Executive Summary

Bartha Maria Knoppers
Project Director
Centre de recherche en Droit Public
Université de Montréal
3101, Chemin de la tour
Montréal, (QC) H3T 1J7
knoppers@droit.umontreal.ca
Phone: +1 (514) 343-6714
Fax: +1 (514) 343-2122

1. Introduction

The *Public Population Project in Genomics* (P3G) is an international project aimed at creating a publicly accessible database that fosters collaboration between researchers in the field of population genomics. This international project involves a consortium of four leading, complementary research projects. While the specific aims of each project may vary, their common goal is to understand the role of genetics in human health and disease. Access to the data would be contingent upon the legitimacy of the intended use by the researcher. The consortium would provide the four projects, as well as other member countries at a later time, with resources, tools and know-how to perfect data management for improved methods of transfer and sharing.

2. Synopsis of Project Partners

GenomEUtwin. The aims of the project are to study 800,000 twin pairs from a collaborative European pool of registries through a combination of genetic, epidemiology and phenotype data for common multifactorial diseases. The group aims to harmonize an ethical framework and protocol for sampling and constructing a database network for Europe.

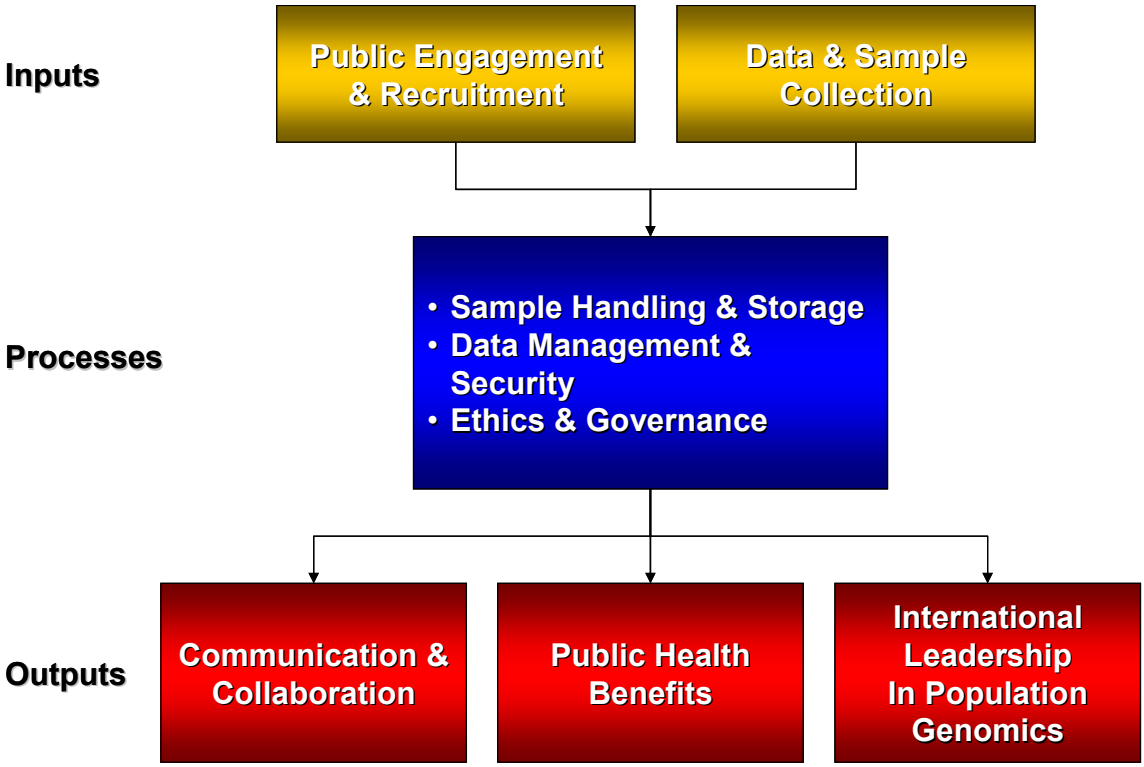
Estonian Genome Project. The goal of the project is to collect data into a database from up to one million participants, including health status, genomic DNA, plasma and genealogical data. The database will be potentially utilized for large case-control association studies and other genetic studies involving the discovery of interactions between genes, diseases and environments.

UK Biobank. The aim of the project is to obtain samples from 500,000 volunteers. The dataset will be utilized to appraise the effects of genetic and environmental factors on the risk of common multifactorial diseases.

CARTaGENE. The objective of the project is to obtain personal, medical and sociological data and biological samples from 60,000 random participants for the study of normal genetic variation. The group also aims to provide leadership on the socio-ethical and legal issues in population research, banking and databases.

See Appendix for complete project descriptions.

3. P3G Overview



3.1 Inputs

3.1.1 Public Engagement and Recruitment

P3G objectives	Proposed actions
• PR and communications strategy in order to engage the public	– Build public awareness and credibility of consortium – Monitor and report results of communication and public consultation practices utilized by each project, such as website, information available to the public, media, focus groups etc. – Continue to develop communication strategy as a consortium and harmonize the P3G message
• Harmonization of recruitment standards for each project	– Define common ethical standards of recruitment (no coercion, overcompensation, misinformation etc.) – Define common ethical standards of consent (no invasion of privacy, not too restrictive, etc.) – Ensure that differences in recruitment methods (random vs. physician vs. twins) do not hinder the sharing of data and results

3.1.2 Data and sample collection

P3G objectives	Proposed actions
• Setting standards for the methods of surveying data	– Ensure any differences in methods do not impede the sharing of data and results – Decide on importance of follow up vs. anonymity and methods of administration (self vs. physician)
• Ensuring common domains in questionnaires	– Standardize scope of questioning – Harmonize domains of crucial information with regard to medical, demographic, environmental and social data – Establish core list of phenotypes, while leaving flexibility for surveying additional traits – Decide on a list of core biomarkers – Develop and utilize common nomenclature
• Validation of questionnaires and data collection	– Establish methods of quality assurance and control in questionnaires, by setting standards and allowing for validation by some means within the consortium
• Standardization of samples and tests taken	– Define core list of samples taken (e.g. blood, urine, others) – Standardize core biochemical measurements to be taken and equipment needed – Set standards of collection methods and measurements and set up quality control mechanisms

3.2 Processes

3.2.1 Sample handling and storage

P3G objectives	Proposed actions
• Harmonization of sample handling and storage	– Set strict guidelines for handling and storage of samples – For DNA samples, set standards of extraction methods that are acceptable to all projects to ensure highest quality samples – Set up standards of DNA sample storage to ensure integrity and security – Establish standards of blood/tissue sample preparation and storage (may be important for future proteome and transcriptome testing)
• Developing standard operating procedures for sample handling and storage	– Set standards for measuring quality and quantity of DNA samples – Create standard procedures for biological measurements (e.g. what can be done on fresh or stored samples)

	– Set standards of equipment and researchers for quality assurance of samples

3.2.2 Data management and security

P3G objectives	**Proposed actions**
• Establishing standards of data collection and analysis	– Establish common standards of DNA data collection e.g. genotyping, epigenetic markers etc. – Set standards of biological measurements e.g. phenotypes, biomarkers etc. – Establish and utilize common nomenclature for data, when possible – Set standards of data analysis, bioinformatics and statistics utilized
• Organization of IT and logistics needed for data management	– Determine how data will be connected (sharing between databases vs. one common database) and technologies needed for data input and transfer – Decide on how data will be presented and organized – Ensure capacity of database(s) is sufficient for amount of data that will be collected and shared – Software utilized by each project should allow for input into shared database or for transfer between databases (depending on organizational infrastructure)
• Management of data access and membership	– Define membership, ensuring the maintenance of scientific and ethical integrity – Establish data release policy – Guard against conflicts of interest – Ensure data remains relevant and does not become too diffuse – Coordinate the sharing of data, expertise and tools between users
• Ensuring data security	– Ensure database is compatible for data sharing while maintaining standards of data integrity, confidentiality and communication – Develop common coding for names and samples – Decide on use of security measures such as software barriers

3.2.3 *Ethics and Governance*

P3G objectives	Proposed actions
• Ensuring transparency and openness of data and results	– Define how access will be granted – Ensure no exclusive access, secrecy of shared results or unnecessary destruction of data – Adhere to set standards of consent, privacy, integrity and data protection
• Managing ownership and copyright issues	– Coordination of data access for clarity on ownership and copy-write issues – Decide on sharing of benefits including intellectual property and public benefits – Protect academic freedom – Ensure that data accessed, used, and enriched will return to central database after IP protection – Create a body to oversee ethics and governance issues

3.3 Outputs

3.3.1 *Communication and Collaboration*

P3G objectives	Anticipated achievements
• Sharing and transfer of data and results between users	– Data will be accessible to public, while ensuring ethics, data integrity and security – Ability to coordinate and connect results from all projects – Will also share experts and researchers

3.3.2 *Public Health benefits*

P3G objectives	Anticipated achievements
• Advancement of the understanding of health and disease	– Increase statistical significance through cross-comparison – Knowledge of the role of genetics on health and disease will be advanced – Large samples sizes and comparative evaluation will increase validity of results – Validated results will be available to public – Increase public awareness of developments and importance of population genetics – Results can be used for the advancement of science world-wide

3.3.3 *International leadership in population genomics*

P3G objectives

- Becoming international leaders in population genomics

Anticipated achievements

- Support transfer of data, results and technology to the international community
- Stimulate new research strategies and innovations world-wide
- Leadership role in the development of ethical and legal standards in population genetics
- Establish an infrastructure of open access to data that will be valuable for organizations world-wide
- Liaison with international and regional organizations on global health issues and initiatives

4. Infrastructure

4.1 Input

4.1.1 *Public Engagement and Recruitment*

Existing infrastructure

- All four projects have websites
- P3G website has been launched

Potential infrastructure objectives

- Possible centralized or linked PR/information group

4.1.2 *Data and Sample Collection*

Existing infrastructure

- UK Biobank has six collaborating centres within universities that recruit participants and collect samples
- Estonian Genome Project has a laboratory in Tartu where samples from 12 Estonian counties are stored and processed
- GenomEUtwin recruitment and collection done in 8 participating countries (Finland, Sweden, Norway, Denmark, the Netherlands, Italy, United Kingdom, Australia)
- CARTaGENE has three sampling centres in Montreal, Quebec and Saguenay

Potential infrastructure objectives

- Storage space with capacity to meet set standards of sample collection

4.2 Processes

4.2.1 Sample handling and storage

Existing infrastructure

- As outlined, UK Biobank's collaborating centres within universities collect samples and Estonian Genome Project have a laboratory for the collection and processing of samples
- GenomEUtwin has core centres for DNA extraction & genotyping in Uppsala and Helsinki (with harmonized protocols)
- CARTaGENE proposes to utilize GenVault Robotic platform for storage, processing and management of samples

Potential infrastructure objectives

- Storage capacity that meets standards set by consortium
- Infrastructure and equipment for established tests and measurements

4.2.2 Data Management and Security

Existing infrastructure

- GenomEUtwin core facilities for database (Stockholm), Statistics (Leiden), Epidemiology (Copenhagen), Ethical issues/consent (Oslo)
- UK Host coordinating centre at the Manchester Science Park
- Estonian genome project coding centre, for security of samples and data
- CARTaGENE utilizes Entrust platform, developed by First Genetics Trust System, a corporation known for solid and secure computer systems

Potential infrastructure objectives

- International core for common database or for linking data
- Centralized or linked IT core to manage database and data transfer
- Common database or compatible software for access between datasets (proposed use of XML system for communication between projects)

4.2.3 Ethics and Governance

Existing infrastructure

- Frameworks for ethics and governance exist for all four projects
- Various advisory boards for each project

Potential infrastructure objectives

- Centralized or linked core for overall management and governance (in Montreal?)

4.3 Outputs: Communication and Collaboration, Public Health Benefits, International Leadership in Population Genomics

- Common or linked infrastructure for the access of data and results internationally

5. Appendix - Project Overviews

eesti geenivaramu

Estonian Genome Project
A 'Medical Care' Model
www.geenivaramu.ee

The idea of a national gene bank-Estonian Genome Project (EGP) - was proposed by Andres Metspalu, professor of biotechnology at the University of Tartu, who observed that Estonians were representative of the European gene pool. Preparations for the Estonian Genome Project were launched in 1999. In December 2000, the Estonian Parliament (*Riigigoku*) passed the Human Genes Research Act, providing for the creation of a central gene bank of the Estonian population. This Act ensures the highest ethical and security principles, regulating the establishment and use of the database. In conformity with this Act, the Estonian Genome Project Foundation (EGPF) was created in March 2001, with the aim of coordinating the creation of a central Gene Bank database of health and genetic data of the Estonian population. The project is based on a public-private partnership between the EGFP, owner of the database and privacy shelter, and EGeen, a public limited company, the exclusive commercial licensee of the database, responsible for the financing of the project.

The underlying idea of the EGP is the establishment of Gene Bank- a database including phenotype and genotype data of the Estonian population- to carry out scientific research, genetic and health studies, in order to find the genes that cause and influence common diseases; and to increase public awareness of developments in the fields of gene technology and biotechnology in general. Another aim of the project is to create a collection of health care status descriptions of a large part of the Estonian population, collect tissue samples of donors, create LD maps of donors, and develop software that enable the data and products of the Genome Project to be marketed.

The long-term goal of the Project is the practical implementation of systematic advances in genomics to public health at a massive scale. It aims to collect the data of up to 1, 000, 000 people.

The pilot phase (from October 2002 to March 2003), funded by the EGeen Inc. and Republic of Estonia, was declared completed in three countries and the main project covering the whole of Estonia started in 2003.

GenomEUtwin Overview
Genetic and life style risks of common diseases
www.genomeutwin.org

The project is underway and its implementation is coordinated by the National Public Health Institute and University of Helsinki under the guidance of Professor Leena Peltonen. It is carried out in several European countries (Finland, Sweden, Norway, Denmark, the Netherlands, Italy, United Kingdom) and Australia. The population cohorts used in the GemonEUtwin study consist of several twin cohorts and the MORGAM population cohort. It has received funding from the European Union for a period of four years.

The GenomEUtwin project has established five core facilities operational since October 2002: Epidemiological (Copenhagen), Database (Stockholm), Ethical Issues/Consent (Oslo), DNA extraction and Genotyping (Uppsala and Helsinki) and Statistical (Leiden).

Studies of large population cohorts are needed to transform genetic information to a detailed understanding of the predisposing factors in diseases affecting most human populations. European twin cohorts provide a unique competitive advantage for investigations of the role of genetics and environment or life style in the etiology of common diseases. This project applies and develops new molecular and statistical strategies to analyze unique European twin and other population cohorts so as to define and characterize the genetic, environmental and life style components in the background of health problems. The target traits are: stature, obesity, migraine, coronary heart disease, stroke and longevity. The project has a foundation on the European strength in genetics, epidemiology and biocomputing and the efficient collaboration of twin researchers, genetic epidemiologists, molecular geneticists and mathematicians.

The GenomEUtwin general research objectives are:

- to develop an intellectual European framework to stimulate inventions and novel strategies in the genetic epidemiological research of common diseases and traits
- to utilize maximally the unique features of population cohorts (longitudinal data and ample information about life-style and environmental factors)
- to guarantee the access of European investigators to the epidemiological, phenotypic and genotypic information on European population cohorts
- to create unique infrastructure for research into common diseases and the training of scientists in quantitative biology

The project contains total of over 800 000 twin pairs on which data has been collected for decades. In the first stage of the study, 10 000 twin pairs are genotyped with genome-wide markers and targeted gene-specific SNP markers.

The UK Biobank

UK Biobank Overview
Genetic Epidemiology for Public Health Purposes
www.ukbiobank.ac.uk

In June 1999, the Wellcome Trust and the MRC committed funds to the UK Biobank (then known as the UK Population Biomedical Collection). An expert working group- chaired by Professor Tom Meade, Director of the MRC Epidemiology and Medical Care Unit, Wolfson Institute, London –was established to discuss the project's feasibility and how the Collection might be organized. Several issues were debated, including the types of information that the project should gather, the diseases that might be investigated, the role of GPs in the study, whether the volunteers should be adults or children, and the size of study needed to obtain statistically significant results.

The UK Biobank resource will be centrally managed from a Coordinating Centre, in Manchester (Hub). It has the overall responsibility for delivering the project including data management and quality assurance, computing and financial management and formal custodianship of the data and biological samples. It coordinates the activities of several Regional Centres (Spokes) who are responsible for participant recruitment and initial data and sample collection. The Hub is set up as a charitable company limited by guarantee and jointly owned by the funding partners.

There is a separate oversight body to oversee this work, independent of both the users of the information and the scientists involved in developing it. This Interim Advisory Group is responsible for ensuring that the samples and the data collected are used responsibly and within the terms of the consent obtained from the participants.

The UK Biobank is committed to developing a public Ethics and Governance Framework (EGF) to set standards for the project and make sure that all necessary safeguards are in place to ensure that the data and samples are only used for scientifically and ethically approved research. A public draft of the EGF has been prepared by the UK Biobank funders - the Department of Health, the Medical Research Council, and The Wellcome Trust - with the advice of an Interim Advisory Group on Ethics and Governance. The framework addresses key issues including enrollment, consent, confidentiality as well as sample and data ownership, access and management. Compliance and governance of these key issues is also addressed in the framework.

The main aim of the study is the investigation of the separate and combined effects of genetic and environmental factors (including lifestyle, physiological and environmental exposures) on the risk of common multifactorial diseases of adult life. The proposed study will address a large number of existing hypotheses regarding the combined effects of genotype and exposure on a range of important health incomes. In practice, the hypotheses to be addressed, will be defined by researchers who are successful in applying to use data from the study, once recruitment and follow-up are sufficiently complete.

The project will involve half a million people of 45-69 years of age. Currently, data collection is being prepared. Collection will begin in 2004.

CARTaGENE Overview
Genetic Variation in Normal Modern Populations
www.cartagene.qc.ca

The project is directed by Dr. Claude Laberge, president and Director of the Network of Applied Genetic Medicine of Quebec (RMGA). A multidisciplinary team of the RMGA has been working since 1999 on the structure of a new and original project in the area of population genomics, CARTaGENE. Several workshops have been organized in order to obtain constructive feedback from ethicists, policy-makers and the public on the initial project. In light of the comments and concerns voiced during those workshops, the team has recently re-evaluated a certain number of elements. The sample of 60 000 participants will be double-coded. Health events will be updated through ICD-9 codes (but not from the participant's medical records) and the possibility of re-contact by questionnaires will be offered in the consent form. There is ongoing consultation of the public.

The CARTaGENE project will consist of: a DNA extraction, storage and management platform; a clinical platform for large-scale human genomics studies; a data centre; and an informatics and data management platform. The Institute for Populations and Genetics, a not-for-profit, independent organization, will be created to control and manage, in the public interest, major population genetics/genomics/proteomics projects, including CARTaGENE project.

The CARTaGENE project will map genetic variation in a large reference population of Quebec. Although sampling will be done only in Quebec, the CARTaGENE team includes researchers from Quebec and the rest of Canada. The CARTaGENE resource will allow large-scale medical, pharmacogenomic and public health studies, including association studies of common diseases or "protective" phenotypes and lead to the discovery of new susceptibility genes. The demographic component of the project will determine mutation frequencies in different regions of the Province, and thus guide the establishment of health programs of medical genetic services tailored to the needs of the regional sub-populations. Globally, the aim is to provide information for the best use of genetic knowledge and technology in the public health system.

The long-term goal is the constitution of an "international" project for integrated population genomics, as well as the providing of well-matched case-control sub-cohorts for finding genes predisposing to common, complex diseases or phenotypes of interest, particularly for high frequency alleles and so expand the development of predictive and diagnostic technologies.

CARTaGENE is a modern population genomics project that integrates and exploits the advantages of the existing Canadian health care and legal system, i.e. universal health care, public health care, public health expertise, unique medical identifier numbers, comprehensive health and environmental record linkage with a genealogical/demographic database, extensive privacy and access legislation and a heterogeneous modern population.
Once funding is secured, samples will be randomly collected from 60 000 individuals aged 25-74 within a period of four years.