Identification of inherited disease genes

Michel Koenig (EA_7042, Institut Universitaire de Recherche Clinique, Mpt)
michel.koenig@inserm.fr
Jean-Baptiste Rivière (Dpt Hospitalo-Universitaire de Génétique Moléculaire, Dijon)

09/14	The Revolution of the Human Genome Project/ Linkage Studies
09/21	Linkage studies for monogenic and multifactorial diseases
09/28	High-throughput sequencing and strategies for monogenic disease gene identification
10/05	Diseases by somatic and germline *de novo* mutations : concepts and investigation strategies (JB Rivière/M Koenig)

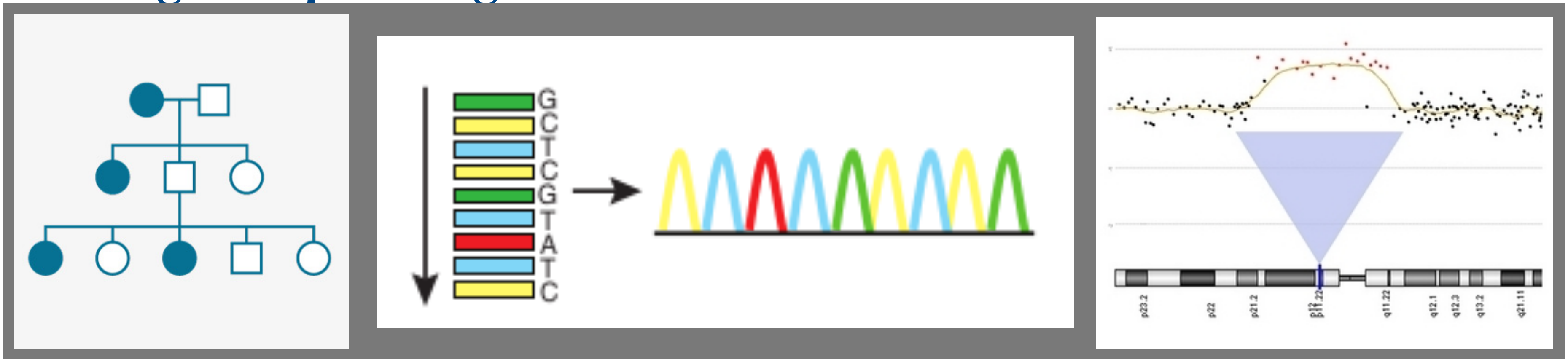
(simplified) history of DNA sequencing (part 1)

- 1977 Fred Sanger, "DNA sequencing with chain-terminating inhibitors” (radioactive electrophoresis on **plate** gels)
- 1984 Sequencing of the Epstein-Barr virus genome, 170,000 nt ...
- 1987 Applied Biosystems launches the first automated plate gel sequencers (**fluorescent** Sanger technique) ABI 370.
- 1995 Craig Venter et coll. (Institute for Genomic Research): first complete genome of a free-living organism, the bacterium Haemophilus influenzae (1,830,137 nt); first use of whole-genome shotgun sequencing.
- 1999 ABI introduces the 96 **capillary** sequencer (ABI Prism 3700) for the Human Genome Project (still fluo. Sanger)
- 2001 1st draft of the Human Genome Sequence (Sanger)

  2003 Completion of the Human Genome Project
- 2007 1st sequencing of an individual human genome Craig Venter (100 millions $) (–> 2 million $ in 2008)

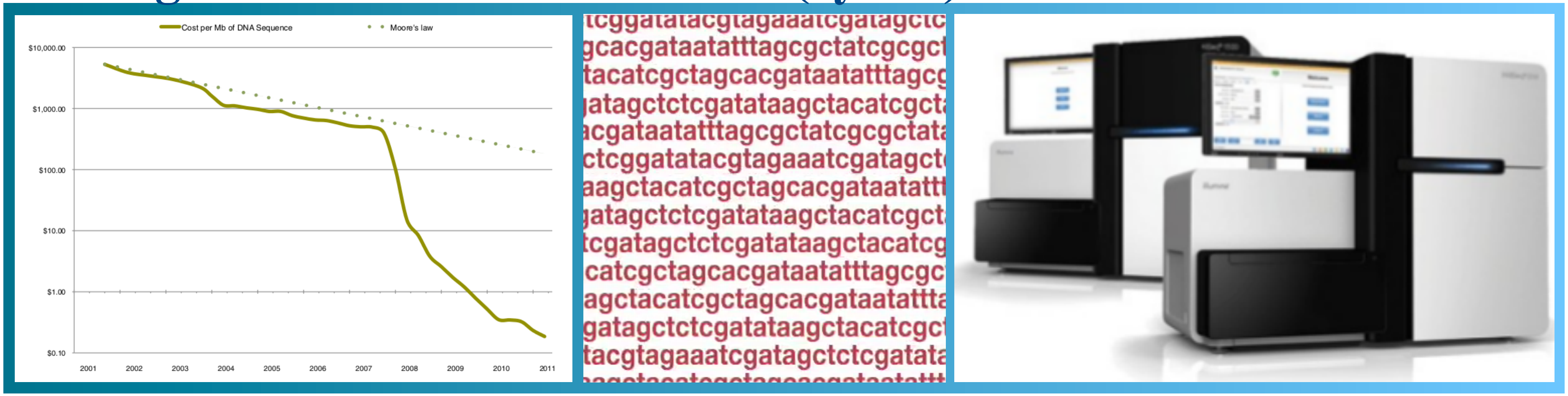
Technological developments

Sanger sequencing :



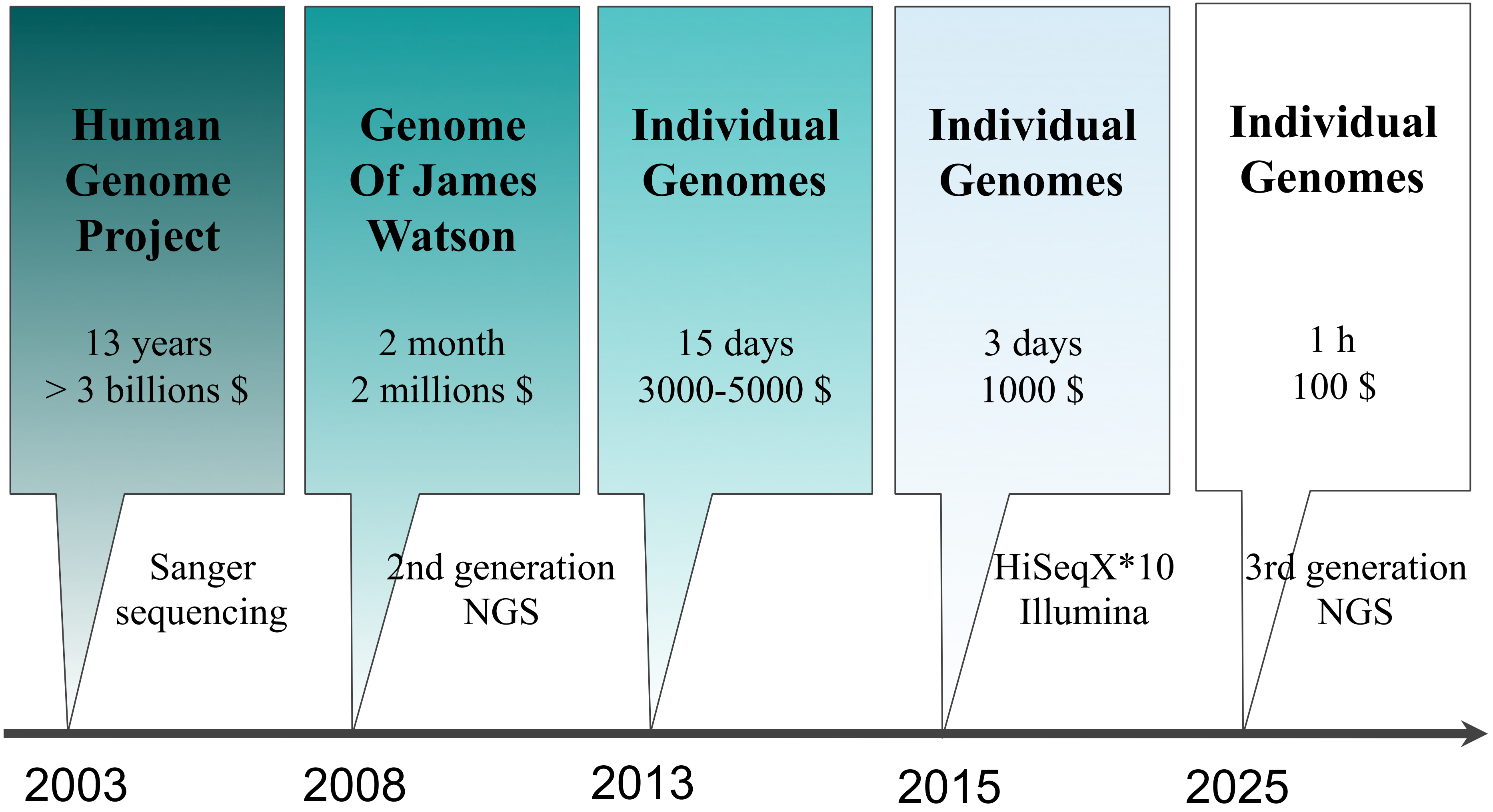
Next generation sequencing or massively parallel sequencing

2nd generation : wash and scan (cycles) since 2004



3nd generation NGS : single molecule sequencing

Towards routine sequencing of entire human genomes (history of sequencing part 2)



Human Genome Project

13 years
> 3 billions $

Sanger sequencing

2003

Genome Of James Watson

2 month
2 millions $

2nd generation NGS

2008

Individual Genomes

15 days
3000-5000 $

2013

Individual Genomes

3 days
1000 $

HiSeqX*10 Illumina

2015

Individual Genomes

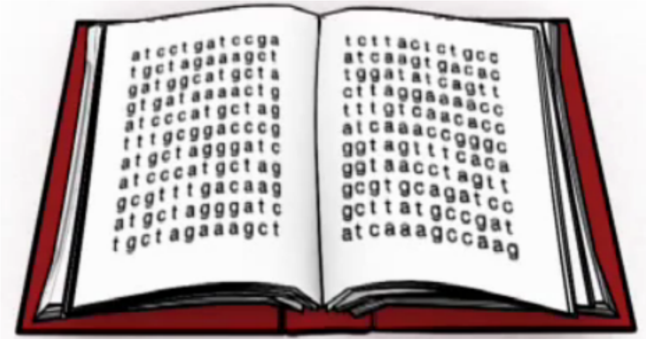
1 h
100 $

3rd generation NGS

2025

The human genome

- Diploïd organism, 23 pairs of chromosomes
- 3.3 billion pairs of nucleotide
- 60 % of repeated sequences
  - Interspersed seq. (retrotransposons [LINEs/SINEs], transposons 45 %
  - Pseudogenes 1 %
  - Simple repeats (microsatellites) 3 %
  - Segmental duplications (non-homologous recombination) 5 %
  - Satellite sequences (recombination, tandem repeats) 6 %
- 2 % of protein coding sequences
- 20,687 genes and a mean 6.3 isoforms per locus.

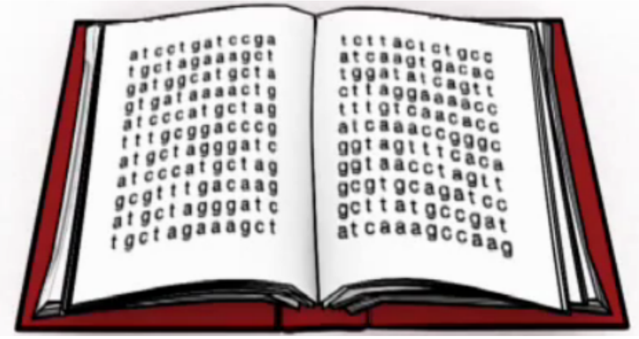


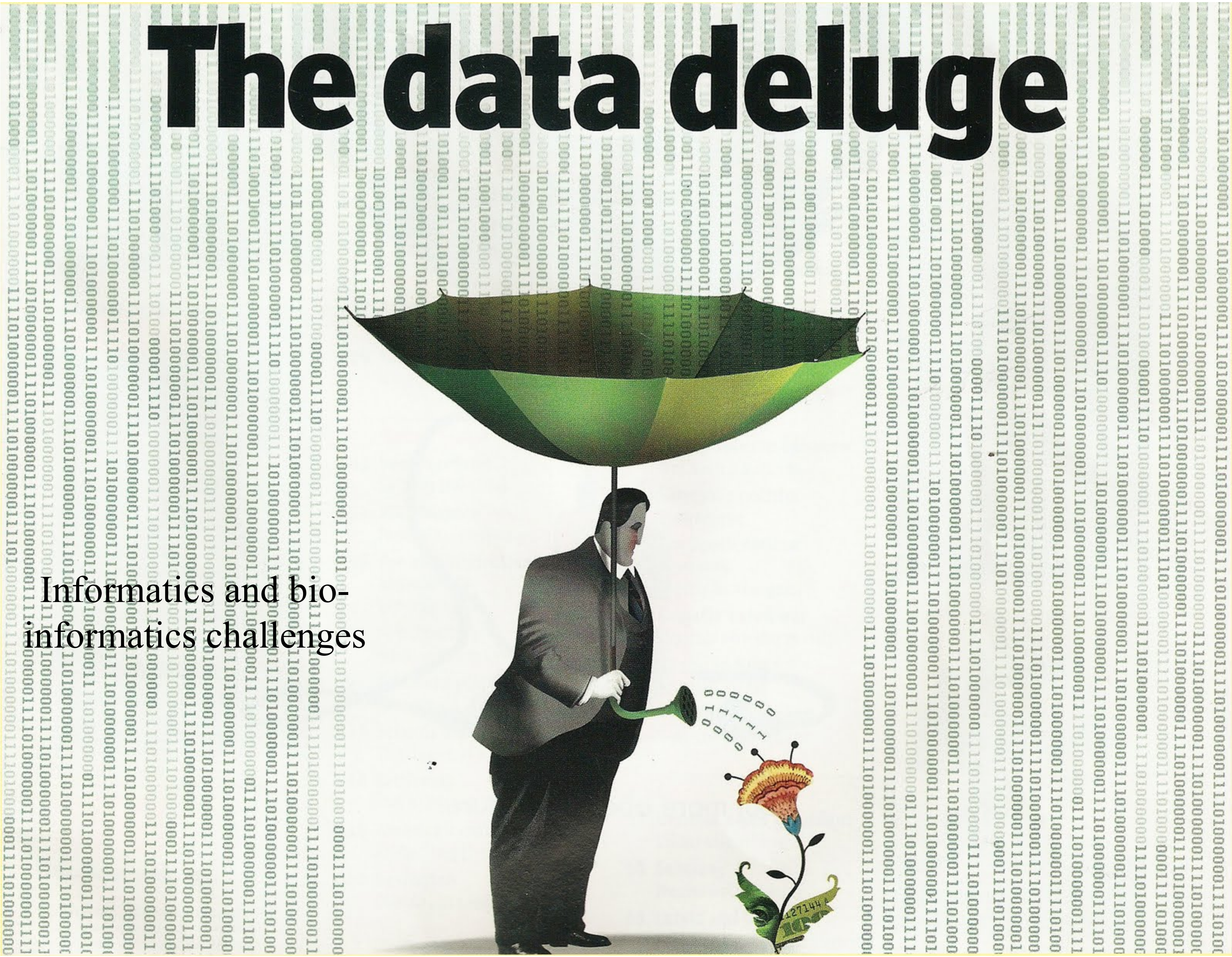
Variation of the human genome

- **Mean variations, per individual :**
  - > 1 000 copy number variations (CNV)
  - 3 to 4 millions single nucleotide variations (SNV)

    including ≈ 20 000 in or near protein coding sequences :
    - 10 000 silent variations
    - 9 000 missense variations
    - 100 nonsense variations
    - 100 splice site variations
- Only about 10 variations are pathogenic = disease causing

  (genetic load)





Informatics and bio-informatics challenges

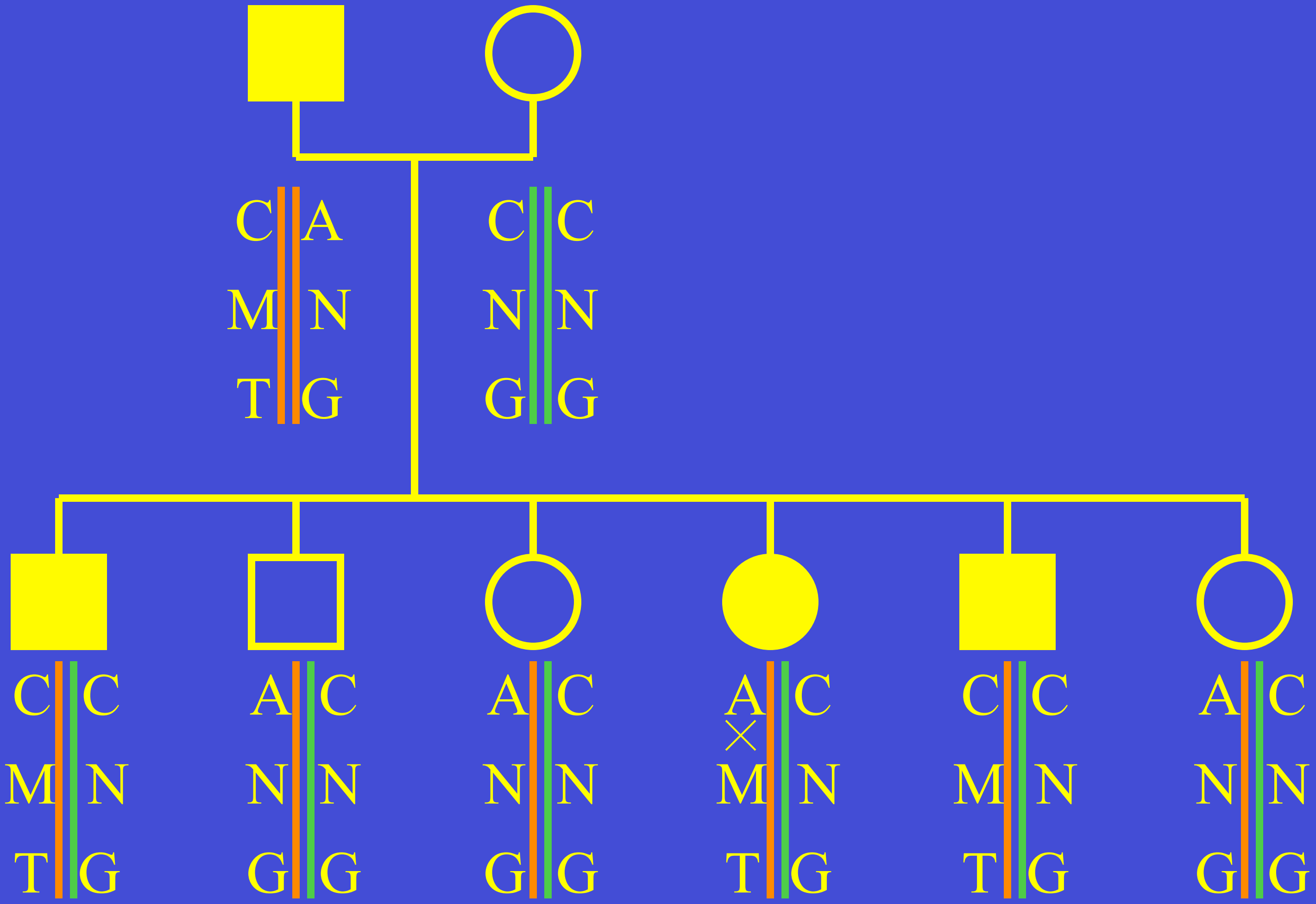
Identification of disease genes by family studies

Inherited diseases :

- Two or more affected per family
- Monogenic / multifactorial diseases
- Localisation of disease genes by linkage studies
- Calculation of likelihood of linkage: LOD score

Topics :

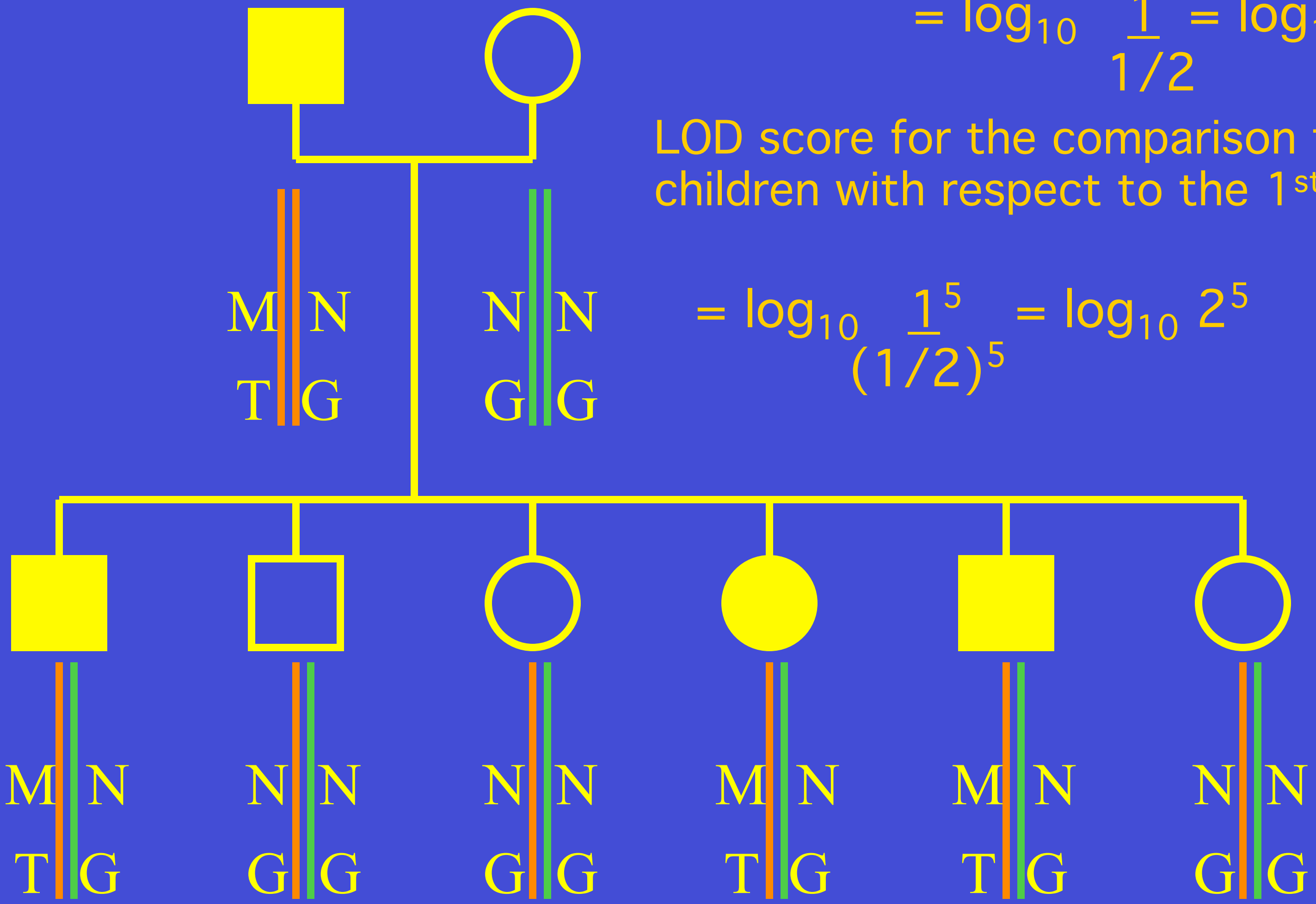
- Dominant diseases
- Recessive diseases
- Homozygosity mapping
- Founder effect


C A
M N
T G
C C
N N
G G
C C
M N
T G
A C
N N
G G
A C
N N
G G
A C
M N
T G
C C
M N
T G
A C
N N
G G

$$\text{LOD score} = \log_{10} \frac{\text{Proba. to observe segregation of marker linked at distance } \theta}{\text{Proba. to observe segregation if not linked (} \theta = 0.50)}$$

For θ = 0, LOD score for the comparison the 2nd child with the 1st child :

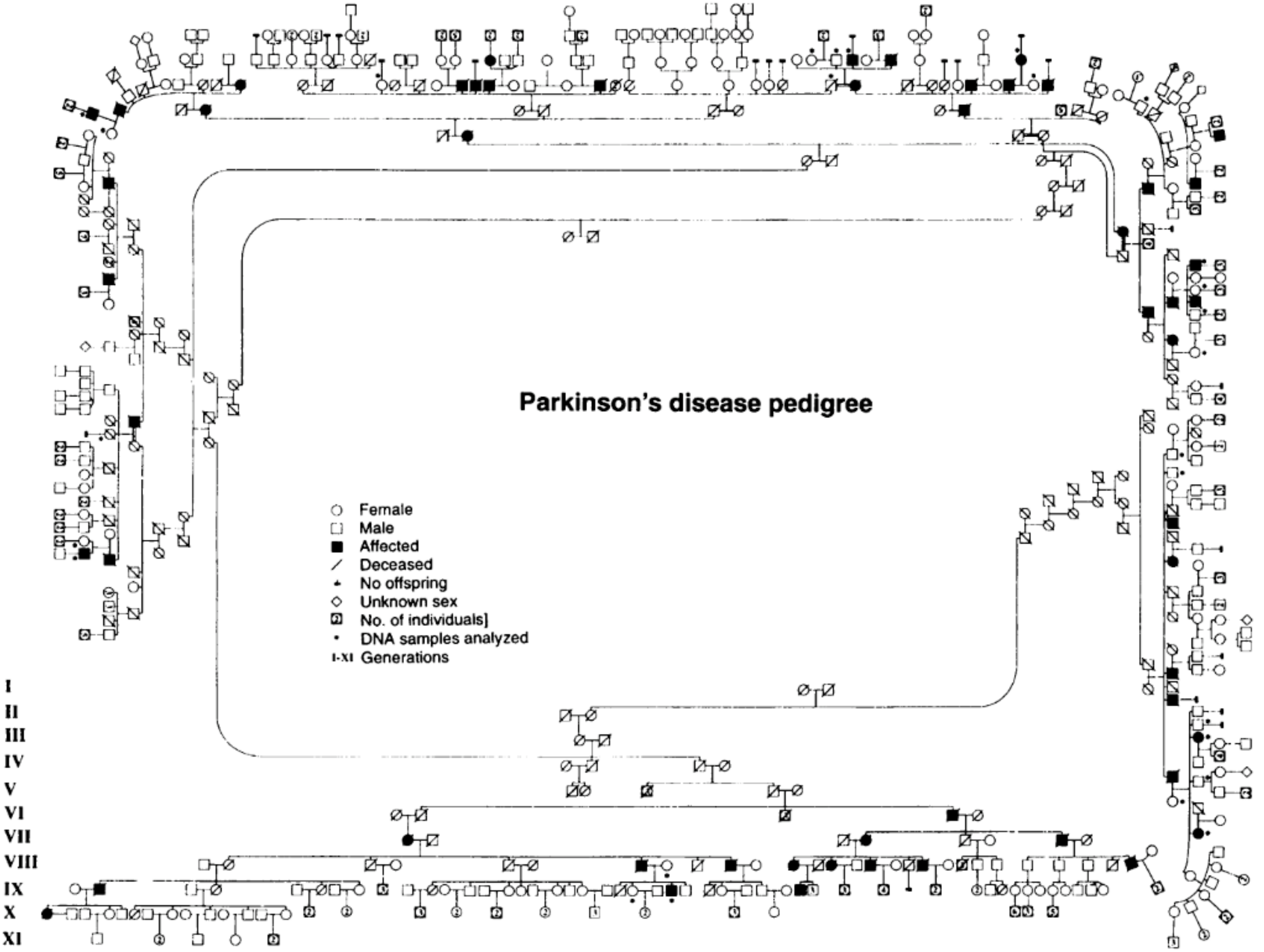
$$= \log_{10} \frac{1}{1/2} = \log_{10} 2 = 0.3$$



LOD score for the comparison the 5 next children with respect to the 1st child:

$$= \log_{10} \frac{1^5}{(1/2)^5} = \log_{10} 2^5 = 5\text{x}0.3 = 1.5$$

	probability in favor of linkage	probability by chance	Bonferroni correction for multiple testing x60
LOD score = 1.5	32 to 1 (2^5)	p=0.03	>1
LOD score = 3	1000 to 1 $(\approx 2^{10})$	p=0.001	p=0.06
LOD score = 4	10 000 to 1 $(>2^{13})$	p=0.0001	p=0.006



Parkinson's disease pedigree

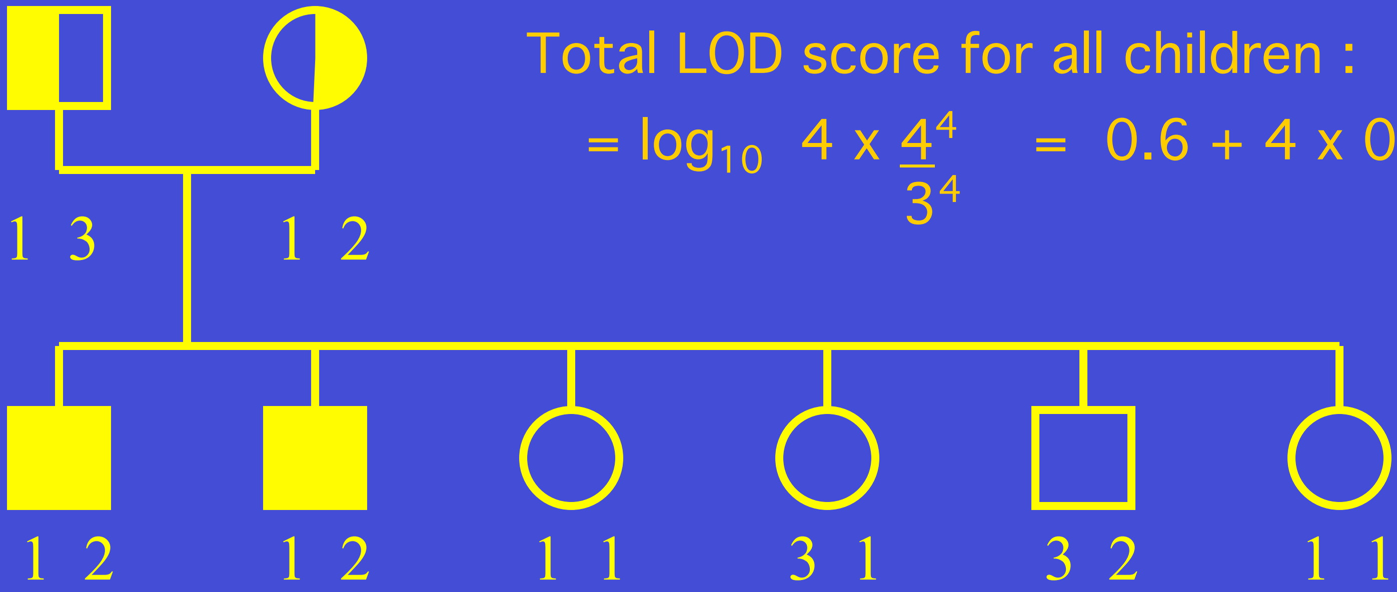
Recessive disease

for the 2nd affected child :

$$\text{LOD score} = \log_{10} \frac{1 \times 1}{1/2 \times 1/2} = \log_{10} 4 = 0.6$$

for the 1st healthy child :

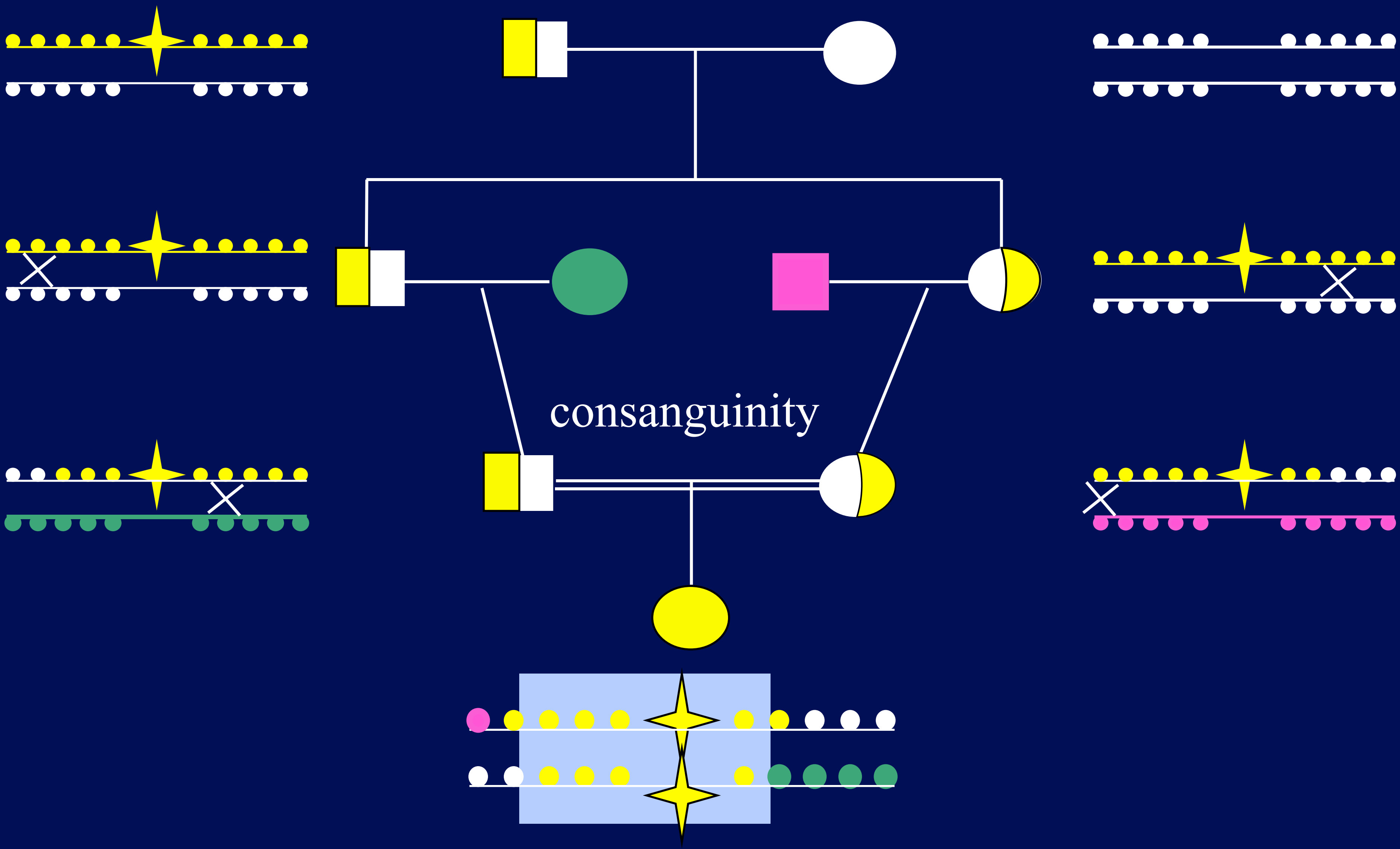
$$\text{LOD score} = \log_{10} \frac{1/3}{1/4} = \log_{10} \frac{4}{3} = 0.125$$



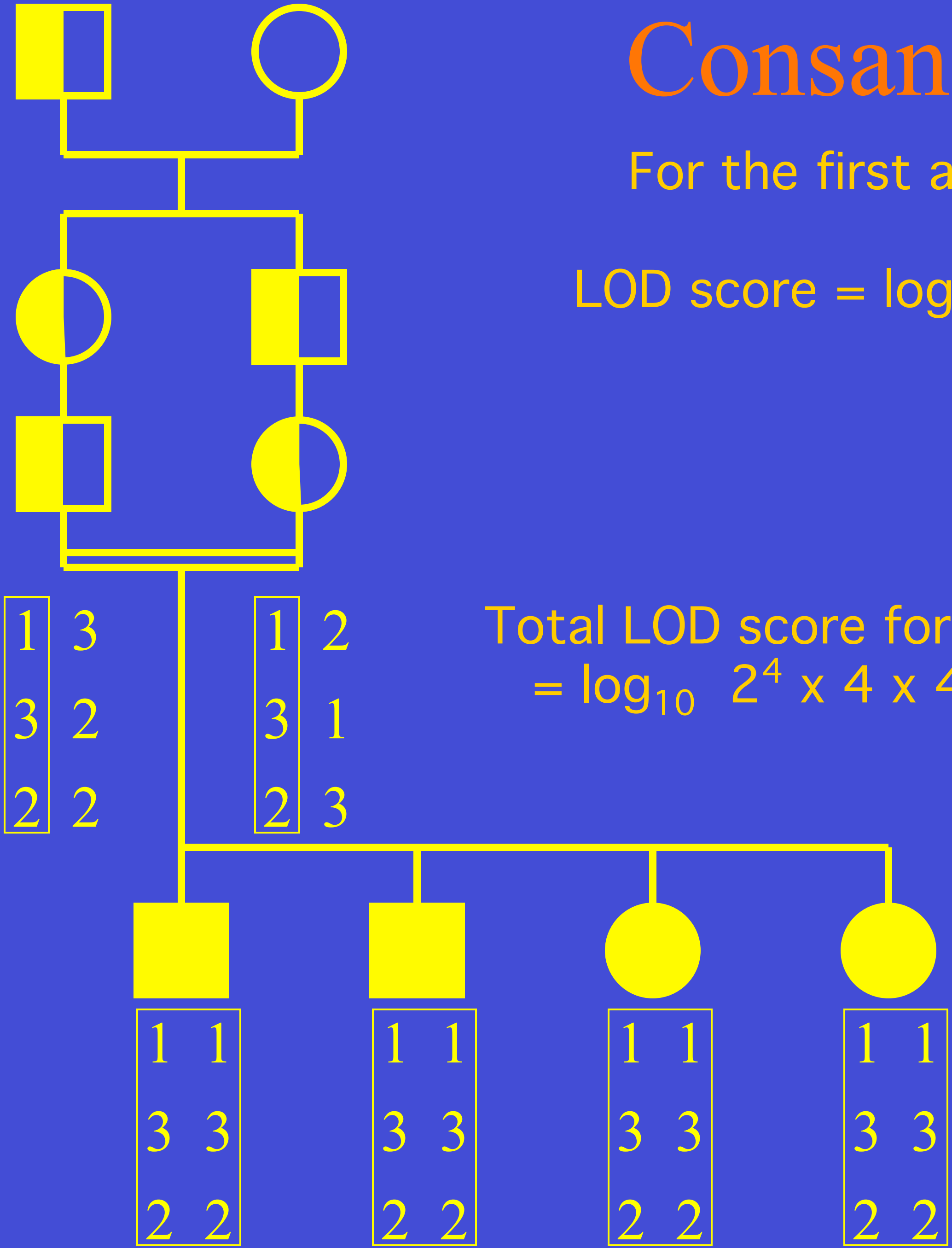
Total LOD score for all children :

$$= \log_{10} 4 \times \frac{4^4}{3^4} = 0.6 + 4 \times 0.125 = 1.1$$

Linkage analysis by homozygosity mapping


consanguinity

Consanguinity



For the first affected child:

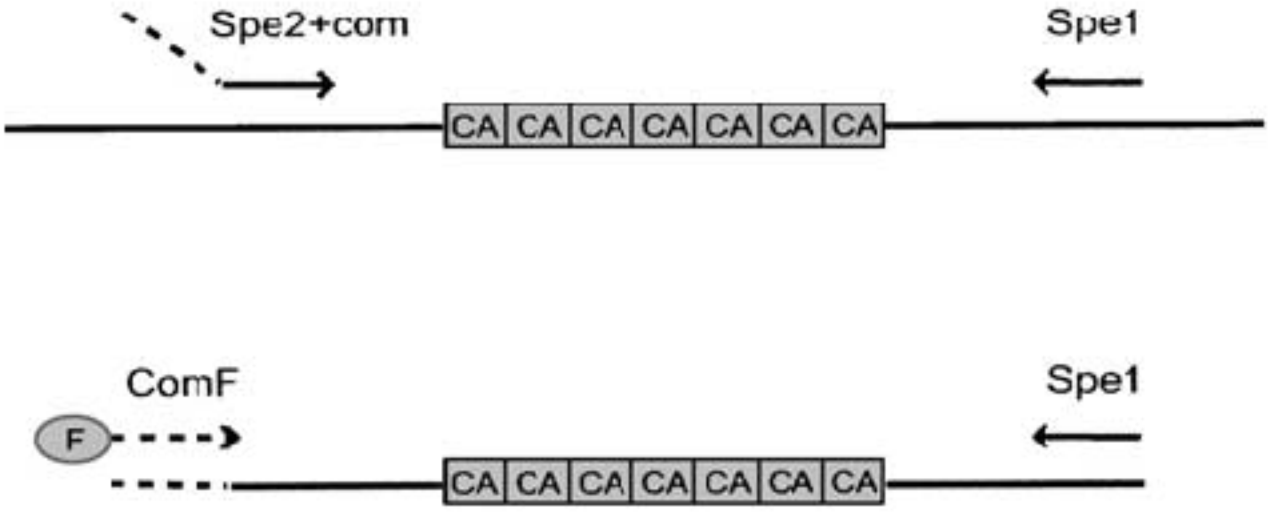
$$\text{LOD score} = \log_{10} \frac{1 \text{ x } 1}{1/2 \text{ x } 1/2} \text{ x } \frac{1 \text{ x } 1}{1/2 \text{ x } 1/2}$$

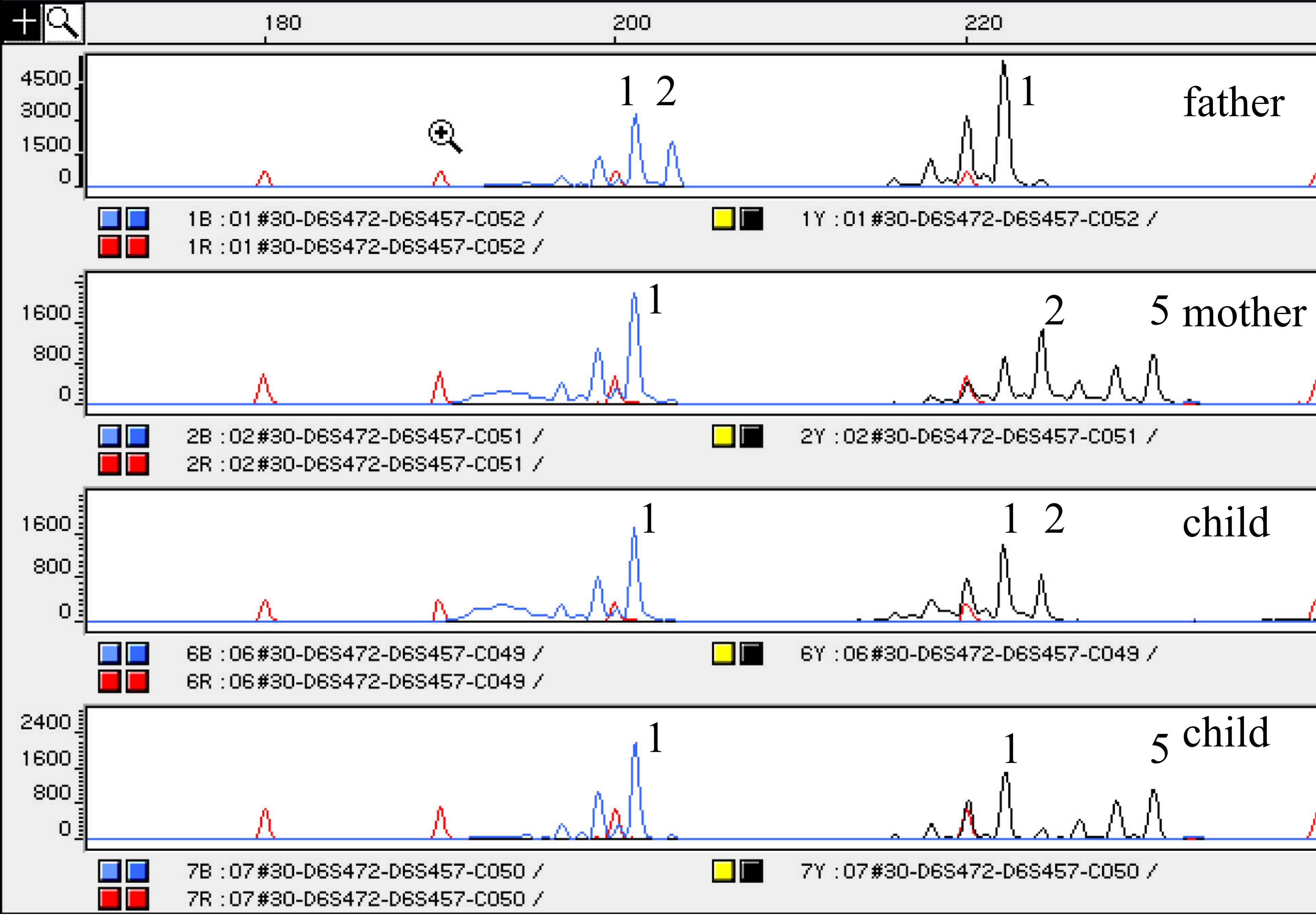
$$= \log_{10} 2^4 = 1.2$$

Total LOD score for all 4 affected children :

$$= \log_{10} 2^4 \text{ x } 4 \text{ x } 4 \text{ x } 4 = 1.2 + 3 \text{ x } 0.6 = 3.0$$

Multiallelic microsatellite marker
DNTR dinucleotide tandem repeat





1B : 01 #30-D6S472-D6S457-C052 /
1R : 01 #30-D6S472-D6S457-C052 /
1Y : 01 #30-D6S472-D6S457-C052 /

2B : 02 #30-D6S472-D6S457-C051 /
2R : 02 #30-D6S472-D6S457-C051 /
2Y : 02 #30-D6S472-D6S457-C051 /

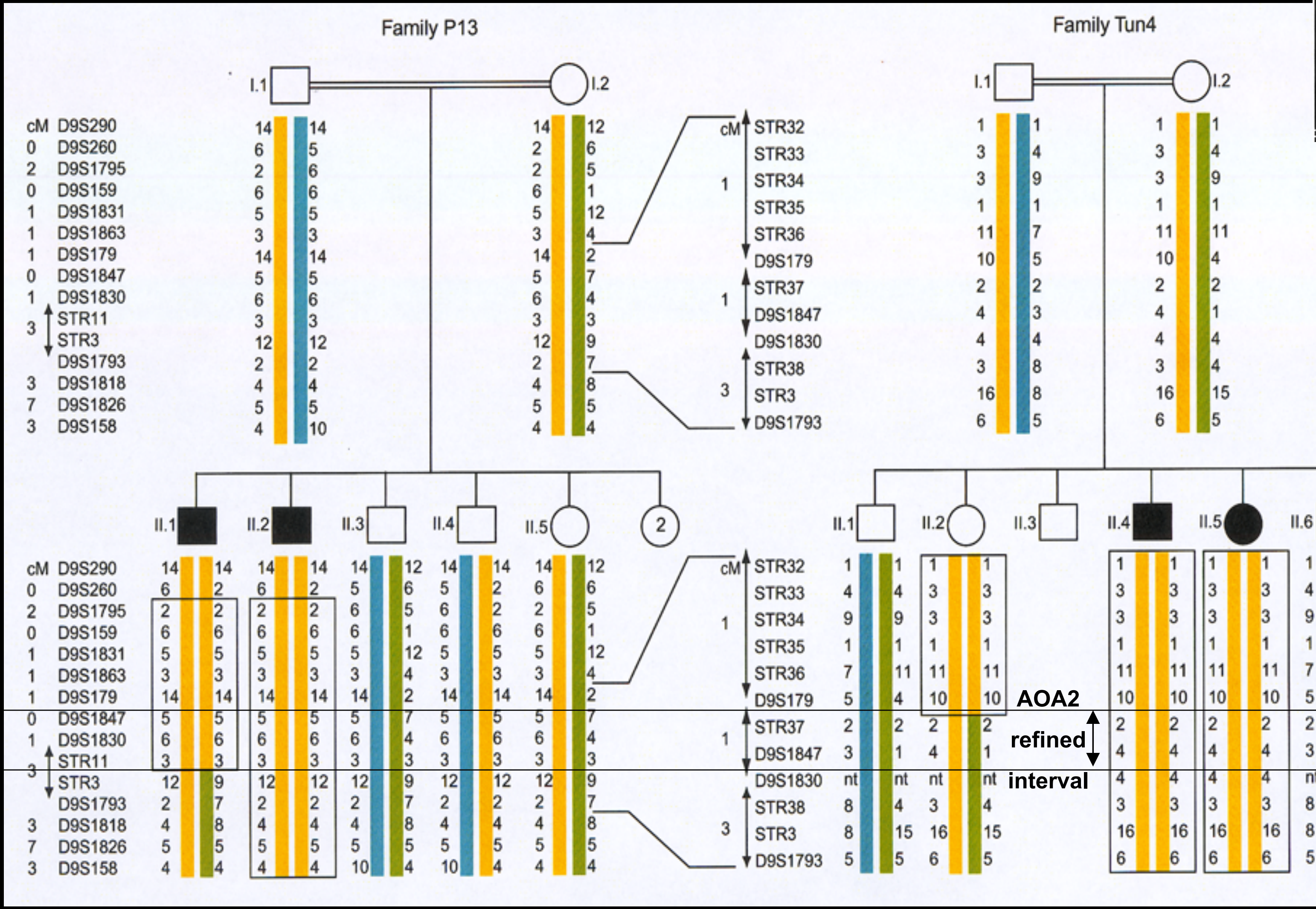
6B : 06 #30-D6S472-D6S457-C049 /
6R : 06 #30-D6S472-D6S457-C049 /
6Y : 06 #30-D6S472-D6S457-C049 /

7B : 07 #30-D6S472-D6S457-C050 /
7R : 07 #30-D6S472-D6S457-C050 /
7Y : 07 #30-D6S472-D6S457-C050 /

Ataxia with oculomotor apraxia type 2

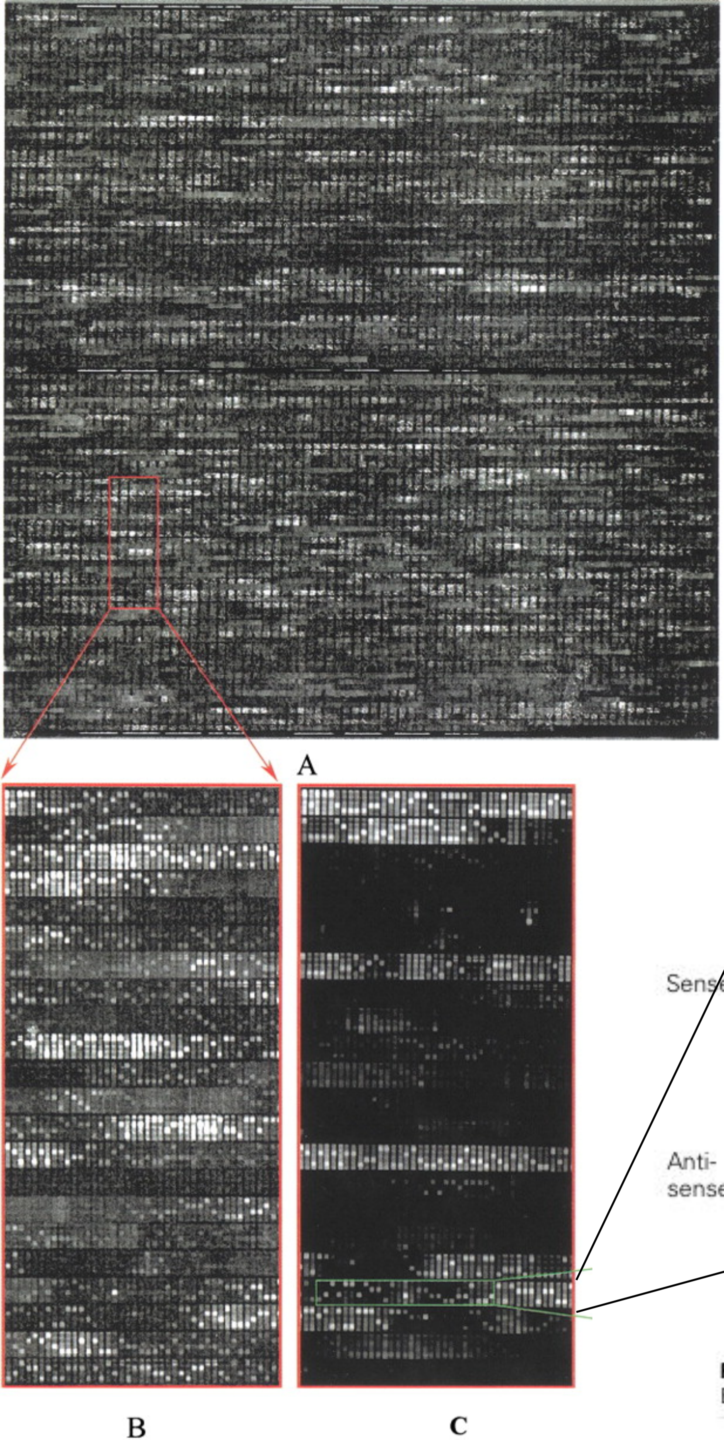
Portuguese family - P13

Tunisian Family - Tun4



Genotyping with GeneChip Array 10K SNP (single nucleotide polymorphisms) Affymetrix

(Actually GeneChip 250K and 906K SNP)



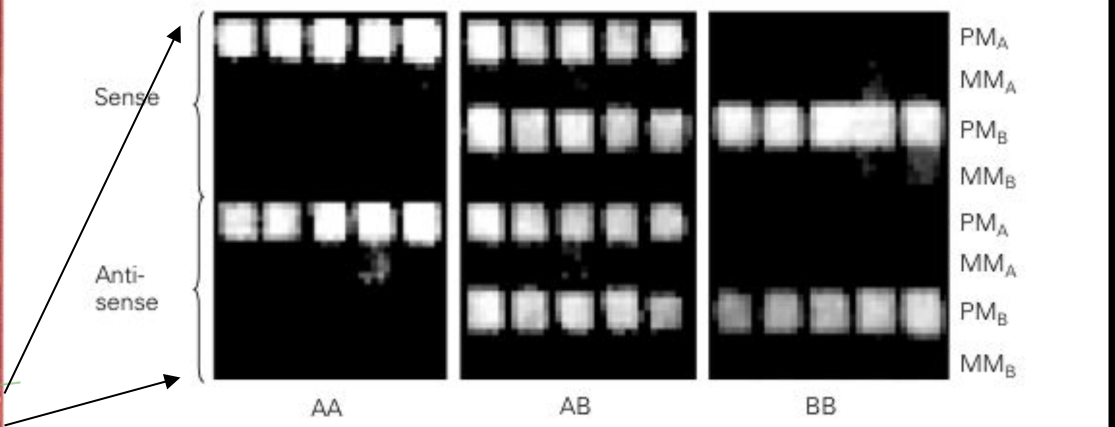
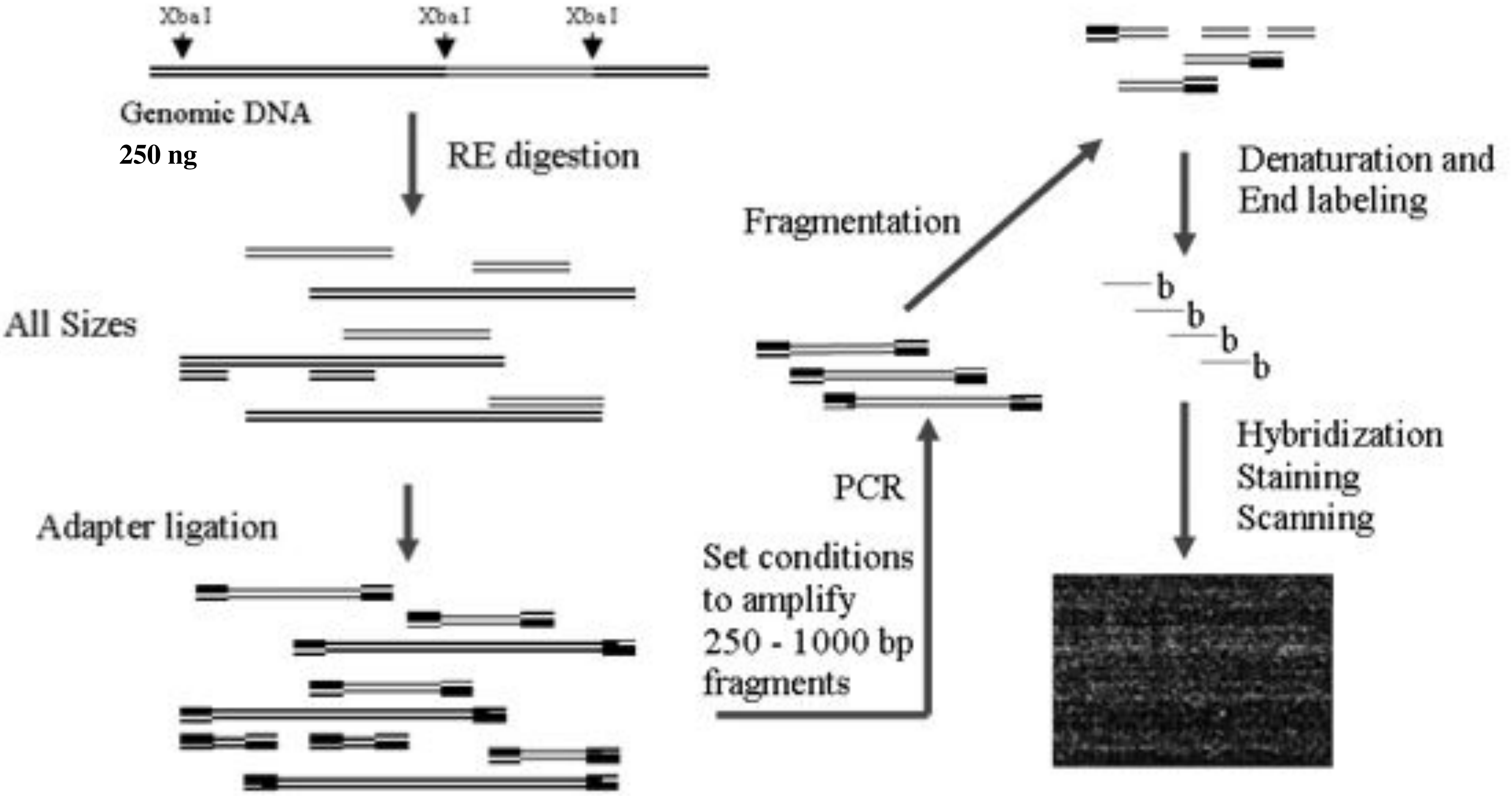
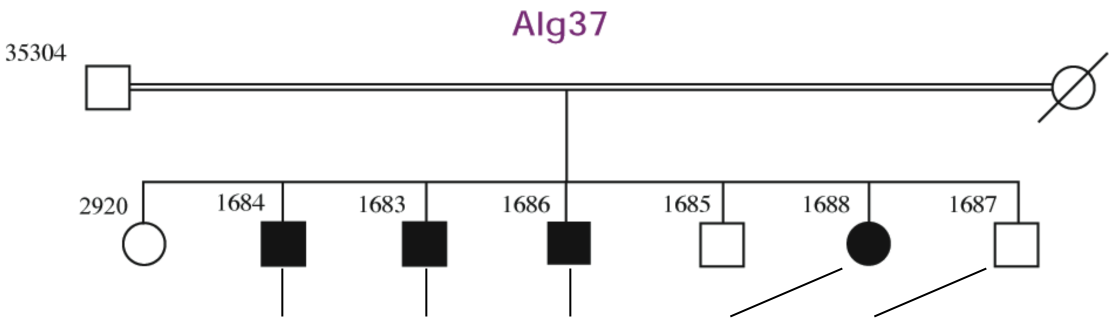


Figure 1.2
Example of Allele-Specific Hybridization with 40 probes / SNP

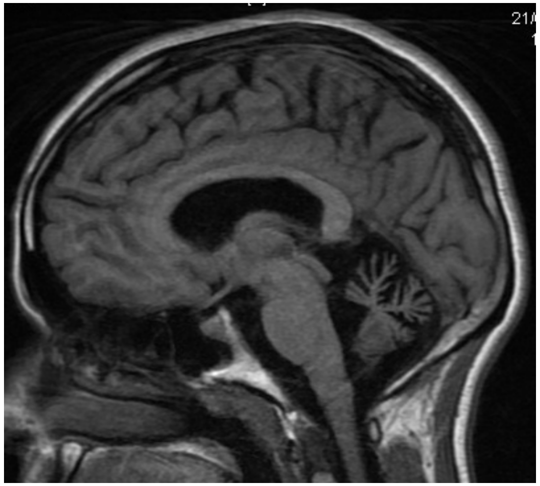
I. Method: GeneChip array Affymetrix

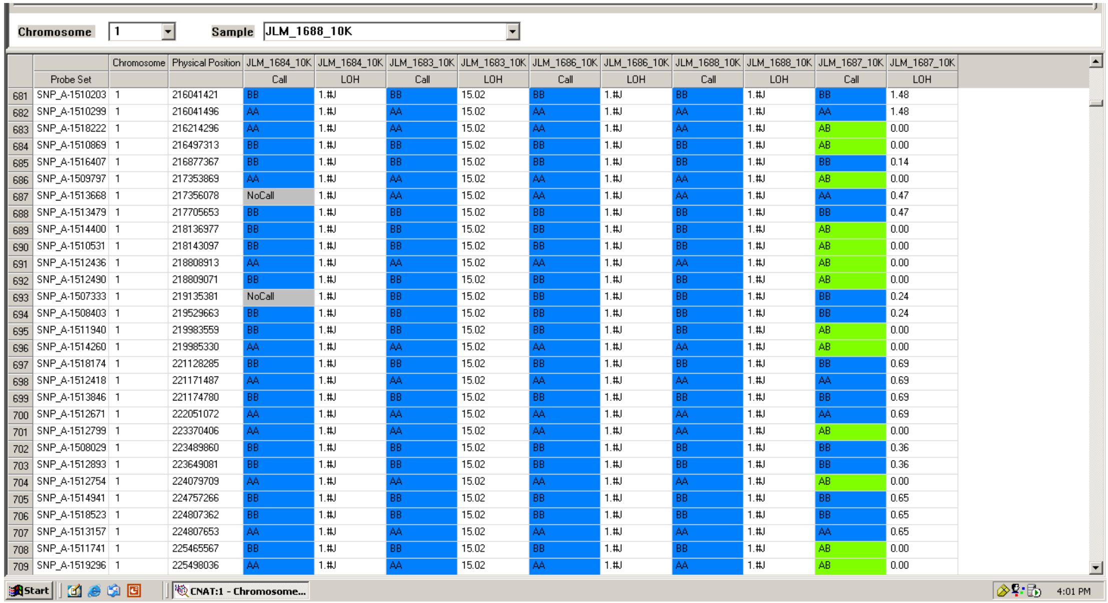


Genomewide scan in a large consanguineous family



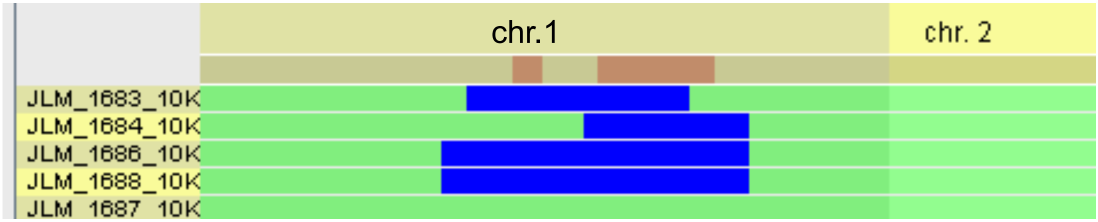
Coenzyme Q10 deficiency





Chromosome 1 Sample JLM_1688_10K

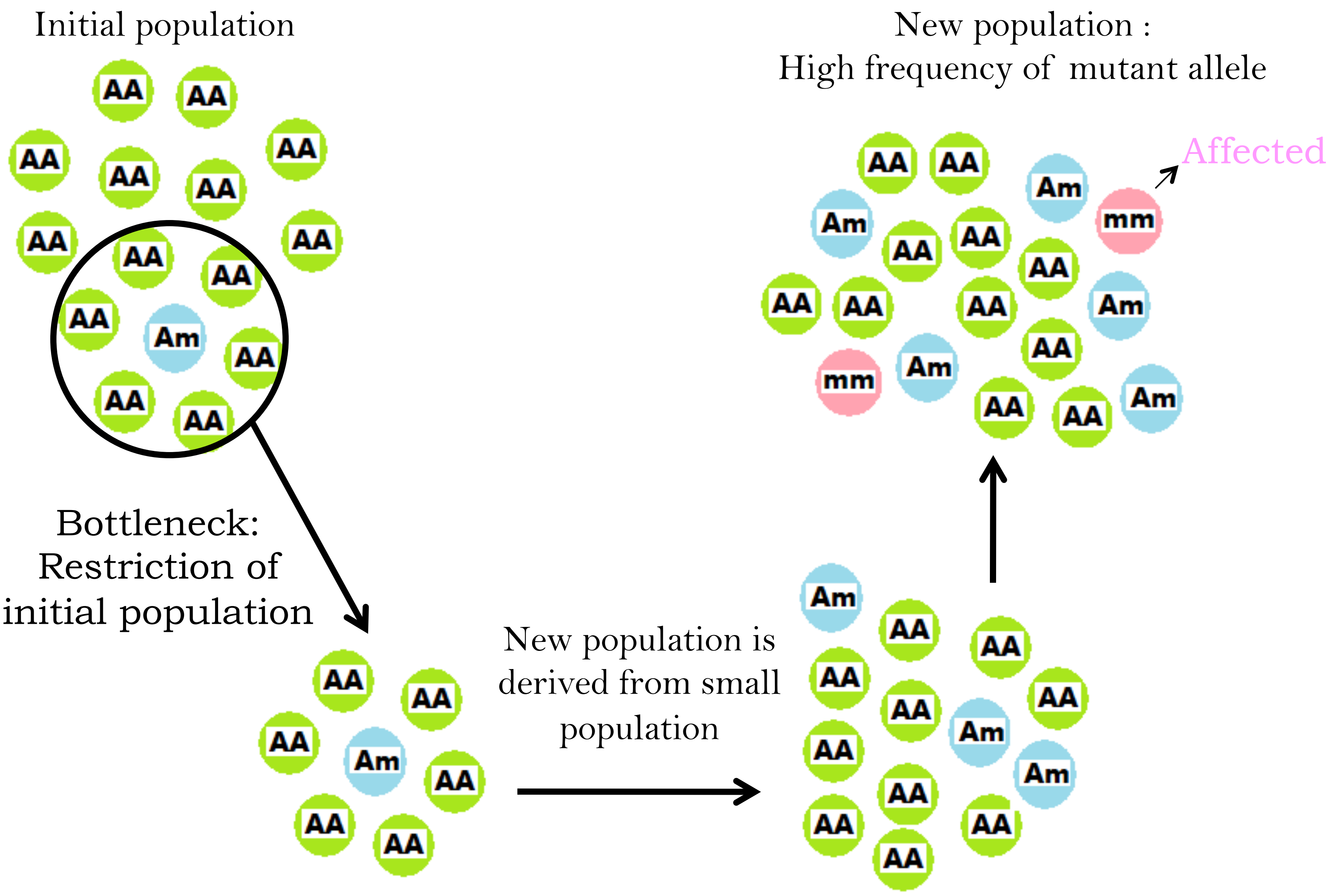
		Chromosome	Physical Position	JLM_1684_10K	JLM_1684_10K	JLM_1683_10K	JLM_1683_10K	JLM_1686_10K	JLM_1686_10K	JLM_1688_10K	JLM_1688_10K	JLM_1687_10K	JLM_1687_10K
	Probe Set			Call	LOH	Call	LOH	Call	LOH	Call	LOH	Call	LOH
681	SNP_A-1510203	1	216041421	BB	1.#J	BB	15.02	BB	1.#J	BB	1.#J	BB	1.48
682	SNP_A-1510299	1	216041496	AA	1.#J	AA	15.02	AA	1.#J	AA	1.#J	AA	1.48
683	SNP_A-1518222	1	216214296	AA	1.#J	AA	15.02	AA	1.#J	AA	1.#J	AB	0.00
684	SNP_A-1510869	1	216497313	BB	1.#J	BB	15.02	BB	1.#J	BB	1.#J	AB	0.00
685	SNP_A-1516407	1	216877367	BB	1.#J	BB	15.02	BB	1.#J	BB	1.#J	BB	0.14
686	SNP_A-1509797	1	217353869	AA	1.#J	AA	15.02	AA	1.#J	AA	1.#J	AB	0.00
687	SNP_A-1513668	1	217356078	NoCall	1.#J	AA	15.02	AA	1.#J	AA	1.#J	AA	0.47
688	SNP_A-1513479	1	217705653	BB	1.#J	BB	15.02	BB	1.#J	BB	1.#J	BB	0.47
689	SNP_A-1514400	1	218136977	BB	1.#J	BB	15.02	BB	1.#J	BB	1.#J	AB	0.00
690	SNP_A-1510531	1	218143097	BB	1.#J	BB	15.02	BB	1.#J	BB	1.#J	AB	0.00
691	SNP_A-1512436	1	218808913	AA	1.#J	AA	15.02	AA	1.#J	AA	1.#J	AB	0.00
692	SNP_A-1512490	1	218809071	BB	1.#J	BB	15.02	BB	1.#J	BB	1.#J	AB	0.00
693	SNP_A-1507333	1	219135381	NoCall	1.#J	BB	15.02	BB	1.#J	BB	1.#J	BB	0.24
694	SNP_A-1508403	1	219529663	BB	1.#J	BB	15.02	BB	1.#J	BB	1.#J	BB	0.24
695	SNP_A-1511940	1	219983559	BB	1.#J	BB	15.02	BB	1.#J	BB	1.#J	AB	0.00
696	SNP_A-1514260	1	219985330	AA	1.#J	AA	15.02	AA	1.#J	AA	1.#J	AB	0.00
697	SNP_A-1518174	1	221128285	BB	1.#J	BB	15.02	BB	1.#J	BB	1.#J	BB	0.69
698	SNP_A-1512418	1	221171487	AA	1.#J	AA	15.02	AA	1.#J	AA	1.#J	AA	0.69
699	SNP_A-1513846	1	221174780	BB	1.#J	BB	15.02	BB	1.#J	BB	1.#J	BB	0.69
700	SNP_A-1512671	1	222051072	AA	1.#J	AA	15.02	AA	1.#J	AA	1.#J	AA	0.69
701	SNP_A-1512799	1	223370406	AA	1.#J	AA	15.02	AA	1.#J	AA	1.#J	AB	0.00
702	SNP_A-1508029	1	223489860	BB	1.#J	BB	15.02	BB	1.#J	BB	1.#J	BB	0.36
703	SNP_A-1512893	1	223649081	BB	1.#J	BB	15.02	BB	1.#J	BB	1.#J	BB	0.36
704	SNP_A-1512754	1	224079709	AA	1.#J	AA	15.02	AA	1.#J	AA	1.#J	AB	0.00
705	SNP_A-1514941	1	224757266	BB	1.#J	BB	15.02	BB	1.#J	BB	1.#J	BB	0.65
706	SNP_A-1518523	1	224807362	BB	1.#J	BB	15.02	BB	1.#J	BB	1.#J	BB	0.65
707	SNP_A-1513157	1	224807653	AA	1.#J	AA	15.02	AA	1.#J	AA	1.#J	AA	0.65
708	SNP_A-1511741	1	225465567	BB	1.#J	BB	15.02	BB	1.#J	BB	1.#J	AB	0.00
709	SNP_A-1519296	1	225498036	AA	1.#J	AA	15.02	AA	1.#J	AA	1.#J	AB	0.00



HomoSNP Program

→ Homozygosity by descent shared by all 4 affected siblings on chromosome 1.

Founder effect (1)



Initial population

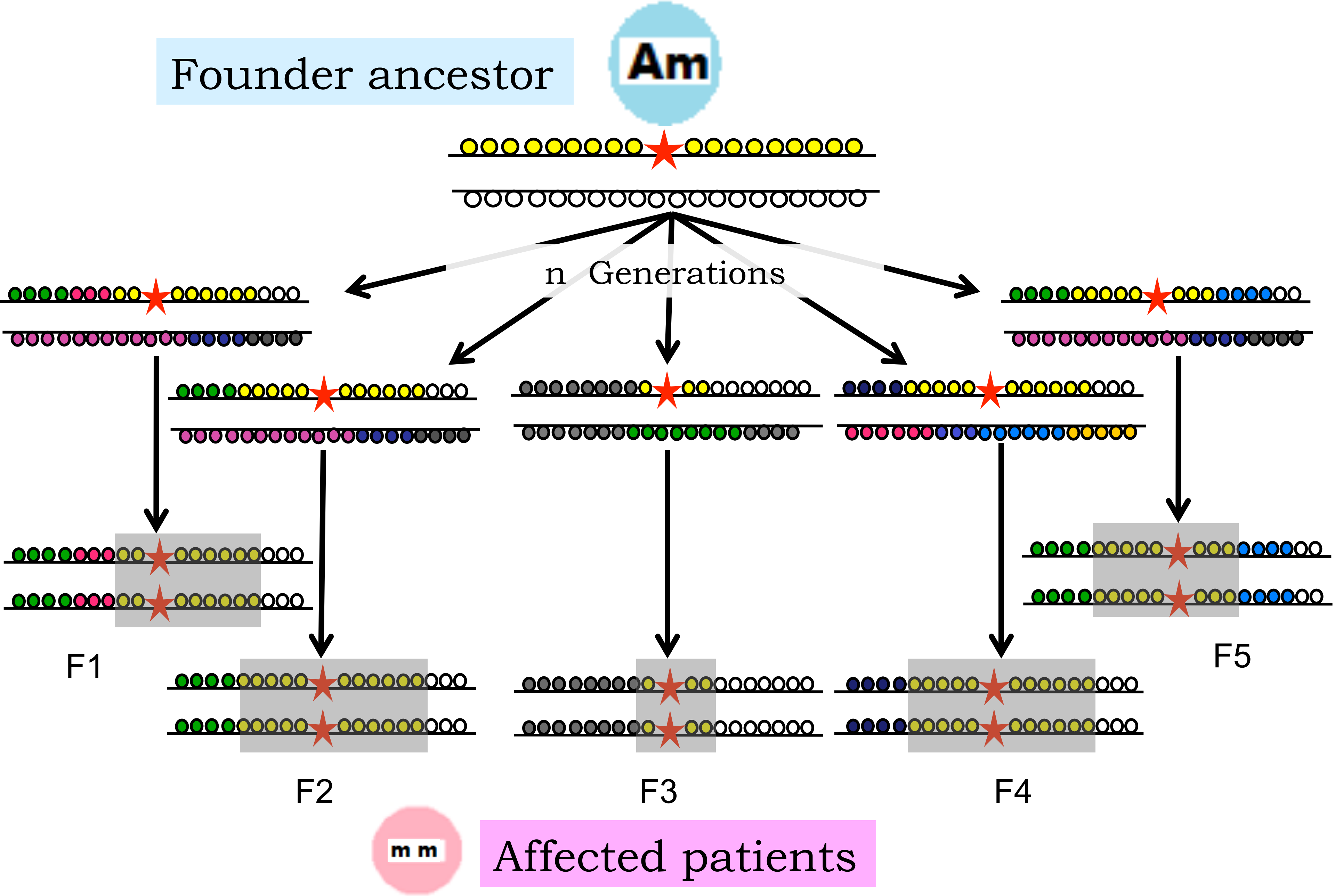
New population :
High frequency of mutant allele

Affected

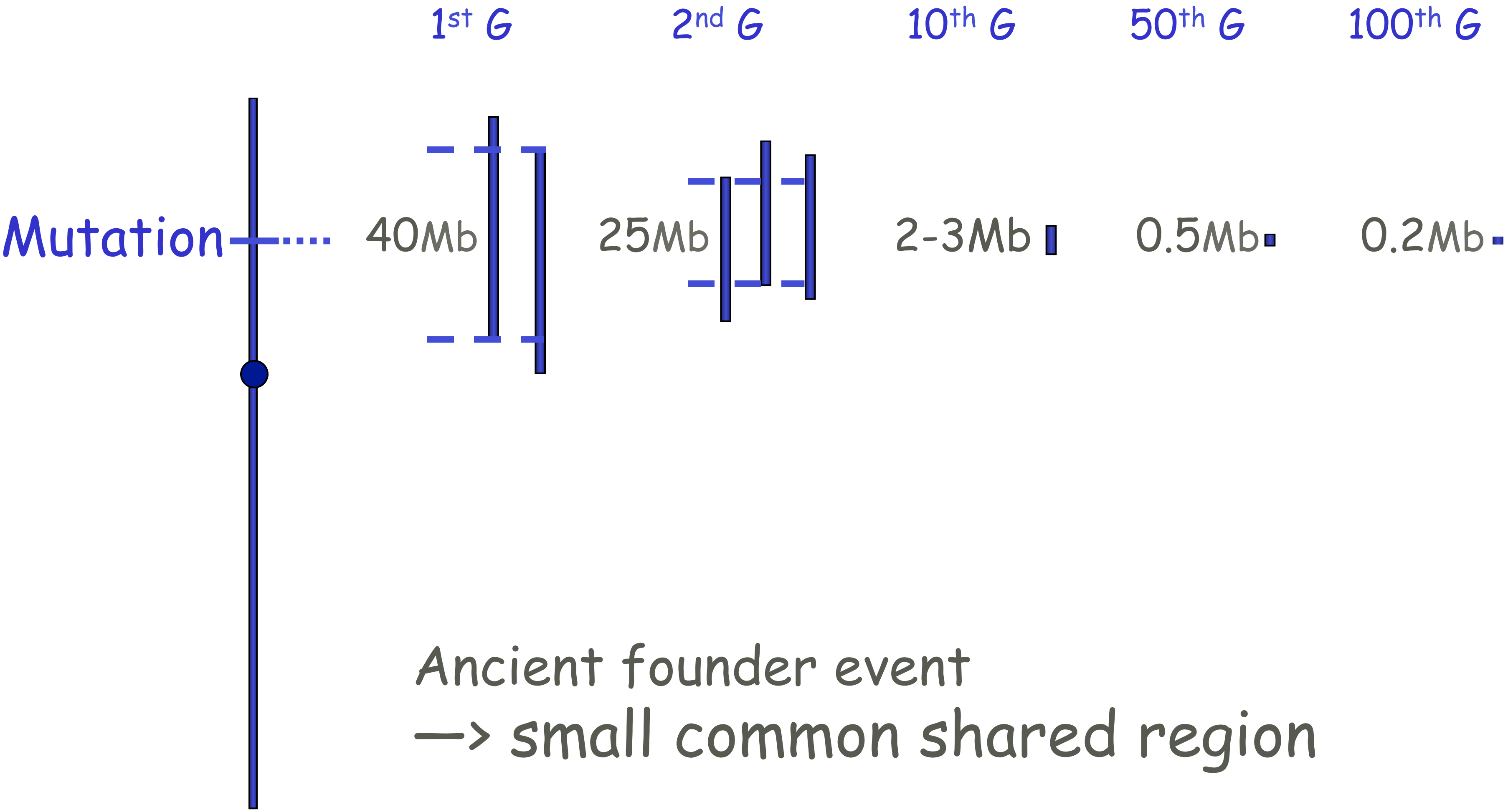
Bottleneck:
Restriction of
initial population

New population is
derived from small
population

Founder effect (2)



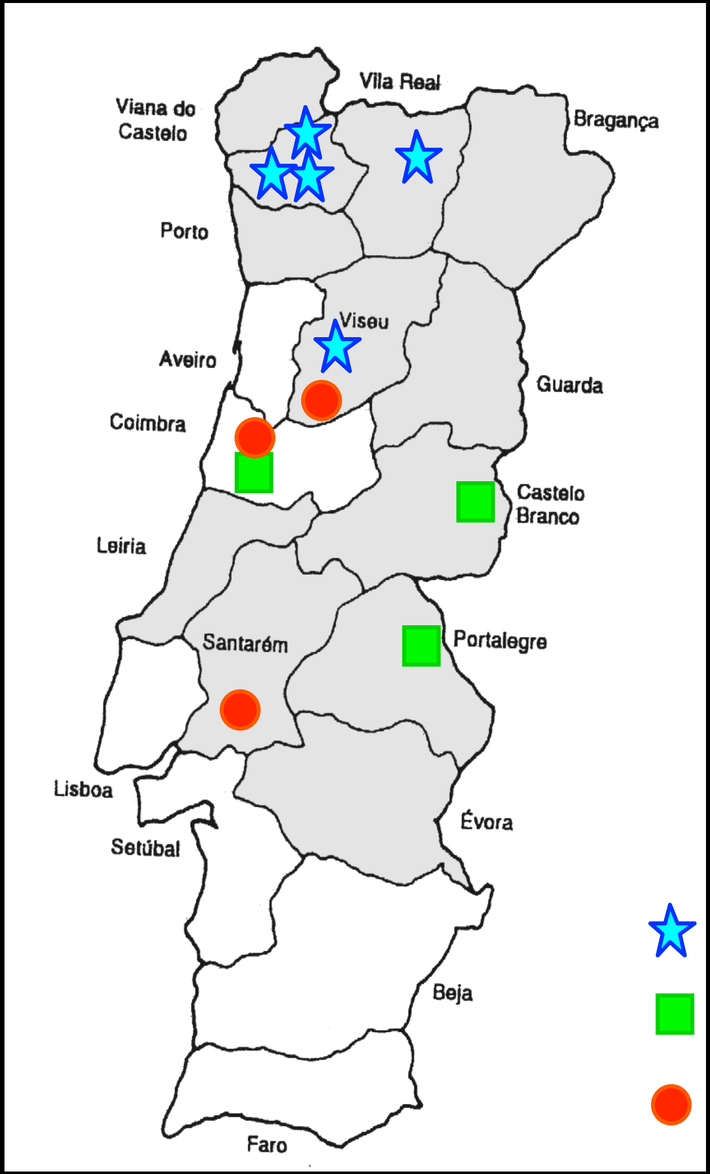
Variation of the length of a common shared region among patients, around the founder mutation

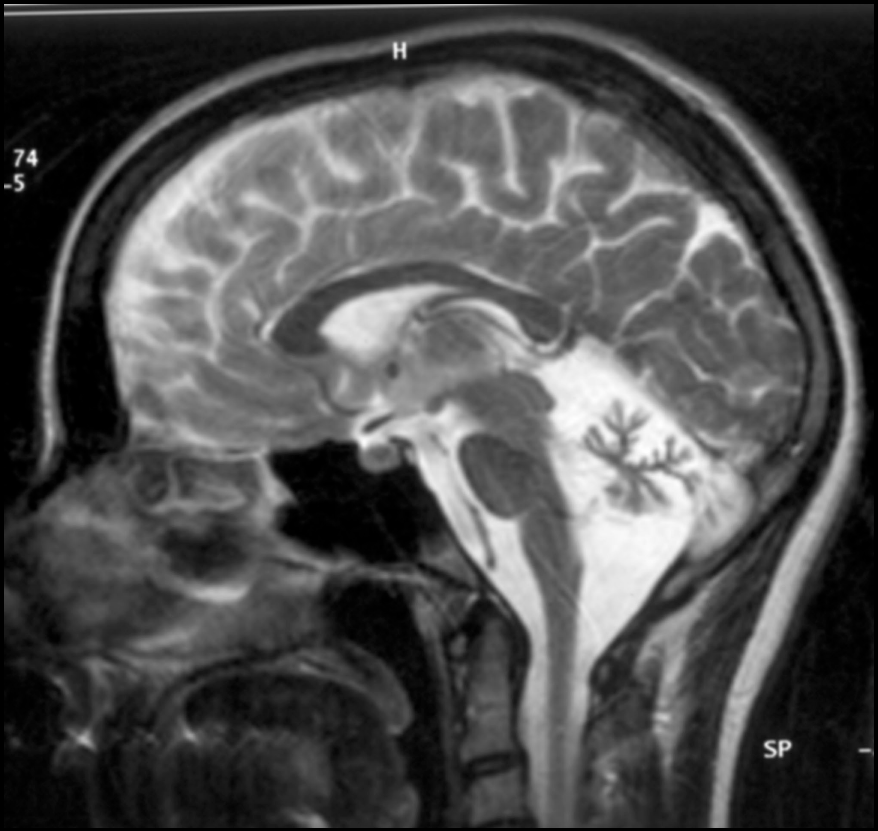


Ancient founder event
—> small common shared region

Founder effect

Geographical distribution, on the Portuguese mainland, of families with AOA1 (linked to 9p13)





★ families linked to 9p13

🟩 families with unknown linkage status

● families not linked to 9p13

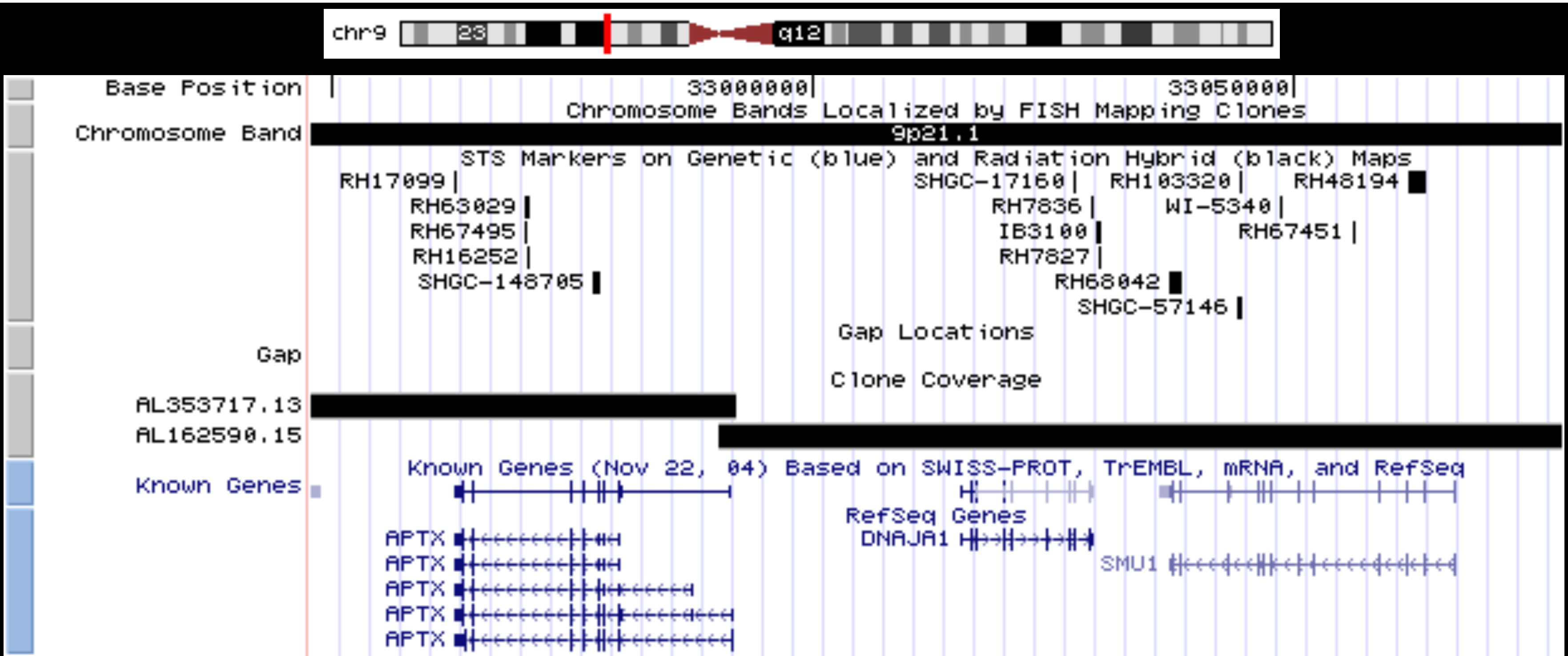
Founder haplotype for AOA1

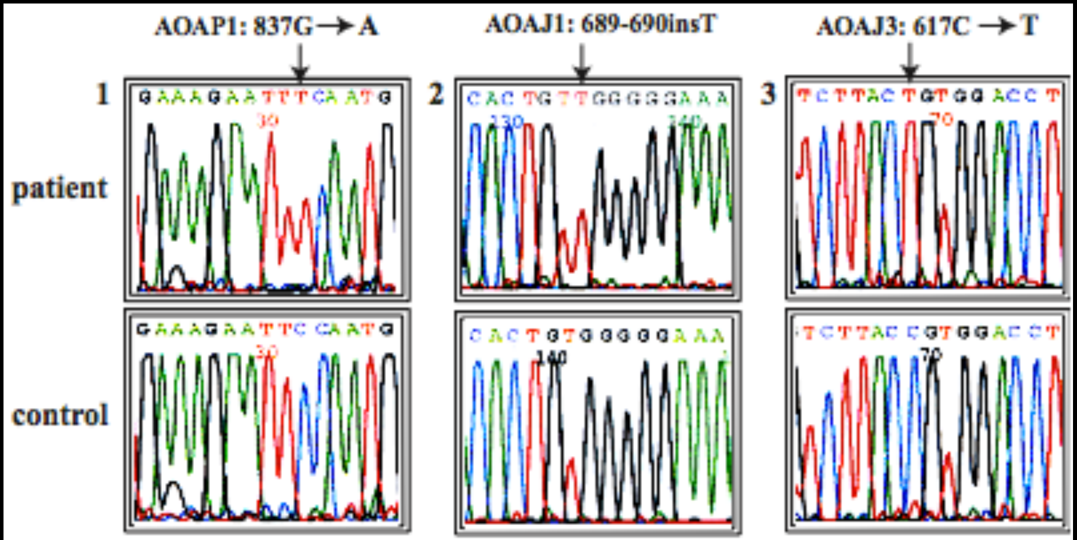
	AOAP4	AOAP11	AOAP7	AOAP1	AOAP5	AOAJ2	AOAJ1	AOAJ3
D9S1853	8 - 1	7 - 8	9	4	1	1	1	1
MS1 IRE-BP1	4 - 3	3 - 2	3	3	3	4	2	3
MS2 IRE-BP1	5 - 1	1 - 5	1	2	2	5	4	2
MS30	8	8 - 7	8	8	7	7	9	7
MS31	3	3 - 3	3	3	5	2	5	2
MS25	4	4 - 9	4	4	4	4	4	4
MD24	2	2	2	2	2	5	5	5
D9S1788	6	6	6	6	6	3	3	3
D9S1845	9	9	9	9	9	16	16	18
D9S165	11	11	11	11	11	10	10	11
MS26	7	7	7	7	7	5	5	5 - 9
MS28	3	3	3	3	3	3	3	3 - 4
MS21	3	3	3	3	2	3	3	2 - 4
D9S1878	14	14	14	14	6	5	5	5 - 14
D9S1817	12	12	12	12	13	16	16 - 13	16 - 15
D9S1805	3	3	3	3	3	4	4 - 4	4 - 5

AOA1 refined interval

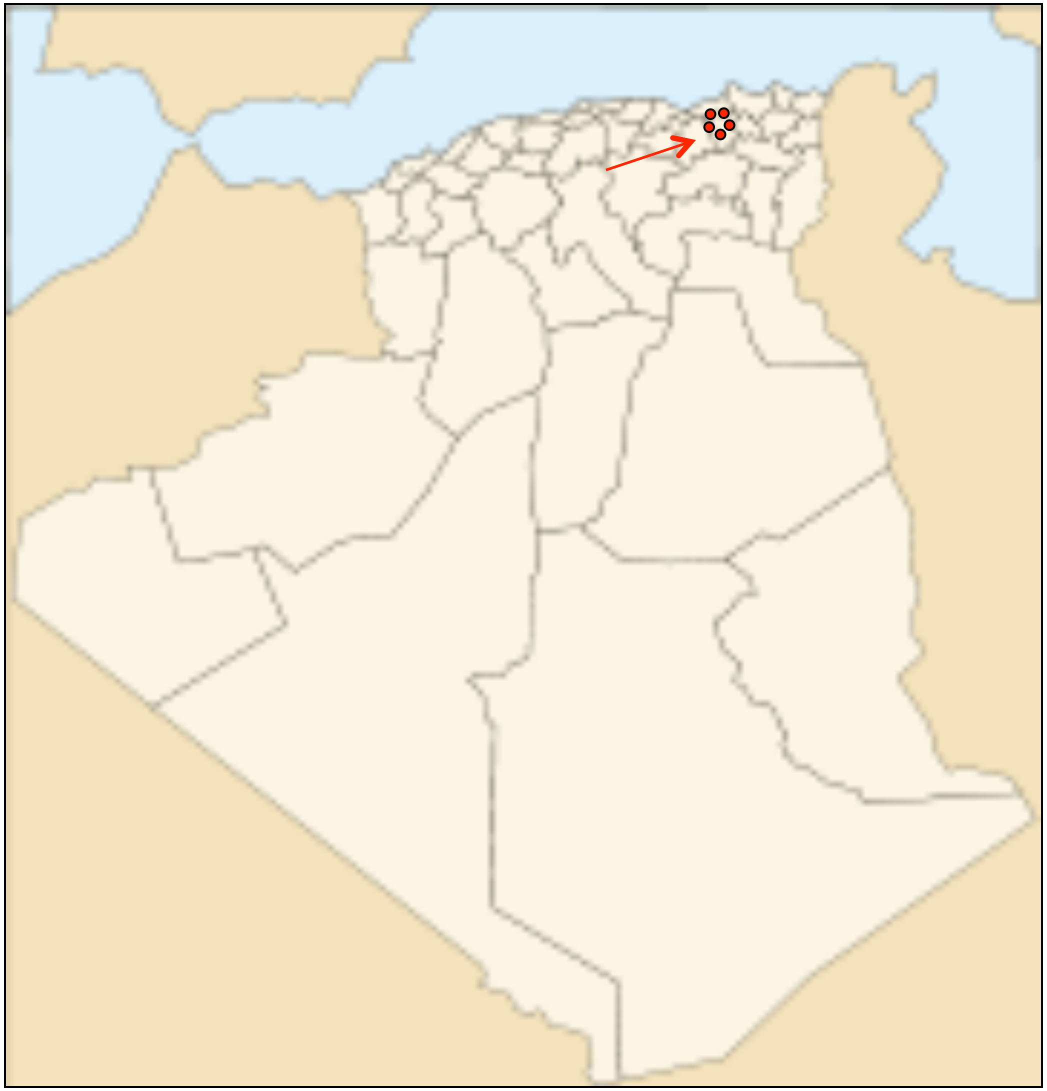
Linkage disequilibrium

UCSC Genome Browser on Human May 2004 Assembly

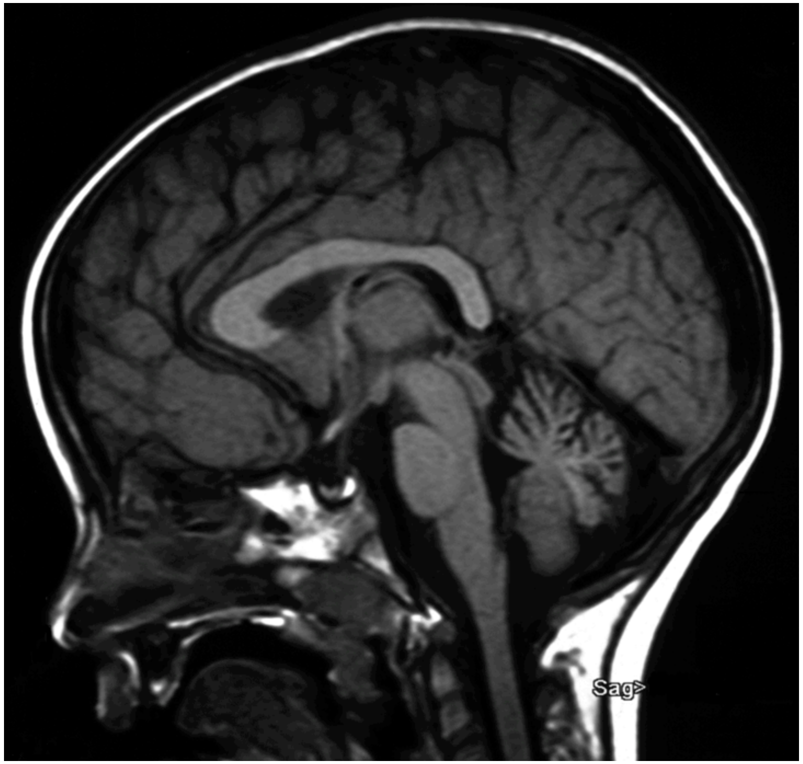




Geographic origin of the 5 Algerian families linked to chr20 North-East of Algeria (Sétif wilaya)



PHARC syndrom (Polyneuropathy, Hypoacousia, Ataxia, Retinitis P., Cataract)



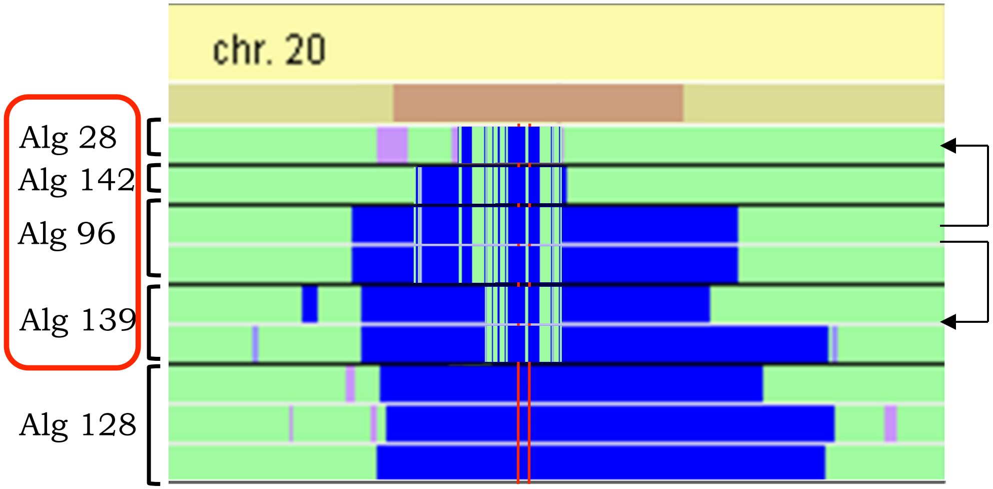
Founder haplotype in 4 out of 5 families linked to chr. 20

p13
p12.1
q12
q13.2

		Alg 128		Alg 139		Alg 96		Alg 142		Alg 28	
19601543	**rs720436**	A	A	A	A	B	B	A	B	A	B
20106062	rs1028434	B	B	ND	ND	B	B	B	B	B	B
20264206	rs928066	B	B	ND	ND	B	B	B	B	B	B
20287310	rs721424	A	A	A	A	A	A	A	A	A	B
20483833	rs1074440	B	B	B	B	A	A	A	A	B	B
20610543	D20S912	301	301	297	297	299	299	299	299	301	301
20826383	rs2038383	B	B	B	B	B	B	B	B	B	A
20930386	rs755963	B	B	B	B	B	B	B	B	B	A
21258407	**D20S190**	253	253	255	255	259	259	259	259	259	261
22854446	rs1888610	B	B	B	B	A	A	A	A	A	A
22923923	rs717756	B	B	ND	ND	A	A	A	A	A	A
22987041	rs2007743	A	A	ND	ND	A	A	A	A	A	A
23271987	rs1112819	B	B	ND	ND	B	B	B	B	B	B
23283569	D20S871	194	194	190	190	194	194	194	194	194	194
23295553	rs999072	B	B	B	B	A	A	A	A	A	A
23532116	rs726217	A	A	A	A	B	B	B	B	B	B
23570758	rs2254635	B	B	ND	ND	A	A	A	A	A	A
23573547	**rs2145231**	B	B	A	A	B	B	B	B	B	B
23937652	rs3843776	B	B	A	A	A	A	A	A	A	A
23937782	rs3843777	A	A	B	B	B	B	B	B	B	B
23937930	rs3848799	B	B	B	B	B	B	B	B	B	B
24033110	rs761863	A	A	B	B	B	B	B	B	B	B
24122028	rs487665	B	B	ND	ND	B	B	B	B	B	B
24371416	rs722834	A	A	A	A	A	A	A	A	A	A
24996940	rs2387577	A	A	A	A	A	A	A	A	A	A
24997283	rs2207631	B	B	B	B	B	B	B	B	B	B
25127155	rs2387733	B	B	ND	ND	A	A	A	A	A	A
26114910	D20S191	227	227	229	229	229	229	229	229	229	229
29310063	rs1474945	B	B	B	B	B	B	B	B	B	B
29469970	rs721220	A	A	A	A	A	A	A	A	A	A
29940293	D20S111	249	249	251	251	251	251	251	251	251	251
30429539	D20S200	283	283	271	271	271	271	271	271	271	271
31123410	**DH1**	263	263	267	267	265	265	265	265	ND	ND
31599059	**D20S890**	199	199	213	213	210	210	210	210	208	210
31929741	D20S878	229	229	225	225	227	227	227	227	227	227
31982015	rs725478	A	A	A	A	A	A	A	A	A	A
32000125	DH2	194	194	202	202	202	202	202	202	ND	ND
32309697	rs2378132	A	A	A	A	A	A	A	A	B	A
32325217	rs819144	A	A	A	A	A	A	A	A	B	A
32325488	rs819145	B	B	B	B	B	B	B	B	A	B
33431481	**rs725908**	A	A	ND	ND	A	A	A	B	B	A
33915657	D20S909	151	151	145	145	153	153	153	153	ND	ND
34316842	D20S847	157	157	153	153	151	151	ND	ND	ND	ND
34166072	rs3850528	A	A	B	B	A	A	A	A	A	A
35310424	rs1073768	A	A	B	B	A	A	A	B	B	A

CDK5RAP1
H470P

Homozygous mutation of the α/β hydrolase 12 (ABHD12) gene in 4 Algerian families

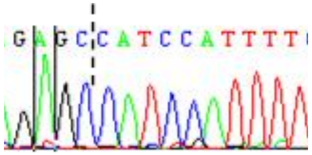


Affymetrix SNP array 6.0
(906.000 SNP) :

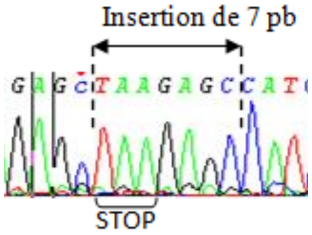
9 Mb/1950 SNP /60 genes

6 Mb/642 SNP /33 genes

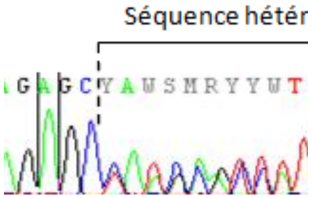
control



patient



carrier



Linkage for multifactorial diseases :

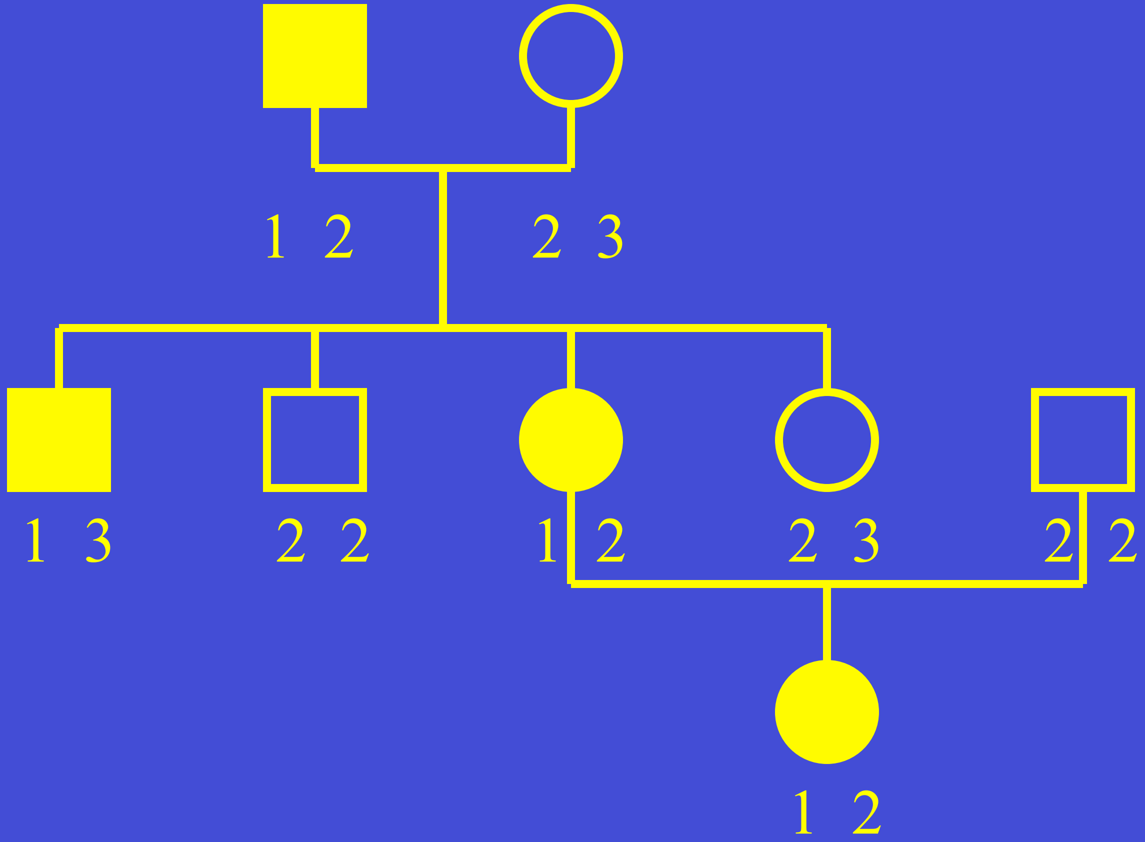
Same principles

(genetic linkage, linkage disequilibrium = association studies)

But very small penetrance

–> small LOD scores despite very large cohorts of patients

LOD score


1 2
2 3
1 3
2 2
1 2
2 3
2 2
1 2