Searching the causes of complex diseases

A.-j. Valleron, G. Biau & P. Bougnères

alain-jacques.valleron@inserm.fr
gerard.biau@upmc.fr
pierre.bougnieres@inserm.fr
agenda

• Two categories of diseases
  – « simple » diseases: Genetically (G) determined.
  – « complex » diseases: Genetic and Environmental € factors.

• The methods and sources of data for searching the G and E causes of complex diseases.

• The ISIS cohort of type 1 diabetes patients
  – Why DT1 is a good model
  – The data available

• Some possible contests.
Two classes of diseases

• « simple » diseases
  also called monogenic, mendelian
  e.g. Cystic Fibrosis
  4000 « simple » diseases

• « complex » diseases
  e.g. : cancer, Alzheimer, autism, obesity,
  diabetes type 2, Diabetes type 1 (juvenile), …
  < 200 complex diseases
« simple » diseases

1 single deleterious mutation in 1 single gene dramatically changes the nature or quantity of a crucial protein. Because this protein acts in a physiological pathway, an important function of the organism is altered, whatever the environment of the individual can be. Mutation is necessary and sufficient to induce the disease.

Familial or sporadic (if mutation just happens de novo).

Most causal genes are known Biology is clear (usually).
Complex diseases: Genetics

(N non deleterious mutations (polymorphisms) in N different genes mildly change the quality or quantity of a set of N proteins. as these proteins acts in various physiological pathways, important functions of the organism are slightly modified.
Complex diseases: environment

Direct evidence that many environmental factors can cause complex diseases:
- Tobacco and lung cancer, cardiovascular diseases
- Social factors and obesity
- Exposure to sun and cancer of the skin
- ...

Indirect evidence
- Incidence of Type 1 Diabetes has doubled in 20 years
- When a twin develops T1D in childhood, his monozygotic (MZ) twin develops T1D in <50% of cases
- ...

©inserm, 2013
the G x E causality of complex diseases

Complex diseases require interaction of genetics and environment to occur.

Different combinations of genetic polymorphisms (« G ») and environmental factors (« E ») can cause the « same » disease (heterogeneity of causal interactions).

Many « associated » genomic variants (frequencies different in cases and controls) are known, but very few causal genes are known: the relationship between associated variants and alteration of gene function is elusive.

Nothing is known about statistical sub-grouping of (nG x nE) interactions: G and E are being studied independently of each other.

Nothing is known about the GxE biological interaction.
Type 1 Diabetes (T1D) as a model of complex diseases

• T1D (aka juvenile diabetes, Insulin Dependent Diabetes)
  – Incidence between 2/100,000 (Japan) and 30/100,000 (Finland)

• Our objectives
  – Find associations of E factors with T1D
  – Find « statistical interactions » of these associated E factors with G
  – Follow the track to find causal E (statistically related with associated E)
  – Understand biology of E and biological interaction of E with G
Briefly: Epidemiological methods

• The search for G factors was « hypothesis driven » up to the 90’s
• After the genomic revolution, the search became « data driven »

• The search for E factors is still predominantly « hypothesis » driven
• We plan to search E and ExG causality with a data driven approach.
Methods to search the genetic causes of complex diseases
Pre(gen)omic era

Family studies: « Pedigrees » are studied and fit to genetical models.

« candidate genes » are searched experimentally, and validated on men.

HYPOTHESIS DRIVEN RESEARCH
Methods to search the genetic causes of complex diseases genomic era

1990+ : launch of the human genome project
2003 : Official presentation of the 1st human genome

1999 : TSP (The SNP consortium)
2005 : 1st GWAS study

DATA DRIVEN RESEARCH
Genotyping

• 3.4 billions of pairs of bases (A, C, G, T)
  ...ATCCACG...

• The genomes of two persons are identical in >99.9% positions.

• In 10-15 millions positions (the Single Nucleotid Polymorphisms, SNP), the genotypes may differ

• Chips sample from 300,000 to several millions of these SNPs (all of them: « full exome »)
Principle of a GWAS study

Genotyping of
• N cases of the disease
• N « controls »

The distributions of the genotypes are compared at each position with a statistical test.
The discovery of a « significant » SNP is not the discovery of a gene. It opens a path towards the identification of genes.

(miillions of tests : false discovery rates...)

©inserm, 2013
Example
(data from ISIS : comparison of 1200 T1D to controls)
The proportion of attributable risk

- GWAS study give for each SNP odds ratio OR (which approximate relative risks RR)
- RR = risk of having disease if (e.g.) genotype AT/ risk of having disease if their genotype is not AT.
- Assume that this genetic factor is causal
- Assume that the fraction of subjects with the genetic risk factor is f, the proportion AR of cases explained by the factor is

\[ AR = f \frac{RR - 1}{f( RR - 1) + 1} \]
In most of the GWAS studies of complex diseases, the RR are small (typically 1.5)

If $f = 25\%$, $AR \approx 12 \%$

If $f = 5\%$, $AR \approx 2.5 \%$
Published Genome-Wide Associations through 06/2011, 1,449 published GWA at p≤5x10⁻⁸ for 237 traits

NHGRI GWA Catalog
www.genome.gov/GWAStrudies
The classical study of environmental factors (EF)

- In general « hypothesis driven »:
  (EF are the pesticides, the radiations, smoking, etc)

- Two designs:
  - Cohort studies (two groups, one exposed to the EF, the other not exposed are followed up and compared in terms of incidence of the disease)
  - Case-control studies (one group of patients with disease, is compared to an « appropriate » control group in term of their past exposure to EF.)
The « environnementome » project:
a data driven approach to study EF

• Parallel the genomic approaches
• Scan all possible markers of environmental variability
• Use high throughput methods of analyse to identify markers associated to disease
• Search the « real factor » behind the marker

E.g. : the marker is the proximity of a vineyard at the place of birth, which occurred in the fall.
The « real factor » is a pesticide used in the vineyard just before the harvest
The ISIS project
ISIS

- Multicenter cohort
- 6000 T1D patients, 75 centres
- C: bioclinical followup (e-CPR)
- G: genotyping of a subset of 2500 patients
- E: 800 questions questionnaire (subset of 1500 patients) + mapping of geocoded addresses of the patients to environmental databases.
Public WebSite: www.isis-diab.org
The « environnementome »

Address is geolocalized
(latitude, longitude)

Mapping of the address with environmental databases

Population density
Known on a 1km x 1km grid

Climate
15 daily parameters over the last 30 years
(grid 25km x 25km)

Carine Land Cover
Grid of 250m x 250m ©Inserm, 2013

Solid waste incinerators
(18,000 x N pollutants)

Weekly local exposure to infectious diseases after 1984 sentiweb.org
One example of environmental database: Corine Land Cover

Grid 250m*250 m or 100m*100 m

Soil Occupation (44 classes):
Data from photos satellites
1990, 2000

Source:
European Environment Agency (EEA)
## 2.2 Cultures permanentes

### 2.2.1 Vignobles

Parcelles plantées d'arbres fruitiers ou d'arbustes fruitiers : cultures pures ou mélange d'espèces fruitières, arbres fruitiers en association avec des surfaces toujours en herbe. Y compris les châtaigneraies et les noiseraies.

### 2.2.2 Vergers et petits fruits

Surfaces plantées d'oliviers, y compris oliviers et vignes sur la même parcelle.

### 2.2.3 Oliveraies

Surfaces plantées d'oliviers, y compris oliviers et vignes sur la même parcelle.
Environmental infectious exposures at birth
(the three younger daughters of AJV)

Measles

Influenza Like Illnesses
measure of environmental exposures example of infections

1 patient born on 15 december, 1995
Diagnosed with Diabetes Type 1 on 23 May, 2004

Many possible measures, exposure windows

- Cumulative sum of the local infectious exposures during the last 6 months of pregnancy, during the 1st year, the 2nd year, the 3rd year, .., between birth and diagnosis...
- Number of times when the patient has experienced the outbreak of one of the studied diseases during the last 6 months of pregnancy, etc...
- .....
Case control comparison of E exposures

• « virtual controls » are defined:
  For each case, a « virtual » control of same age (± 6 months) is randomly chosen on the map, in regions where patients have been recruited, proportionally to the population density of 0-15 aged. Exposures are estimated similarly to those of patients.

• « physical controls » (real patients with another disease also available)
How to study the G x E interaction

• Limitation:
G information is available only for patients

• Method: The « case-only » method
(see e.g. Khoury et al., Am. J. of Epidemiology, 1996)
Principle: at each of the 300,000 SNPs, the exposures of the patients are compared on the three genotypes at this SNP (e.g. CC, TC, TT)
The GxE relation
(example of one infectious E factor)

(En cliquant sur la case, on accède au graphique correspondant)

<table>
<thead>
<tr>
<th>nom</th>
<th>plus3plus18</th>
<th>moins3plus18</th>
<th>plus0plus12</th>
<th>plus0plus6</th>
<th>plus0plus36</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1004446</td>
<td>NS</td>
<td>NS</td>
<td>S</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>rs10501022</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>rs1064205</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>rs10743152</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>rs11557467</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>rs12463137</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>rs12849510</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>rs17400329</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>rs1810636</td>
<td>NS</td>
<td>S</td>
<td>NS</td>
<td>NS</td>
<td>S</td>
</tr>
<tr>
<td>rs2290400</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>rs2406233</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>rs2476601</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>rs249240</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
</tbody>
</table>
SNP : rs2406233  Maladie : vari

Variable : cumul

Taille ech :
C/C = 298  T/C = 291  T/T = 452

Mean ech (sd) :
C/C=888(623)  T/C=979(677)  T/T=1067(673)

Shapiro Test pval
C/C = 2.55e-08  T/C = 1.13e-13  T/T = 4.83e-11

kruskal.test  p.value = 0.000963
Bonferroni cor  p.value = 0.0337
Our assets

• The blindness of our data-driven search for E: no favorite hypotheses, our only limits are the available data bases

• Several environmental data bases have been worked out for geographic mapping at geocoded patients adress (nationwide is easiest in France than in USA states or other countries)

• There is large variety of space-time exposures in rural and urban populations of studied patients and controls

• The studied period is the period when T1D incidence has increased, most likely due to E forces
The objectives

• Find associations of E factors with T1D: some exposures are more (less) common in T1D than in controls: likely heterogeneous across patients (??)
• Explore interactions of these « associated E factors » with G
• Find the track from these « associated E » to the « causal » E
• Later (at laboratory) : Understand biology of E and biological interaction of E with G
The data that could be provided

• Characterization of 2500 patients genomes through genotyping 300,000 genomic polymorphisms (most of them common SNPs)
• Characterization of environmental exposures in these 2500 + 2000 patients
• Characterization of environmental exposures in virtual (or physical) controls
Examples of possible contests (supervised)

• **Supervised 1-** data set patients + controls. Both with E info.
  
  **Problem:** Classify new subjects as patients or controls.

• **Supervised 2a-** data set patients with E info only.
  
  Patients are classified in 3 clinical categories A, B and C
  
  **Problem:** classify new patients as A, B or C.

• **Supervised 2b-** data set patients with G, E and C info
  
  **Problem:** same as above
Examples of possible contests
(unsupervised)

• **Unsupervised:** data set 2500 patients with G and E info. Some C.

  **Problem:** Cluster the observations into k homogen groups with the following constraints:
  
  – The parameter k is a positive integer which should be determined by the applicant.
  
  – Each found group should be defined in terms of a function of the original features.
  
  – The most significant features of each group should be reported.
Ethics, Confidentiality

• Patients have given an informed consent
• Data provided will be anonymized. No indirect info such as birth place + birth date.
• The national agency for protection of privacy has given an approval, under conditions that must be respected (e.g. the dataset can only be available to the researchers working on the project and duly authorized by the PI: this implies that the dataset can only be on a password protected website, that the researchers working on it must register and that they must agree to do not share data with unauthorized persons)