i-BALSAC: A multisectoral infrastructure for the study of the Franco-Canadian population

Hélène Vézina[1], Jean-Sébastien Bournival[1], Simon Girard[1], Marc St-Hilaire[2]

[1] Projet BALSAC, Université du Québec à Chicoutimi

[2] Centre interuniversitaire d'études québécoises, Université Laval

Paper presented in the session *Large linkage projects: New opportunities* at the 43rd Annual meeting of the Social Science History Association, November 9, 2018.

Abstract

For almost 50 years, BALSAC has been reconstructing the genealogical lines of the Quebec population of French Canadian descent. Five years ago thanks to a major grant from the Canadian Foundation for Innovation a partnership was established with the PRDH at the Université de Montréal and the Centre interuniversitaire d'études québécoises. Together these groups have created the IMPQ (Integrated infrastructure of the Quebec population historical microdata). The IMPQ rests on the fusion of the PRDH and BALSAC database and on the linkage of Canadian census data (1851-1911) to the fused dataset.

In this presentation we describe i-BALSAC a new initiative aimed at developing an infrastructure to study the French Canadian population through a joint genealogical, genomic and geographical approach. Concerning the genealogical data, we intend to rely on a methodology based on handwriting recognition to add birth and death data to the existing IMPQ data to expand the time period where complete family reconstitution is available. This data will be linked to genetic information on participants from various biomedical projects conducted in Quebec. Lastly, we will develop a historical GIS in order to provide a tool for the analysis and interpretation of the spatial dimension of genealogical and genetic data.

Introduction

The BALSAC Population Database has been in development since 1971 at the Université du Québec à Chicoutimi (UQAC). It now includes over 3 million Quebec civil records, including all Catholic marriages (almost 85% of all marriages) from 1621 to 1965[1]. For over 40 years, BALSAC has inspired and empowered innovative multidisciplinary studies in fields like genetic epidemiology, historical geography, population genetics, evolutionary biology, demography and social history. In collaboration with the researchers who use the data, the BALSAC team closely follows theoretical, methodological and technological developments in these various fields to ensure the database remains a relevant infrastructure that responds to the needs and demands of the scientific community and contributes to the progress of knowledge.

In this optic and thanks to funding from the Canada Foundation for Innovation (CFI), BALSAC was involved in the construction of the Integrated Infrastructure of Historical Microdata of the Quebec Population (IMPQ)[2] from 2013 to 2017 in partnership with the Centre universitaire d'études québécoises (CIEQ) and the Research Program in Historical Demography (PRDH). The PRDH initiated in 1966 the Registre de la population du Québec ancien (RPQA) which contains the entire Catholic population of Quebec from the beginning of French settlement in Canada in 1608 (first record in 1621) to 1799, comprising 700,000 birth, marriage and death records. Thanks to the CIEQ and the PRDH a good share of Quebec data from the Canadian historical censuses prior to 1921 has also been digitized and formatted in databases. The construction of the IMPQ included three components: a full fusion of the BALSAC and RPQA databases, the harmonization of existing census data series and expansion of the geographical coverage, the linkage between civil records found in BALSAC and census data (households and individuals) and linkage across censuses (1851-1911). The IMPQ portal was officially launched in October 2018 (Figure 1).

Today BALSAC intends to pursue its development in response to three types of requests expressed in recent years. First, researchers in genomics who work with genealogical information want a formal and sustainable structure that integrates both genealogical and genetic data; currently, those data are only linked for specific projects for a limited time. Second, extending our exhaustive reconstruction of the Quebec population to the 20th century by integrating the totality of available civil records (birth, marriage and death) is of great interest to researchers in both social and biological sciences, notably for research in evolutionary biology. Finally, given the importance of the geographical dimension in social and biomedical research, and the enormous potential offered by geographic information systems, it appears highly relevant to create a framework to analyze and interpret the spatial components of genealogical and genetic data.

[1] For additional information on the BALSAC database, refer to http://balsac.uqac.ca/english/
[2] For more info on IMPQ : https://impq.uqtr.ca:8082/fmi/webd/IMPQ_PORTAIL (in French only)

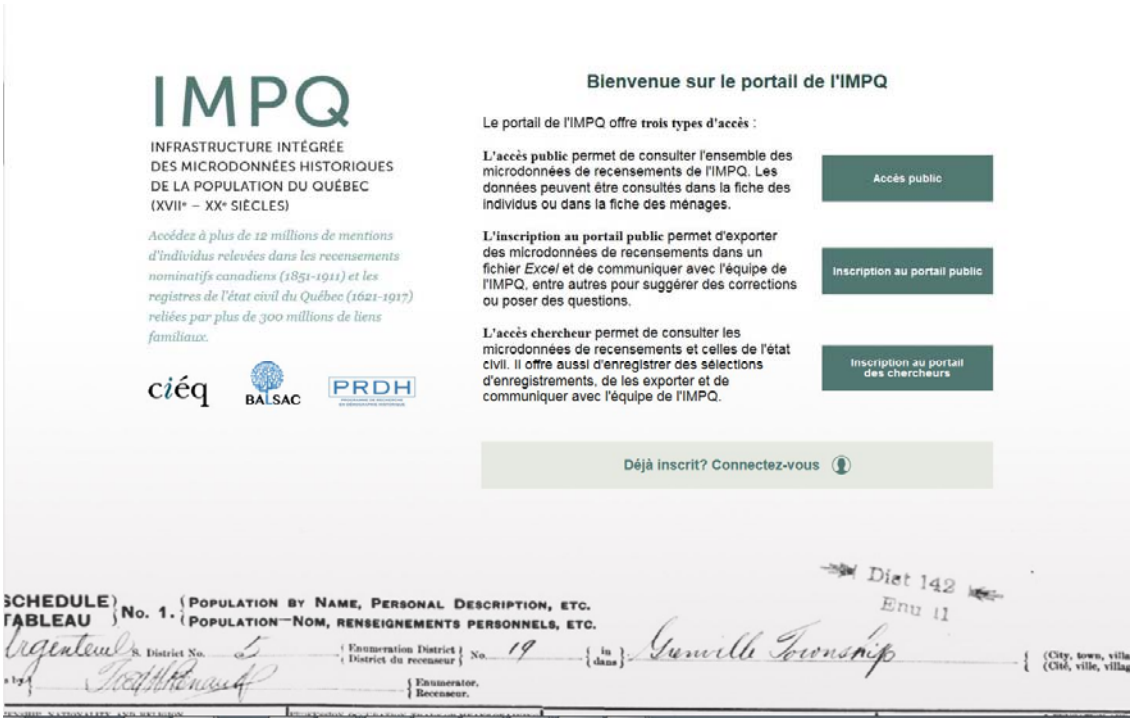

Figure 1: The IMPQ portal

In order to meet these needs, we aim to create the BALSAC infrastructure (i-BALSAC), a dynamic and versatile multisectoral platform for cutting edge research in biological and social sciences. The BALSAC database is the core of this infrastructure which will bring together genealogical, genetic and geographic data. The project also includes the development of analytical and statistical tools (enabling joint exploration of genealogical and genetic data) and mapping (historical geographic information system (HGIS)) to optimize exploitation of these data sets. The development of i-BALSAC rests on the continuation of partnerships with the CIEQ, the PRDH and the CARTaGENE research platform, and will pave the way to new collaborations, notably with the Bibliothèque et Archives nationales du Québec (BAnQ) for the demographic component and Génome Québec for the genetic component.

In this presentation, we will define the five major objectives that will guide the development of i-BALSAC, from the integration of new data to the launch of the Web portal. We will also discuss research opportunities as well as expected impact and foreseen challenges of the project.

Development of the infrastructure

Figure 2 presents the conceptual diagram for the construction of i-BALSAC. Each of the five objectives are described in the sections that follow.

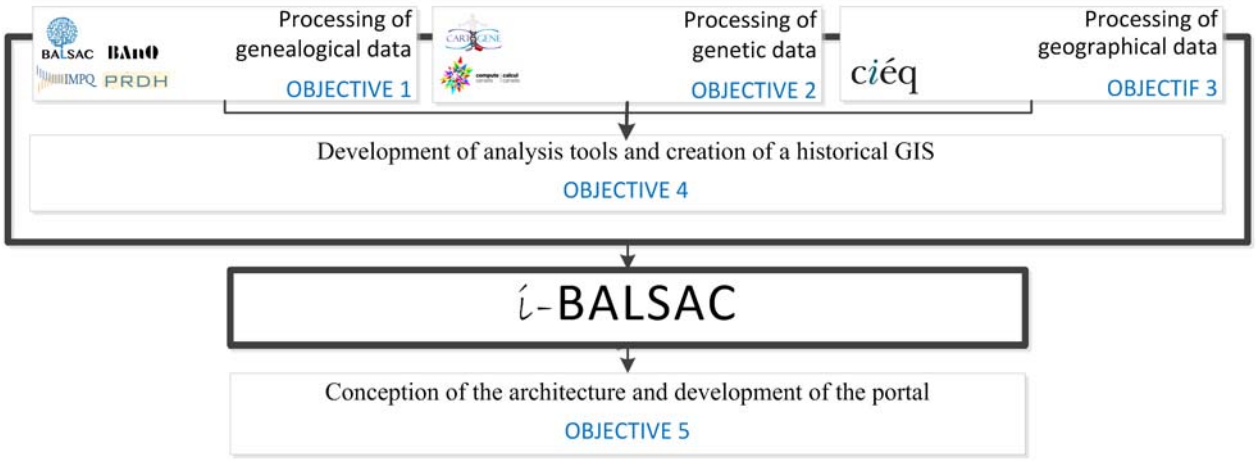

Figure 2: Conceptual diagram of the creation of i-BALSAC

Objective 1: Genealogical Data Processing

The BALSAC database contains marriage records from the early 17th century until 1965 for all Quebec regions[3]. The only exception is for Saguenay-Lac-Saint-Jean, where the three types of data (birth, marriage and death) are available until 1971. Through its participation in the creation of the IMPQ, BALSAC also has access to Quebec birth and death records up to 1849 (for 1621-1799, the records in the IMPQ come from the RPQA). All these data sets (first four lines in Table 1) are linked together, so genealogical lines can be reconstructed over almost 350 years, and family histories from the beginning of European settlement to the mid-19th century.

Table 1: Genealogical data to be included in i-BALSAC. In black, already existing data and in red, new data to be transcribed.

Type of records	Geographical coverage	Period	Source	Number
BMD	Québec	1621 - 1799	RPQA (IMPQ)	640,000
M	Québec	1800 - 1965	BALSAC	2,000,000
BD	Saguenay	1838 - 1971	BALSAC	550,000
BD	Québec	1800 - 1849	IMPQ	1,600,000
BD	Québec	1850 - 1920	BALSAC (BAnQ)	6,000,000

B: Births; M: Marriages; D: Deaths

[3] Records prior to 1800 (N = 69 000) come for the RPQA and were obtained through a collaborative agreement with the PRDH.

The first objective will be to integrate into BALSAC the birth and death records for the Quebec population from 1850 to 1920 (last line in Table 1). It is estimated that six million records will be added to BALSAC, doubling in size the current database. A partnership has been established with BAnQ who will provide nearly 2 million images covering this period, most of them in a high-resolution digital format. The important volume of data to be extracted from the records has led us to make the decision to use optical handwriting recognition technology to transcribe a maximum of records at minimum cost. To transcribe birth and death records we will work with an external organization specialized in the development of optical recognition tools.

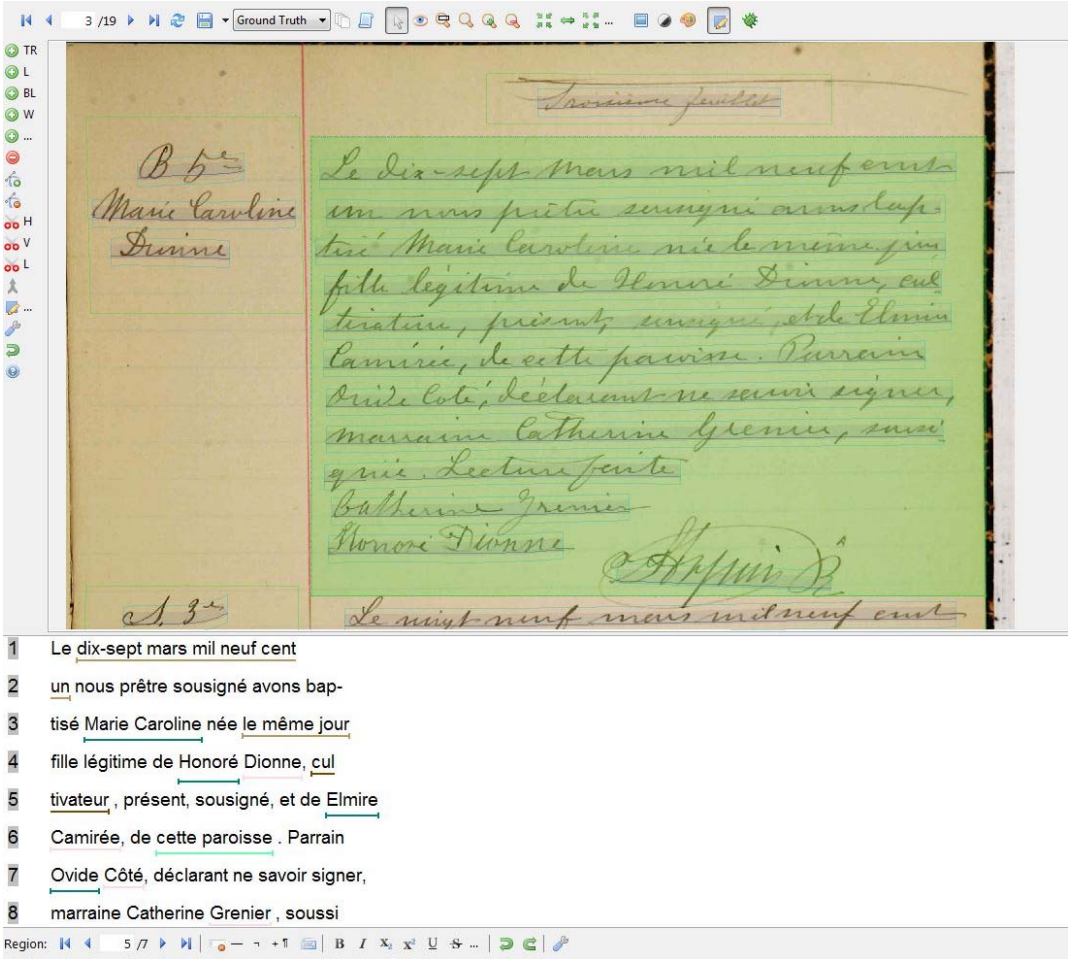

Figure 3: An example of the transcription of an act with Transkribus

A handwriting recognition software will be adapted to read civil records that, despite great variations in calligraphies, have a structure which remains fairly constant over time. For the machine learning phase, we will rely first on complete transcriptions of several acts that will serve as ground truth for the language model to be developed. As shown in Figure 3, for the transcription phase we are currently working with Transkribus[4], a free and powerful handwritten

[4] See https://transkribus.eu/Transkribus/

text recognition software that allows a thorough layout analysis of historical documents. In addition, the marriage records already in BALSAC as well as various dictionaries (names, occupations, residences) compiled at BALSAC will enrich the model, allowing the recognition process to identify and extract the main entities (names, dates, places, etc.) as easily as possible. We are fully aware that limitations such as the quality of original documents or the quality of the digitization will prevent the recognition process to attain the same results as a human eye. But again, the expected gains in terms of costs and time for integrating such a large volume of data greatly justifies using innovative solutions.

The massive volume of birth and death records will also require adapted tools to import them into the existing BALSAC database. The linkage of new data will rely entirely on the adaptation of existing linkage tools at BALSAC that were designed to handle marriage records. The new data will be linked to the marriage records already in BALSAC to reconstruct families and, ultimately, the population as rigorously as possible. Lastly, a final stage of semi-automatic integration will allow both the treatment of complex situations and the validation of the linkage of births and deaths through a sampling process. Once integrated, these records will go through the coherence queries implemented in BALSAC (over 170) to ensure and maintain the integrity of the database.

Objective 2: Genetic Data Processing

The second objective will be to integrate into a single set genetic data collected on individuals recruited in the French-Canadian population. This will be achieved through a partnership with the CARTaGENE project[5], which has collected biological samples as well as health, lifestyle and sociodemographic data from 43,000 Quebec residents aged 40 to 69 at time of sampling. About 25% of participants have filled a genealogical questionnaire for inclusion in BALSAC. Aside from CARTaGENE, i-BALSAC will also rely on several collaborations with biomedical scientists who have agreed to share genetic data from participants to their research projects (given appropriate consent obtained from participants). This will also substantially increase the number of individuals from which genetic data is available.

Genetic data will be assembled, standardized and integrated in i-BALSAC. For each individual, information such as postal code or residential address will also be preserved in order to make the link with the geographical data (see Objective 3). Furthermore, for individuals with genealogical information, haplotypes (sequence of genetic markers on a chromosome) will be reconstructed and integrated in the database. Using bioinformatics methods relying on kinship links among individuals, it will be possible to impute complete sequences from available genetic data and to project them in the ascending genealogical lines of contemporary individuals (Figure 4). When

[5] For more information on CARTaGENE, see https://www.cartagene.qc.ca/en

completed, this will constitute one of the most detailed and structured genetic database available for a single population.

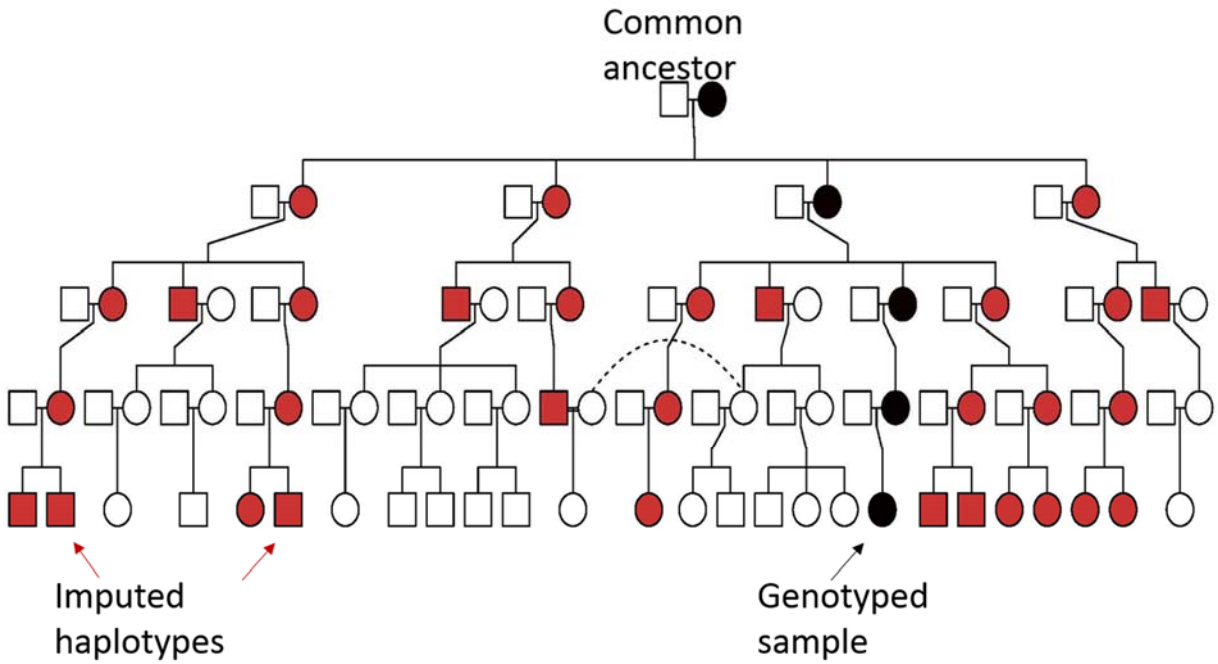

Figure 4: Maternal haplotype (mitochondrial DNA) imputation in genealogical lines

Objective 3: Geographic and Contextual Data Processing

Biological characteristics of a population are related to settlement processes on the territory and demographic relationships across communities over time. The creation of i-BALSAC thus includes the development of a framework for analyzing and interpreting the spatial dimension of genealogical and genetic data. Thanks to the integration of contextual data on the local populations of Quebec, it will be possible to contextualize genetic and genealogical data from the point of view of the demographic composition and of various characteristics related to the socio-economic development of localities.

The geographical component aims to provide i-BALSAC with the necessary resources to fully take into account the spatial dimension of the formation and evolution of local populations, for the entire Quebec territory. To do this: 1) we will reconstitute and integrate the geographies of the vital records registration and the Canadian censuses of the 18th, 19th and 20th centuries into a HGIS (see Objective 4); 2) we will characterize the local spaces from a socioeconomic point of view in order to use geographic analysis tools.

Geographic Data

Two sets of spatial data will be considered: the residences declared in the civil records as well as the places of registration of these records on the one hand, and the geographic information from the Canadian censuses on the other. To process geographic information from the civil records, BALSAC developed over the years an original location system based on the basic residential unit (*Unité résidentielle de base* URB). This unit fulfills two criteria: 1) it regroups under one code the localities that broadly correspond to the municipalities that existed at the end of the 1980s; 2) it contains at least one Catholic parish with records. This unit is also transhistorical: from the moment the URB exists, it does not change. Once geolocated by geographic coordinates, it makes it possible to locate the demographic events consigned in the civil records. The URB is thus useful for cartographic representation (location) of demographic phenomena distributed at regional or larger scales. However, it does not allow us to qualify these spaces, as it is not linked to any contextual data that would establish the economic bases or the degree of development of the local territory (population size, growth, maturation for rural spaces, economic sectors or level of specialization for urban spaces). To do so, census data is needed.

Consequently, the second set of spatial data to be included in the HGIS is based on aggregate census data published by Statistics Canada and formatted by the CIEQ over the last 30 years. A vast set of geolocated data was compiled for all Census subdivisions (CSDs) in Quebec from the 15 censuses published between 1790 and 1971, containing the headcount and distribution by sex, religion and ethnic origins. This data will give i-BALSAC the basic information to contextualize genealogical and genetic data and will serve to establish the level of development of local communities (see next section). Furthermore, this data is associated to geographic areas, either CSD areas for fourteen censuses or points for the last one (1790). To allow the processing of data from the French Regime, we will construct polygons for parishes and seigneuries in existence in 1722. We will then create a concordance table between the URBs in BALSAC and the census subdivisions. For geographic information outside of Quebec, the cartography will be based solely on the georeferencing of the declared residences, without the addition of contextual data (see below).

Context of Development of Local Spaces

Having established concordance between geographic data from civil records and censuses, we will be able to add to the HGIS additional layers of geographic information already available at the CIEQ, including: 1) railway construction; 2) topography, strongly associated with agricultural potential; 3) location of every post office in Quebec; 4) economic data (businesses) from Dun & Bradstreet credit records. To track the geographic evolution of civil records, a layer will be added for each parish indicating its religious status (mission, parish with open records) and the year of any change in status. As territorial occupation was based primarily on agriculture, which was highly differentiated from one place to another, a layer will be created with aggregate data from

agricultural censuses (1831-1971). Finally, we will use civil records (declared occupations) to establish socioprofessional profiles for each location and decade. Combined with aggregate census data, this will help us to establish the local socioeconomic context for any demographic event (pioneer fringe, occupied or saturated rural, specialized or general urban by size, etc.). i-BALSAC users will be able to account for the socioeconomic logic that fuels demographic dynamics, which in turn have an impact on the genetic composition of populations at every spatial scale.

Objective 4: Development of Tools for Combined Analysis of Genealogical and Genetic Data and Creation of HGIS

The previous objectives describe how the genealogical, genetic and geographic data will be assembled, standardized and integrated into i-BALSAC. To facilitate exploitation of the data and given the unique nature of the infrastructure, we also intend to develop new spatial analysis tools and statistical methods.

Toolkit for the Combined Analysis of Genealogical and Genetic Data

Genealogical data has the potential to greatly improve the study of genetic data[6] as it provides information on kinship links and on the geographic origins of ancestors. This information can be used to improve methods for imputing missing genetic data, methods for estimating haplotypes and genomic segments shared among individuals or within a single individual (homozygous segments), as well as methods for genetic linkage and association studies. However, there are very few statistical and bioinformatic tools that can exploit very large genealogical structures like those available at BALSAC. As we build i-BALSAC, we intend to implement original methodological tools to facilitate the integration of genomic and genealogical data in the statistical module R GENLIB[7]. GENLIB is a library of genealogical analytical functions originally designed at BALSAC to process and analyze genealogical data. These tools will be available on the portal (see below) for users, and they will also be used to infer the genetic structure of the French-Canadian population. This structure will be added to i-BALSAC to follow gene transmission in the genealogical lines thus providing an exceptional canvas for studies on the genetic determinants of health.

Development of the Historical Geographic Information System

The implementation of the on-line HGIS rests on the integration of two spatial datasets described in the previous objective and an original Web JavaScript application to interpret and display

[6] Speed D, Balding DJ (2015). Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*, *16*(1), 33–44. doi: 10.1038/nrg3821

[7] Gauvin et al. (2015). GENLIB: an R package for the analysis of genealogical data. *BMC bioinformatics*, *16*, 160. doi:10.1186/s12859-015-0581-5

mapping queries for researchers. The application will be an integral part of the i-BALSAC interface. Its primary functions will be: 1) the cartographic selection of locations for analysis; 2) the management of map layers; 3) the representation of data in thematic maps (points, segments or surfaces) using proportional symbols or colors; 4) and the exporting of maps as PDFs. The portal will also be used to aggregate results by location and to create thematic maps.

Over the years, the CIEQ has collected a wealth of resources related to HGIS employed in various projects, like a collection of historical Quebec atlases[8] The online mapping application will be based on proven technologies like *OpenLayers,* a library of JavaScript functions for high-performance online mapping applications, and will be used to develop an interactive approach to Web cartography.

Objective 5: Database Architecture and Portal Design

As i-BALSAC is intended for interdisciplinary research, the database architecture must make it possible to work in an environment that facilitates access to each of the datasets by users whether their focus is on demographic, genetic or geographic aspects. For instance, in order to facilitate use of demographic data in a genetic context, the latter will be converted to, but will not be limited to, a genealogical format. This will allow users to work with the population genealogical structure unfolding over successive generations and spanning almost four centuries. Synchronisation routines will ensure to maintain correspondence between genetic and genealogical data for any given individual for research purposes. Geographic information will be transmitted in real time to users on the i-BALSAC portal on a direct secure link between CIEQ and BALSAC. These environments will be dynamic and allow for new data to be added at any time.

A user-friendly interface will be developed to access the data, tools and documentation. It will also facilitate navigation between the different resources (i-BALSAC, BALSAC, IMPQ, CARTaGENE, CIEQ). Furthermore, this system will let us closely manage permissions and filters for data transfers, preventing the transfer of unauthorized data into i-BALSAC. Levels of access will be defined to take into account the diversity of research projects to be conducted and, more importantly, the limitations due to the confidential nature of civil records less than a hundred years old. Indexing will allow queries based on different levels and criteria, for example to work on groups of individuals, families, geographical locations and carriers of specific genetic mutations. The new tools will be entirely integrated in the main database to allow users to create complex and diverse queries.

[8] See espace.cieq.ca

Research opportunities

Developing the infrastructure will offer a wide range of research opportunities. For the social sciences, the addition to the existing data of all birth and death records until 1920 will extend by almost a century the period for which we know the full demographic parameters of the Quebec population (1621-1920). This addition will allow for historical analyses of a depth unprecedented in Quebec and exceptional on the international stage. It will open new avenues to understand the evolution of social structures in Quebec, as well as the interactions between the social and cultural groups that compose the population. Thanks to i-BALSAC it will be possible to study individuals in their family and community networks, in the meshing of individual-family-community relations where changes are lived and shared, where socio-professional or geographical mobility operates, and where people meet and identities shift throughout lifecycle stages and across generations. Geographic visualisation modules will allow researchers to use historical information to represent migration or evolution of populations over the last centuries.

In health-related fields, i-BALSAC will provide original data that can serve as a foundation to study genetic determinants of health. Analyzing cohorts using both genetic and genealogical data can help us understand the consequences of demographic history on the genetic structure of a population and measure its impacts on genetic epidemiology studies. Some findings could facilitate the development of study protocols to identify genetic variations associated with complex diseases, which eventually could lead to better strategies for treatment, diagnostic and prevention. Moreover, the integration of contextual data will broaden the interpretative framework of analyses and may contribute to the identification of risk factors linked to physical, social or family environments. Visualization and geographic representation tools will help researchers conduct stratification studies and verify if associations are specific to sub-populations or geographically linked individuals. Furthermore, linking genetic information with geographic data will help clinicians better understand the location and distribution of patients carrying the same mutation, and establish models for the distribution of detrimental mutations that could be of great interest for public health. For researchers in evolutionary biology, adding birth and death records will help establish selective variables of fitness for each haplotype (genomic region) in the population. Then, it will be possible to evaluate if haplotypes with a diminished fitness value compared to the average are transmitted along with traits that can explain that diminution. Thanks to the data in i-BALSAC, the French-Canadian population can become a reference population where models can be developed and tested before being validated on other populations.

Expected impacts and foreseen challenges

The creation of i-BALSAC will have positive impacts in many areas. Our first goal, which is a foundation for all the others, lays in the conservation, valorisation and evolution of an exceptional scientific, historical and biological legacy. Second, and as shown above, the ability to contextualize analyses in space and time, the access to linked genealogical and genetic data and to development of tools to facilitate joint analysis of these data, will be the foundation for innovative scientific research and methodological advances. To maximise utilisation of i-BALSAC, access to data will be free through a Web portal. The only limitations will be due to the confidential nature of some of the data collected from civil records (for events dating back less than a century) and the necessity to protect the identity of participants in research projects as defined in the the Canadian Tri-Council Policy Statement: *Ethical Conduct for Research Involving Humans*[9]. BALSAC Research Services will manage the terms and procedures to access the genealogical and genetic data and offer support to researchers who wish to use the infrastructure.

A third goal is to make i-BALSAC an open and dynamic infrastructure that the scientific community can continue to enrich with new data and tools, even after completion. The interface will be designed to receive data from collaborators who wish to contribute to the collective project. Open access to all resources and analytical tools will facilitate exchanges among researchers and establish a network based on sharing and communication. Moeover, i-BALSAC will help reinforce the synergy between existing research initiatives that collect data from different sources and organise, exploit and analyze them in a collaborative approach. As an example, civil records from i-BALSAC will also be integrated in IMPQ, increasing its spatial and temporal reach for researchers in social sciences and history.

We foresee challenges related to the optical character recognition of vital records as we will be using this technology for the first time. Our goal is to obtain the best possible results given the quality of data and the capacity of the technology. We are conscious that is a major undertaking but we are confident that the experience of the BALSAC team who will work in close collaboration with BAnQ and with our partner in charge of OCR will allow us to overcome these challenges. There are also issues surrounding data confidentiality and the integration of genetic data will make it necessary to adapt BALSAC confidentiality requirements. Again, on top of the BALSAC experience in such matters, we will rely on the support of our partners (in particular the CARTaGENE population project) to address this issue successfully.

[9] Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada, Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans, December 2014.

Conclusion

Building i-BALSAC will make it possible to achieve high-resolution mapping of the French-Canadian population in an approach that is at once genomic, genealogical and geographic, and offer a historical perspective over four centuries. It will make available a broad set of biographical information that is situated in time and space, enabling the study of populations based on individual trajectories within families and communities in a multigenerational perspective. To our knowledge, this level of precision, completeness and temporal depth has not been achieved for any population. The idea of such an infrastructure is particularly appropriate for the French-Canadian population, particularly because of the modalities of its formation (initial founding effect, presence of admixture, diversity of regional settlement histories), the resulting genetic structure and the exceptional quality of genealogical data. Ultimately, i-BALSAC will become a laboratory for population studies and a space for researchers to address questions through innovative multidisciplinary studies, a much needed approach in a world faced with increasingly complex issues.

Appendix: The i-BALSAC team

The i-BALSAC team regroups three specialists in historical and genetic demography (Lisa Dillon, Marc Tremblay, Hélène Vézina), one historical geographer (Marc St-Hilaire), two historians (Yves Frenette and Claude Bellavance), three genomics researchers (Simon Girard, Simon Gravel, Sébastien Jacquemont) and one genetic epidemiologist (Marie-Hélène Roy-Gagnon). The role of each member of the team is defined as follows: Vézina is the project's scientific director and will take the lead on the first objective (genealogical data processing) supported by Tremblay, Dillon and Frenette; Girard will supervise the second objective (genetic data processing) supported by Gravel and Jacquemont; St-Hilaire will be responsible for Objective 3 (geographic data processing); for Objective 4, tools for combined analysis of genealogical and genomic data will be developed in the laboratories of Roy-Gagnon and Gravel while Bellavance will be responsible for building the historical HGIS; finally, Objective 5 (database architecture and portal design) will be led jointly by Vézina and Girard, with participation from Bellavance for the HGIS.