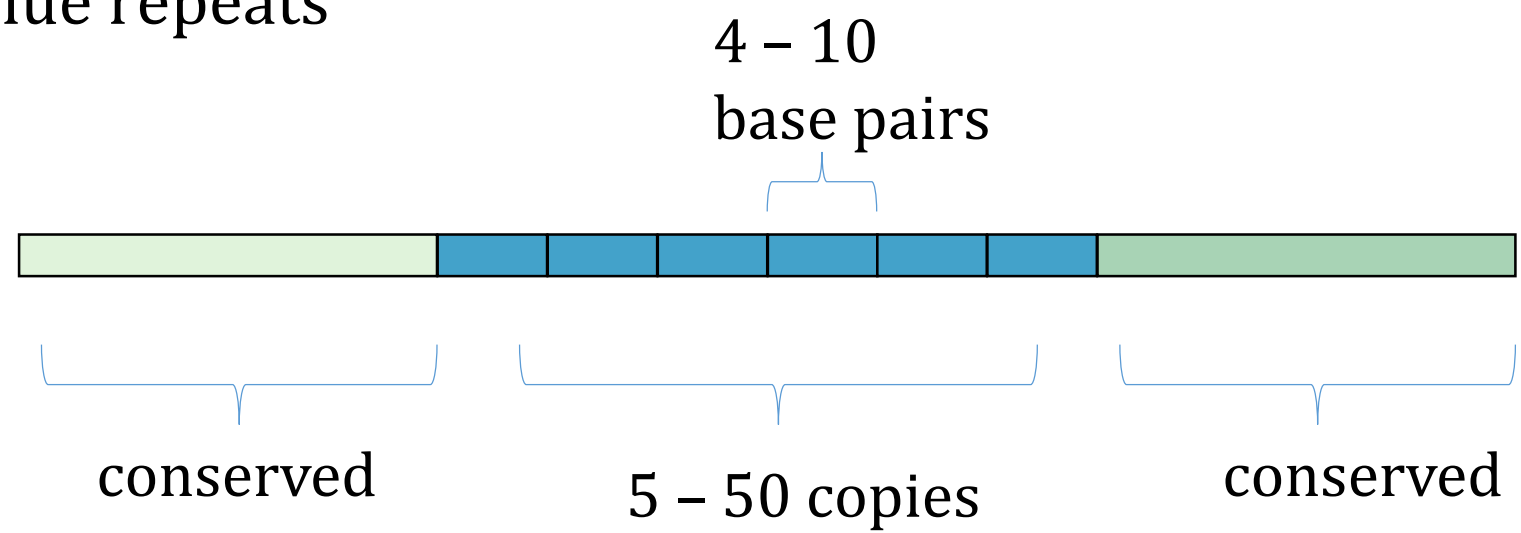
- very discrete – you have 10, 11, 12… of something
- not just two possibilities
- should be present in everybody
- can be quickly measured
- can be standardised
  - courts can be given agreed statistics
  - databases



STR general structure

Conserved regions can be used for primers

- copy / amplify just these pieces
- count blue repeats



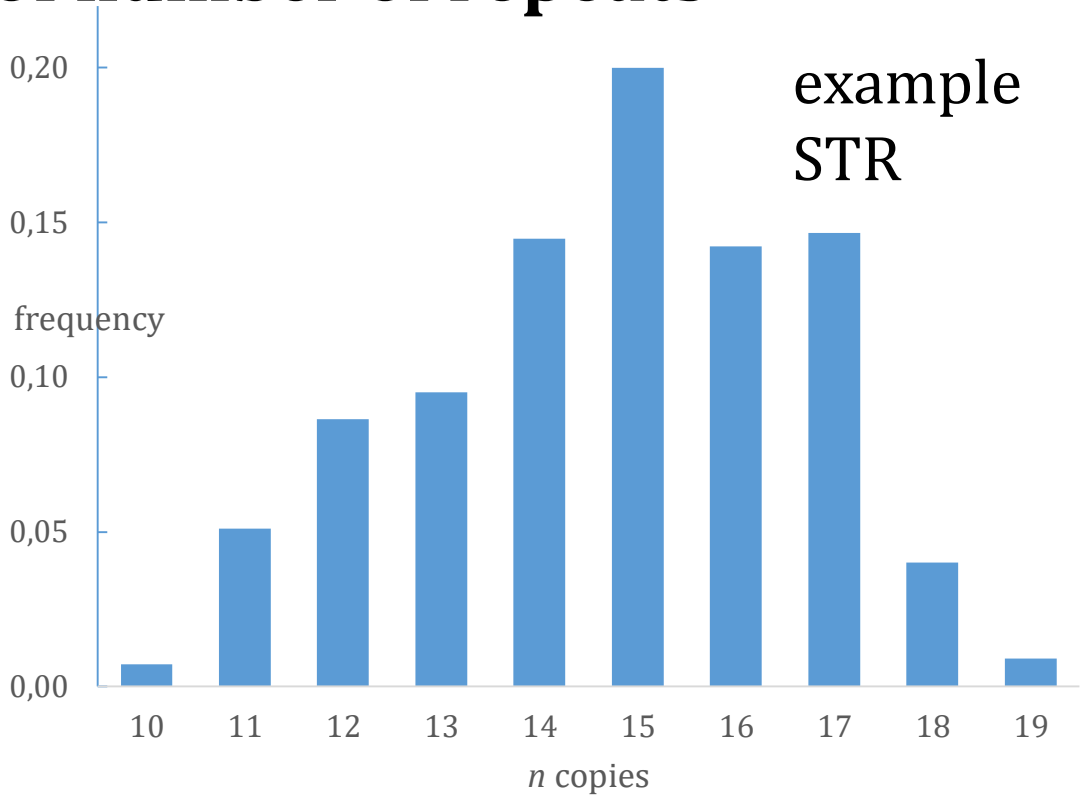
Typical length of amplified regions

- 100 – 300

Example…

Distribution of number of repeats

Similar properties for other STR's



Where do numbers come from ?

From many samples collect frequencies of copy numbers

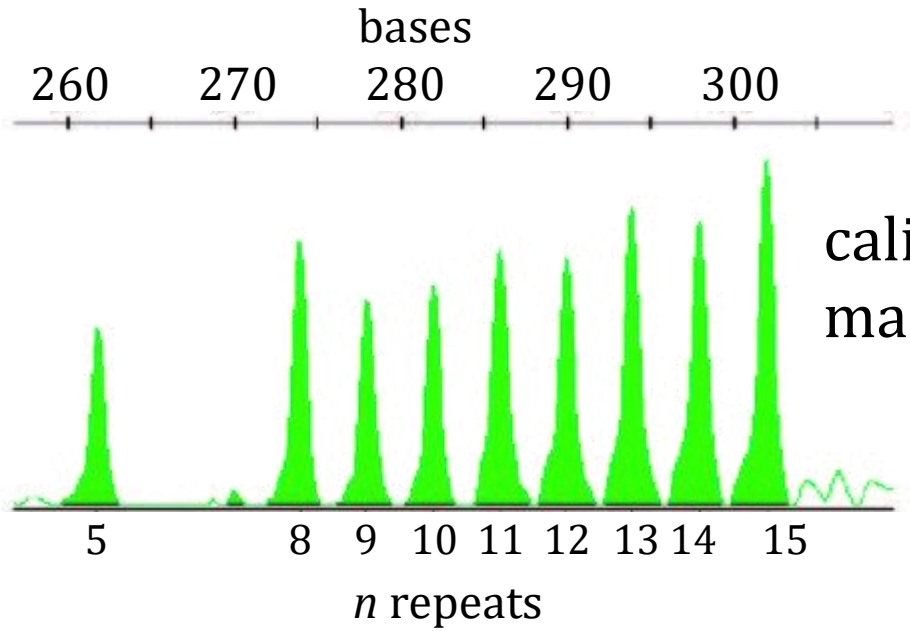
- purely empirical

How easy is it to count the copies ?

Counting repeats

GGAGGATGACTGTGTTCCCACTCTCAGTCCTGCCGAGGTGCCTGACAGCCCTGCACCCAGG
AGCTGGGGGTCTAAGAGCTTGTAAAAAGTGTACAAGTGCCAGATGCTCGTTGTGCACAAAT
CTAAATGCAGAAAAGCACTGAAAGAAGAATCCCGAAAACCACAGTTCCCATTTTTATATGG
GAGCAAACAAAGCAGATCCCAAGCTCTTCCTCTTCCCTAGATCAATACAGACAGACAGACA
GGTG/**gata/gata/gata/gata/gata/gata/gata/gata/gata/gata/gata**/T
CATTGAAAGACAAAACAGAGATGGATGATAGATACATGCTTACAGATGCACACACAAACGC
TAAATGGTATAAAAATGGAATCACTCTGTAGGCTGTTTTACCACCTACTTTACTAAATTAA
TGAGTTATTGAGTATAATTTAATTTTATATACTAATTTGAAACTGTGTCATTAGGTTTTTA
AGT

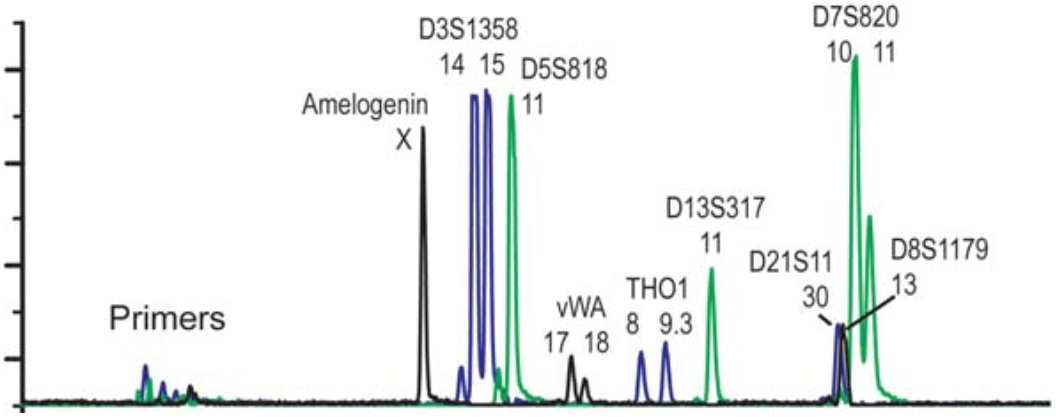
DNA → amplified fragments → counts of STRs



calibration markers

a real test does not rely on just one STR…

Example result



For the STR "D3S1358",

- an allele with 14 copies
- an allele with 15 copies

For THO1

- 1 × 8 copies + 1 × a variant of 9 copies
- for …

What if we have a list of STR's ? Put in a table

Liu,P, Li, X…. Mathies, R.A. Lab Chip, 2011, 11, 1041–1048

Database entry

Example for one person

- D13S317
  - 11 and 14 repeats

Marker	Allele	
	1	2
AMEL	X	Y
CSF1PO	10	10
D13S317	11	14
D16S539	9	11
D18S51	14	16
D21S11	28	30
D3S1358	16	17
D5S818	12	13
D7S820	9	9
D8S1179	12	14
FGA	21	22
TH01	6	6
TPOX	8	8
VWA	17	18

sex

13 standard loci

Federal bureau of investigation (FBI) 13 standard loci

- minimum – may include more

European databases also include these 13

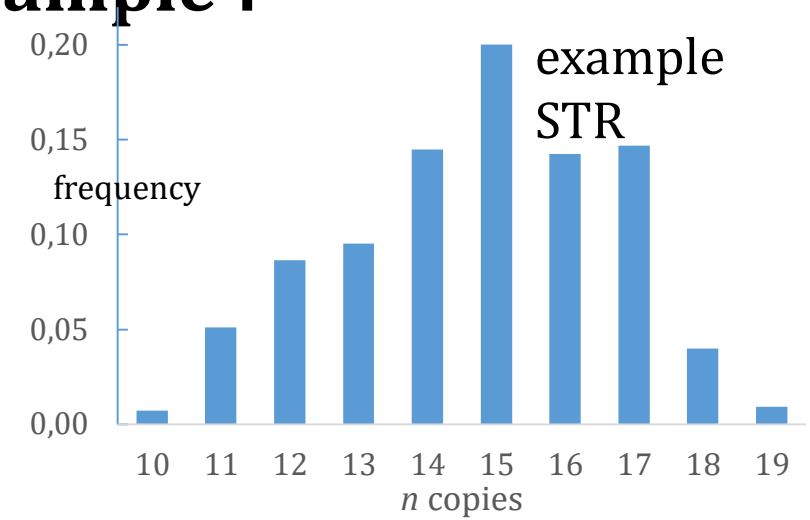
Commercial kits include more sites

Sex marker

- amelogenin is 6 bases shorter on X Chromosome

Statistics – different questions…

How unique is a sample ?



For "example STR"

$p(\text{match})$ depends on number of copies

$$n = 15 \qquad p \approx 0.2$$

$$n = 17 \qquad q \approx 0.15$$

$0.2 \times 0.15 = 0.03$, but we have two Würfel (alleles)

$$\mathrm{P}(\text{match}) = 2pq = 0.06$$

What if locus is homozygous ? $\mathrm{P}(\text{match}) = p^2$

Many N_l loci ? Match probability is the product

$$\prod_{i=1}^{N_l} \mathrm{P}(\text{match}_i)$$

$$\prod_{i=1}^{N_l} \mathrm{P}(\mathrm{match}_i)$$

For each locus i

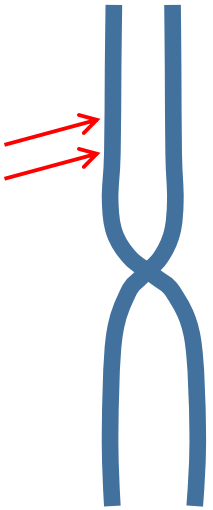
look up frequency of the allele in the sample at i

Assumption

- one can multiply the probability of independent events

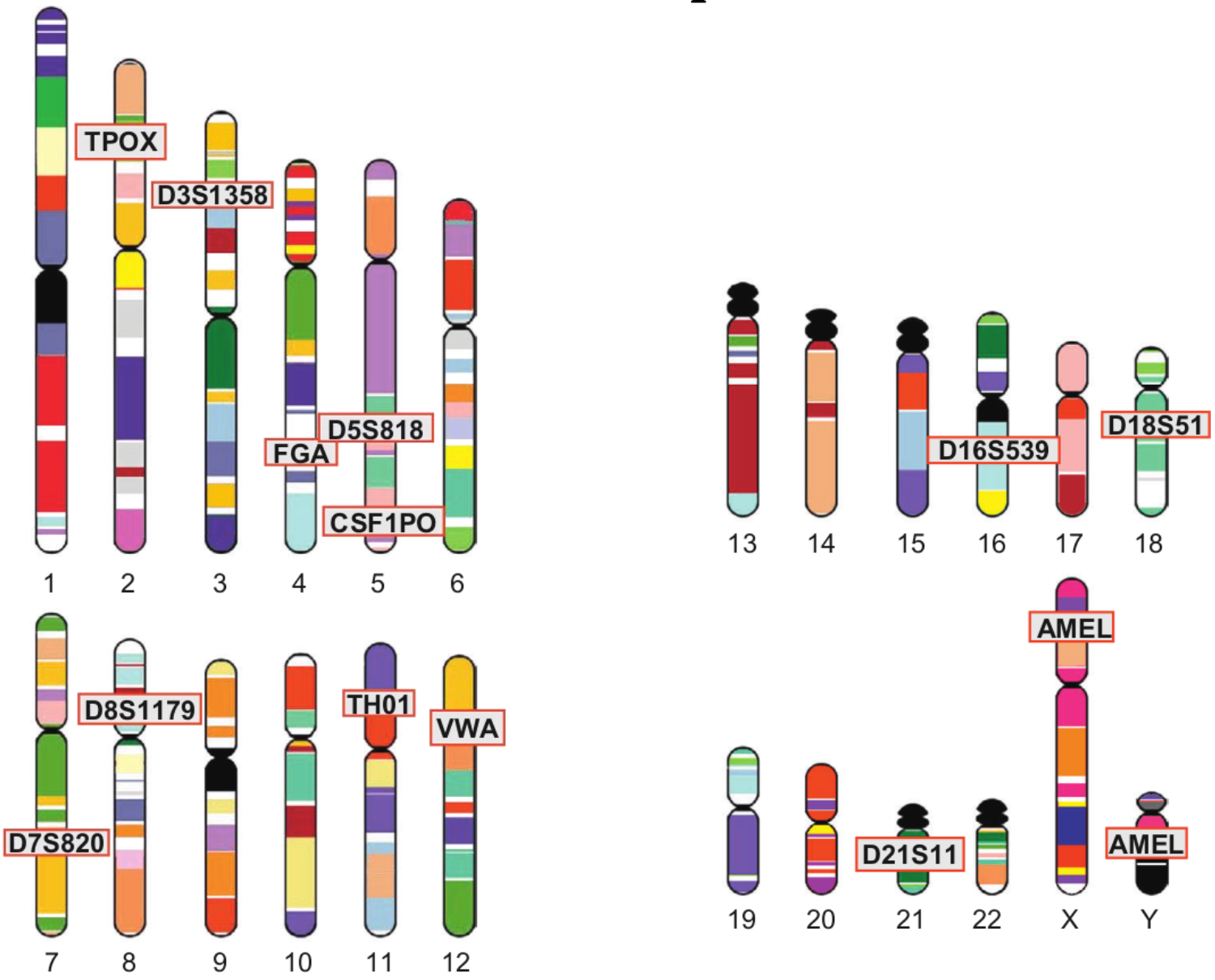
What if two loci are near each other on some chromosome ?

- allele 1 and 2 are not independent of each other (linkage)



Is this a problem ? …

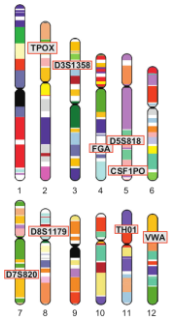
location of important loci



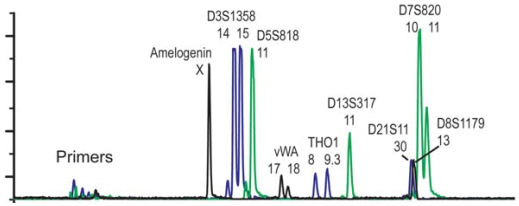
Godbey, W.T., An Introduction to Biotechnology, Academic Press (2014)

Criteria for STR loci

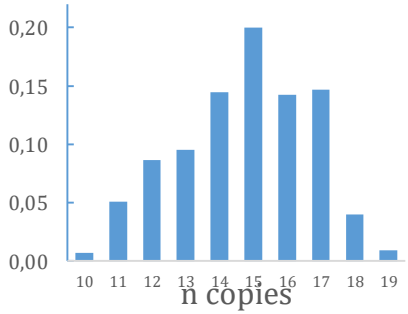
1. Distributed over chromosomes and within chromosomes
- avoid linkage



2. Repeat length 4 or 5
- separable on electrophoresis



3. Distribution of number of repeats within population
- more unique possibilities
- lower probability of random match



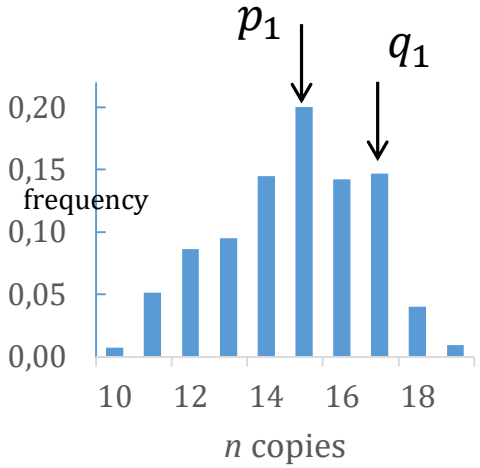
statistics questions

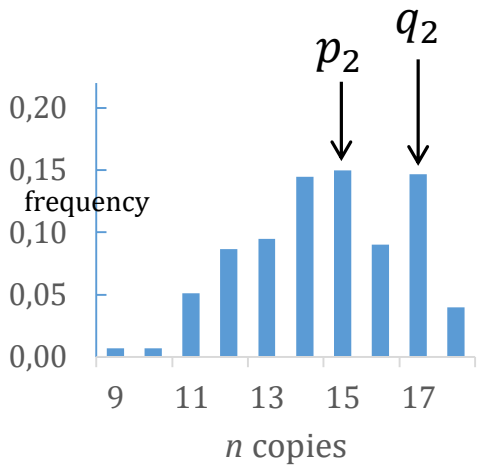
- how unique is a sample ?
- how good is a test kit ?
- what is the probability that I committed a crime ?
- can I be ruled out ?
- how likely is a database hit ?

Prosecutor or general advice to court

Prosecutor has one sample – gets $P(\text{match})$ for this sample

Can we be more general ? What is the worst case ?





… next STR's

least informative would be most common alleles at each STR

$$P(\text{match}) = \prod_{i=1}^{n_{STR}} 2p_i q_i$$

where each pq is the most common in population

Probability that you committed crime

Propose two hypotheses for the evidence

1. you are guilty / "inclusion", prosecutor's hypothesis H_p
2. defendant's hypothesis H_d
   crime was committed by random man

Probability of evidence $P(E)$ given a hypothesis $P(E|H)$

- likelihood ratio $\frac{P(E|H_p)}{P(E|H_d)}$

Advantage of writing in this way ?

$$\frac{P(E|H_p)}{P(E|H_d)}$$

Simplest – perfect match

- prosecutors hypothesis explains evidence $P(E|H_p) = 1$
- defense hypothesis is probability of someone else in the population $P(E|H_d) =$ random match (very small)

Mixed samples – two person's DNA found – one matches

- H_p - you committed crime, H_d the other committed crime

  $\frac{P(E|H_p)}{P(E|H_d)} = 1$, both explanations are equally likely

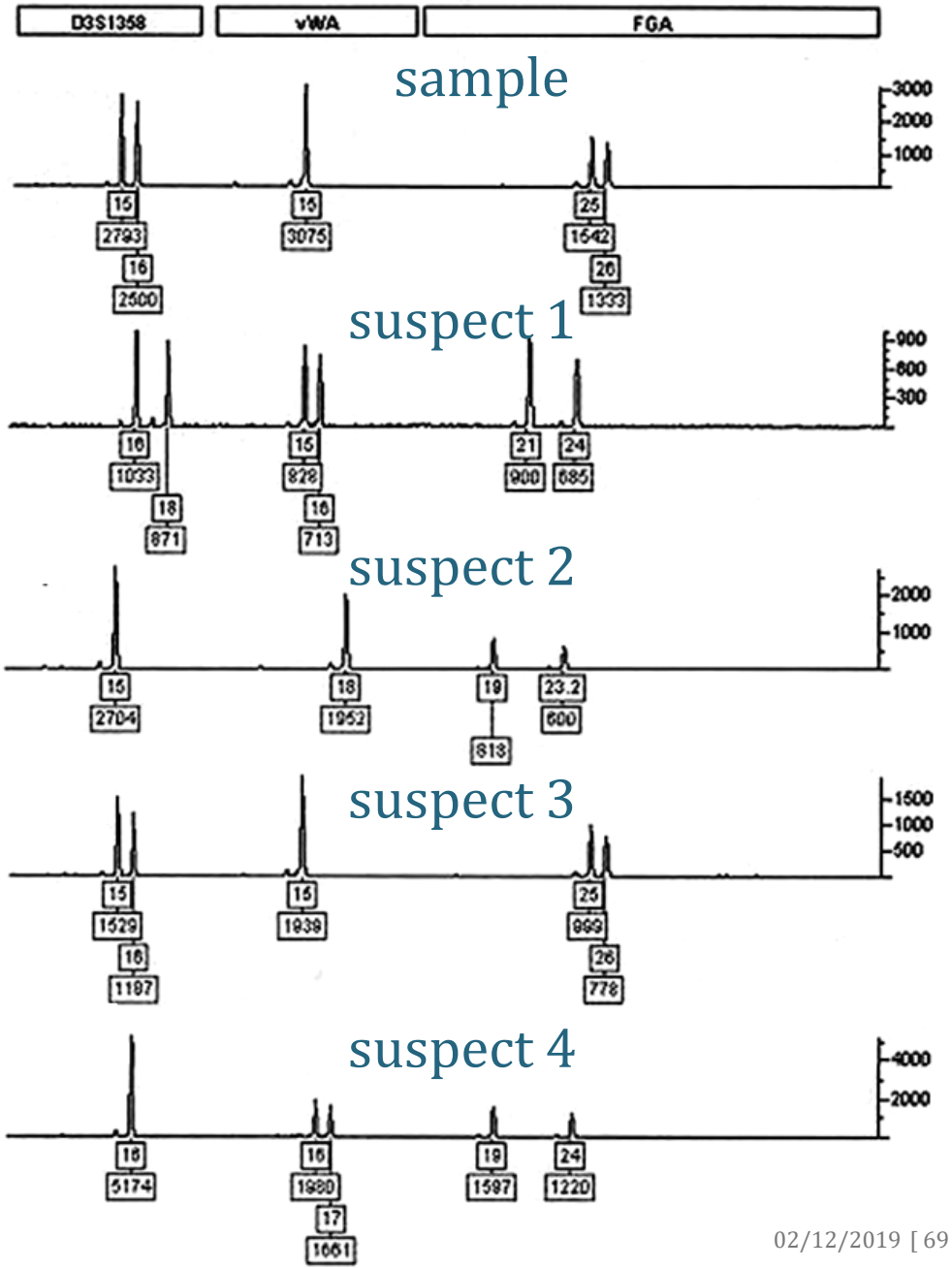
Extra evidence

- Criminal had red car, you have red car 10% of cars are red
- H_d (random match) is 10 % as likely, $\frac{P(E|H_p)}{P(E|H_d)}$ is 10× likelier

When do we not need probabilities ?

Suspects 1, 2 and 4 can be ruled out

No need to calculate probabilities



Conley, J.M., Moriarty, J.C. Scientific and expert evidence, Aspen Publishers, 2007

Database hits - corrections

DNA from crime scene checked against database

- you are found - is this fair ?
- imagine $P(\text{match}) = 10^{-9}$
- database with 10^7 people
  - 1 / 100 searches will find somebody by chance

Two answers

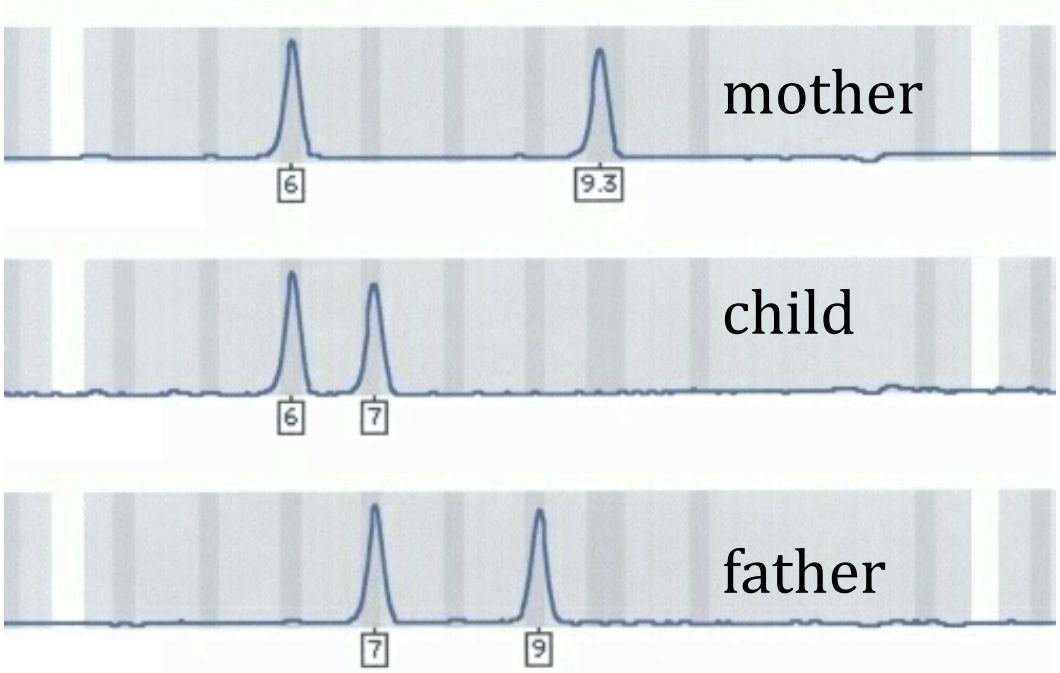
- divide $P(\text{match})$ by n in database (10^7)
- calculate $P(\text{match})$ using STRs not in database
  database has about 13 core loci, kits have around 20

Paternity

Are you my parent ?
samples are usually good quality – no degradation

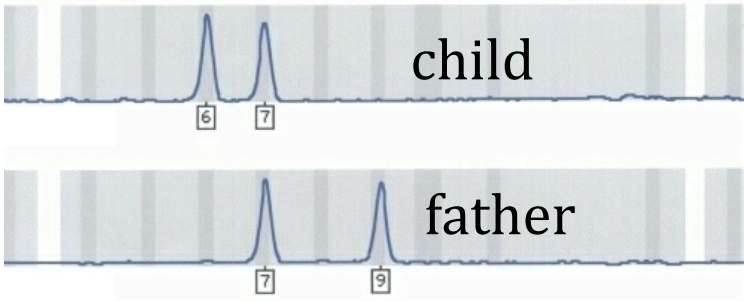
At each STR

- the two alleles must come from mother or father



Adams, J. (2008) Paternity testing: blood types and DNA. Nature Education 1(1):146

Paternity statistics



Statistics at one site / STR

- previously $2pq$

  numbers like $2 \times 0.1 \times 0.2$

Now we have two Würfel

$P(\text{match}) = p + q$ imagine a number like 0.2+0.1

Over all STR's random match is $\prod_{i=1}^{n_{\text{STR}}}(p_i + q_i)$

much higher than $\prod_{i=1}^{n_{str}} 2p_i q_i$

Chances of random match are higher with parenthood, but

- a full DNA test kit might have > 20 STRs

Final problems and issues

All probability calculations are based on population samples

- what if sample + suspect come from tiny village in Bayern?
- you are part of Corleone family ? Mafia
  related DNA on crime scenes