Genetics and Molecular Biology, 37, 1 (suppl), 171-185 (2014)

www.sbg.org.br

Genealogical data in population medical genetics: Field guidelines

Fernando A. Poletta[1,2], Ieda M. Orioli[1,3] and Eduardo E. Castilla[1,2]

[1]*Estudio Colaborativo Latinoamericano de Malformaciones Congénitas, Instituto Nacional de Genética Médica Populacional, Rio de Janeiro, RJ, Brazil.*
[2]*Latin American Collaborative Study of Congenital Malformations, Center for Medical Education and Clinical Research, Buenos Aires, Argentina.*
[3]*Departamento de Genética, Instituto de Biologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil.*

Abstract

This is a guide for fieldwork in Population Medical Genetics research projects. Data collection, handling, and analysis from large pedigrees require the use of specific tools and methods not widely familiar to human geneticists, unfortunately leading to ineffective graphic pedigrees. Initially, the objective of the pedigree must be decided, and the available information sources need to be identified and validated. Data collection and recording by the tabulated method is advocated, and the involved techniques are presented. Genealogical and personal information are the two main components of pedigree data. While the latter is unique to each investigation project, the former is solely represented by gametic links between persons. The triad of a given pedigree member and its two parents constitutes the building unit of a genealogy. Likewise, three ID numbers representing those three elements of the triad is the record field required for any pedigree analysis. Pedigree construction, as well as pedigree and population data analysis, varies according to the pre-established objectives, the existing information, and the available resources.

Keywords: medical genetics, population medical genetics, geographic clusters, isolates, rare diseases.

Introduction

When the medical genetics of very large families, demes, or populations need to be studied, the field of Population Medical Genetics is utilized. Population Medical Genetics is a medical genetics subspecialty which main concern is to understand the causes of prevalence as well as the maintenance of unusual diseases in populations, generally, rural, isolated, and small, as conceptually defined and discussed elsewhere (Castilla and Orioli, unpublished data).

Data collection, handling, and analysis of large pedigrees require the use of specific tools and methods not widely known to human geneticists. Unfortunately, the graphic pedigree is the most popular tool in medical genetics in spite of its severe limitations in terms of family data recording, storage, and analysis.

This paper is intended to serve as a guide for fieldwork in population medical genetics research projects. To facilitate the understanding of the recommendations provided and their underlying concepts, a real population named Aicuña is offered as an example throughout this paper in italic text. Further information about Aicuña has been published elsewhere (Castilla and Adams, 1990, 1996; Bailliet *et al*., 2001).

Send corresponce to Fernando A. Poletta. Estudio Colaborativo Latinoamericano de Malformaciones Congénitas, Centro de Educación Médica e Investigaciones Clínicas, Dirección de Investigación. Galván 4102, Buenos Aires (1431), Argentina. E- mail: fpoletta@eclamc.org.

Field Guidelines for Collection, Handling, and Analysis of Genealogica Data

Pedigree objectives

Before you start interviewing people and recording a family history, reflect until you have a precise idea of why are you going to do it as well as an approximate notion of the expected workload.

The objective of a pedigree could be the recording of a breeding structure, family or population in which a proven or suspected mutant gene is segregating or some degree of endogamy might be present. While the former in most cases is applied to families in which the type of inheritance is to be proven and/or individuals at risk are to be identified, the latter is usually the case in populations with a higher than expected frequency of some deleterious phenotype.

Even though there is not a clear-cut distinction between a family and a population, some idea of the number

of individuals (pedigree members) to be recorded is needed in advance, or at least the order of magnitude. For instance, a large family seldom reaches one hundred persons, even including deaths, while the genealogical size of a population will usually be greater than one thousand and could even be as high as ten thousand.

When dealing with a population, be sure you actually need a pedigree because, depending upon your objective, performing a census or a register will be a better approach in many situations.

Initially, briefly describe the population to be studied, its size, and relevant characteristics. Also describe the geographic area where the population is located with precise limits, preferably within an administrative unit if it can be identified (municipality, region, state, etc.) in order to obtain background information when required.

Nature of collected information

As any living organism, families are as dynamic as the individuals composing them. Therefore, their information is neither complete, nor conclusive, and all collected data are to be taken as provisional and conditional.

Information components: Informer and individual

In practical terms, any pedigree must identify the informer and the individual or pedigree member. For instance, a person numbered 001_017 is understood as individual # 017 provided by informer # 001. Because different collected pedigrees of a given family or population do overlap or even repeat the same information, a single individual can have more than one number, and the hypothetical person 001_017 may simultaneously be person 023_133.

These provisional numbers primarily serve to organize field notes. Regardless of the numbering, the identity of a given individual is based on its genealogical links. Different numbers with the same parents and descendants can only refer to the same individual, and any information at odds with this assumption must be investigated and corrected.

Once the identity of individuals within the population is established, the next step is to resolve any differences in the information contained on separate forms and finally to assign an identification number to the individual in question.

Each individual record contains two distinct blocks of information concerning genealogical and personal data.

Genealogical data

The first block is comprised of a unique number identifying the nuclear family unit or basic triad, namely, the individual and his/her father and mother. Each number cannot be given to more than one individual. Nevertheless, consecutivity, or sequentiality, is not necessary, and numbering can be discontinuous.

Personal data

The second block contains the individual's personal information, such as, for instance, gender, full name, date and place of birth, marriage and death, phenotype, etc. Likewise, its length is unlimited because and each researcher can structure it as preferred, and links with other databases is both possible and desirable.

Tabulated data recording

The problem of graphic pedigree recording

Graphic pedigrees are very limited from all points of view. Their capacity for individuals is limited to less than one hundred, while genealogical or historical populations in most isolates are in the order of tens of thousands. The gametic links among individuals very soon become an incomprehensive set of interlocking and overlapping lines. The allowed personal information is limited to gender, indicated by squares or circles, plus less than ten similar data points, such as live-dead status, birth-death years, a couple of genomic indicators, and nothing else. Furthermore, the most limited value of pictorial genealogies is not related primarily to its use as a database but rather to its inefficiency for use in data analysis. As a matter of fact, the only utility of pictorial representation of family trees is to display a family genetic structure and/or disease distribution that has already been studied and proven by other methods. Even this depiction is bound to appropriate signs and nomenclature. Graphic symbols and keys follow preestablished standards provided by specialized journals or societies as summarized by Bennett *et al*. (2008) for the use of genetic counselors.

Data collection form

The sample page shown in Table 1 illustrates the process of data recording during the inquiry process. General information at the top of the page includes the identification of the Informer and the Operator, both by means of a code number, plus the date and place of data collection. The form contains the following two distinct parts: **Genealogy**, comprised by a triad of ID numbers for the person in a given row and the father and mother; and **Personal Information**, which here, solely for demonstration, includes gender, life or death status, year of birth, death, and emigration, and an amputated field for first name, indicating the continuity of the form following the right border. Some usual and also special situations are illustrated, namely the following:

Numbering of individuals within the pedigree is absolutely free and up to the operator decision. The only rule is not to use the same number for more than one pedigree member, since each individual number is an identity number. Conversely, a skipped or unused number creates no problem in pedigree analysis.

Some pieces of information require specific working definitions. For instance, for migration information, a local territory needs definition first, as in, for instance, a munici-

Table 1 - Tabulated genealogy: Data collection form.

Informer ID: ###			Operator ID: ###		Date/S: - - / - - / - -		Place/S:	
Pedigree ID: ###			Personal information					
Individ	Father	Mother	G	L	BIR-YR	DTH-YR	Name	AFF
001	004	003	F					2
002	004	003	M					2
003	005	006	F					1
004	007	008	M					1
005	014	009	M					1
006	999	999	F					1
007	017	016	F					1
008	013	012	M					1
009	013	012	F					1
010	013	012	F					1
011	013	012	M					1
012	999	999	F					1
013	024	023	M					1
014	999	999	M					1
015	017	016	M					1
016	021	020	F					1
017	999	999	M					1
018	021	020	M					1
019	021	020	F					1
020	024	023	F					1
021	999	999	M					1
022	024	023	F					1
023	999	999	F					1
024	999	999	M					1

G: Gender: M male, F female, I indetermined, blank: unkown.
L: Live or death status; BIR-YR: Year of birth; DTH-YR: Year of death (if dead); AFF: Affected status 1 (No); 2 (Yes).

pality. Moving into a region, if born outside, or out of a region, if born inside, define patterns of migration (immigration or emigration) constituting important components of the genetic structure and dynamics of a given population.

The correspondence between tabulated and graphical genealogy is shown in Table 1 and Figure 1. In the example presented here, the number 001 individual seems to be the index case because their ancestors and sibs are recorded.

Individual # 002 is also affected and is the brother of the index case because both have the same parents (# 004 and # 003).

Individuals # 005 and # 006 are the maternal grandparents, while # 007 and # 008 are the paternal grandparents.

Individual # 006 was fathered by an unknown person; however, it is known that these parents were members of the isolated community and were therefore numbered 999.

The relevance of this distinction is that for inbreeding calculations, while members of the 000 sibship are assumed to be outbred, those of the 999 group will have the average expected inbreeding coefficient of his/her generation.

In the absence of a specially designed data-recording form, a simple sheet of squared paper can very well do the job. Likewise, a simple spreadsheet application can be easily formatted to serve as a data-recording form if direct data keying is thought to be suitable for the planned fieldwork process.

Information sources: Oral and written

In general terms, the reach of well-informed collaborative oral informers reaches back to ancestors born in the mid 19th century. Considering that we are working at the beginning of the 21st century, that a current ancient informer could had be born in the 1930s, and that generation length is estimated to be approximately 30 or 35 years, the informer parents were born around the year 1900, grandparents around 1870, and great-grandparents about 1840.

Any information previous to that time limit must be restricted to written informers.

Information on extended genealogy is limited to a small number of informants who have a certain vocation for family history and moreover enjoy recounting information that they received from their parents. Nevertheless, an in-

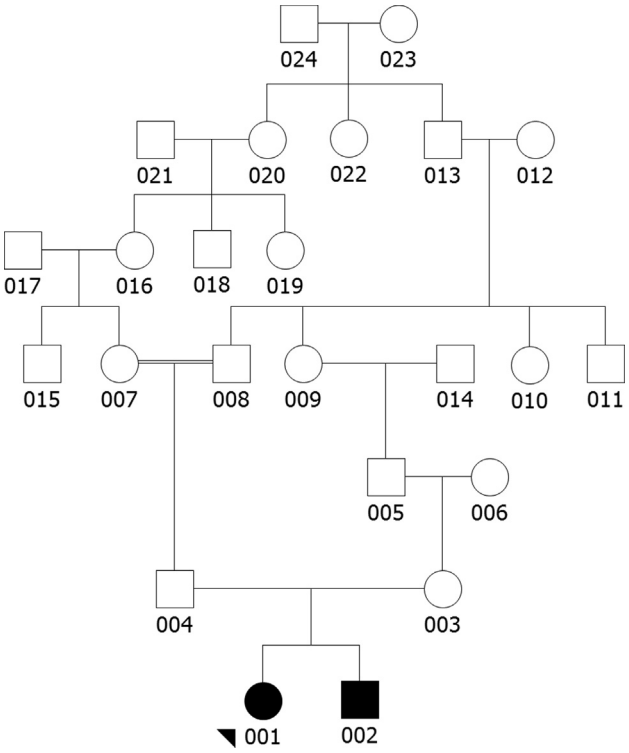

Figure 1 - Graphical pedigree of the tabulated genealogy shown in Table 1.

terest in history may not be consonant with an interest in genealogy. Usually, a small number of "amateur genealogists" is sufficient to maintain oral traditions and to prick the "genealogical conscience" of the community at large.

Conscious or not, genealogical interest cannot be solely based on a pure sense of belonging, or even of eternity, because a family spreads over time, making a predominant last name almost immortal, as in the case of the Buendias of Macondo, the Terra-Cambarás of Santa Fe, or the Ormeños of Aicuña (Castilla and Adams, 1996).

Genealogical interest could be directly linked to land and property rights, which, whether effective or not from the legal standpoint, provides the members of a genealogy with a perhaps unperceived but attractive sense of aristocracy. This was the reason for the genealogy of the pioneer generations in New England (Greven, 1972), the Petrorca Valley in Chile (McCaa, 1983), and also in Aicuña (Castilla and Adams, 1990). The influence of land property on the population genetic structure has been demonstrated because family branches with more land tended to emigrate less, therefore increasing the mean inbreeding coefficient across generations in the 19th and 20th centuries (Castilla and Adams, 1990).

Natural, or illegitimate, offspring might be hided or even masked by attributing a wrong paternity. The latter is particular of out of the wedlock conceptions. Since the impact of this type of errors is difficult to validate, even if suspected, coefficients of relationships and inbreeding for a given population pedigree must be understood as a minimal estimate of reality. This issue will be further covered in section "Illegitimacy, Confidentiality, and other Ethical issues".

Oral sources

For family groups, the best informer is a knowledgeable member of that group. For more remote genealogical data, usually related to the population founders and early generations, the best informers could be just one or a few well-informed persons, usually well known by all community members.

In isolated populations, the reliability of the informants is reinforced by endogamy, which reduces the number of ancestors (*loss of predecessors*), as well as by a feeling that could be called the "genealogical conscience".

Evoked persons are not fully dead

Even though the knowledge of genetic links among members of past generations may be relatively limited, a general knowledge of them is shared by all. There is a particular awareness of and sensitivity towards the dead that may not be exclusive to just one population; rather, this sentiment so eloquently evoked by Gabriel García Márquez for many of the characters in *Cien Años de Soledad* (1973), may be common to all rural communities in Latin America. In Aicuña, this sentiment results in the impression that no one ever really dies; many decades after physiological death, an individual lives on in legends and anecdotes such as "the stone of Don Rosario", "the spring of the Olivas" or "the footpath of Severino".

Each place named after a deceased person usually comes with a story that revives him/her, as well as the lifestyle and conditions at that time, helping the whole process of memory recall needed during the building of a large pedigree. Furthermore, this preeminence of the dead in the collective consciousness of the community greatly facilitates the task of oral informers.

Close and remote informants

Obviously, people know more of their family group than others. Different degrees of extension can be established between the nuclear family, or parental triad, and the whole kinship composing the whole community at the level of household, grandparents, descendants, etc. The recommended strategy to generate a large pedigree will vary from case to case, largely depending upon the research objectives and population size and distribution.

Current and ancient informants

The majority of informers are able to provide reliable information on their own generation and the subsequent one, namely their own offspring. Youth in general demonstrate very limited interest in their ancestors, and elders do not know much about infants and children.

Aicuña (example 1): At the end of a week of interviews, in spite of frequent digressions, two of the informants had provided both genealogical and demographic information on a total of 795 present and past members of the community. Don Julián's parents were born in 1863, and his great-grandparents, who were born between 1790 and 1801, were the offspring of persons censed in Aicuña in 1810 and signatories of the legal action creating the two estancias of Aicuña and Tambillos in 1833. Thus, it is hardly surprising that such an informant was able to provide reliable data from the beginning of the 19th century on, especially considering that the inhabitants of small consanguineous communities, with little immigration, usually possess good information on their ancestors.

Uni and multi personal sources

The sources of oral information include isolated individuals or groups of two or three persons, whether related or not and residing in town or not. Within a group of two or three informers, one usually is the primary source of information, while the others corroborate the information or provide additional information on a particular member of the pedigree.

In all other cases, groups of two or three informers are preferable to single individuals because discussion of a certain individual or ancestor frequently serves to stimulate the memory, often with the result of giving more credibility to the data.

Aicuña (example 2): The 49 informants utilized included single individuals such as informant #008, groups of two or three persons, such as informants #003, Don Julían and Doña Luisa. These individuals were located in the village of Aicuña (e.g., #003, and #008) in the city of Chilecito, such as informer #012, or even 1,300 km away in Buenos Aires, such as informer #1. This last informant was the first one interviewed because he lived in the city where the population survey was designed. He served as a "test" for the data forms used in the field before the final versions were being printed. Informer #001, aged 62 when the interviews took place, was able to provide genealogical and demographic information concerning 182 members of his immediate family and descendants.

The most prolific oral source of information was most likely group #003, which was already mentioned. Don Julián was 81 years old when he was interviewed, and Doña Luisa was 72. Julián and Luisa were much more informative as a couple than on their own. When talking about a branch of the family tree that was not very well known, in general concerning families that had emigrated, they would typically begin to digress, telling stories that eventually sparked their memories, allowing them to retrieve dates and genealogical information. They themselves would say that it was preferable to have each other present "para no dejarlo mentir" (to stop them from lying).

Although Julián and Luisa were one of the best sources of information in Aicuña, there were at least 10 others with similar characteristics. While they were being interviewed, the village nurse listened because the interviews took place in a small clinic in the village that was under her responsibility. She herself, 35 years old at the time, had a special interest in the contemporary generations of Aicuña and was able to provide information on 1036 individuals. When the data were updated, some 15 years later, after the death of Julián and Luisa, the nurse had become the best informant in the village for the past generations, having acquired a considerable amount of information in the intervening years from the elders in the village. Interestingly, a new nurse in the village had taken on the role of being the informant for the contemporary population. Thus, a new generation of informants had arisen in Aicuña to guarantee the continuity of the oral tradition, which was a responsibility, according to standards of the community, which, although not written, were no less rigid that those that were.

Illegitimacy, confidentiality and other ethical issues

Everyone is capable of providing information on their immediate family, and no one can provide better information on a given paternity than the mother of the child, provided she is willing to do so. Further information on these issues is provided below in Section "Ethics for field work".

Willingness

The reluctance to cooperate in providing genealogical information may have several underlying reasons. Usually they are not voiced or may be unclear for the person approached. However, that unwillingness should be respected without putting any pressure on the individual, not only because of ethical principles but also in order not to enter incorrect gametic links into the pedigree in question.

Because a 0000, or even better, a 9999, parenthood is largely preferable to a mistaken number given, thanks should be expressed for the refusal, even when a final, more vague question could close the conversation, as described in the next section below.

Ilegitimacy

While all biological offspring are natural , irrespective of their legal or religious status; illegitimate refers to being illegal, illicit, or even criminal, in reference to a given set of civil or religious rules that are valid for a given culture at a given time and place. Natural should be preferred to illegitimate under the assumption that the real meaning is natural and illegitimate. Out of wedlock, meaning outside of matrimony, is another way to designate a conception occurred during a valid matrimony but was seeded by a person not the formal spouse.

It is wise to investigate the terms and their real meanings employed in the community. Likewise, the general attitude toward parenthood should be understood. So-called

illegitimate paternities are not necessarily sinful or shameful.

Nevertheless, illegal paternities are rarely recorded in written sources of information such as vital statistics, church records, or legal documents.

The effort to be invested to define a given gametic link largely depends upon its genealogical value, measured as the contribution of the unknown person to the genome of subsequent generations or even to the genetic pool of the current demographic population.

Usually, crucial genetic inputs to the genealogy occurred several generations back, and it is very unlikely to depend upon the unrevealed paternity of a recently born child from an adolescent mother.

Confidentiality

Single informers serving as oral sources of information are generally preferable when information on parentage is concerned because such individuals are more likely to disclose information on certain paternities that was concealed from the rest of the community. This also, naturally, allows this information to remain confidential.

Nevertheless, it is unrealistic to promise full confidentiality of the collected pedigree information. Even if names are concealed, the genealogical links of a given member within a defined pedigree, finite no matter how large it may be, makes it easy to be identified by other pedigree members. Furthermore, in medical genetics, when the ultimate aim is the prevention of a disease, anonymity could curb the possibility to benefit persons at risk from suffering severe illnesses.

Other ethical issues

Sometimes, a person refuses to speak about their ancestors because they wish to let their ancestors to rest in peace. This could reflect an unwillingness to collaborate with the interrogators instead of a true expression of their beliefs. The investigator must be aware of this type of attitude because if we ignore the reason for a certain reproductive pattern in a given family or population, we could also be unaware of the possible reasons for concealing information.

Written sources

Two different procedures can be applied to work with written records: 1. A search for a genealogical path: a string of gametic links following the procedures elaborated in 6.4.: ascent, descent, sideways. 2. A search for a person: a genealogy member listed in the alphabetical indices and confirmed by genealogical links, in order to complete or check personal information such as names, dates, places, etc.

At a first site visit, identify the available written records and the year of the first registration in each log book.

Ecclesiastic log books

In Latin America, church records refer exclusively to the Roman Catholic Church, primarily if working with log books from the 19th century and earlier. These books are very similar in different countries and regions, as well as in different languages, such as Spanish, Portuguese, or others.

Parish records have considerable genealogical and demographic interest because they contain the following four different categories of information: baptisms, marriages, confirmations and deaths. Three of them reflect vital events (birth, marriage, and death), while one (confirmation) could be a proxy for a population census.

Church books are usually well kept and stored in good conditions although some of them may be three or four centuries old. The reading quality is contingent on the preservation of paper and ink and largely depends upon the regional climate; the dryer, the better. Desert regions are ideal for the preservation of books and for any object of archeological value.

Nineteenth and even 18th century handwriting is easy to read after one day of self-taught training, and the few abbreviations used are easily decipherable – for example Ma. for Maria, Fsco. for Fransisco, v. for *viuda* (widow) or *viudo* (widower), and so on.

In general, the quality of the parish registers decline significantly in both organization and detail after the year 1900, most likely because of the increasing presence of the Provincial and State governments and the accompanying civil records, which relieved the Church of the responsibility for recoding vital statistics. Church death recording is practically nonexistent because deaths are registered by governmental civil registry, and burial is no longer limited to the churchyard (camposanto).

Baptism

Each baptism (Christening) record contains the following information: date of birth, sex, first and last name, first and last names of both parents, or just of the mother in the case of a natural birth.

Records are written in chronological order in several handwritten volumes.

In addition, there usually is an alphabetical index, which facilitates any searches for specific members of the pedigree.

Baptism records provide genealogical information for a single unit in two successive generations, a child and his/her parents.

Marriage

Each marriage record contains the following information: first and last names and ages of the couple and the first and last names of both sets of parents.

These records are also written in chronological order in several handwritten volumes.

An alphabetical index exists but only for the male partners.

Marriage records provide genealogical information for two linked units in two successive generations, both spouses and the four parents. This is the most valuable church record for genealogical assembly.

In the case of consanguineous marriages, the appropriate ecclesiastical dispensation is recorded with a complete description of the type and level of consanguinity. This is hardly surprising because dispensation was required in the Catholic Church until 1917 for marriages between partners as distantly related as third cousins (Royo-Marín, 1958).

Death

The nature of the death records varies substantially over time. First and last names and age are always present, although names of the partners and cause and place of death appear inconsistently. These records are listed chronologically in several handwritten volumes and, as for baptism and marriage records, an alphabetical index also exists.

Death registries usually state the cause of death, which is generally useless for medical purposes, except to differentiate violent from natural causes ("*Morte matada e morte morrida*" in Portuguese). Obviously, folkloric or cultural cause of death could be useful for other non-medical specialties.

Confirmation

The significance of the confirmation records to the reconstruction of the demographic and genealogical history of the population is somewhat different from the other records because the date of confirmation does not correspond to any important demographic variable, and confirmations are administered to any baptized person of sufficient age.

Confirmations are only performed by the bishop, who is usually established in a large town far from the locality in which the genealogy originated. A rural region would rarely be visited more than once in the bishop's lifetime. Because these visits occurred so rarely, very few of the inhabitants would have been confirmed during a previous visit. Therefore, all the inhabitants were confirmed at once, and the few that had already been confirmed on a previous visit acted as godfathers and mothers. Thus, the confirmation records show that a small group of godparents, which included members of the Episcopal cortege, served repeatedly for the confirmation of one or two hundred persons. In this way, many of the inhabitants of a community and the surrounding areas were registered at every visit by the bishop as confirmees, godparents, or in a few cases as both. In other words, the confirmation records for a given place and date can be considered much like a population census.

Because the entry for each confirmee provides information on the first and last name, sex, age and parents' names, these confirmation records contain important genealogical as well as demographic information. As to the former, like baptism, confirmation records provide genealogical information for a single unit in two successive generations, a child and his/her parents. As to the latter, each person is registered with names, gender, and age, allowing for a population census under these variables, assuming total coverage between confirmees and godparents.

Civil

Civil registry

Because civil registers in Latin America were established at the start of the 20th century, a period easily reached by oral informers active one hundred years later, they have rarely been used until now. However, those records could be useful to validate personal data for pedigree members, such as year and place of birth or death.

Analogous to parish records, civil registries also cover the three basic vital events: birth, marriage, and death.

Elector registry

Voter registries are currently available in electronic form for many populations. One use for them in human genetics is in the study of last names as phenotypic markers. Several analytical methods have been described for the use of last names as indicators of ethnicity or breeding structure not only in isolated groups but in large populations as well (Bronberg *et al*., 2009).

Military service records

Given that the existence and characteristics of military service records varies among countries and periods, it is wise to investigate the situation for the population group in consideration.

In some places, medical examination for military service is the only cross sectional population sampling that exists after birth, even if it is limited to the male gender in most Latin American countries.

Another use of these records is in the historical study of anthropometric indexes indicating the nutritional status of a given population through certain time period (Vargas *et al*., 2010).

Personal

A variety of other documents such as wills, property deeds, and various publications concerning important individuals in the family may also exist, which also contain valuable information for your research project.

Historical

Considering that a large family or population pedigree usually extends over a long time period, and it is inserted into a larger human body such as a country or region, the local history should be investigated. Cultural or physical catastrophes could influence the breeding pattern with further genetic effects such as drift. For instance, in the south Atlantic island of Tristan da Cunha, there were two unusual events, emigration of over half the population in a single year in the mid-19th century, and a major boat disaster, both of which significantly affected its demographic history (Thompson, 1978).

Nevertheless, isolated populations could also be unaffected to the events of the general population. For instance, no reference was found in Aicuña to the long civil war occurred in that Argentine province, La Rioja, during the 19th century (Castilla and Adams, 1996), or to the Guerra de Canudos in the military garrison, Monte Santo, in the late 1890s (Manzoli *et al*., 2013).

Aicuña (example 3): The sudden flood in 1952, described in detail in the historical book of Aicuña, merits comment because it was the largest catastrophe described in the history of Aicuña. It occurred around sunset on the 19th of January and was caused by torrential rain in the hills surrounding Aicuña. The torrents of water and large rocks destroyed plantations, irrigation works and even houses, although no one was injured or killed. The significance of the record of this catastrophe is that it serves as a point of reference from which we deduce that nothing more serious occurred in Aicuña during 350 years of well-documented history. Thus, the history of this isolate differs in this respect, for example, from that of Tristan da Cunha (Thompson, 1978). From a genetic structure standpoint, catastrophes are relevant as long as they substantially reduce the effective population size, creating conditions for genetic drift to occur and affecting one or more mutated genes.

The most exceptional characteristic of the population of Aicuña is the existence of a large and complete genealogy extending over many consecutive generations through three centuries because the founding of the community. Table 2 summarizes the sequence of seven landmark events in the history of Aicuña occurring in 1674, 1723, 1833, 1866, 1912, 1955, and 1971 for which good written documentation exists. The information provided in these documents allowed the construction of the basic genealogy of Aicuña since 1610, the approximate date of birth of the recognized founder of the kinship, General Pedro Nicolás de Brizuela. With the exception of the interval between the judgment of 1723 and the successor agreement of 1833, the intervals between all other dates are less than 50 years, or one or two generations. Consequently, much of the information obtained from consecutive events overlaps. This not only facilitated the construction of the genealogy but also ensured its accuracy. The interval of 110 years between 1723 and 1833 includes four generations, but at the time the population size was very small.

Fiction

Some families and populations are described or incorporated into fictional literature, either directly or indirectly. Among the former, the population of Ricaurte, in the Valle del Cauca Department, Colombia, a cluster for fragile X (Wilmar Saldarriaga, personal communication), was the or-

Table 2 - Historical facts in the Aicuña population ensuring a recorded population pedigree extended through 15 generations, along 297 years.

Year	Fact
1674	The land of Aicuna was purchased by the founder of this kinship, who immigrated from Spain in 1630, and given as a deed of gift to his illegitimate son. The scripture is filed in the court house of La Rioja. It covers the first two generations: birth years 1610 and 1650.
GAP (Years: 49 / Generations: 2)	
1723	A granddaughter of the founder and her husband won a trial against influential people trying to take the land away from the legal heirs, whose rights are based on pedigree data. The trial record is filed in the court house of La Rioja. It includes a third generation, with its birth-year about 1680.
GAP (Years: 110 / Generations: 4)	
1833	The land was divided in two parts, Aicuña and Tambillos, through a successory agreement signed by members of the seventh generation of the kinship. This issue was based on well recorded detailed genealogical data. The agreement record is filed in the court house of La Rioja. It extends the pedigree up to a 7th. generation, with a birth-year about 1780.
GAP (Years: 33 / Generations: 2)	
1866	For an unknown reason, a local priest named Fray Aymon wrote a detailed pedigree expanding the one made in 1833 a couple of generations further. This document is kept in the parish of Villa Union. The pedigree is extended to a 9th generation, with a birth-year about 1840.
GAP (Years: 46 / Generations: 2)	
1912	A new successory agreement was signed among the heirs of the land of Aicuna, including further genealogical data. The agreement record is filed in the court house of Chilecito. The pedigree reaches now to an 11th generation, with a birth-year about 1890.
GAP (Years: 43 / Generations: 2)	
1955	A commission was created to update the pedigree, to list out the declared heirs and to carry on a new successory agreement, which was finally performed in 1963. Pedigrees and declaration of heirs is kept by commission members of Aicuña. Pedigree extended to a 13th generation, with a birth-year around 1920.
GAP (Years: 16 / Generations: 1)	
1971	The authors expanded the pedigree up to births occurred in 1971, which do approximately correspond to members of the 15th generation of the kinship. Both, written and oral sources of data were used.

igin of El Divino (Alvarez-Gardeazábal, 1986), a novel also produced for TV. Seemingly fictional genealogies, such as the Terras-Cambarás in "O Tempo e o Vento" by Erico Verissimo (1984), and the Buendias in *Cien Años de Soledad* by Gabriel García Márquez (1973) most likely represent the reality of whole populations (Castilla and Adams, 1996). Even outside of Latin America, large pedigrees served as inspiration for well-known novels as *The Tin Drum* by Günter Grass (1999).

Pedigree construction

Depending upon the programmed plan, one or more operators might be simultaneously be collecting genealogical data from one or more informers. In a simple example, while one person is working at the parish house searching for specific information concerning consanguinity type and degree of specific marriages, a second colleague may be collecting the pedigree of a family group from two oral informers.

Field methods

These are tasks to be organized for performance at the population site, that is, in the field.

Preliminary consultations

Search for suitable informers, including available documentation, and suitable verbal informers.

Prioritize well-informed elders because they may not be available at your next site visit.

Working definitions for space, time, and persons

Space, time, and persons are the classical three pillars of epidemiology (Porta, 2008).

SPACE. Define areas for demographic purposes of the population, as well as for related definition such as migration patterns. Covering progressively concentric areas can be a useful practical approach.

TIME. Do not register ages but rather year of birth because age is relative to the time of data recording. Families and populations as breeding units are nearly eternal. Thus, define the intended time extension ahead of time, for instance, the 18th century if the founder generation is estimated to have entered the region around the year 1700.

PERSONS. Decide the type of individuals or pedigree members to be included. BY SURVIVAL: induced abortions, spontaneous miscarriages, stillbirths, neonatal deaths, etc. BY DISEASE: affected, unaffected, or unknown. Special attention must be given to the unknown, which can be easily confused with unaffected when the disease is not conspicuous or congenital, is highly lethal, or is a local taboo. In any of these situations, the estimated penetrance of a mutant gene can be underestimated or even undetected. OCA (Oculo-cutaneous albinism) is a good example of an easily recognizable phenotype because it is obvious to any person without requiring a special examination; it is present at birth, it survives to adulthood, and it is usually not linked to any cultural anathema and so does not need to be hidden in Latin America at least (Jeambrun and Sergent, 1991). The alternative situations presented by diseases as galactosemia, ataxia-telangiectasia, Huntington's chorea, etc. are easy to identify.

BY PLACE OF RESIDENCE OR DEATH: inclusion or elimination of emigrated family branches.

Search for sources of information

List available informers, oral and written, their localization, validation of reliability, and expected usefulness of each source for specific sections and procedures in constructing the genealogy.

Steps: Initial, extension, repairs

Define specific steps to be taken, assigned operators, estimate timing, decide operative places, select dates, etc.

Initial

Decide on your strategy for index cases. They could be persons affected by a certain disease, the youngest member, or the eldest member of that family group, etc. Feel free to decide according to your preset objectives (see Pedigree objectives), and general planning, but make a decision and stick to it.

Extension

Decide on the direction (see Section Pedigree expansion, below) and extent of pedigree progression. For instance, you can work with an oral informer to generate pedigree data by ascent until the eldest ancestors this informer knows, leaving these data to be linked with information on elder generations obtained or to be obtained from written documents. The linkage, if successful, is defined as a junction or intersection (*entronque*).

Repairs

Review the collected information and search for omissions and inconsistencies to be repaired in the next encounter with the same informers. Some omissions may require the interview of other or more informers. The affected information can be genealogical (parentage; gametic links) or personal, the former being more critical because it does not affect only a single individual. Compare data from different informers (See Section Pedigree assembling, below).

Pedigree assembling

Perhaps the most time consuming component of the assembly of the data set involving the linking of multiple records is putting together information gathered by different operators from different informers. This work involves three aspects, resolution of different sources of information, linkage of genealogical information with demographic information, and linkage of records of nuclear families.

The existence of multiple sources of information makes it possible to construct an unusually reliable and

complete data set. However, this inevitably means that for some individuals, discrepancies and inconsistencies have to be solved. While many discrepancies disappear after careful review of the records and are revealed as simple errors in transcription, others are not so easily worked out. In such cases, further sources of information should be sought out and investigated. If the discrepancy remains, which occurs in only a small minority of cases, the data must be discarded and/or recorded simply as missing or unknown.

To validate means to ratify or rectify the genealogical information regarding the identity of a person, his/her mother and father; whenever possible, at least three independent sources of information for each gametic link (mother-child and father-child) are to be collected.

Pedigree expansion

Ascent

This phase of the survey follows the family in a direct line through an ancestor to the oldest known genealogy members. Ancestors are followed upwards in a direct line, without sidewise extensions.

This procedure intends to disclose possible genealogical links among not obviously related family groups. Furthermore, it simultaneously searches for genealogy founders.

One of the effects of inbreeding is reduced ancestors. "Reducing families" is another way of seeing the same reduction. The more families are linked in their origin, the easier it will be to define the breeding structure for that population or family (See Section Four basic sub-routines, below).

Descent

This procedure is used when one or more crucial breeders are identified. A crucial breeder can be a person identified for a given pedigree as the founder, a migrant to another location from whom a family branch and a new locality is derived, a survivor of a catastrophe seeding many offspring, etc.

Sideways

Sideways genealogical data collection, that is, extending it to siblings of the ancestors and their descendants, is mainly used in order to identify related persons with a given phenotype, such as a disease for instance. Otherwise, those data can be summarized, leaving them for later data completion or not.

Linkage of nuclear families within the genealogy of the population

Linkage among nuclear families is usually performed manually, although when a very large pedigree is to be handled, the process can be programmed and executed electronically. Nevertheless, manual terminating is recommended.

Aicuña (example 4): The linkage of nuclear families in the genealogy was performed manually. The basic genealogy elaborated during the 18th and 19th centuries was used as the basic frame of reference because it contained the largest numbers of individuals. However, "biological" logic required that links were established descending the pedigree in contrast to the usual approach of ascending the pedigree from the contemporary population toward the ancestral founders. Initially, a branch of the genealogy was ascertained by descent and then verified by ascent, preferably using independent sources of information. The overwhelming predominance of a few last names and a tendency to repeat first names in a certain branch or the popularity of certain first names in a given generation occasionally made the task of record linking quite difficult. In such cases, the use of nicknames as identifiers proved to be invaluable. These nicknames were often derived from some unusual physical characteristic, such as La Rubia (the Blonde), El Turco (literally the Turk, but also an epithet commonly used in the area to refer to someone with a foreign appearance), a particular handicap, such as El Rengo (the limper), some official capacity the individual may have, such as El Aguatero (the Waterer), or El Caminero (the road worker).

Biological validation of pedigree data

Because pedigree data are based on feeble sources, the resulting genealogies should be validated with biological facts, whenever possible. In the case of the Aicuña population, an 85 % agreement was found between the biological and genealogical data (Bailliet *et al*., 2001).

Databases with genealogical data

Database construction

As described in the previous sections, genealogical data could be collected by the tabulated method using paper recording forms. These genealogical data should then be keyed into databases for electronic storage, quality control, and analysis. A simple spreadsheet application such as Calc, Excel, Lotus 1-2-3, etc. or a database management program such as Base, MySQL, Fox-Pro, Access, or similar, could be used to design these databases.

When a single pedigree is considered, each record represent one pedigree member that is identified by an individual ID, while the triad IDs (individual, father, and mother) and the connections among nuclear families define the degree of relationship among all individuals in the pedigree (See Table 1 and Figure 1).

However, these individual IDs may not be adequate as unique identifiers when several families are recruited into the study and are recorded in the genealogical database. Thus, another unique identifier, called the "key variable" is then necessary to identify each individual within the overall databases. This "key" identifier could be, for example, a combination of the pedigree ID and the Individual

ID (PID+IID), or a unique alphanumeric identifier that is not related to the pedigree IDs. The latter would be recommended to maintain a higher level of con?dentiality with respect to certain labeled materials and within databases that include names and disease and genotype information.

In addition, this key variable is useful to interconnect data from several databases associated with a research project including personal information, namely, demographics, medical examinations, phenotypes, family medical history, perinatal history, genotypes, etc. Information concerning each individual stored as single records in different databases could then be linked using this key variable, bringing different data sets together into customized job files according to specific requirements for queries and data analysis (Figure 2).

Four basic sub-routines

Four main outputs indicating different aspects of the pedigree breeding structure can be produced, as indicated by McLean (1969):

- By ascent: tracing all direct ascendants from any given person in the pedigree, designed as ancestry, kinship, origin, parentage, extraction, or root.
- By descent: providing all direct descendants from any given person in the pedigree: clone, heritage, legacy, or lineage.
- Inbreeding: calculating the inbreeding coefficient for any person in the pedigree, including parental consanguinity.
- Relationship: computing the coefficient of relationship between any pair of persons in the pedigree, also known as parentage or family links.

Pedigree outputs

Pedigree and directory

Dataset records ordered numerically by ID numbers are a Pedigree, and when they are ordered alphabetically by name, a Directory. Having this printed output on hand makes the reconstruction of a given person's genealogy simple and fast even in the absence of an electronic device. The names of the required person are sought in the Directory, the ID number notated, and then the full record is searched in the Pedigree, where the parental IDs are also written. Pedigree ascent can then be easily performed following successive parental numbers. If descent is desired, the original person's ID number is pursued in the corresponding column, Father or Mother, according to the gender. This procedure can be more troublesome when working manually if individuals have not been identified with consecutively numbered IDs and the parent columns are not ordered numerically. With the help of known family groups, last names, or localities, this search is possible, although not easy. Nevertheless, this procedure is facilitated with the use of available computational resources, some of which are described below in Section Software for pedigree analysis (below).

Population genetic structure

Two different population units that partially overlap should be considered, namely, the Genealogical and the Demographic. The first one, also designated as the "historical" population, can be defined as all the descendants of a given founding person, couple, or group of persons, dead or alive, emigrated or not, while the second is defined as all persons living within a defined geographic area at a given time, including but not exclusive of the descendants of a given founding person, couple, or group of persons. For instance, in the case of Aicuña, the genealogical population included over 8,000 persons, while the demographic population at the time of the study comprised approximately 400 inhabitants for the strict area, the village, and 2,000 for the broader territory, the "estancia" (Castilla and Adams, 1990).

Database A

	A	B	C	D	E	F	G	H	I
1	key	Pedigree name	Individual ID	Father ID	Mother ID	Gender	Affected	Place_of_res	Place_of_birth
2	aic_121	aic_subped2	1	4	3	F	2	Aicuña	Aicuña
3	aic_122	aic_subped2	2	4	3	M	2	Aicuña	Aicuña
4	aic_123	aic_subped2	3	5	6	F	1	Aicuña	Aicuña
5	aic_124	aic_subped2	4	7	8	M	1	Aicuña	La Rioja
6	aic_125	aic_subped2	5	14	9	M	1	Aicuña	La Aguada
7	aic_126	aic_subped2	6	999	999	F	1	Aicuña	Aicuña
8	aic_127	aic_subped2	7	17	16	F	1	Aicuña	Aicuña

Database B

	A	B	C	D	E	F	G	H	I
1	key	rs10213286-A1	rs10213286-A2	rs3775261-A1	rs3775261-A2	rs1042484-A1	rs1042484-A2	rs12532-A1	rs12532-A2
2	aic_121	2	1	1	1	1	1	2	2
3	aic_122	1	1	1	1	1	1	2	2
4	aic_123	1	1	1	2	1	1	1	2
5	aic_124	2	1	1	1	1	1	2	1
6	aic_205	1	1	1	2	1	1	1	2
7	aic_206	1	1	1	1	1	1	1	1
8	aic_207	1	2	1	1	1	2	0	0

Figure 2 - A "key" variable enables the merging of different data sets into customized job files according to specific requirements for data analysis.

Pedigree analysis

The approaches chosen for the analysis of genealogical data will depend on the questions being asked in each aim of the research study. Thus, with appropriate design and analysis, it would be possible estimate the population breeding structure; evaluate the existence of genetic factors associated with disease susceptibility in these families; and in combination with genotyping data, detect causative genes through linkage and family-based association analysis.

In a descriptive analysis of the reproductive structure of the population, we could trace and list all direct ***ascendants*** or ***descendants*** from any given person in the pedigree, select a subset of individuals who have a ***common ancestor***, or find individuals with ***multiple mates*** or ***consanguineous mating pairs***. Moreover, it would be interesting to estimate some indicators such as the ***inbreeding coefficient*** (*f* or *F*). It could be defined as the coefficient of correlation between alleles of uniting gametes for a individual (Wright, 1922), relative to gametes drawn at random from a population; or as the probability that two homologous alleles in an individual are identical by descent (IBD) (Malécot, 1948), that is to say, they are derived from one allele of a common ancestor. This could be estimated from pedigree data using Wright's path method (Wright, 1934). Additionally, the ***coefficient of relationship*** (*r*) (Wright, 1922), also known as ***kinship coefficient*** or ***relatedness*** between two individuals, is the probability that genes randomly selected from an individual "A" and from an individual "B" are IBD (Falconer and Mackay, 1996), so it is equal to twice the inbreeding coefficient of their possible offspring ($r = 2f$). These indicators are frequently used as a measure of the level of homozygosity and are useful predictors of covariance and phenotypic correlation between relatives. They can be expressed per individual, as the population average, or as the average for some particular groups such as affected and unaffected, founders and non-founders, or by generations.

When the disease under study has a complex etiology, one of the first steps in the approach is to determine whether there is evidence of familial aggregation and a possible effect of genes. Pedigree data can be used to calculate the frequency of the disease by the degree of relationship with the probands and estimate ***recurrence patterns*** (Davie, 1979; Khoury *et al*., 1993) and ***familial relative risks*** (Risch, 1990) before genotyping is done. Familial relative risks (FRRs) or relative recurrence risk (RRR) for a relative of type R (λR) are calculated as the ratio of the risk of recurrence to a relative, R, and the risk in the population (baseline prevalence). The patterns of familial recurrences and the drop off in the observed FRR from the first ($\lambda 1$), to second ($\lambda 2$), and third-degree ($\lambda 3$) relatives can potentially suggest different models of inheritance based on the expected values for a single major locus, different multiplicative multilocus (epistatic) models (Risch, 1990), and for the multifactorial threshold model of inheritance (Reich *et al*., 1972).

Pedigree data can also be used to measure phenotypic correlations by degree of relationship (pairs of relatives with concordant and discordant phenotypes). This allows us to estimate the proportion of the total phenotypic variance that is due to inheritance of genetic factors (***heritability***) (Emery, 1986) and evaluate whether this pattern of correlation is consistent with a possible genetic effect.

If positive evidence of familial aggregation and a pattern of correlation consistent with a possible effect of genes has been found, the next step is to search for the effect of a major gene segregating in these families. This kind of statistical analysis can also be performed using pedigree data, which is known as ***complex segregation analysis*** (***CSA***) (Lalouel *et al*., 1983). Briefly, CSA is a more elaborate method to compare the statistical fit of the observed pedigree data with alternative and more complex models of inheritance, such as sporadic, environmental, multifactorial/polygenic, Mendelian major gene, and several mixed models. Different parameters that define each alternative model are estimated through the maximum-likelihood method: the frequency of the high-risk allele (q) at the major gene (inferred from phenotype segregation in pedigrees), the allele transmission probabilities (1.0; 0.5; and 0 for the three genotypes in Mendelian models), and the multifactorial heritability (the additive influence of polygenic background different from the major gene effect). These parameters can then be utilized in parametric linkage analysis or to used estimate the power and sample size in family-based association studies. In CSA, there are specific analytical approaches based on the characteristics of the trait under study (for example, a qualitative or quantitative trait); two of the most employed approaches for the study of diseases are the unified mixed model (Lalouel and Morton, 1981) and the regressive multivariate logistic model for binary traits (Karunaratne and Elston, 1998).

Starting from databases, as defined in Section Databases for genealogical analysis (above), genealogical data will be set up for these types of queries and statistical analysis. As outlined below, several computational resources to carry out pedigree analyses are available.

Software for pedigree data handling and analysis

We briefly describe here some software and computational packages for data handling and pedigree analysis. In each case, the formats of input and output files are shown.

As an example, data files from the *Aicuña research project* will be used. A subpedigree with 196 individuals has been extracted from the complete genealogy of Aicuña (8,696 individuals) and is available for download as supplementary material in an ASCII text file (File S1).

Pedigree drawing

Several pedigree-drawing software programs are available for use in medical population genetics, including the following: *COPE* (COllaborative Pedigree drawing Environment) (Brun-Samarcq *et al*., 1999), *Cyrillic* (Chapman, 1990), functions run in R such as *plot.pedigree* in the package *kinship* (Sinnwell and Atkinson, 2013), *pedigreemm*, *pedigree*, and *pedtodot* in the package *gap* (Zhao, 2007), *Haplopainter* (Thiele and Nürnberg, 2005), *Madeline* (Trager *et al*., 2007), *Pedigree/Draw* (Mamelka *et al*., 1987), *PedigreeQuery* (Kirichenko, 2004), *PedHunter* (Agarwala *et al*., 1998), and *Progeny* (Progeny Software LLC, South Bend, Indiana, USA), among several others.

As an example, the process for plotting the genealogy in File S1 using Progeny is as follows: (1) Download the tab-delimited text File S1; (2) Unfold the menu options and select: *Pedigrees* > *Import...* >; (3) In the Import module, browse the file and then select the File Delimiter ("tab"); the Import Type ("Pedigrees"), and enter the values that used in the database for Male, Female, and Deceased; and select the columns names. The pedigree will automatically be drawn, adding missing parents. For complex pedigrees with loops, multiple matings and consanguineous matings, manual edition of the pedigree could be necessary. The graphical pedigree corresponding to the tabulated genealogy in File S1 is shown in Figure S1.

An alternative to commercial software is to use functions for pedigree drawing under the GNU General PublicLicense, such as packages available in R. For example, the package called kinship (and kinship2) can be used to plot the tabulated genealogy in File S1. The format of the input file should be as shown in File S2. Call the package by typing the following at the shell prompt:

```
 > R
 > install.packages("kinship2")
 > library(kinship2)
 > aicped < - read.table("C:/your_current_direc-
tory/file_s2.txt", header=T, sep="\t")
 > attach(aicped)
 > ped < - pedigree(id, fid, mid, sex, aff)
 > par(xpd=T)
 > plot.pedigree(ped, cex=0.45, symbolsize=1.5)
```

This function gives a graphic object (Plots) that can be saved as image format and is shown in Figure S2. A detailed description of the arguments and options available in this package can be obtained typing "?plot.pedigree" or "?kinship2" at the shell prompt.

More information about other software for drawing pedigrees is available in "An alphabetic list of Genetic Analysis Software" (http://www.jurgott.org/linkage/ListSoftware.html).

Pedigree description

As detailed in Section Pedigree analysis (above), several statistics can be calculated to describe a pedigree. The classical four outputs "Ascent", "Descent", "Inbreeding", and "Relationship" originally implemented in a computer program by McLean (1969) can be now calculated from several available and user-friendly software programs.

Table 3 shows a detailed description of pedigree statistics and breeding structure of the genealogy with 196 individuals in File S1 and plotted in Figure S1. We will briefly describe here some available software programs that can be used to calculate these and other indicators.

A general description of the pedigree can be obtained by running the program PEDINFO available in S.A.G.E. version 6.2 (2012). To import the data to S.A.G.E., the input file must have the format shown in File S3. Programs in S.A.G.E. may be executed by means of a command line directive or from the provided Graphical User Interface (GUI) as follows:

Table 3 - Analysis of pedigree structure and inbreeding of a subpedigree with 196 individuals from the *Aicuña research project*.

	Affected	Unaffected	Total[1]
Pedigrees	-	-	1
Individuals	19	145	196
Gender: Male	9	71	94
Female	10	74	102
Founder	0	29	61
Nonfounder	19	116	135
Consanguineous mating pairs	-	-	11
Inbreeds	9	17	26
Average inbreeding coeff.	0.016	0.034	0.028
Average inbreeding coeff. (in the inbreds)	0.0075	0.0039	0.0037
Max inbreeding coeff.	0.0376	0.0625	0.0625
Min inbreeding coeff.	0.0039	0.0078	0.0039
Inbreeding coeff. (*f*) distribution:			
0.00 < f < = 0.05	-	-	23
0.05 < f < = 0.10	-	-	3
0.10 < f < = 1.00	-	-	0
Longest ancestral path (LAP): 0	-	-	60
1 to 5	-	-	27
6 to 10	-	-	104
11 to 12	-	-	5
Pairs[2]:	Concordant Aff.	Concordant Unaff.	
Parent/Off	0	201	270
Sib/Sib	14	204	218
Grandparents	0	228	318
Avuncular	0	438	438
Half Sib	0	12	12
Cousin	0	912	912

(1) Includes missing/unknown data; (2) Pairs of relatives that show concordant or discordant phenotype (affected or unaffected).

First, in the "Setup Dialog" window, select "Create new project", and choose a project name. In the next step, select the option "I have all data required by S.A.G.E. pedigree but not parameter file". Browse the directory to where the input file is located in your computer, select the format file ("tab" and "single" in the File S3), and import the data file. The data file is now available as S.A.G.E. internal data (see folder "Data" on the left) and can be used to run PEDINFO and other S.A.G.E. programs such as SEGREG for complex segregation analysis or the LODLINK and MLOD programs for model-based linkage analysis.

To run PEDINFO, in the menu window, select Analysis > Summary Statistics > PEDINFO, and then complete the directory path of data file and the name of output file. In "Analysis Definition", select whether you want statistics for a particular trait or covariate and/or statistics for each pedigree. Run it. An output folder will be created in the left-hand section of the windows that includes the output files with the results of the analysis. Supplementary material File S4 shows the PEDINFO results for the same pedigree analyzed as an example.

Another user-friendly program used in population genetics for pedigree analysis is CFC (Colleau, 2002), which can provide general information on the structure of pedigrees, check for pedigree errors, extract a list of ancestors and descendants, draw the paths that connect relatives of inbred individuals, and compute indicators such as inbreeding coefficients, coefficient of relationships, ancestral decomposition of inbreeding coefficients, probabilities of gene origin, etc.

When starting to use CFC, a pedigree input file should be formatted as shown in Supplementary Material File S5. To open the pedigree file, select File > Open, and select the input file. The pedigree file will be read, checked for errors, and saved in "pgd" extension.

It is possible to obtain a brief structure of the pedigree, including the distribution according to the inbreeding coefficients, the longest ancestral path (LAP), number of founders, number of inbreeds, average of inbreeding coefficient, etc., by selecting Tools > Pedigree Structure.

The results of these analyses can be exported as a .txt file, and the output file for the example analyzed is shown in Supplementary Material File S6. All other analyses can be run from the "Tools" menu.

Others sophisticated statistical packages for pedigree analysis are available, including Genetic Analysis Package (GAP) (Zhao, 2007), PedHunter (Agarwala *et al*., 1998), or Madeline (Trager *et al*., 2007), but the description of the mandatory formats for input files and the arguments to execute the specific commands are beyond the aims of this manuscript.

We hope that this fieldwork guide for collecting, handling and analyzing genealogical data will help to design research projects in population medical genetics and promote the study of isolated populations in Latin America.

Ethics for field work

As discussed before (Section Illegitimacy, confidentiality and other ethical issues), the right of a proband to confidentiality conflicts with the carrier's right to be informed when seeking prospective counseling. Discussions and recommendations on these issues can be found in the WHO report of a consultation on community genetics in low- and middle-income countries, section 7.2 Confidentiality issues, (WHO, 2010, pp 16).

Concerning the use of written genealogical sources such as churches, civil, electorate, military, etc., Brazilian legislation (Resolução CNS 466/2012) does not require IRB approval for data obtained from public databases.

Acknowledgments

This project was supported by grants from INCT-INAGEMP; Ministry of Science and Technology/CNPq, Brazil, grant no. 573993/2008; CNPq, Brazil, grants # 402045/2010-6, 481069/2012-7, 306396/2013-0; CNPq, Programa Ciência sem Fronteiras (CSF), modalidade AJT, Brazil, processo # 370799/2013-5; FAPERJ, Brazil grant # E-26/102.797/2012; CONICET, National Research Council of Argentina; ANPCyT, Argentina, grant # PICT 2010-2798. No funding bodies had any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Agarwala R, Biesecker LG, Hopkins KA, Francomano CA and Schaffer AA (1998) Software for constructing and verifying pedigrees within large genealogies and an application to the Old Order Amish of Lancaster County. Genome Res 8:211-221.

Alvarez-Gardeazábal G (1986) El Divino, Plaza & Janés. Bogotá, 233 pp.

Bailliet G, Castilla EE, Adams JP, Orioli I, Martínez-Marignac V, Richard SM and Bianchi N (2001) Correlation between molecular and conventional genealogies in Aicuña: A rural population from Northwestern Argentina. Hum Hered 51:150-159.

Bennett RL, French KS, Resta RG and Doyle DL (2008) Standardized human pedigree nomenclature: Update and assessment of the recommendations of the National Society of Genetic Counselors. J Genet Counsel 17:424-433.

Bronberg RA, Dipierri JE, Alfaro EL, Barrai I, Rodriguez-Larralde A, Castilla EE, Colonna V, Rodriguez-Arroyo G and Bailliet G (2009) Isonymy structure of Buenos Aires city. Hum Biol 81:447-461.

Brun-Samarcq L, Gallina S, Philippi A, Demenais F, Vaysseix G and Barillot E (1999) CoPE: A collaborative pedigree drawing environment. Bioinformatics 15:345-346.

Castilla EE and Adams J (1990) Migration and genetic structure in an isolated population in Argentina: Aicuña. In: Adams JP (ed) Proceeding of Convergent Issues in Genetics and Demography. Oxford University Press, New York, pp 45-62.

Castilla EE and Adams J (1996) Genealogical information and the structure of rural Latin-American populations: Reality and fantasy. Hum Hered 46:241-255.

Colleau JJ (2002) An indirect approach to the extensive calculation of relationship coefficients. Genet Sel Evol 34:409-422.

Chapman CJ (1990) A visual interface to computer programs for linkage analysis. Am J Med Genet 36:155-160.

Davie AM (1979) The 'singles' method for segregation analysis under incomplete ascertainment. Ann Hum Genet 42:507-512.

Emery A (1986) Methodology in Medical Genetics: An Introduction to Statistical Methods. Churchill Livingstone, London, 197 pp.

Falconer D and Mackay T (1996) Introduction to Quantitative Genetics. Longman, London, New York, 480 pp.

García Márquez G (1973) Cien Años de Soledad. Editorial Sudamericana, Buenos Aires.

Grass GW (1999) The Tin Drum. Pantheon, New York.

Greven PJ (1972) Four Generations: Population, Land and Family in Colonial Andover, Massachusetts. Cornell University Press, Ithaca, 349 pp.

Jeambrun P and Sergent B (1991) Les Enfants de la Lune: L'albinisme chez les Amérindiens. IRD Editions, Montpellier, 351 pp.

Karunaratne PM and Elston RC (1998) A multivariate logistic model (MLM) for analyzing binary family data. Am J Med Genet 76:428-437.

Khoury MJ, Botto L, Waters GD, Mastroiacovo P, Castilla E and Erickson JD (1993) Monitoring for new multiple congenital anomalies in the search for human teratogens. Am J Med Genet 46:460-466.

Kirichenko A (2004) An algorithm of step-by-step pedigree drawing. Russ J Genet 40:1176-1178.

Lalouel JM and Morton NE (1981) Complex segregation analysis with pointers. Hum Hered 31:312-321.

Lalouel JM, Rao DC, Morton NE and Elston RC (1983) A unified model for complex segregation analysis. Am J Hum Genet 35:816-826.

MacLean C (1969) Computer analysis of pedigree data. In: Morton NE (ed) Computer Applications in Genetics. University of Hawaii Press, Honolulu, pp 82-86.

Malécot G (1948) Les Mathématiques de l'Hérédité. Masson, Paris, 63 pp.

Mamelka P, Dyke B and MacCluerJ (1987) Pedigree/Draw for the Apple Macintosh. Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio.

Manzoli GN, Abe-Sandes K, Bittles AH, Da Silva DS, Fernandes LDC, Paulon R, De Castro ICS, Padovani CM and Acosta AX (2013) Non-syndromic hearing impairment in a multiethnic population of Northeastern Brazil. Int J Pediatr Otorhinolaryngol 77:1011-1082.

McCaa R (1983) Marriage and fertility in Chile: Demographic turning points in the Petorca Valley, 1840-1976. Westview Press, Boulder, 207 pp.

Porta M (2008) A Dictionary of Epidemiology. 5th edition. Oxford University Press, New York, 316 pp.

Reich T, James JW and Morris CA (1972) The use of multiple thresholds in determining the mode of transmission of semi-continuous traits. Ann Hum Genet, 36:163-184.

Risch N (1990) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222-2228.

Royo-Marín A (1958) Teología Moral para Seglares. Tomo II: Los Sacramentos, Biblioteca de Autores Cristianos, Madrid, 731 pp.

Sinnwell J and Atkinson B (2013) Kinship2 package vignette Version 1.5. 0.

Thiele H and Nürnberg P (2005) HaploPainter: A tool for drawing pedigrees with complex haplotypes. Bioinformatics 21:1730-1732.

Thompson E (1978) Ancestral inference II. The founders of Tristan da Cunha. Ann Hum Genet 42:239-253.

Trager EH, Khanna R, Marrs A, Siden L, Branham KE, Swaroop A and Richards JE (2007) Madeline 2.0 PDE: A new program for local and web-based pedigree drawing. Bioinformatics 23:1854-1856.

Vargas DM, Arena LFGL and Soncini AS (2010) Secular trends in stature growth in Blumenau-Brazil in relation to human development index (HDI). Rev Assoc Med Bras 56:304-308.

Veríssimo É (1984) O Tempo e O Vento. Editora Globo, Porto Alegre.

WHO (2010) World Health Organization: Community genetics services: Report of a WHO consultation on community genetics in low-and middle-income countries. World Health Organization, Geneva.

Wright S (1922) Coefficients of inbreeding and relationship. Am Nat 56:330-338.

Wright S (1934) On the method of path coef?cients. Ann Math Stat 5:161-215.

Zhao JH (2007) gap: Genetic analysis package. J Stat Softw 23:1-18.

Internet Resources

S.A.G.E. (2012) Statistical Analysis for Genetic Epidemiology. Release 6.3. http://darwin.cwru.edu/.

Supplementary material

The following online material is available for this article:

- Figure S1 - Graphical pedigree with 196 individuals corresponding to the tabulated genealogy in File S1.
- Figure S2 - Graphical pedigree that is the output file of *kinship2* (R package).
- File S1 - Tabulated genealogy of subpedigree with 196 individuals extracted from the complete genealogy of Aicuña (8,696 individuals).
- File S2 - Input file format for drawing pedigrees using *kinship2* (R package).
- File S3 - Input file format for pedigree analysis using S.A.G.E.
- File S4 - Output file of PEDINFO program included in S.A.G.E.
- File S5 - Input file format for pedigree analysis using CFC.
- File S6 - Output file with the "Pedigree structure" using CFC software.