



# Towards harmonization of biobanking initiatives, a comparative analysis of population-based biobanks

Isabel Fortier, Ph.D.

ESF-UB Conference in biomedicine  
Biobanks: Introduction and Next Steps

1-6 November 2008

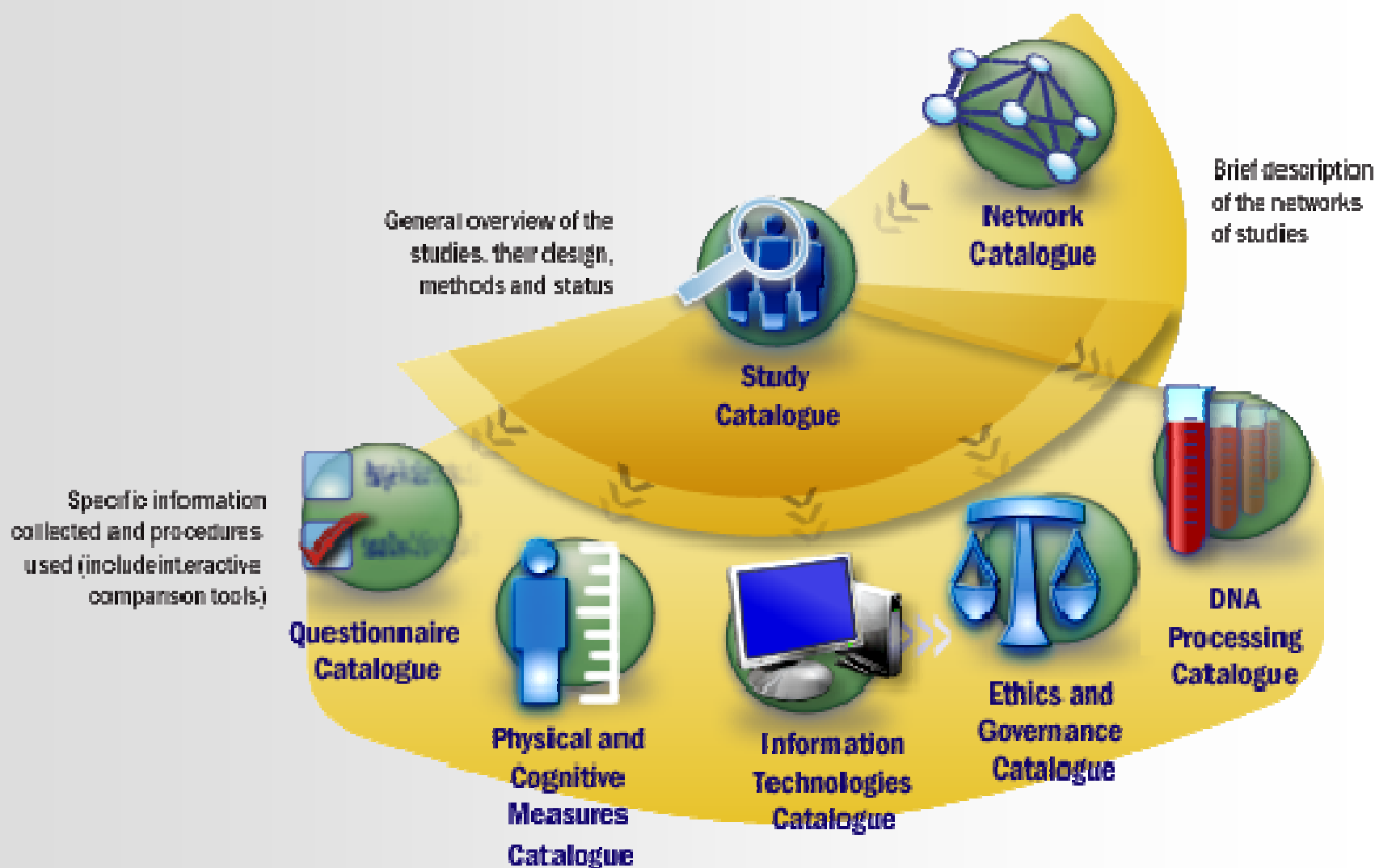




# **1. Overview of large population-based biobanks worldwide**

# P<sup>3</sup>G Observatory Catalogues

## Large population-based biobanks



## Study Catalogue: current contents

**123 large (>10 000 healthy participants)  
population-based studies**  
(P<sup>3</sup>G members and non-members)

- 58 studies with complete information
- 65 studies with summary information

Study design	Number of studies	Number of Participants TARGETED
Cohort	102	9,740,000
Case-control	3	120,000
Clinical trial	6	500,000
Cross-sectional	9	140,000
Others	3	50,000

## Number of participants targeted (recruited or to be recruited) (N=122)

<b>Number of participants</b>	<b>Number of studies</b>	<b>Number of participants TARGETED</b>
<b>Less than 50 000</b>	<b>73</b>	<b>1,300,000</b>
<b>50 000 to 99 999</b>	<b>20</b>	<b>1,400,000</b>
<b>100 000 to 499 999</b>	<b>22</b>	<b>3,400,000</b>
<b>500 000 and more</b>	<b>7</b>	<b>4,400,000</b>
		<b>Total: 10,500,000</b>

## Current status of the studies (N=121)

<b>Current status</b>	<b>Number of studies</b>	<b>Number of participants TARGETED</b>
<b>Study ended</b>	<b>5</b>	<b>310,000</b>
<b>Recruitment ended, follow-up progressing</b>	<b>85</b>	<b>7,140,000</b>
<b>Recruitment of participants progressing</b>	<b>20</b>	<b>1,490,000</b>
<b>Pilot / preparation phase progressing</b>	<b>11</b>	<b>1,400,000</b>

**Slide 6**

---

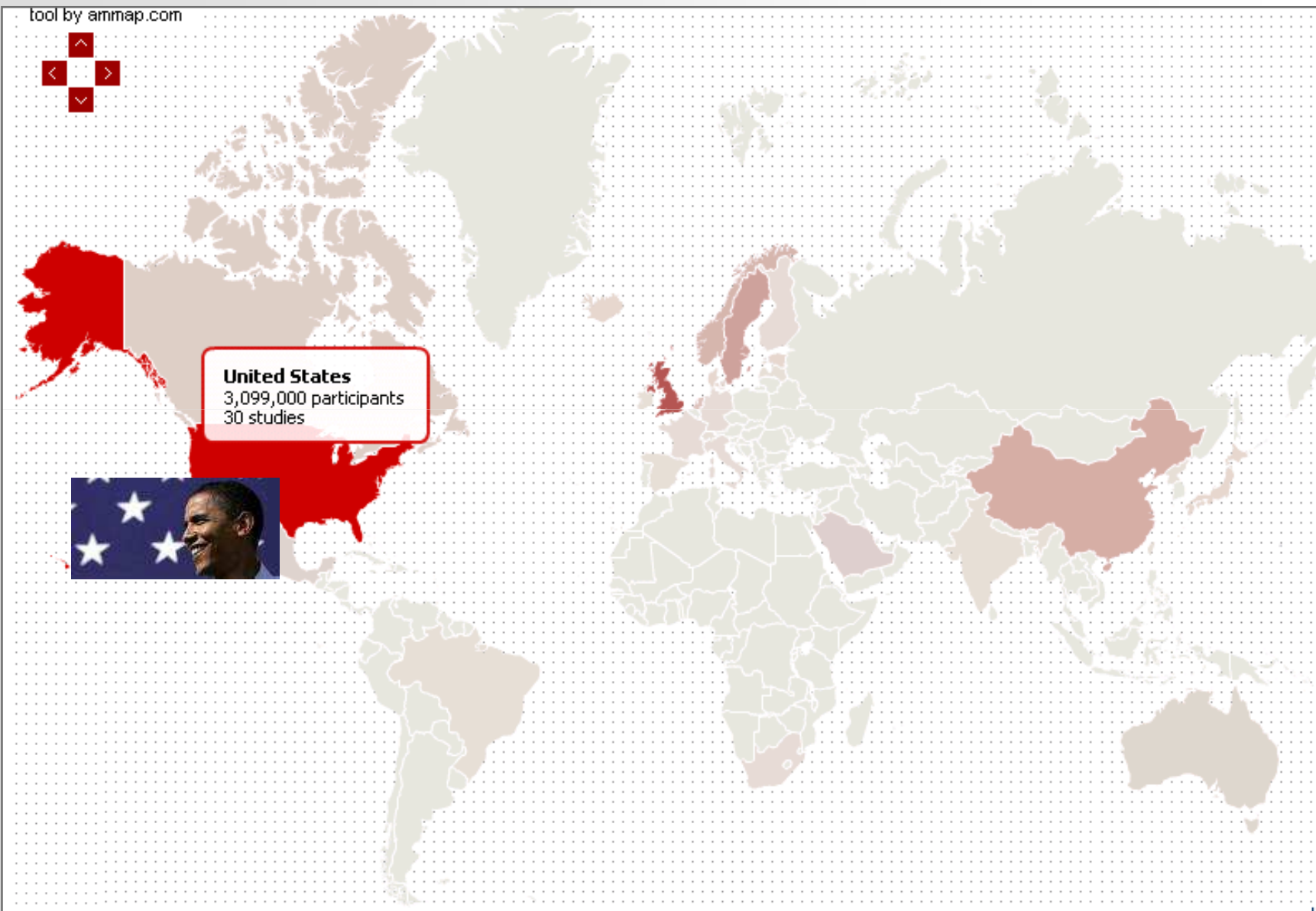
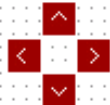
**MSOffice1** ; 05/11/2008

## Selection criteria: Country of residence (N=122)

		Number of studies	Number of participants <b>TARGETED</b>
<b>Single-country</b>	<b>Europe</b>	<b>60</b>	<b>4,600,000</b>
	United Kingdom	11	2,000,000
	Scandinavian countries	31	1,900,000
	Others	18	600,000
	<b>America</b>	<b>38</b>	<b>3,500,000</b>
	United-States	30	3,100,000
	Others	8	400,000
	<b>Australia/New Zealand</b>	<b>5</b>	<b>200,000</b>
	<b>Asia</b>	<b>13</b>	<b>1,400,000</b>
<b>Several countries</b>	<b>Europe, America, Australia</b>	<b>6</b>	<b>800,000</b>



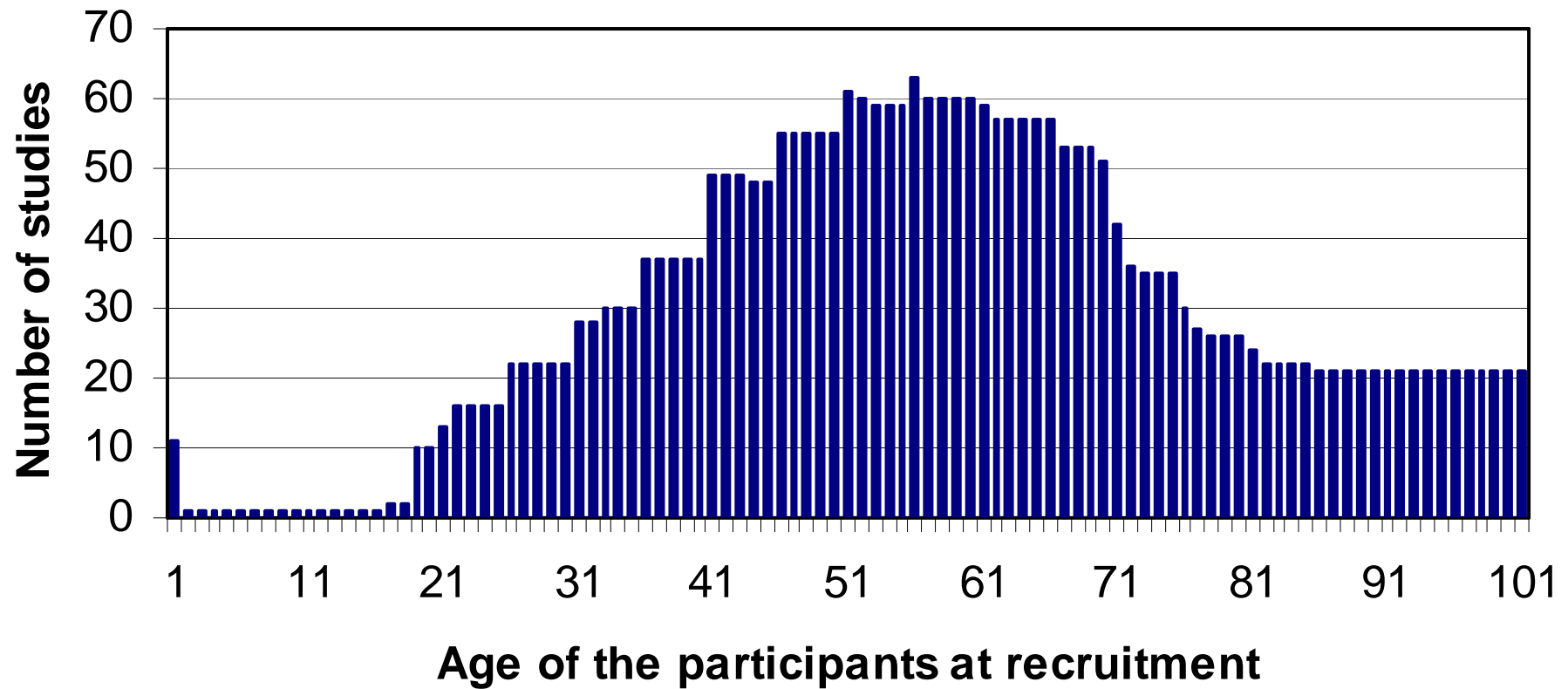
tool by ammap.com



**United States**  
3,099,000 participants  
30 studies



# Selection criteria: Age distribution at recruitment (N=78)



## Disease history at recruitment (ICD10)

Keywords (ICD10)	% of studies
Endocrine, nutritional and metabolic diseases (IV;E00-E90)	89 %
Diseases of the circulatory system (IX;I00-I99)	89 %
Diseases of the respiratory system (X;J00-J99)	89 %
Neoplasms (II;C00-D48)	83 %
Diseases of the musculoskeletal system and connective tissue (XIII;M00-M99)	78 %
Diseases of the digestive system (XI;K00-K93)	61 %
Diseases of the genitourinary system (XIV;N00-N99)	61 %

\*Total number of participants targeted within 18 studies: 3 428 006

# Life habits and environmental exposures at recruitment

Life Habits/behaviours	
Keywords	% of studies
Smoking/tobacco use	94 %
Alcohol use	89 %
Nutrition	89 %
Physical activity	83 %
Sleep patterns	50 %
Physical environment	
Keywords	% of studies
Passive smoking exposure	61 %
Chemical exposures at work	44 %

\*Total number of participants targeted within 18 studies: 3 428 006

## Socio-demographic characteristics at recruitment

Keywords	% of studies
Education level	83 %
Working status	78 %
Birth location	67 %
Marital status	61 %
Income	39 %

\*Total number of participants targeted within 18 studies: 3 428 006

## Physical and cognitive measures at recruitment

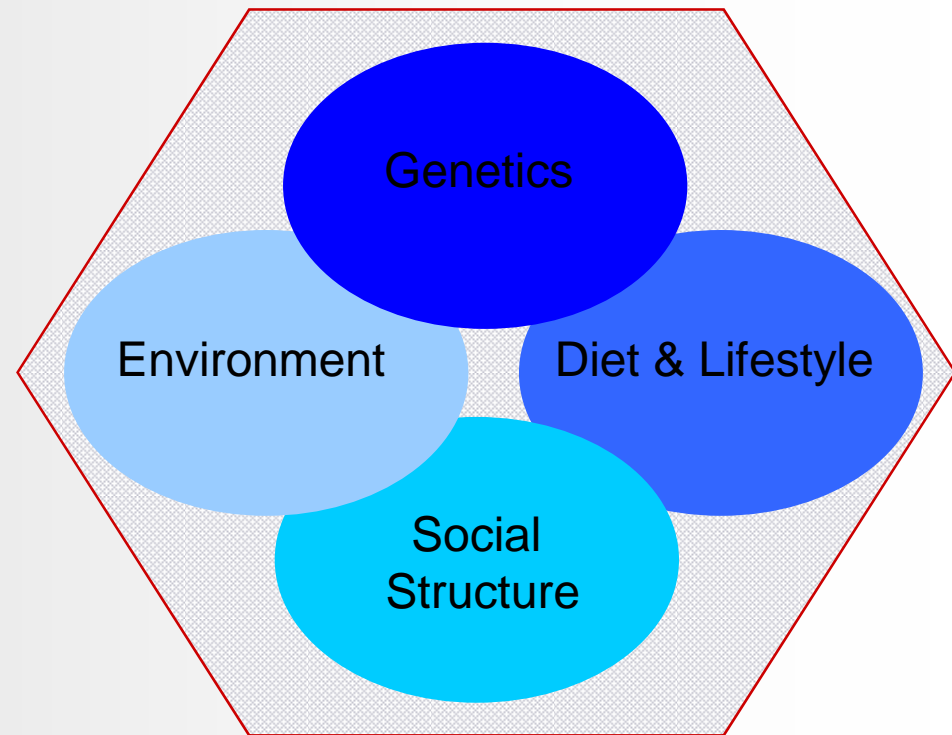
Physical and cognitive measures	% of studies
Weight	100 %
Standing height	100 %
Blood pressure	90 %
Heart rate	70 %
Body circumferences	70 %
Waist circumference	70 %
Hip circumference	50 %
Respiration functions	50 %
Mental functions	50 %
Vision	40 %
Electrical activity	40 %
Bone density	30 %
Bioimpedance	20 %



## **2. Need for harmonization**

# The Causal Complexity of Chronic Diseases

Diabetes  
Asthma  
Heart Disease  
Schizophrenia  
Cancer  
Multiple Sclerosis  
Obesity  
Arthritis



“webs of causation”

- BUT: serious difficulty to identify associations that can consistently be replicated





## Why do we face such difficulty to identify and replicate genetic associations?

It can be explained in many ways including:

1. The fundamental complexity of the expression and aetiology of the disorders of interest
2. The need to tease out small biological effects from within this complexity
3. The heterogeneity of study designs and methods
4. The challenge of designing and conducting optimal studies in genomic epidemiology

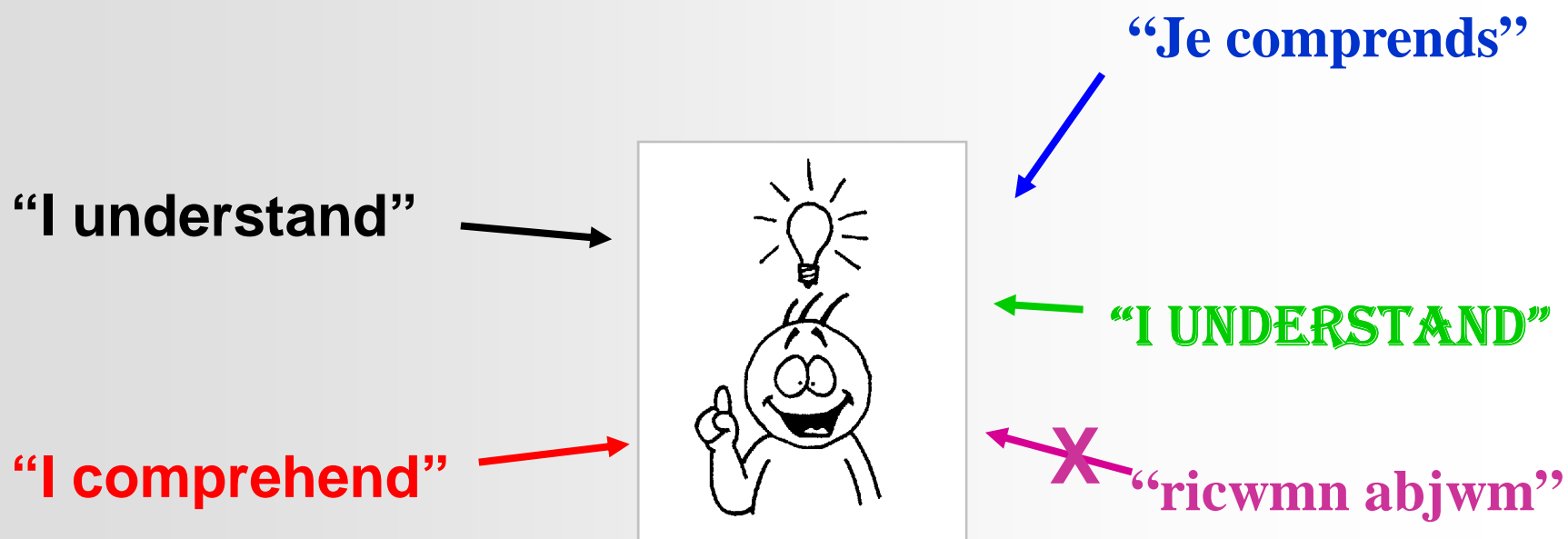
But, there is no doubt that a major contributor to the problem is the lack of statistical power.


## How are we responding?

- **Conduct studies with optimal designs**
  - Increase the quality of individual studies.
  - When relevant, use continuous disease-related traits and health determinants.
- **Increase the size of individual studies**
- **Promote the conduct of meta-analysis**
  - **Sharing results:** Ideally based on published and unpublished results.
  - **Sharing raw data and samples:** Need to promote harmonization between biobanks to enable pooling of raw information.

# Biobank harmonization

- “A set of procedures that promote, both now and in the future, the effective interchange of valid information and samples between a number of studies or biobanks, accepting that there may be important differences between those studies”

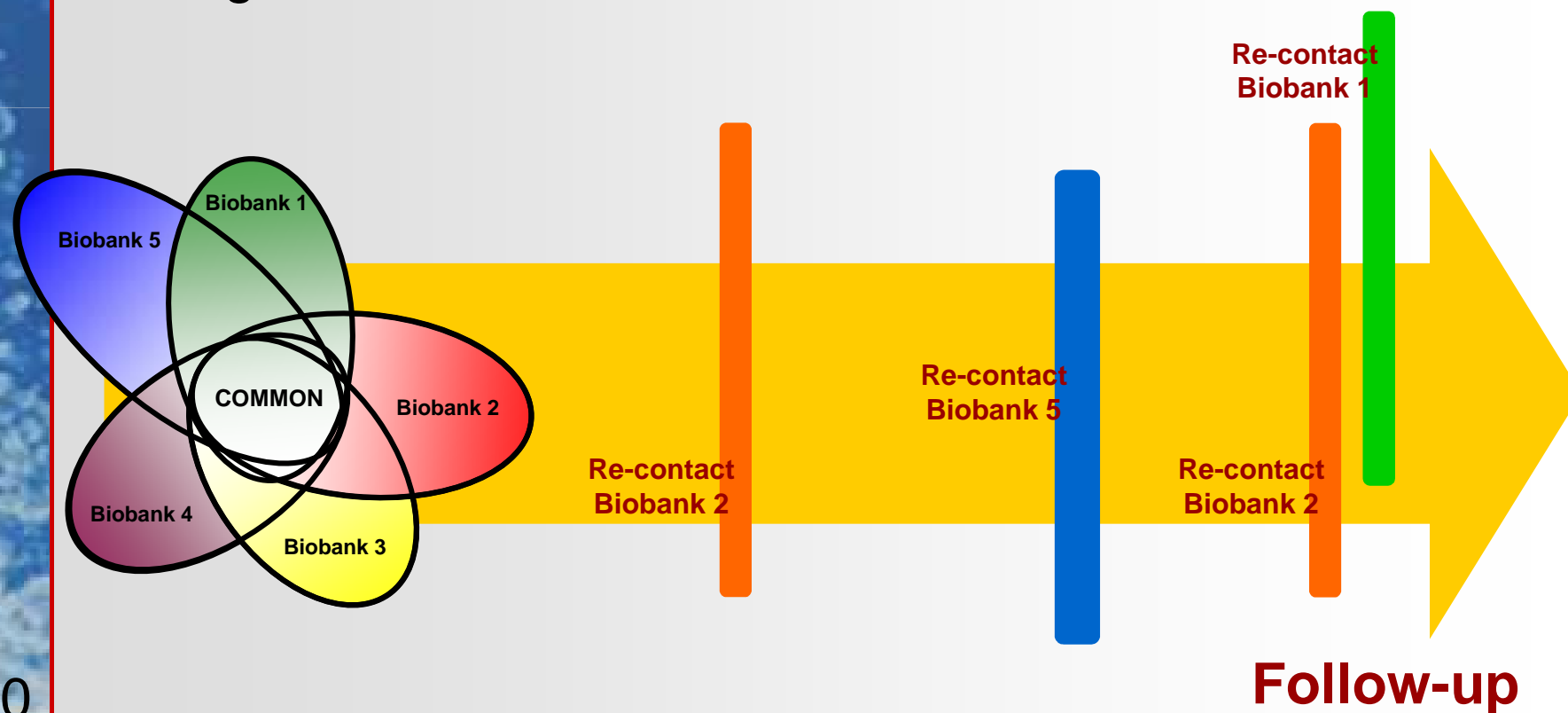


The background of the slide is a complex, dense network of lines. The most prominent feature is a thick, tangled web of yellow lines that crisscross the entire frame. Interspersed among these yellow lines are thinner, more delicate lines in shades of red and grey. The overall effect is one of extreme complexity and interconnectedness, resembling a large-scale network graph or a dense web of data connections. The lines are irregular in thickness and direction, creating a sense of chaotic but structured connectivity.

**Pooling data?  
OK, but what are you  
looking for?**

## Harmonize yes, but what?

- Questionnaire
  - Physical and cognitive measures
  - Biochemical measures
  - Registries



## We must take into account (1)

- **Design of the studies**
  - Important variations between biobanks in their: designs, populations targeted, sampling frames, selection criteria, biases, etc.
    - Leads to major difficulties in identifying a common framework
  
- **Data and sample collection and processing**
  - Increasing complexity of the information collected but lack of common standards and common procedures.
  - The specific challenges of prospective and retrospective harmonization.

## Harmonize the past and the future

- **Retrospective harmonization**
  - Pool information that has already been collected.
  - Needed, but the quantity and quality of information that can be shared is limited by heterogeneity.
  
- **Prospective harmonization**
  - Develop, ahead of time, common methods to collect and store information.
  - Subsequent pooling more efficient, but difficult to define the future needs, obtain agreement on common standards and to implement these standards in current practice.

## We must take into account (2)

- **Ethics and governance**
  - Need to share data/samples between studies/countries under different jurisdictions
  - Agreement to pool or exchange data/samples not necessarily included in the consent
  - Intellectual property and rules of access to information
  
- **Information technology**
  - The need to develop IT systems allowing secure integration of information under **varying formats** and potentially incompatible systems.



## What needs to happen?

**1. FOSTER COLLABORATION**

**2. OPTIMIZE DESIGN**

**3. PROMOTE HARMONIZATION**

**4. FACILITATE KNOWLEDGE TRANSFER**



## Public Population Project in Genomics

### P<sup>3</sup>G Working groups:

- Genomics and Biochemical Investigations
  - Comparative analysis of major guidelines (IARC, OECD, ISBER, etc.)
- Knowledge Curation and Information Technology
  - Open source IT management system for biobanks
- Ethics, Governance and Public Engagement
  - Generic Consent Form
- Epidemiology and Biostatistics
  - Data Schema and Harmonization Platform for Epidemiological Research (DataSHaPER)

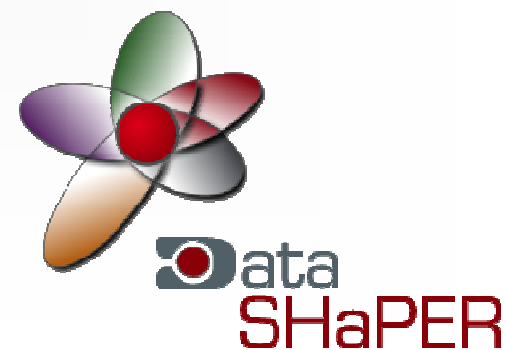




### **3. Example of new tools supporting harmonization**

***Data Schema and Harmonization Platform  
for Epidemiological Research (DataSHaPER)***

***New update soon!***



## Three steps toward harmonization

Identify core sets of information to be shared  
(selection and definition of the variables)

Assess potential to share the core set of information  
between a group of biobanks

Achieve processing and pooling of information  
(Real Data)

## The *Generic* DataSHAPER: Data Schema

- Core set of variables identified by experts from more than 25 biobanks.
- Supports the construction of **cross-sectional baseline questionnaires** for general purpose biobanks enrolling middle-aged participants.
- List simple enough to be used in a variety of contexts
- Set of variables that is comprehensive enough to ensure the realization of valid research
- NOT a prescriptive list of all the variables to be collected by a biobank!
- Complementary to development of specialized datasets for particular interests (e.g. particular diseases, environmental exposures, etc.).

## Examples of domains covered

### Health outcomes

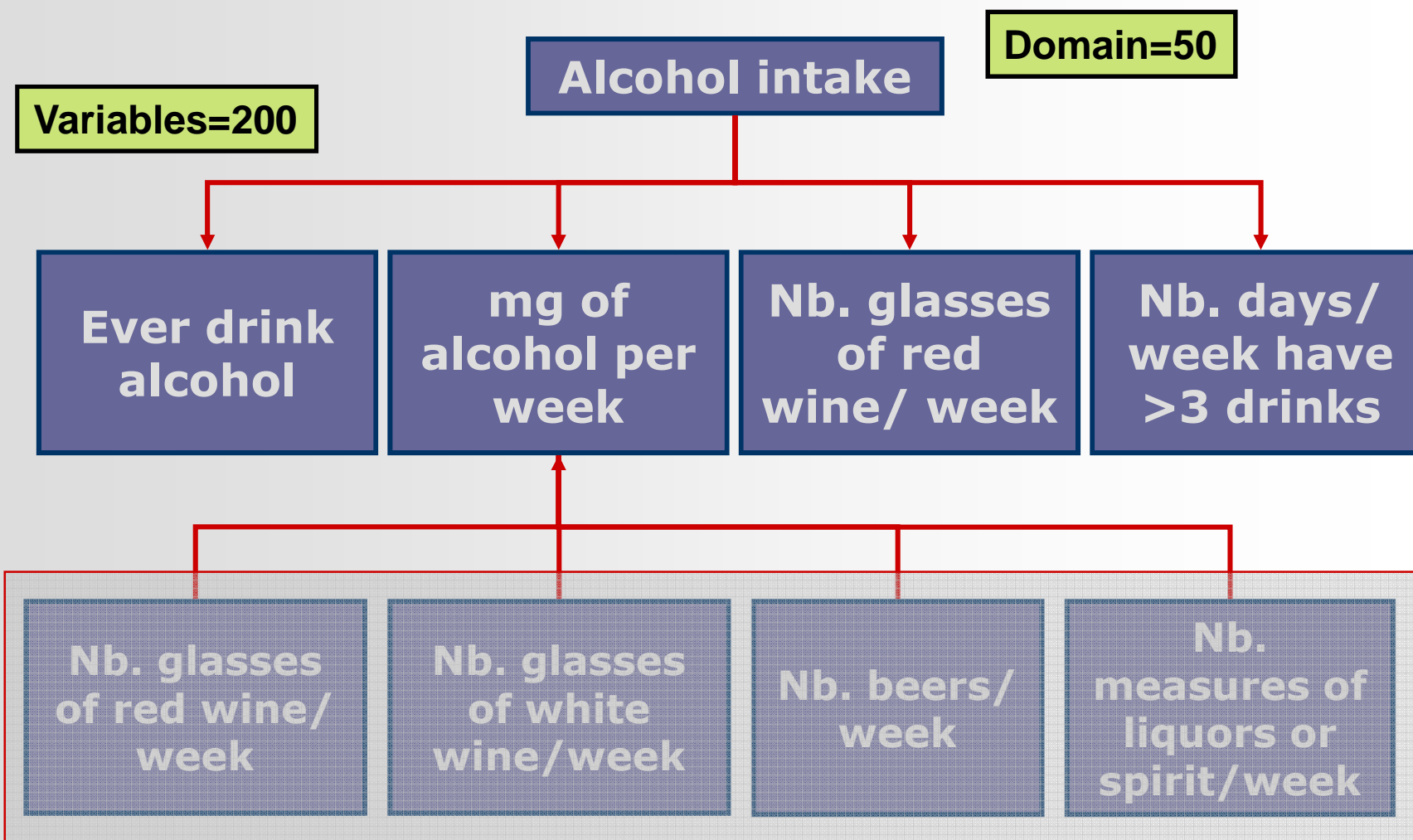
- Cancer; diabetes; stroke; myocardial infarction; familial history of cancer, etc.

### Health determinants

- Smoking, alcohol intake, birth location (subjects, parents and grand-parents), education, income, passive smoking exposure, working status, physical activity

### Physical measures

- Anthropometric measures, resting heart rate, blood pressure



## DataSHaPER: Harmonization platform

- Support the evaluation of the potential to share individual items of information between biobanks.
  - Provide a structure to define the level of matching and the algorithms to be applied to the data of a study to create the variables of the DataSHaPER.

	Study 1	Study 2	Study 3	Study ...
Variable 1	Red	Yellow	Green	Green
Variable 2	Green	Green	Green	Green
Variable 3	Yellow	Red	Red	Red
Variable ...				



# Data SHaPER: Data Processing and Pooling Platform

- Collaboration with different organizations



The screenshot shows the Obiba website, which is an open source software platform for biobanks. The website features a navigation menu with links to 'Obiba', 'Software Directory', and 'P3G Observatory'. The main content area is divided into three columns: 'Obiba' (with links to Overview, Vision, Why open source?, P3G Core, Online Survey, Software Directory, Mailing List, and Partners), 'Overview' (with text about the project's aim and current development areas), and 'Tools' (with links to Software Directory and Online Survey). There are also sections for 'Projects' (with links to GenoByte, Sample Management, and Federation of Biobanks), 'Events & News' (with sections for New Release, Upcoming Event, and Past Event), and a footer with logos for P3G and Genome Québec.



**THANKS!**

P<sup>3</sup>G Observatory  
[www.p3gobservatory.org](http://www.p3gobservatory.org)



Genome Québec



Genome Canada

PROMOTING  
HARMONISATION OF  
**PHOEBE**  
EPIDEMIOLOGICAL  
BIOBANKS IN EUROPE