L'information familiale est elle encore utile à l'ère de la génomique?

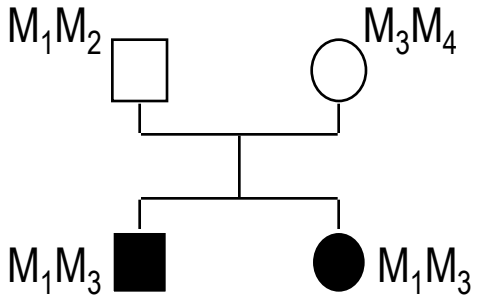
Françoise Clerget-Darpoux
INSERM U781 - Paris Necker

Genome-wide linkage studies

Twenty years ago, genome-wide linkage study on affected sib pairs was a very popular strategy for searching genetic risk factors involved in a multifactorial disease.

Affected Sib Pair method

Sample of affected sibs typed for genetic markers



IBD variable measures, for each marker, the number of parental alleles Identical By Descent

	observed proportion	expected if no linkage
IBD = 2	z_2	1/4
IBD = 1	z_1	1/2
IBD = 0	z_0	1/4

Some successes...

Mapping a susceptibility gene for Crohn's disease on 16q

Hugot et al,1996

→ ***NOD2***

Genome-wide linkage studies

For most factors :

Low power of detection

Low power of replication

For factors which were detected :

large uncertainty on their location

A second generation screen of the human genome for susceptibility to insulin dependent diabetes mellitus

Concannon et al (Nature Genetics, 1998)

Suceptibility loci	Localisation	# de paires	MLS
IDDM1 (HLA)	***6p21***	618	**32.5**
IDDM2 (Insulin)	***11p25***	607	0.60
IDDM3	15q26	506	0.03
IDDM4	11q13	778	0.43
IDDM5	6q25	852	1.46
IDDM6	18q21	302	0.00
IDDM7	2q31	653	0.72
IDDM8	6q27	730	1.14
IDDM9	3q21	543	0.23
IDDM10	10 cen	609	0.40
IDDM11	14q24	433	0.28
IDDM12 (CTLA4)	***2q33***	585	0.84
IDDM13	2q34	418	0.36
IDDM15	6q21	772	**3.51**
IDDM --	*1q*	*798*	*2.84*

Genome-wide linkage studies

It turns out to be like «a manic depressive history »

The euphoria of linkage findings being replaced by the dysphoria of non-replication has become a regular pattern in genetic studies. (N. Risch & D. Botstein, 1996)

« *We argue that [the linkage] method has limited power to detect* ***genes of modest effect*** *but that a different approach that utilizes candidate genes has far greater power, even if one needs to test every gene in the genome. Thus, the future of the genetics of complex diseases is likely to require large-scale testing by association analysis* »

Risch et Merikangas, Science, 1996

What is a gene of modest effect ?

Let assume the effect is due to **a single di-allelic variant** (s,S) in the gene:

- two alleles with frequencies **q** , 1-q.
- three genotypes ss sS SS
  with penetrances f **GRR_1**f **GRR_2**f

GRR_1 and **GRR_2** are the relative penetrances the effect of the gene is most often reported in terms of a single parameter value **OR = GRR_1** in assuming **$GRR_2 = (GRR_1)^2$**

Power of linkage detection in affected sib pairs

The distribution z2, z1, z0 and consequently,

the power of linkage detection,

depend on q, GRR_1, GRR_2

Effect of a genetic factor in linkage analysis

The greater the deviation from **(0.25, 0.50, 0.25),** the higher the power of linkage detection.

q	z2	z1	z0
0.00	1.00	0.00	0.00
0.10	**0.83**	**0.17**	**0.01**
0.20	0.69	0.28	0.03
0.35	0.56	0.38	0.06
0.55	0.42	0.46	0.13
0.75	0.33	0.49	0.18
0.90	0.28	0.50	0.22
1.00	0.25	0.50	0.25

allele with recessive effect

q	z2	z1	z0
0.00	0.50	0.50	0.00
0.10	**0.43**	**0.49**	**0.08**
0.20	0.38	0.49	0.14
0.35	0.33	0.49	0.19
0.55	0.29	0.49	0.22
0.75	**0.26**	**0.50**	**0.24**
0.90	0.25	0.50	0.25
1.00	0.25	0.50	0.25

allele with dominant effect

(Thomson et Bodmer, 1977)

Peu de puissance pour détecter un allèle à risque fréquent

Comparing the power of association and linkage tests: insulin gene and type 1 diabetes

Association with an allele (short length) of a VNTR in the insulin gene promotor

allele frequency : 85% in cases versus 75% in controls *(Bell et al, 1984).*

No evidence for linkage in 95 affected sibpairs
Spielman, Baur & Clerget-Darpoux (GAW5 ,Chantilly,1987)

It exists genetic models for which a factor cannot be detected through linkage but can be detected through association. Particularly true for frequent risk allele.

« We argue that [the linkage] method has limited power to detect genes of modest effect but that a different approach that utilizes ***candidate genes*** *has far greater power, even if one needs to test every gene in the genome. Thus, the future of the genetics of complex diseases is likely to require large-scale testing by* ***association analysis*** *»*

Risch et Merikangas, Science, 1996

⇨⇨Projet Hapmap

Etudes d'association pangénomiques avec des tagSNPs

Etudes d'association pangénomiques des diabètes de type 1 et 2

gène	Région	Freq all	OR
HLA	*6 p21*		
Insuline	*11p15*	0.38	1.25
PTPN22	*1p13*	0.09	1.98
CTLA4	*2q33*	0.55	1.18
ILR2	*10p15*	0.87	1.33
KIAA0350	16p13	0.68	1.20
?	12q24	0.42	1.22
ERRB3	12q13	0.34	1.28
PTPN2	18p11	0.16	1.30

Diabète de type 1 *Nature Genetics, 2007*
(500K SNPs sur 6000 cas/ 6200 témoins)

gène	region	Freq all	OR
TCF7L2	10	0.26	1.37
CDKN2B	9	0.83	1.20
IGF2BP2	3	0.30	1.14
FTO	16	0.38	1.17
KCNJ11	11	0.47	1.14
CDKAL1	6	0.31	1.12
HHEX	10	0.53	1.13
SLC30A8	8	0.65	1.12
PPARG	*3*	*0.86*	*1.14*

Diabète de type 2 *Science 2008*
(500K SNPs sur 14.586 cas/17.968 témoins)

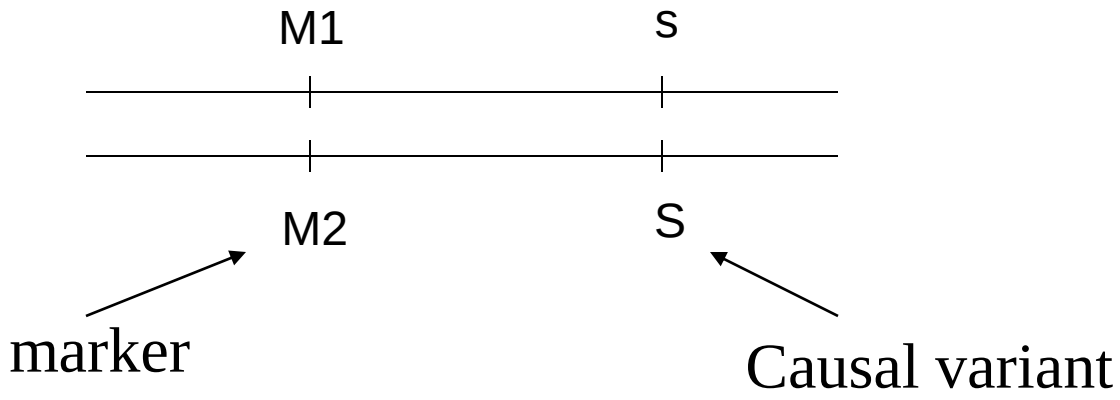
Candidate gene strategy

The aim is to get information on the candidate gene in order to show and model its effect (measurement of the differential risks)

It cannot generally be achieved by using association information only.

To model the effect of a candidate gene

Let first consider a very simple situation where the gene effect is only due to a single causal variant with two alleles s,S of frequency q, 1-q. Let assume that we have the typing of an intragenic marker with two alleles M1 et M2 of frequency p, 1-p.



Can we infer from the marker observation the role of the causal variant ?

Is the the marker genotype distribution unambiguous for the modeling?

To a same marker genotypic distribution in case and control samples may correspond different models for the causal variant(s)

Example : the class1 allele (M1) of the VNTR flanking the Insulin gene in Type I Diabetes

	M1M1	M1M2	M2M2
Controls (171) *Bell et al, 1984*	74	82	15
Patients (77) *GAW5,1987*	58	18	1

Modeling the Insulin gene effect in Type I diabetes

The observed distributions of the VNTR are compatible :

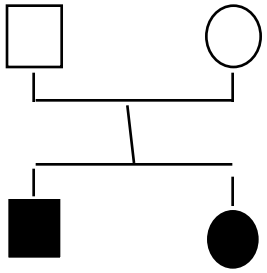
- with a direct role of the VNTR (0.75, 0.25)
  (situation 1)

- with the role of a biallelic variant (0.14, 0.86)
  in LD with the VNTR
  (situation 2)

Under such hypotheses, the sharing of affected sib pairs for the insulin gene are not the same

Affected Sib Pair (type 1 diabetes) IBD sharing on the insulin gene

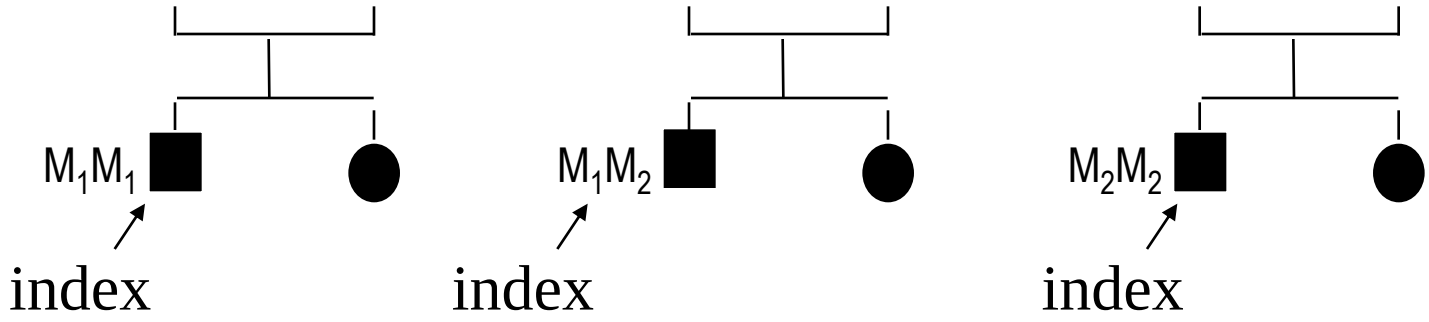
IBD distribution depends on the underlying model



	IBD =2	IBD=1	IBD=0
Expected in situation 1	0.27	0.49	0.24
Expected in situation 2	0.47	0.46	0.07
observed	**0.26**	**0.50**	**0.24**

Situation 1 = role of the VNTR

Utilisation simultanée des informations liaison et association stratification des paires de germains



Si pour le gène candidat, les génotypes confèrent des risques différents, les vecteurs IBD des paires de germains seront différents selon le génotype de l'index

⇨ **vecteurs IBD stratifiés** *(méthode MASC; Clerget-Darpoux et al, 1988)*

IBD stratifié sur la longueur du VNTR du promoteur de l'insuline

M1 = longueur courte = allèle à risque

Index genotype	IBD= 2	IBD=1	IBD =0
M1M1	.29	.50	.11
M1M2 and M2M2	.10	.50	.40
all	.26	.50	.24

Utilisation de l'information simultanée de liaison et d'association

Arthrite rhumatoïde et PTPN22

Association entre arthrite rhumatoïde et un tagSNP de PTPN22 : rs2476601 (OR = 1,66).
Modélisation de PTN22 sur 511 paires de germains atteints
Bourgey M, Perdry H, Clerget-Darpoux F

- IBD stratifié incompatible avec effet du tagSNP
- IBD stratifié compatible avec effet d'une combinaison de 3 SNPs

Différentiel de risque entre génotype le plus à risque et le moins à risque est de **2,7** pour le tag SNP et de **4,68** pour la combinaison des trois SNPs

Rheumatoid Arthritis and PTPN22

Rs2476601	GRR	3 SNPs	GRR
CC	1	CC-AA-AA	1.60
		CC-AA-AG	1.76
		CC-AA-GG	**3.60**
		CC-AG-AA	1.73
		CC-*G-AG	**2.35**
		CC-GG-AA	1
CT	1.66	CT-AA-AA	2.88
		CT-AA-AG	3.11
		CT-AG-AA	2.61
TT	2.7	TT-A*-AA	**4.68**

*, either the A or the G alleles of rs12730735.

Autres exemples

IL2RA et Sclérose en plaques

Différentiel de risque entre génotype le plus à risque et le moins à risque est de :

1,56 pour le tag SNP retenu dans l'étude pangénomique
4 pour la combinaison de SNPs obtenu par modélisation d'IL2RA

Insuline et Diabète de type 1

OR de 1,25 pour tag SNP et de 3,4 pour le VNTR

HLA et Maladie Coeliaque

OR de 7 pour tag SNP et de 25 pour un hétérodimère DQA-DQB spécifique

Rare variants
and multifactorial diseases

Rare variants may explain some of the missing heritability

From Maher, 2008

The case of the missing heritability

Heritability not interpretable unless unrealistic hypothesis

Modeste by André Franquin

Advantage of using patients with familial history

Variant with a risk allele frequency q=1% and OR= 4

	Genotypic distribution for this variant (a, A)	
	aA	AA
General population	2%	98%
Patient population	8%	92%
Population of patients having an affected sib	17%	83%

Coll: H Perdry, B Müller-Myhsok

Sample size to sequence to show the variant involvement

In a genome wide search ($\alpha = 5 \times 10^{-8}$) when comparing patients to 1000 control genomes

S1 = Unrelated patients	2030 (P= 80%)
S2 = Patients having an affected sib	200 (P > 99%)

Large benefit in terms of cost sequencing

For an OR of 3 instead of 4, the variant will be also detected in S2 with a good power (P~80%) but not in S1 (P~20%).

Additional advantage information provided by the affected sibs

The proportion of affected sibs
sharing 2, 1 or 0 parental haplotypes
Identical By Descent (IBD = 2,1,0)
depends on the genotype of the sequenced patients

200 sequenced patients (S2)		IBD=2	IBD=1	IBD=0
34	**aA**	**.39**	**.51**	**.10**
166	AA	.24	.50	.26
All 200 sib pairs		.26	.50	.24

Demonstrating the variant involvement through linkage information

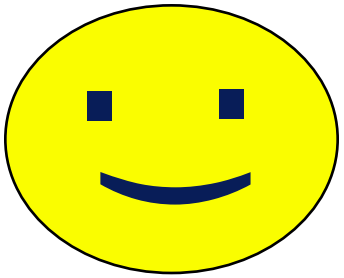
After detection of a rare risk allele through the sequencing of patients having an affected sib, its involvement can be **further confirmed** by comparing, in the subset of patients carrying the rare allele, the numbers of sibs IBD=2 and IBD = 0

In the previous example: **≈ 34 sibs**

→ power >90%

Advantages of studying patients with familial history

Two advantages:



- cost of sequencing
- demonstration of the variant involvement

Familial information is crucial

- ✓ to prove the involvement of rare variants
- ✓ to distinguish transmitted and de novo mutations (CNV)
- ✓ to model the causal variation and to correctly estimate the genotypic relative risks
- ✓ to have power for demonstrating GxG interaction

Take Home Message

Contrasting the genome
of unrelated patients and healthy controls
will not face the complexity of human diseases.

So many different biological mechanisms are possible
that it would be foolhardy to neglect
any kind of information, in particular, familial information

Bourgain C, Génin E, Cox N, Clerget-Darpoux F. Are genome-wide association studies all that we need to dissect the genetic component of complex human diseases? Eur J Hum Genet, 2007,

Clerget-Darpoux F & Elston R : Are Linkage Analysis and the Collection of Family Data Dead? Prospects for Family Studies in the Age of Genome-Wide Association. Human Genetics, 2007

Perdry H, Muller- Myshok B, Clerget-Darpoux F A new analytical approach to prove the involvement of a rare variant in disease susceptibility. in preparation