Titre: Title:	Layered founders : a novel approach to investigate the ancestral transmission of complex trait
Auteurs: Authors:	Ettore Merlo, Benoit Deslauriers, Giuliano Antoniol, Pierre-Luc Brunelle, Michèle Jomphe, Gérard Bouchard, Johanne Tremblay and Pavel Hamet
Date:	2004
Type:	Rapport / Report
Référence: Citation:	Merlo, Ettore, Deslauriers, Benoit, Antoniol, Giuliano, Brunelle, Pierre-Luc, Jomphe, Michèle, Bouchard, Gérard, Tremblay, Johanne et Hamet, Pavel (2004). Layered founders : a novel approach to investigate the ancestral transmission of complex traits. Rapport technique. EPM-RT-2004-05.

EPM–RT–2004-05

LAYARED FOUNDERS : A NOVEL APPROACH TO INGESTIGATE THE ANCESTRAL TRANSMISSION OF COMPLEX TRAITS

Ettore Merlo, Benoit Deslauriers, Giuliano Antoniol, Piere-Luc Brunelle, Michèle Jomphe, Gérard Bouchard, Johanne Tremblay and Pavel Hamet
Département de Génie informatique
École Polytechnique de Montréal

Juin 2004

Poly

ÉCOLE POLYTECHNIQUE MONTRÉAL

EMP-RT-2004-05

Layered Founders : A Novel Approach to Investigate the Ancestral Transmission of Complex Traits

Ettore Merlo(1), Benoit Deslauriers(2), Giuliano Antoniol(3), Pierre-Luc Brunelle(2), Michèle Jomphe(4), Gérard Bouchard(4), Johanne Tremblay(2) and Pavel Hamet(2)

1Département de génie informatique, École Polytechnique de Montréal, 2Centre hospitalier de l'Université de Montréal (CHUM), 3Université de Sannio, 4Université de Québec à Chicoutimi.

Juin 2004

Dépôt légal :
Bibliothèque nationale du Québec, 2003
Bibliothèque nationale du Canada, 2003

EPM-RT-2004-05
Layered Founders: A Novel Approach to Investigate the Ancestral Transmission of Complex Traits
par : Ettore Merlo[(1)], Benoit Deslauriers[(2)], Giuliano Antonio[(3)], Pierre-Luc Brunelle[(2)], Michèle Jomphe[(4)], Gérard Bouchard[(4)], Johanne Tremblay[(2)] et Paval Hamet[(2)]
[1]Département de génie informatique, École Polytechnique de Montréal, [2]Centre hospitalier de l'Université de Montréal (CHUM), [3]Université de Sannio, [4]Université de Québec à Chicoutimi

Abstract (technical report)

A novel approach based on graph theory is presented to reason about the genetic contribution of ancestors at different genealogical distances from today's individuals (different definitions of layers and distances are proposed and discussed). It allows the maximum likelihood classification of \textit{specific founders} who predominantly contribute to one class of individuals and the analysis of \textit{separability} of specific founders with respect to two classes of individuals that have been selected based on LOD (logarithm of odds) score determined by a total genome scan and on ScaI marker genotype of a candidate gene of hypertension, ANP. Several experiments have been performed on a genealogy comprising more than 40,000 people and spanning 17 generations from the Saguenay-Lac-Saint-Jean population. We have computed: the founders obtained by using different definitions of layers and distances, the contribution of specific and unique founders, and the separability of specific founders. The results indicate that most definitions of layers of founders show a similar trend over layers for size and content, that specific and unique genetic contributions are very high for recent generations and, as expected, decrease for older generations, and, also, that separability is higher for recent generations than for older ones. The presented approach allows a much finer grain analysis of genetic contribution of founders than previously-reported approaches.

Layered Founders: A Novel Approach to Investigate the Ancestral Transmission of Complex Traits

Ettore Merlo[1], Benoit Deslauriers[2], Giuliano Antoniol[3], Pierre-Luc Brunelle[2], Michèle Jomphe[4], Gérard Bouchard[4], Johanne Tremblay[2] and Pavel Hamet[2]

[1]Department of Electrical and Computer Engineering, École Polytechnique de Montréal, P.O. Box 6079, Downtown Station, Montreal, Quebec, H3C 3A7, Canada
e-mail: ettore.merlo@polymtl.ca

[2] Centre hospitalier de l'Université de Montréal (CHUM) [3] University of Sannio
[4] Université du Québec à Chicoutimi

ABSTRACT

A novel approach based on graph theory is presented to reason about the genetic contribution of ancestors at different genealogical distances from today's individuals (different definition of layers and distances are proposed and discussed). It allows the maximum likelihood classificatio of *specifi founders* who predominantly contribute to one class of individuals and the analysis of *separability* of specifi founders with respect to two classes of individuals that have been selected based on LOD (logarithm of odds) score determined by a total genome scan and on ScaI marker genotype of a candidate gene of hypertension, ANP.

Several experiments have been performed on a genealogy comprising more than 40,000 people and spanning 17 generations from the Saguenay-Lac-Saint-Jean population. We have computed: the founders obtained by using different definition of layers and distances, the contribution of specifi and unique founders, and the separability of specifi founders.

The results indicate that most definition of layers of founders show a similar trend over layers for size and content, that specifi and unique genetic contributions are very high for recent generations and, as expected, decrease for older generations, and, also, that separability is higher for recent generations than for older ones.

The presented approach allows a much fine grain analysis of genetic contribution of founders than previously-reported approaches.

1 Introduction

The Saguenay-Lac-Saint-Jean (SLSJ) region (Province of Quebec, Canada) is located on the north shore of the St. Lawrence River, about 200 km north-east of Quebec City. Its territory covers some 11,000 km^2. Settlement began in this region in the mid-1800's, originating mostly from the relatively small border region called Charlevoix; from 1840 to 1870, 80% of the Saguenay settlers were born in Charlevoix [1]. Historically and genetically speaking, these two regions have maintained a close relationship. Even today, nearly 90% of individuals of the Saguenay population (which approaches 300,000 inhabitants) born between 1950 and 1970 have these firs settlers as ancestors [2]. Both populations are characterized by a relatively high frequency of some rare hereditary diseases, mainly recessive ones [3].

In the historical context of the SLSJ population, we have to look at its formation from the Charlevoix population. There are at least two reasons to be interested in the formation of this population. First, the Charlevoix population is at the base of the foundation of the population of the Saguenay in the XIX century. The second reason is the presence of specifi genetic diseases common to these regions. We have to go back to the history of the Charlevoix population to be able to understand the situation that prevails today in SLSJ.

Charlevoix lies on the north shore of the St. Lawrence River, approximately 100 km north-east of Quebec City. This region covers 5,700 km^2, but habitable space is much less than that. The population is concentrated on the edge of the coast which is 100 km long by approximately 10-25 km wide. The firs founders of the Charlevoix region came from the population of Quebec City. At the end of the XVII

century, there were 200 people. In 1831, the population of Charlevoix was more than 8,000 people, and demographic pressure began to build up. This pressure started a migration from Charlevoix to the Saguenay. The genetic pool that we observe today in the Charlevoix and the SLSJ regions came from a relatively small number of faraway ancestors who produced a founder effect [4, 5]. By definition a founder effect designates a migration phenomenon from a mother population which settled in a territory where they can reproduce themselves and give birth to a new population. This founder effect produces genetic homogeneity at least at some loci. We can say that the founder effect of the population of Charlevoix is a result of two factors: the firs is the relatively low immigration that followed settlement of the population, and the second is the selective geographic origin of the founders of Charlevoix. More than half of the nucleus of founders were united at the firs degree level (sister and brother), and 45% of this nucleus came from the same region at the border of the old province of Perch and Maine in France [3]. Thus, it is possible to say that the present genetic pool of Charlevoix was formed mainly from a nucleus of ancestors who came from these regions of France.

All these facts are concordant with the thesis of an important founder effect. The present population of Charlevoix would have been generated by a relatively small number of ancestors who would have transmitted, to their descendants, some mutated genes [2].

It was in 1838 that the migration started from Charlevoix to Saguenay. Family analysis from the region's historical register showed that until 1870, more than 80% of migration to the Saguenay came from the Charlevoix region [6]. In fact, despite its 280,000 inhabitants, the population of the Saguenay distinguishes itself by cultural homogeneity. We can see some evidence of this genetic homogeneity (for example, the high incidence of some uncommon genopathies). Because the Saguenay population was created by a migration stream that came mostly from Charlevoix, we can specificall talk about a founder effect, at least in a broad sense. Nevertheless, data on the establishment and reproduction of pioneers from Charlevoix showed that a kind of multiplier social effect was added to the founder effect, because the founders who came from Charlevoix profite by some economical and social conditions which let them reproduce much faster and diffuse their genes much more than other pioneers [3].

2 Research Context

For more than 20 years, the studies of the SLSJ population enriched our knowledge on genetic causes of monogenic disorders [7]. Most of these studies were conducted on recessive diseases. The high frequency of some recessive diseases in this population has been attributed to high fertility rate and founder effect [2, 7].

This population derives from three migration waves. The firs was the migration from France to "Nouvelle France" in the XVII and XVIII centuries. The second was the sub-population that migrated from the region of Quebec to Charlevoix at the beginning of the XVIII century. The third wave started in 1840 and was generated by a migration process from Charlevoix settling into the SLSJ region [1, 5].

More recently, researchers began to investigate, with this population, multifactorial diseases like Alzheimer's, manic depression and lymphomas [8, 9].

A founder effect has already been suggested for several monogenic diseases in the SLSJ population [4, 5]. Our group investigated, in the SLSJ population, the genetic determinants of hypertension, a highly multifactorial and polygenic disease [10], for which the classical "founder effect" is difficul to demonstrate. The present study explores the concept of founder effect further by determining separable sets of founders most likely responsible for genetic differences of separable classes of descendants. This approach allowed us to develop the concept of "quantitative founder effect", which we have applied in a parallel study on metabolic components of hypertension [11]. Here, we report the equations behind this quantitative founder effect in complex traits using a candidate gene of hypertension and obesity, the atrial natriuretic peptide (ANP), and microsatellite markers from chromosomes 1 and 3 [11].

2.1 Candidate Gene Approach

ANP (atrial natriuretic peptide) [12, 13] is a member of the natriuretic peptide family involved in blood pressure regulation and volume homeostasis [14–16]. Its receptors are highly expressed in the vasculature, the kidney glomeruli and the adrenal cortex [17]. ANP receptors have been also identifie in adipocytes and a lipolytic pathway involving natriuretic peptides has recently been discovered in human fat cells [18]. The ANP gene is located on human chromosome 1 at position 1p36.2. We have genotyped the ANP gene in 696 subjects using two markers, ScaI and BstXI, and performed linkage analyses with SIBPAL software from the S.A.G.E. package in normotensive and hypertensive sibpairs from the SLSJ population. We have observed highly significan linkage of the ANP locus to many phenotypes related to blood pressure, sodium excretion and fat distribution [19]. ANOVA analyses suggested a strong genotypic effect of the ANP gene on wake and sleep systolic blood pressure and on sodium excretion in hypertensive individuals. The genotypic effect of the ANP gene on systolic and diastolic blood pressure is dependent on body mass index (BMI) in both normotensive and hypertensive individuals [19], suggesting an interaction between hyper-

tension and obesity.

For the BstXI marker (restriction sites in the firs intron and second exon of the ANP gene), we performed polymerase chain reaction (PCR) of 35 cycles (5 sec at 94^{o}C, 5 sec at 61^{o}C and 10 sec at 72^{o}C) with sense primer AGA CAG AGC AGC AAG CAG TG and antisense primer CAT TCC ATC CCC AGT TCC with an initial denaturation of 5 min at 94^{o}C and fina extension of 7 min at 72^{o}C, followed by BstXI digestion (2 hrs at 45^{o}C) with 1U of enzyme in a volume of 10 μL. For the ScaI marker (located in exon 3 caused by a mutation of the firs nucleotide of the stop codon), 30 cycles of PCR (5 sec at 94^{o}C, 5 sec at 63^{o}C, 5 sec at 72^{o}C), with an initial denaturation of 5 min at 94^{o}C and fina extension of 7 min at 72^{o}C, were performed with sense GGC ACA CTC ATA CAT GAA GCT TTT T and antisense primer GCA GTC TGT CCC TAG GCC CA, followed by ScaI digestion (2 hrs at 37^{o}C) with 5U of enzyme in a 10 μL sample reaction. PCR and enzymatic digestion were followed by electrophoresis on agarose gel and the genotype of each sample was visualized under UV light.

2.2 Total Genome Scan Genotyping

The DNA of 500 subjects was genotyped at the Broad Institute, using 377 microsatellite markers with a modifie version of the Cooperative Human Linkage Center Screening Set, version 6.0, that included Genethon markers for an average of 9.1 cM coverage of the entire human genome [20]. Additional markers (Genethon) at a 5cM density were used on chromosome 1q and 3q.

3 Computational Method

Motivation for the present study stemmed from the desire to explicitly take into consideration several factors that influenc the analysis of founders in an isolated population and that have not been explicitly addressed in past literature.

Therefore, we have developed original equations based on graph theory [21] that help better defin and investigate the founder effect.

We wanted to take into consideration the shape of the genealogy – whatever it was. We discovered that the shape of the frontier of the genealogy was quite irregular. In the ascending direction, founders without recorded parents already appeared three generations ago from today's individuals. Others had a genealogical depth of up to 17 generations.

It seems counter-intuitive to associate both types of founders with the phenomenon of a few groups of immigrants coming from France two centuries ago. Possibly, information incompleteness in the databases may also help to explain such diversity in the length of genealogical paths from today's individuals.

This difference in the length of genealogical paths has an impact on the computation of a founder's genetic contribution, which exponentially decreases with genealogical path length. If the number of paths from founders are comparable, we would tend to attribute more genetic weight to a founder closer to today's individuals.

We would expect the number of paths from a distant founder to be exponentially larger than that of closer founders, so that comparability is preserved with respect to distance of founders from today's individuals. Unfortunately, we fin individuals with no children at several different genealogical levels that are not those of today's individuals. We are currently unable to distinguish between individuals who really did not have children in their lives and missing information from the databases.

These arguments suggest the possible existence of some sort of numerical distortion in the application of conventional genetic contribution equations and founder effect approaches caused by the available genealogical data.

Another dimension to be taken into consideration is the criss-crossing of genealogical paths in time. The often misleading intuition is to think of genealogies in terms of trees. This is definitel not true in the investigated population, in which genealogical paths indeed split at every generation, but very often merged again some generations later.

Interestingly enough, there is a difference between real time measured in time units like years, and logical genealogical time measured in generations or in number of path splits. It seems that along some genealogical subpaths, people get married and have children at a faster pace than along some other later merging paths. The counter-intuitive result is that if we take two distinct grandparents at exactly n generations ago from today's individual, they may happen to be one child of the other, despite the identical number of generations separating them from today's individual. An example is depicted in Figure 1.

For this purpose, a graph theoretical approach [22] has been followed to be able to rigorously reason about the properties of gene transmission probability along genealogical paths. Graph-based approaches are also used in other areas of biology [23–26].

4 Genealogical Graph

Genealogical information about populations can be represented in graph theory by a directed graph in which nodes represent individuals from a $POPULATION$ set and directed edges connecting two individuals represent parental relationships. In particular, since no cycle is present in a genealogical graph because it is impossible to have children who are at the same time one's ancestors, genealogical graphs are indeed directed acyclic graphs (DAG).

In general, a graph G is composed of a set V_G of vertices

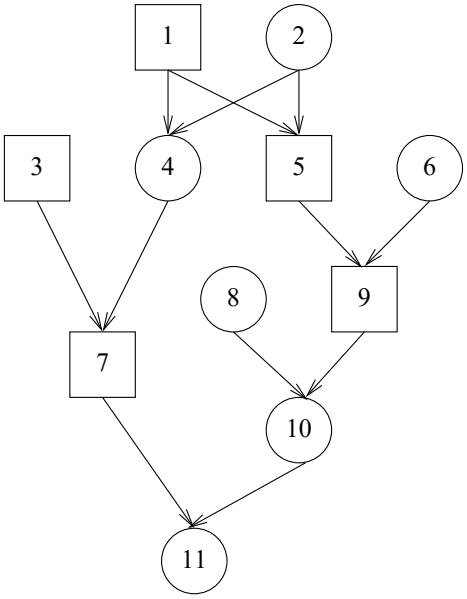

Figure 1. Ancestors reached 3 generations ago are in fact parent and child. Consider individuals 1, 5 and 11. There are two paths from 1 to 11, namely path (1, 4, 7, 11) shown on the left hand side of the genealogical graph and path (1, 5, 9, 10, 11) on the right hand side. Hence, both individuals 1 and 5 can be reached 3 generations ago from 11, and 5 is the child of 1. Males are represented by squares, females by circles.

(nodes) and a set E_G of edges. In our interpretation, nodes correspond to individuals, and edges correspond to mother and father relations to children, as follows:

$$
\begin{gathered}
G = (V_G,\ E_G) \\
V_G = POPULATION \\
E_G = \{(v_i,\ v_j) \mid \\
(v_i,\ v_j \in V_G) \wedge \\
((v_i = father(v_j)) \vee \\
(v_i = mother(v_j)))\}
\end{gathered}
\tag{1}
$$

A graph representation of genealogies can easily be constructed from the parental and genealogical information contained in tabular form. Let T be a table which carries genealogical information about parents and children as follows:

$$
T\ =\ (entry_1,\ \ldots,\ entry_i,\ \ldots,\ entry_n)
$$

$$
\begin{gathered}
entry_1\ =\ (c_{1,1},\ \ldots,\ c_{1,j},\ \ldots,\ c_{1,m}) \\
\ldots \\
entry_i\ =\ (c_{i,1},\ \ldots,\ c_{i,j},\ \ldots,\ c_{i,m}) \\
\ldots \\
entry_n\ =\ (c_{n,1},\ \ldots,\ c_{n,j},\ \ldots,\ c_{n,m})
\end{gathered}
\tag{2}
$$

where some columns represent the mother's and father's relations:

$$
\begin{gathered}
c_{i,r} = child\_id_i \\
c_{i,s} = mother\_id_i \\
c_{i,t} = father\_id_i
\end{gathered}
\tag{3}
$$

Nodes and edges in the genealogical graph can easily be constructed from T as follows:

$$
\begin{gathered}
V_G\ =\ \{v \mid (\exists entry_i \in T \mid \\
(ID(v)\ =\ child\_id_i) \vee \\
(ID(v)\ =\ mother\_id_i) \vee \\
(ID(v)\ =\ father\_id_i)\ \}
\end{gathered}
$$

$$
\begin{gathered}
E_G\ =\ \{\ (v_1,\ v_2) \mid \\
(v_1 \in V_G) \wedge (v_2 \in V_G) \wedge \\
(\exists entry_j \in T \mid \\
(ID(v_2)\ =\ child\_id_j) \wedge \\
((ID(v_1)\ =\ mother\_id_j) \vee \\
(ID(v_1)\ =\ father\_id_j))\ \}
\end{gathered}
\tag{4}
$$

where $ID(v)$ is the identifie in T associated with node v.

4.1 Genealogical Graph Features

The genealogical graph, obtained from data on the SLSJ population, presents the characteristics indicated in Table 1. Part of these genealogical data have been obtained from the BALSAC database [27].

Number of nodes	43,605
Number of edges	60,866
Number of sources	5,125
Number of sinks	16,099
Maximum depth	17
Disk space (bytes)	1,661,297

Table 1. Genealogical Graph Features

Motivation for the present study can be re-interpreted in graph theory as follows. Topological features of genealogical graphs represent genealogical concepts. For example, *founders* correspond to graph *sources*, which are define as those nodes whose $indegree$ is zero. Graph *sinks* are those nodes whose $outdegree$ is zero and they correspond to the most updated current genealogical information. Today's individuals are in general close to sinks, many are indeed sinks, and so on.

As mentioned in Section 1, sources appear at different distances from sinks, some of them even appear as early as three edges in the reverse direction from sinks.

In this perspective, the set of sources has a fairly irregular shape which has an impact on the computation of founders separability with respect to descendant characteristics.

Also, if we take two nodes $v,\ w \in V_G$, where v is a parent of w, i.e. $(v,\ w) \in E_G$, paths may exist such that their length d between v and today's individual z, who is often a sink, is also observed between some children w of v and z.

5 Layered Approach

We have define the *layered founders* approach to overcome and normalize the above-mentioned possible distortions in the computation of founders. The underlying idea is to compare and reason about founders on the basis of a common distance from today's reference individuals. Intuitively, we name the set of founders who share a common distance from a reference set of individuals as a *layer* of founders in analogy to the geometric image of the concept. Ancestors who share some common feature related to the genealogical distance with respect to candidate individuals are put in the same ancestral layer. In effect, in a genealogy, we will defin several genealogical layers corresponding to the different values of genealogical distance from reference individuals. The firs layer is the set of parents of reference individuals, the second layer is that of grandparents and so on, while information is available.

Since sources appear at different levels in the genealogy, and parents and children nodes may reach a reference individual through a path of the same length, the formal defini tions of layers and genealogical distances have to be carefully conceived to overcome the mentioned limitations.

In the following sub-sections, we will formally defin distances and layers and reason about their usage to compute founders and their separability with respect to today's individuals' features.

5.1 Genealogical Paths

The set of all paths in the genealogical graph G from v to all individuals in S, which can be reference individuals, can be define as follows:

$$paths:\ V \times \mathcal{P}(V) \rightarrow V \cup V^2 \cup \ldots \cup V^n, n \in \mathcal{N}^+$$

$$\begin{gathered} paths(v,\ S) = \{p = (w_0,\ w_1, \ldots, w_n) \mid \\ (\forall i \in [0,\ n], w_i \in V_G) \wedge \\ (w_0\ =\ v) \wedge \\ (w_n \in\ S) \wedge \\ (\forall j \in [1,\ n], (w_{j-1},\ w_j) \in E_G)\} \end{gathered} \tag{5}$$

$paths(v,\ S)$ is the set of any sequence of any finit integer length n of edges in the genealogical graph G, which starts by v and ends in node w_n belonging to S.

Other additional path-related functions can be define for ease of later definitio of layers. They are the *Minimum Genealogical Path Length*, and the *Average Genealogical Path Length*. In formulae:

$$mgpl:\ V \times \mathcal{P}(V) \rightarrow \mathcal{N}_0^+$$

$$mgpl(v,\ S)\ = \begin{cases} \min\limits_{p \in\ paths(v,\ S)} \mid p \mid & if\ paths(v,S) \neq \emptyset \\ undefined & otherwise \end{cases} \tag{6}$$

$mgpl(v,\ S)$ is the length of the shortest genealogical paths between an ancestor v and a set S of descendants, if such paths exist.

$$agpl : V \times \mathcal{P}(V) \to \mathcal{R}$$

$$agpl(v,\ S)\ =$$

$$= \begin{cases} \dfrac{\sum\limits_{p \in paths(v,\ S)} |p|}{|paths(v,\ S)|} & if\ paths(v,S) \neq \emptyset \\ undefined & otherwise \end{cases} \tag{7}$$

$agpl(v,\ S)$ is the average length of genealogical paths between an ancestor v and a set S of descendants, if such paths exist.

Several definition of layers can be conceived on the basis of the notion of genealogical path lengths. We have originally define the layers of *Distance-Based*, *Shortest Distance*, and *Average Distance* founders.

5.2 Distance-Based Layered Founders

$$L : \mathcal{P}(V) \times \mathcal{N}_0^+ \to \mathcal{P}(V)$$

$$\begin{aligned} L(d,\ S)\ = \{v \in V \mid (\exists p \in paths(v,S),\ \ |p| = d)\ \} \cup \\ \{w \in V \mid (\nexists\ (u,\ w) \in E) \wedge \\ (\forall p \in paths(w,S),\ \ |p| < d)\ \} \end{aligned} \tag{8}$$

$L(d,\ S)$ is the set of ancestors v such that the length of some genealogical paths between v and any individuals in S is exactly d together with the set of ancestors w who do not have ascendants in the genealogical graph and whose genealogical paths are all smaller than d in length.

5.3 Shortest Distance Layered Founders

$$L : \mathcal{P}(V) \times \mathcal{N}_0^+ \to \mathcal{P}(V)$$

$$\begin{aligned} L(d,\ S)\ = \{v \in V \mid mgpl(v,\ S)\ =\ d\ \} \cup \\ \{w \in V \mid (\nexists\ (u,\ w) \in E) \wedge \\ (mgpl(w,\ S) < d)\ \} \end{aligned} \tag{9}$$

$L(d,\ S)$ is the set of ancestors v such that the length of the shortest genealogical path between v and individuals in S is exactly d together with the set of ancestors w who do not have ascendants in the genealogical graph and whose shortest genealogical paths are all smaller than d in length.

5.4 Average Distance Layered Founders

$$L : \mathcal{P}(V) \times \mathcal{N}_0^+ \to \mathcal{P}(V)$$

$$\begin{aligned} L(d,\ S)\ = \\ = \left\{v \in V \mid \left(d - \frac{1}{2}\right) \leq agpl(v,\ S) < \left(d + \frac{1}{2}\right)\right\} \cup \\ \left\{w \in V \mid (\nexists\ (u,\ w) \in E)\ \wedge agpl(w,\ S) < \left(d - \frac{1}{2}\right)\right\} \end{aligned} \tag{10}$$

$L(d,\ S)$ is the set of ancestors v such that the length of the average genealogical path between v and individuals in S is exactly d together with the set of ancestors w who do not have ascendants in the genealogical graph and whose average genealogical paths are all smaller than d in length.

5.5 Constrained Layers

$$c\_layer : \mathcal{P}(V) \to \mathcal{P}(V)$$

$$\begin{aligned} c\_layer(L) = \{w \in L \mid \\ (\forall v \in L, \\ (\nexists p = (w_0,\ w_1, \ldots, w_n) \mid \\ (w_0\ =\ v) \wedge \\ (w_n\ =\ w) \wedge \\ (\forall j \in [1,\ n], \\ (w_{j-1},\ w_j) \in E_G))) \} \end{aligned} \tag{11}$$

Constrained layers are composed of those founders who are not a descendant of another founder in the same layer. Constrained layers contain founders whose sharing probabilities with descendants in classes S_A and S_B are independent.

6 Genetic Contribution of Layered Founders

Let us reason about the genetic contribution of ancestors that belong to some layers. The genetic contribution of an ancestor v to a set of descendants S, as reported in [9], is:

$$gc : V \times \mathcal{P}(V) \to [0,\ 1] \cap \mathcal{R}$$

$$gc\ (v,\ S)\ = \begin{cases} \sum\limits_{p \in paths(v,S)} \left(\frac{1}{2}\right)^{|p|} & if\ paths(v,S) \neq \emptyset \\ 0 & otherwise \end{cases} \tag{12}$$

The probability that a descendant belonging to S shares the same gene than ancestor v belonging to layer L_k can be computed from the genetic contribution gc, provided that layer founders probabilities are statistically independent. This probability is called the *identical by descent* (IBD) probability and is define as follows:

$$idbp: V \times \mathcal{P}^2(V) \rightarrow [0,\ 1] \cap \mathcal{R}$$

$$idbp(v,\ S,\ L_k) = \frac{gc(v,\ S)}{\sum\limits_{w \in L_k} gc(w,\ S)} \tag{13}$$

The gene-sharing probability can also be computed on a set of founders:

$$idbp: \mathcal{P}^3(V) \rightarrow (0,\ 1] \cap \mathcal{R}$$

$$idbp(F,\ S,\ L_k) = \sum_{f \in F} ibdp(f,\ S,\ L_k) \tag{14}$$

$idbp(F,\ S,\ L_k)$ represents the probability that a descendant belonging to S shares genes with any ancestor v belonging to F included in L_k. In other words, it represents the probability that descendant v_1 shares genes with founder f_1 or f_2 or ... $f_n \in F$ rather than with other founders in L_k, or that v_2 does that, and so on.

Gene-sharing probability can be computed for the two sets of reference individuals S_A and S_B:

$$idbp: \mathcal{P}^3(V) \rightarrow \mathcal{P}(\mathcal{R}^2)$$

$$idbp(S_A,\ S_B,\ L_k) = \{(x,\ y) \mid (\forall\ f \in\ L_k, \\ (x\ = idbp(f,\ S_A,\ L_k)) \wedge (y\ = idbp(f,\ S_B,\ L_k)))\ \} \tag{15}$$

Elements of $idbp(S_A,\ S_B,\ d)$ correspond to founders f in layer L_k and represent the probability that an individual respectively in S_A and S_B shares genes with a founder f.

6.1 Specificit of Founders

Not all founders in a given layer equally contribute to both classes of reference individuals. Some founders contribute to one class only, while some others' contribution is very similar. Individuals contributing to one class only are aligned on the diagram axes (x for class S_A and y for class S_B), while founders supplying a similar contribution are closer to the diagram diagonal through the origin. To characterize founders contribution to the two classes, we have define the concept of *specificit of founders* as follows:

$$SP: V \times \mathcal{P}^3(V) \rightarrow [0,\ 1] \cap \mathcal{R}$$

$$SP(v,\ S_A,\ S_B,\ L_k) =$$

$$= \begin{cases} 1 - \frac{4}{\pi} \cdot arctan\left(\frac{idbp(v,\ S_A,\ L_k)}{idbp(v,\ S_B,\ L_k)}\right) \\ \qquad\qquad \text{if } idbp(v, S_B,\ L_k) \neq 0 \\ 1 \qquad\qquad \text{otherwise} \end{cases} \tag{16}$$

The specificit of founders is define as the normalized ratio between the gene-sharing probability with class S_A and class S_B over an arc of $\frac{\pi}{2}$ rad. The specificit of founders indicates the probability that founders share their genes more with one class than the other. The higher the specificit , the higher the probability. Specificit is one if a founder contributes to class S_A only; it is zero if a founder equally contributes to both classes; and it is -1 if a founder contributes to class S_B only. Note that it is not possible that $idbp(v, S_A,\ L_k)$ and $idbp(v, S_B,\ L_k)$ are both zero at the same time because of the construction of layers define in equations 8, 9, 10, and 11. Specificit is not tight to genetic contribution. A founder may be highly specifi to a class, while being a small contributor to that class. Conversely, a great genetic contributor to a class may be specifi to the opposite class of reference individuals. Great genetic contributors may also happen to contribute almost equally to both classes and the same may happen to small contributors. In general, specificit definitio emphasizes the differential contribution to two classes, rather than focusing on the absolute level of genetic contribution.

Three classes of founders can be define based on founder specificity *Specifi Layered Founders*, *Unique Layered Founders*, and *Ambiguous Layered Founders*.

6.2 Specifi Layered Founders

$SLF(S_A, S_B, L_k)$ is the set of layered founders belonging to L_k whose probability of sharing genes with individuals in S_A is higher than with individuals in S_B. Therefore, founders v in layer L_k are also specifi to class S_A if their specificit $SP(v,\ S_A,\ S_B,\ L_k)$ is greater than zero. Formally:

$$SLF: \mathcal{P}^3(V) \rightarrow \mathcal{P}(V)$$

$$SLF(S_A, S_B, L_k) = \\ = \{v \in L_k \mid idbp(v,\ S_A,\ L_k) > idbp(v,\ S_B,\ L_k)\} \tag{17}$$

6.3 Unique Layered Founders

$ULF(S_A, S_B, L_k)$ is the set of specifi layered founders belonging to L_k who share genes with individuals in S_A only. Unique founders are the most specifi founders ($SP(v,\ S_A,\ S_B,\ L_k) = 1$).

$$ULF : \mathcal{P}^3(V) \rightarrow \mathcal{P}(V)$$

$$\begin{aligned} ULF(S_A, S_B, L_k) = \{v \in L_k \mid & \\ & (idbp(v,\ S_A,\ L_k) > 0) \wedge \\ & (idbp(v,\ S_B,\ L_k) = 0)\ \} \end{aligned} \tag{18}$$

6.4 Ambiguous Layered Founders

$ALF(S_A, S_B, L_k)$ is the set of layered founders belonging to L_k whose probability of sharing genes with individuals in S_A is identical to that with individuals in S_B. Ambiguous founders are not specifi to any class ($SP(v,\ S_A,\ S_B,\ L_k) = 0$).

$$ALF : \mathcal{P}^3(V) \rightarrow \mathcal{P}(V)$$

$$\begin{aligned} & ALF(S_A, S_B, L_k) = \\ & = \{v \in L_k \mid\ idbp(v,\ S_A,\ L_k) = idbp(v,\ S_B,\ L_k)\} \end{aligned} \tag{19}$$

6.5 Specific Unique, and Ambiguous IBD Probabilities

Gene-sharing probabilities can be computed between classes S_A and S_B and the three above-mentioned classes of layered founders:

$$\begin{aligned} S\_P_{A,A,k} &= idbp(SLF(S_A,\ S_B,\ L_k),\ L_k,\ S_A) \\ S\_P_{A,B,k} &= idbp(SLF(S_A,\ S_B,\ L_k),\ L_k,\ S_B) \\ U\_P_{A,A,k} &= idbp(ULF(S_A,\ S_B,\ L_k),\ L_k,\ S_A) \\ A\_P_{A,A,k} &= idbp(ALF(S_A,\ S_B,\ L_k),\ L_k,\ S_A) \end{aligned} \tag{20}$$

Note that $U\_P_{A,B,k}$ is not relevant, due to the definitio of unique layered founders (equation 18).

Based on the definition of specifi and ambiguous layered founders (equations 17 and 19), the following holds:

$$S\_P_{A,A,k}\ + S\_P_{B,A,k}\ + A\_P_{A,A,k}\ = 1 \tag{21}$$

By definitio of ambiguous layered founders (equation 19), the following holds:

$$A\_P_{A,A,k}\ = A\_P_{B,B,k} \tag{22}$$

The probabilities are depicted in Figure 2.

6.6 Separability of Layered Founders

Separability of layered founders can be define as a measure of the difference in gene-sharing probability between sets of founders. We have define separability as the average difference of $idbp$ probability of a set of founders F_A, at layer L_k, with respect to classes S_A and S_B and of $idbp$ probability of a set of founders F_B with respect to the same classes. Formally:

$$SEP : \mathcal{P}^5(V) \rightarrow [-1,\ 1] \cap \mathcal{R}$$

$$\begin{aligned} & SEP(F_A,\ F_B,\ S_A,\ S_B,\ L_k)\ = \\ & \quad ((idbp(F_A,\ L_k,\ S_A) - \\ & \quad\ idbp(F_A,\ L_k,\ S_B)) + \\ & \quad (idbp(F_B,\ L_k,\ S_B) - \\ & \quad\ idbp(F_B,\ L_k,\ S_A))) \ /\ 2 \end{aligned} \tag{23}$$

Separability is commutative:

$$\begin{aligned} & SEP(F_A,\ F_B,\ S_A,\ S_B,\ L_k)\ = \\ & \quad SEP(F_B,\ F_A,\ S_B,\ S_A,\ L_k) \end{aligned} \tag{24}$$

Simplifying equation 23, we obtain:

$$\begin{aligned} & SEP(F_A,\ F_B,\ S_A,\ S_B,\ L_k)\ = \\ & \quad idbp(F_A,\ L_k,\ S_A)\ - idbp(F_A,\ L_k,\ S_B) \end{aligned} \tag{25}$$

Although separability can be computed on any two sets of founders, it is interesting to measure the separability of specifi founders, as shown in Figure 3. In that case, separability varies between 0 and 1.

7 Results and Discussion

Four sets of mutually-exclusive classes have been used for the experiments performed to illustrate the layered founders approach.

- The firs two are sets of contributing (CF) and anti-contributing (ACF) families for systolic blood pressure (bpday1_glucose_sys_2) on chromosome 1 at 195 cM and diastolic blood pressure (ave_bpd1_sup_dia) on chromosome 3 at 180 cM as described in [11].
- The last two sets are based on the bi-allelic ScaI marker located in the ANP gene. One is the set of CF and ACF families for 24 hours urine sodium excretion (u24hrs_na), obtained from a twopoint linkage analysis on ANP-ScaI and ANP-BstXI, ranked according to their contribution to ANP-ScaI LOD score. The other are individuals who have ANP-ScaI allele 1 (genotypes 1/1 and 1/2) and those who don't (genotype 2/2).

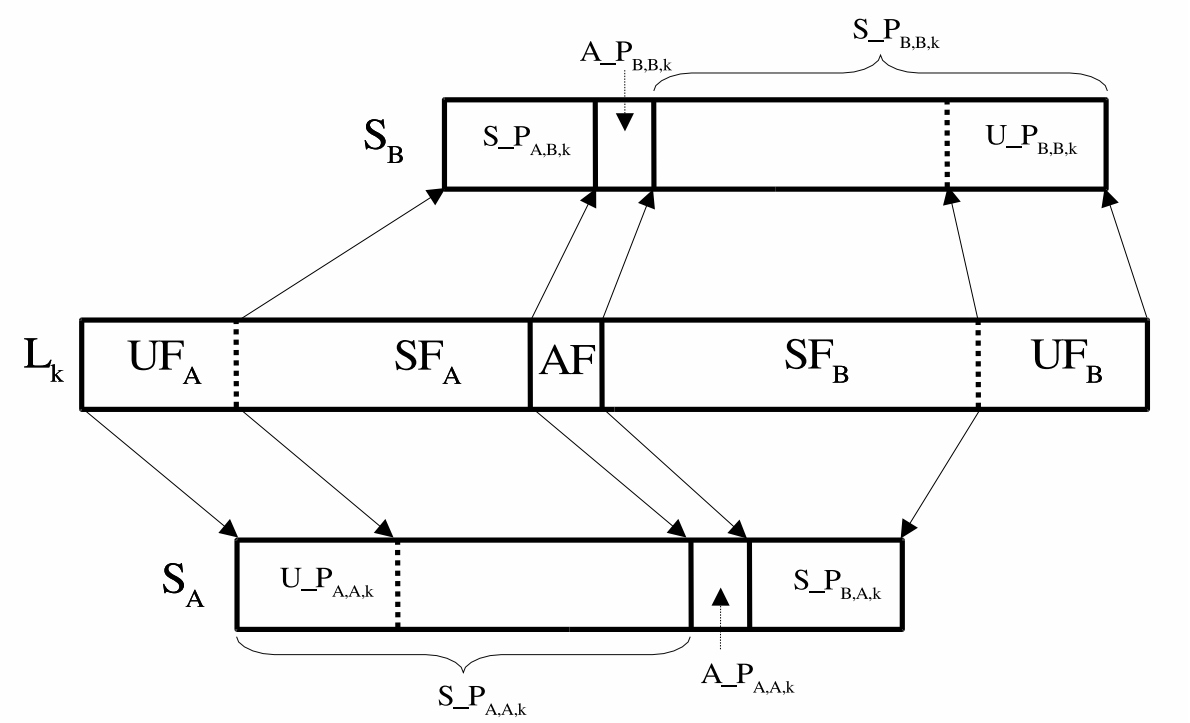

Figure 2. IBD probabilities. The three rows shown represent the $ibdp$ probabilities to class S_B (fi st row), the founders at layer L_k (second row) and the $ibdp$ probabilities to class S_A (third row). Here we assume that the two reference classes are mutually exclusive, i.e. they do not share any individual. The length of the fi st and third rows is 1 ($ibdp$ probability from all founders in L_k) and the length of the second row is the number of founders in L_k. The arrows show the $ibdp$ probability of some set of founders to one of the two classes. The probabilities are named according to equation 20. The layered founders have been categorized as unique founders of S_A (UF_A), specifi founders of S_A (SF_A), ambiguous founders (AF), specifi founders of S_B (SF_B) and unique founders of S_B (UF_B). Note that founders unique to a given class are a subset of founders specifi to that class: $UF_A \in SF_A$ and $UF_B \in SF_B$. Also, the $ibdp$ probability from some class unique founders to that class is included in the $ibdp$ probability from the class specifi founders to that class. As mentioned before, ancestors unique to a class do not contribute genetically to the other class.

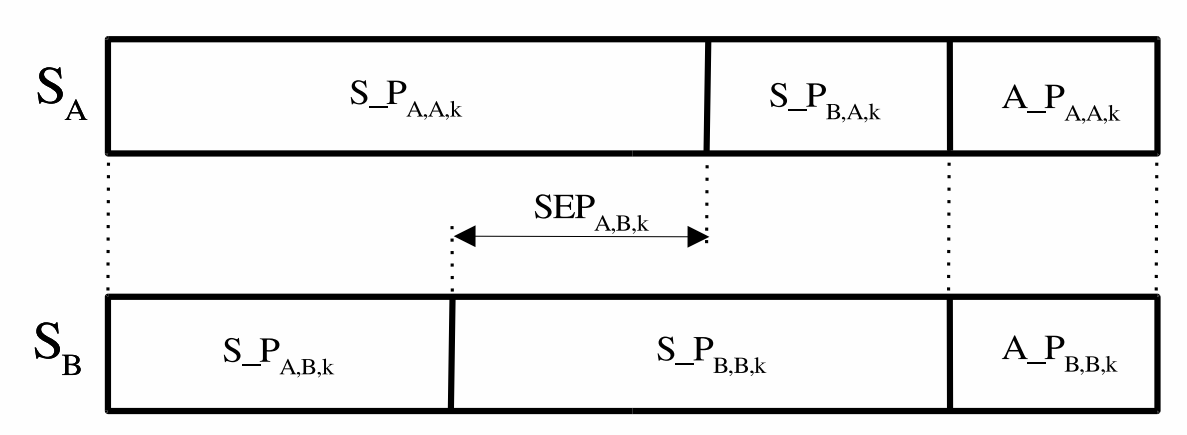

Figure 3. Separability of specifi founders. The $ibdp$ probabilities to classes S_A and S_B are shown in the upper and lower rows, respectively. Equation 25 can be understood graphically: The separability is equal to the length of the upper left block ($ibdp$ probability of S_A-specifi founders to S_A) minus the length of lower left block ($ibdp$ probability of S_A-specifi founders to S_B). Equivalently, it is equal to the length of the lower middle block minus the length of the upper middle block. Note also that as the $ibdp$ probability of the ambiguous founders increases, the separability decreases. If all genetic contribution comes from ambiguous founders, then the separability is 0. On the other hand, it can easily be seen that the maximum separability of 1 is reached when all genetic contribution of both classes comes from their respective specifi founders.

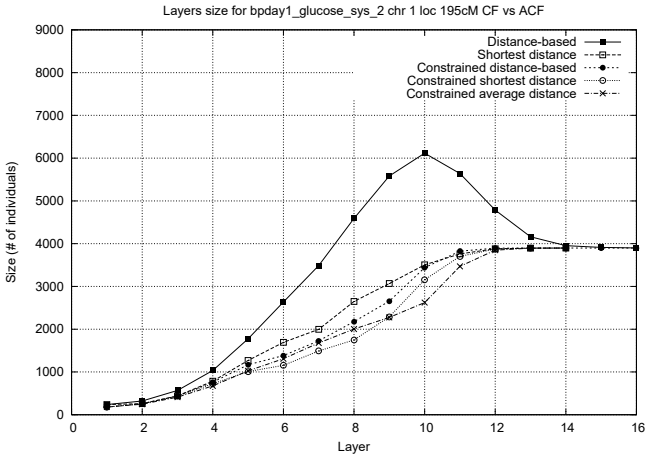

Figure 4. Chromosome 1 - size comparison of layers

Figure 4 shows the size of layers computed according to f ve different layer definitions for chromosome 1.

The distance-based approach presents a peak at an intermediate level (layer 10) for all experiments. Indeed, this approach is unconstrained, so in the central part of the genealogy one founder can reach today's individuals with several paths of different lengths. This peak disappears when the distance-based approach is constrained. Absolute layer sizes are different in the four experiments, but the trend persists. This is understandable, since in distance-based layers, all individuals at all path distances are put into layers, so central parts of the genealogy offer a larger number of paths of different lengths, as expected. Shortest distance, constrained shortest distance, constrained average distance and constrained distance-based layers present a smoother trend. Another observation that can be made is that the layers span is not the same for all definitions and this follows from the definition themselves.

Individuals belonging to layers obtained by different definitions have been compared by the Jacquard coefficien computed as follows:

$$JC(A,\ B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid} \tag{26}$$

Figure 5 reports the Jacquard coefficien for all layers of comparison of the different layer definition for CF and ACF families on chromosome 1. The Jacquard coefficient for the other three experiments are similar and therefore not shown.

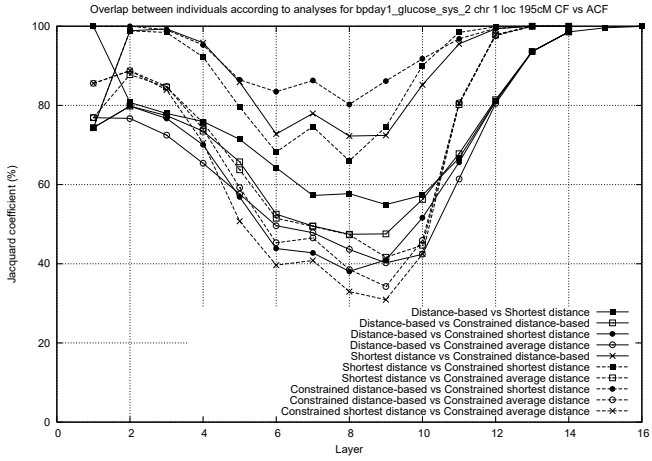

Figure 5. Chromosome 1 - Jacquard comparison of layers

It can be noted that, from a Jacquard coefficien point of view, layers are generally more similar at recent and oldest generations, but differ more at intermediate layers. Indeed, because of the bottle shape of the genealogy, variability tends to be higher at intermediate generations and to concentrate at recent and oldest generations. Also, we see that the Jacquard coefficient at the oldest layers increase sharply and become close to 100%. Indeed, the number of sources from previous layers become increasingly important as compared to the number of individuals that are actually reached at those layers. In the investigated cases, shortest distance and constrained shortest distance are most similar under the Jacquard coefficient

Specifi and unique founders' genetic contributions are presented in Figures 6 and 7 for chromosomes 1 and 3, and in Figures 8 and 9 for ANP-ScaI.

The reported data have been computed using the constrained average layer definition because of its smoother trend, together with satisfaction of the statistical independence property of the founders' genetic transmission.

Table 2 summarizes the extremes of the specifi genetic contribution diagrams.

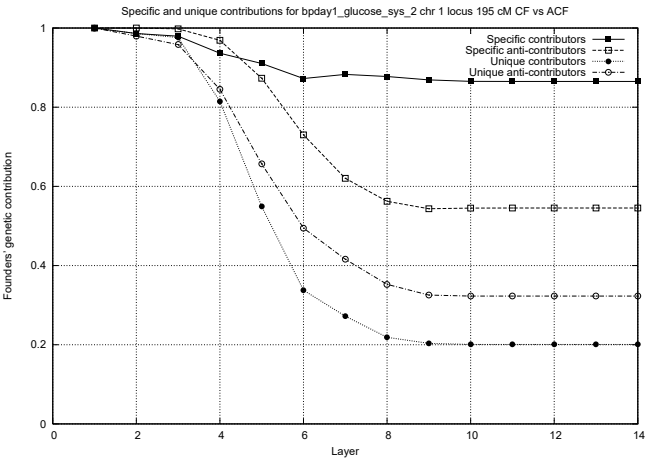

Figure 6. Chromosome 1 - CF/ACF specifi and unique founders' genetic contribution

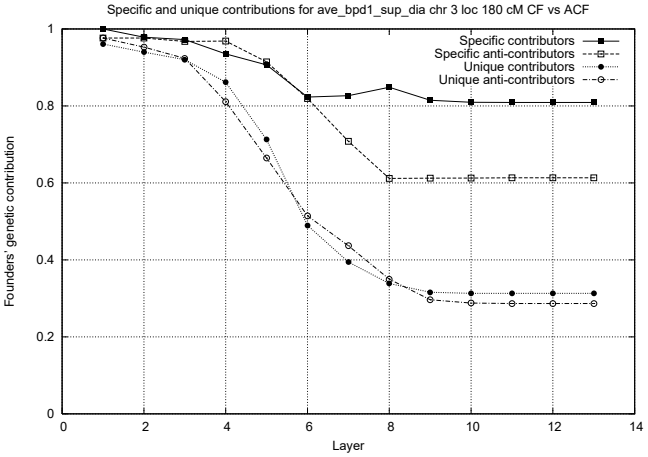

Figure 7. Chromosome 3 - CF/ACF specifi and unique founders' genetic contribution

Class	First layer	Last layer
Chromosome 1 CF	1.0	.86
Chromosome 1 ACF	1.0	.55
Chromosome 3 CF	1.0	.81
Chromosome 3 ACF	.98	.61
ANP-ScaI CF	1.0	.84
ANP-ScaI ACF	1.0	.56
ANP-ScaI 2/2	.71	.41
ANP-ScaI 1/1 or 1/2	.90	.77

Table 2. Specifi genetic contribution

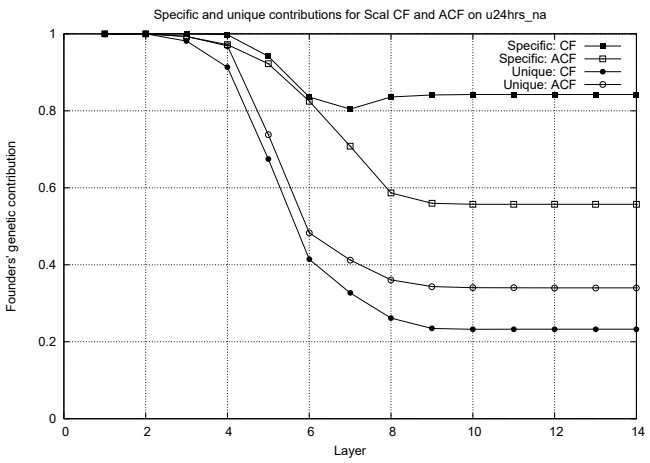

Figure 8. ANP-ScaI contributors - specifi and unique founders' genetic contribution

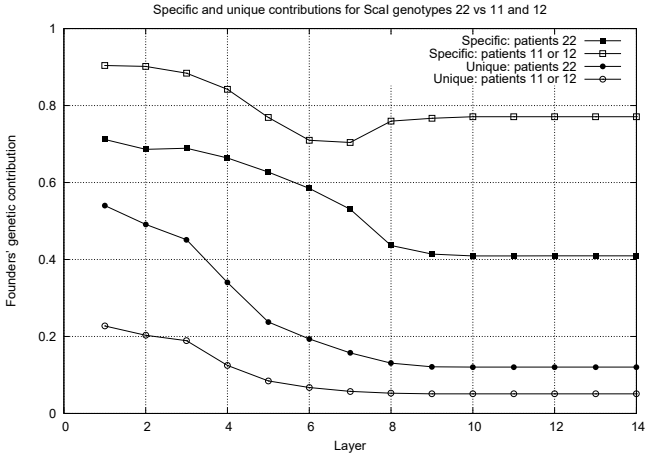

Figure 9. ANP-ScaI genoptype - specifi and unique founders' genetic contribution

It can be observed that in all but one investigated case, specifi founders account for more than 50% of the genetic pool of their class, even at the oldest generations.

Specifi founders of CF in chromosomes 1 and 3 account for 100% of genetic contribution from the most recent layer, and 86% and 81% respectively from the oldest layer. Comparing the specifi founders' contribution of the four experiments, we observe that the three experiments comparing classes obtained on a family basis (i.e. chromosome 1, chromosome 3 and ANP-ScaI contributors) show higher specifi founders' contributions at the firs few layers than the experiment comparing classes obtained on an individual basis (i.e. ANP-ScaI alleles).

Table 3 reports the extremes of the unique genetic contribution diagrams.

Class	First layer	Last layer
Chromosome 1 CF	1.0	.20
Chromosome 1 ACF	1.0	.32
Chromosome 3 CF	.96	.31
Chromosome 3 ACF	.98	.29
ANP-ScaI CF	1.0	.23
ANP-ScaI ACF	1.0	.34
ANP-ScaI 2/2	.54	.12
ANP-ScaI 1/1 or 1/2	.23	.05

Table 3. Unique genetic contribution

Unique founders' genetic contributions to chromosomes 1 and 3 CF and ACF are very high at recent generations (96% to 100% unique contribution at layer 1) and are still relevant at oldest generations (20% to 32% at layers 14 and 13). Unique founders' genetic contributions for ANP-ScaI genotypes are smaller: at layer 1, only 54% and 23% of the genetic pool come from unique founders, and contributions are fairly low at oldest generations (12% and 5%).

Although specifi founders contribute by definitio more to a class than to the other mutually exclusive class, they may however largely contribute to the other class. Separability has been define to measure differences in average genetic contributions of sets of founders to both classes. Figure 10 presents the computed separability for specifi founders for all layers.

Extremes of separability are summarized in Table 4. It appears that chromosome 1 and 3 specifi founders are still highly separable at the oldest generations. The average difference in the genetic contribution of specifi founders to both classes is higher than 97% at layer 1, and higher than 40% at layers 14 and 13. Specifi founders of ANP-ScaI genotype are less separable. The average difference in the

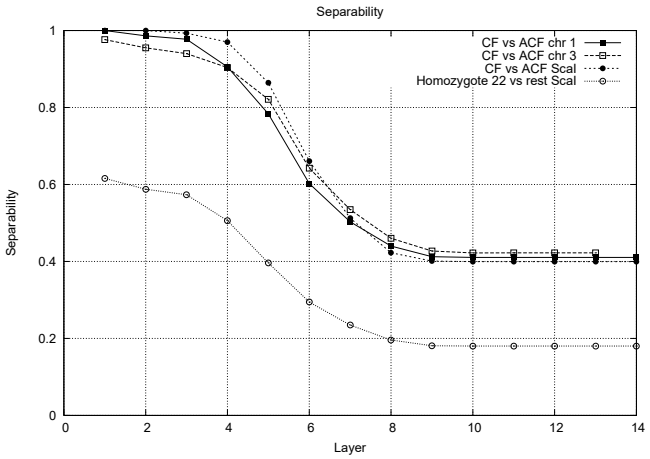

Figure 10. Separability of specifi founders

Class	First layer	Last layer
Chromosome 1	1.0	.41
Chromosome 3	.98	.42
ANP-ScaI contributors	1.0	.40
ANP-ScaI genotype	.61	.18

Table 4. Separability of specifi founders

genetic contribution of specifi founders to both classes of individuals who have allele 1 or who do not is about 61% at layer 1 (fair separability), but the difference is about 18% at layer 14.

It should be remarked that, at all layers, founders from classes that were obtained on a family basis are more separable than those who have been obtained on an individual basis. Also, the separability of the "family-based" classes are very similar (nearly 100% at the firs layer, about 40% at the last layer).

To assess the statistical significanc of separability of specifi founders, a simulation of 10^6 cases has been run. The simulation results are reported in Figure 11 for chromosome 1. In all simulated cases, simulated separability is significantl far from the one computed on specifi founders. Further details on simulated separability can be found in the Appendix.

8 Conclusions

The present experiments show that the layered founders approach is feasible on graph representations of genealogies.

Layered founders can be analyzed at intermediate levels between the current generation and the frontier of the

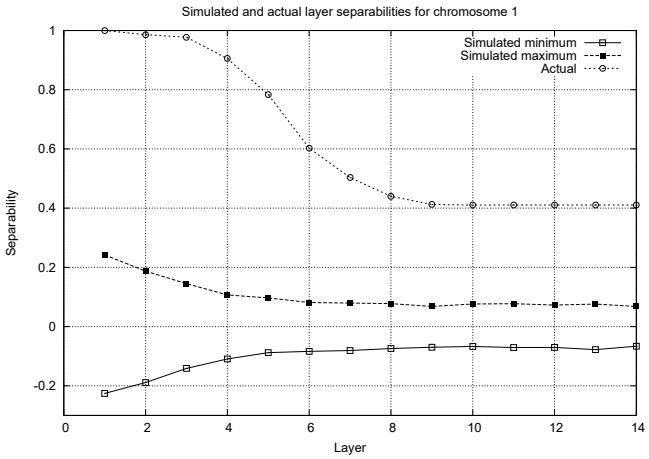

Figure 11. Chromosome 1 - extremes of separability simulation

genealogy, which was the classical, previously-reported approach. Alternative definition of layers have been investigated, and their values provide insight into the founders' genetic contributions at different distances from today's generations.

Founders that contribute more to the genetic pool of a class than to another can and have been identifie and reported at different distances from current generations; unique founders who uniquely contribute to the genetic pool of a class have been identifie too.

A measure of distinct genetic contribution between specifi founders of two classes is captured by the separability definitio which gives the average difference in genetic contribution between two classes of founders. Separability has been measured on different sets of individuals. The statistical significanc of founders separability in the investigated cases has been validated by random simulation.

The layered founders approach allows a fine grain analysis of founders' genetic contribution than previously-reported approaches.

References

[1] Pouyez C, Lavoie Y, Bouchard G, and Roy R. *Introduction à l'histoire des populations du Saguenay, XVIe-XXe siècles*. Presses de l'Université du Québec, 1983.

[2] Bouchard G, Laberge C, and Scriver C. Reproduction démographique et transmission génétique dans le nord-est de la province du Québec (18ieme-20ieme siècle). *Eur J Population*, 4:39–67, 1988.

[3] Bouchard G and De Braekeleer M. *Histoire d'un génome*. Presses de l'Université du Québec, 1991.

[4] Labuda D, Zietkiewicz E, and Labuda M. The genetic clock and the age of the founder effect in growing populations: A lesson from French Canadians and Ashkenazim. *Am J Hum Genet*, 61:768–771, 1997.

[5] Labuda M, Labuda D, Korab-Laskowska M, Cole D, Zietkiewicz E, Weissenbach J, Popowska E, Pronicka E, Root A, and Glorieux F. Linkage disequilibrium analysis in young populations: Pseudo-vitamin d-deficien y rickets and the founder effect in French Canadians. *Am J Hum Genet*, 59:633–643, 1996.

[6] Gauvreau D and Bourque M. Mouvements migratoires et familles: le peuplement du Saguenay avant 1911. *Rev histoire Amérique fr*, 42(2):167–192, 1988.

[7] Scriver CR. Human genetics: lessons from Quebec populations. *Ann Rev Genomics Hum Genet*, 2:69–101, 2001.

[8] Heyer E, Toupance B, Perri C, De Vito O, Foncin J-F, and Bruni AC. Manic depressive illness in a founder population. *Eur J Hum Genet*, 11(8):597–602, 2003.

[9] Heyer E and Tremblay M. Variability of the genetic contribution of Quebec population founders associated to some deleterious genes. *Am J Hum Genet*, 56(4):970–978, 1995.

[10] Kotchen TA, Kotchen JM, Grim CE, George V, Kaldunski ML, Cowley AW, Hamet P, and Chelius TH. Genetic determinants of hypertension: identificatio of candidate phenotypes. *Hypertension*, 36(1):7–13, 2000.

[11] Hamet P, Merlo E, Seda O, Broeckel U, Tremblay J, Kaldunski M, Gaudet D, Bouchard G, Deslauriers B, Gagnon F, Antoniol G, Kotchen TA, Pausova Z, Labuda M, Jomphe M, Gossard F, Kirova R, Tonellato P, Orlov SN, Pintos J, Rioux JD, Platko J, Hudson TJ, Lander E, and Cowley AW. Quantitative founder effect analysis in French-Canadian families contributing to intermediate phenotypes of hypertension. *Parallele submission to Nat Gen*, 2004.

[12] De Bold AJ, Borenstein HB, Veress AT, and Sonnenberg H. A rapid and potent natriuretic response to intravenous injection of atrial myocardial extract in rats. *Life Sci*, 28(1):89–94, 1981.

[13] Lang RE, Tholken H, Ganten D, Luft FC, Ruskoaho H, and Unger T. Atrial natriuretic factor – a circulating hormone stimulated by volume loading. *Nature*, 314(6008):264–266, 1985.

[14] Tremblay J, Desjardins R, Hum D, Gutkowska J, and Hamet P. Biochemistry and physiology of the natriuretic peptide receptor guanylyl cyclases. *Mol Cell Biochem*, 230:31–47, 2002.

[15] Hamet P, Tremblay J, Pang SC, Garcia R, Thibault G, Gutkowska J, Cantin M, and Genest J. Effect of native and synthetic atrial natriuretic factor on cyclic GMP. *Biochem Biophys Res Commun*, 123(2):515–527, 1984.

[16] Weidmann P, Hasler L, Gnadinger MP, Lang RE, Uehlinger DE, Shaw S, Rascher W, and Reubi FC. Blood levels and renal effects of atrial natriuretic peptide in normal man. *J Clin Invest*, 77(3):734–42, 1986.

[17] Tremblay J, Gerzer R, Vinay P, Pang SC, Beliveau R, and Hamet P. The increase of cGMP by atrial natriuretic factor correlates with the distribution of particulate guanylate cyclase. *FEBS Lett*, 181(1):17–22, 1985.

[18] Sengenes C, Berlan M, De Glisezinski I, Lafontan M, and Galitzky J. Natriuretic peptides: a new lipolytic pathway in human adipocytes. *FASEB J.*, 14(10):1345–1351, 2000.

[19] Tremblay J, Faldik K, Podjaski C, Brunelle P-L, Pausova Z, Gaudet D, Tremblay G, Kotchen TA, Cowley AW, and Hamet P. Loci of ANP system as genetic factors in obesity-related hypertension. *Intl J Obesity*, 28:S112, 2004.

[20] Rioux JD, Stone VA, Daly MJ, Cargill M, Green T, Nguyen H, Nutman T, Zimmerman PA, Tucker MA, Hudson T, Goldstein AM, Lander E, and Lin AY. Familial eosinophilia maps to the cytokine gene cluster on human chromosomal region 5q31-q33. *Am J Hum Genet*, 63:1086–1094, 1998.

[21] Cormen TH, Leiserson CE, and Rivest RL. *Introduction to Algorithms*. The MIT Press, Cambridge, 1990.

[22] Rosen KH. *Discrete Mathematics and Its Applications, Fourth Edition*. McGraw-Hill Book Company, 1999.

[23] Barabási A-L and Oltvai Z. Network biology: Understanding the cell's functional organization. *Nat Rev Gen*, 5:101–113, 2004.

[24] Stewart I. Networking opportunity. *Nature*, 427:601–604, 2004.

[25] Subdø J, Bankfalvi A, Bryne M, Marcelpoil R, Boysen M, Piffko J, Hemmer J, Kraft K, and Reith A. Prognostic value of graph theory-based tissue architecture analysis in carcinomas of the tongue. *Lab Invest*, 80(12):1881–1889, 2000.

[26] Lindorff-Larsen K, Vendruscolo M, Paci E, and Dobson CM. Transition states for protein folding have native topologies despite high structural variability. *Nat Struct Molec Biol*, 11(5):443–449, 2004.

[27] Bouchard G, Roy R, Casgrain B, and Hubert M. [Population file and database management: the BALSAC database and the INGRES/INGRID system]. *Hist Mes*, 4:39–57, 1989.

[28] Silverman BW. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, UK, 1986.

[29] Ihaka R and Gentleman R. R: A language for data analysis and graphics. *J Comput Graph Stat*, 5(3):299–314, 1996.

9 Appendix

9.1 Statistical Significanc of Separability

To assess the statistical significanc of separability, a simulation of 10^6 cases has been run using the algorithm reported in Figure 12. The simulation results are reported in Figure 13 for chromosome 1 and in Table 5 for all experiments. For chromosome 1, the difference between maximum simulated separability and that observed for specifi founders is higher than 75% at layer one, but it drops to about 30% at layer 14. This drop also depends on the drop in absolute value of separability. The smaller the separability becomes, the closer it gets to randomly-simulated separability. For all experiments, we see that the lowest difference between actual and simulated separability is 15% (ANP-ScaI genotype, layers 9 through 14).

The kernel density estimates [28] depicted in Figure 13 have been computed using function `density` from the open-source statistical package GNU R [29]. The kernel was Gaussian, Silverman's "rule of thumb" was used to select the bandwidth, and the grid consisted of 1,024 points.

Layer	Experiment											
	Chr 1			Chr 3			ANP-ScaI contributors			ANP-ScaI genotype		
	A	C	d	A	C	d	A	C	d	A	C	d
1	1.0	.24	.76	.98	.25	.73	1.0	.22	.78	.62	.11	.51
2	.99	.19	.70	.96	.21	.75	1.0	.21	.79	.59	.08	.51
3	.98	.14	.84	.94	.15	.79	.99	.16	.83	.57	.06	.51
4	.91	.11	.80	.90	.13	.77	.97	.15	.82	.51	.05	.46
5	.78	.10	.68	.82	.10	.72	.86	.14	.72	.39	.04	.35
6	.60	.08	.52	.64	.09	.55	.66	.10	.56	.29	.03	.26
7	.50	.08	.42	.53	.08	.45	.51	.10	.41	.23	.03	.20
8	.44	.08	.37	.46	.07	.39	.42	.08	.34	.20	.03	.17
9	.41	.07	.34	.43	.07	.36	.40	.09	.31	.18	.03	.15
10	.41	.08	.33	.42	.07	.35	.40	.08	.32	.18	.03	.15
11	.41	.08	.33	.42	.07	.35	.40	.08	.32	.18	.03	.15
12	.41	.07	.34	.42	.07	.35	.40	.08	.32	.18	.03	.15
13	.41	.07	.34	.42	.07	.35	.40	.08	.32	.18	.03	.15
14	.41	.07	.34				.40	.08	.32	.18	.03	.15
Minimum			.33			.35			.31			.15

Table 5. Simulation of separability of specifi founders. The letters A, C and d refer respectively to the actual separability, the closest simulated separability and the difference between the two.

1 $forall\ layers L_k$
2 $\quad F_A = SLF(S_A, S_B, L_k)$
3 $\quad F_B = SLF(S_B, S_A, L_k)$
4 $\quad SEP_{A,B} = SEP(F_A, F_B,\ L_k)$
5 $\quad forall\ simulations\ S_i$
6 $\quad\quad (S\_F_{A,i}, S\_F_{B,i}) = (random(L_k) \mid$
$\quad\quad\quad |S\_F_{A,i}| = |F_A| \wedge |S\_F_{B,i}| = |F_B| \wedge$
$\quad\quad\quad (S\_F_{A,i} \cap S\_F_{B,i} = \emptyset))$
7 $\quad\quad S\_SEP_{A,B,i} = SEP(S\_F_{A,i}, S\_F_{B,i},\ L_k)$
8 $compare\ SEP_{A,B}\ with\ distribution\ of\ S\_SEP_{A,B,i}$

Figure 12. Separability simulation strategy

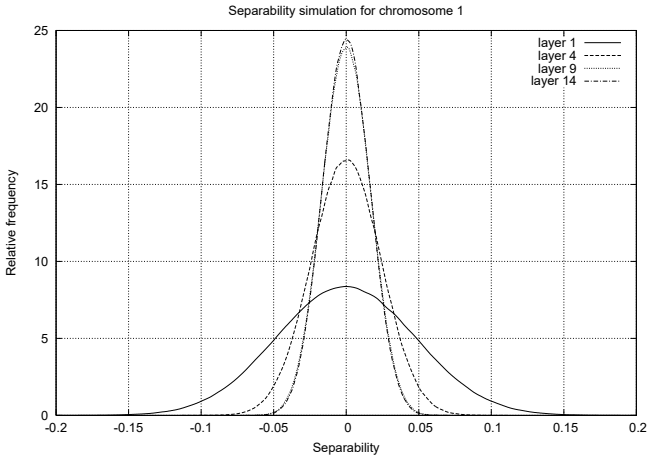

Figure 13. Chromosome 1 - simulation distribution for layers 1, 4, 9 and 14

L'École Polytechnique se spécialise dans la formation d'ingénieurs et la recherche en ingénierie depuis 1873

École Polytechnique de Montréal

École affiliée à l'Université de Montréal

Campus de l'Université de Montréal
C.P. 6079, succ. Centre-ville
Montréal (Québec)
Canada H3C 3A7

www.polymtl.ca

ÉCOLE
POLYTECHNIQUE
MONTRÉAL