LAYARED FOUNDERS: A NOVEL APPROACH TO INVESTIGATE 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
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
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 specific founders who predominantly contribute to one class of individuals and the analysis of 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 Scal 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 Merlo¹, Benoit Deslauriers², Giuliano Antoniol³, Pierre-Luc Brunelle², Michèle Jomphe⁴, Gérard Bouchard⁴, Johanne Tremblay⁴ and Pavel Hamet²

¹Department 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
²Centre hospitalier de l’Université de Montréal (CHUM)
³University of Sannio
⁴Université 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 definitions of layers and distances are proposed and discussed). It allows the maximum likelihood classification of specific founders who predominantly contribute to one class of individuals and the analysis of 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 Scal 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.

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 km². 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 first 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 specific 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 km², 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 first 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 first 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 first 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 specifically 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 profited 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 first 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 difficult 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 identified 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, Scal 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 significant 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 first intron of the ANP gene), we performed poly-
merase chain reaction (PCR) of 35 cycles (5 sec at 94°C, 5 sec at 61°C and 10 sec at 72°C) with sense primer AGA CAG AGC AAC AGG TGG and antisense primer CAT TCC ATC CCC AGT TCC with an initial denaturation of 5 min at 94°C and final extension of 7 min at 72°C, followed by BstXI digestion (2 hrs at 45°C) with 1U of enzyme in a volume of 10 μL. For the Scal marker (located in exon 3 caused by a mutation of the first nucleotide of the stop codon), 30 cycles of PCR (5 sec at 94°C, 5 sec at 63°C, 5 sec at 72°C), with an initial denaturation of 5 min at 94°C and final extension of 7 min at 72°C, were performed with sense GGC ACA CTC ATA CAT GAA GCT TTT T and antisense primer GCA GTC TGT CCC TGG GCC CA, followed by Scal digestion (2 hrs at 37°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 modified 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 influence 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 define 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 find 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 definitely 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.

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:

\[
T = (entry_1, ..., entry_i, ..., entry_n)
\]

\[
entry_1 = (c_{1,1}, ..., c_{1,j}, ..., c_{1,m}) \\
entry_i = (c_{i,1}, ..., c_{i,j}, ..., c_{i,m}) \\
entry_n = (c_{n,1}, ..., c_{n,j}, ..., c_{n,m})
\]

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

\[
c_{i,r} = \text{child}_{jd_i} \\
c_{i,s} = \text{mother}_{jd_i} \\
c_{i,t} = \text{father}_{jd_i}
\]

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

\[
V_G = \{ v | (\exists entry_i \in T) \} \\
(ID(v) = \text{child}_{jd_i}) \lor \\
(ID(v) = \text{mother}_{jd_i}) \lor \\
(ID(v) = \text{father}_{jd_i})
\]

\[
E_G = \{ (v_1, v_2) | \\
(v_1 \in V_G \land v_2 \in V_G) \land \\
(\exists entry_j \in T) \\
(ID(v_1) = \text{child}_{jd_j}) \land \\
((ID(v_1) = \text{mother}_{jd_j}) \lor \\
(ID(v_1) = \text{father}_{jd_j})) \}
\]

where \( ID(v) \) is the identifier 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 defined 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 ∈ V, where v is a parent of w, i.e. (v, w) ∈ E, 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 defined 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 define several genealogical layers corresponding to the different values of genealogical distance from reference individuals. The first 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 definitions of layers and genealogical distances have to be carefully conceived to overcome the mentioned limitations.

In the following sub-sections, we will formally define 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 defined as follows:

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

\[ \text{paths}(v, S) = \{ p = (w_0, w_1, \ldots, w_n) \mid (\forall i \in [0, n], w_i \in V) \land (w_0 = v) \land (w_n \in S) \land (\forall j \in [1, n], (w_{j-1}, w_j) \in E) \} \]  \hspace{1cm} (5)

\[ \text{paths}(v, S) \] is the set of any sequence of any finite 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 defined for ease of later definition of layers. They are the Minimum Genealogical Path Length, and the Average Genealogical Path Length. In formulae:

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

\[ mgpl(v, S) = \begin{cases} \min |p| & \text{if } \text{paths}(v, S) \neq \emptyset \\ \text{undefined} & \text{otherwise} \end{cases} \]  \hspace{1cm} (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.
\[ agpl : V \times \mathcal{P}(V) \rightarrow \mathcal{R} \]
\[ agpl(v, S) = \begin{cases} 
\sum_{p \in \text{paths}(v, S)} |p| \quad & \text{if } \text{paths}(v, S) \neq \emptyset \\
[\text{undefined}] \quad & \text{otherwise}
\end{cases} \quad (7) \]

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

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

### 5.2 Distance-Based Layered Founders

\[ L : \mathcal{P}(V) \times \mathcal{N}_0^+ \rightarrow \mathcal{P}(V) \]
\[ L(d, S) = \{ v \in V \mid (\exists p \in \text{paths}(v, S), \ |p|=d \} \cup \ \\
\{ w \in V \mid (\exists (u, w) \in E) \land \ \\
(\forall p \in \text{paths}(w, S), \ |p|<d) \} \quad (8) \]

\( L(d, S) \) 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

\[ L : \mathcal{P}(V) \times \mathcal{N}_0^+ \rightarrow \mathcal{P}(V) \]
\[ L(d, S) = \{ v \in V \mid mGPL(v, S) = d \} \cup \ \\
\{ w \in V \mid (\exists (u, w) \in E) \land \ \\
(mGPL(w, S) < d) \} \quad (9) \]

\( L(d, S) \) 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

\[ L(d, S) = \begin{cases} 
\{ v \in V \mid (d - \frac{1}{2}) \leq agpl(v, S) < (d + \frac{1}{2}) \} \cup \\
\{ w \in V \mid (\exists (u, w) \in E) \land agpl(w, S) < (d - \frac{1}{2}) \} \quad (10) \]

\( L(d, S) \) 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

\[ cLayer : \mathcal{P}(V) \rightarrow \mathcal{P}(V) \]
\[ cLayer(L) = \{ w \in L \mid \ \\
(\forall v \in L, \ \\
(\exists p = (w_0, w_1, \ldots, w_n) \} \land \ \\
(w_0 = v) \land \ \\
(w_n = w) \land \ \\
(\forall j \in [1, n], \ \\
(w_{j-1}, w_j) \in E_G)) \} \quad (11) \]

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:

\[ gc : V \times \mathcal{P}(V) \rightarrow [0, 1] \cap \mathcal{R} \]
\[ gc(v, S) = \begin{cases} 
\sum_{p \in \text{paths}(v, S)} \left( \frac{1}{2} \right)^{|p|} \quad & \text{if } \text{paths}(v, S) \neq \emptyset \\
0 \quad & \text{otherwise} \quad (12) \]

\[ gc \ (v, S) = \begin{cases} 
\sum_{p \in \text{paths}(v, S)} \left( \frac{1}{2} \right)^{|p|} \quad & \text{if } \text{paths}(v, S) \neq \emptyset \\
0 \quad & \text{otherwise} \quad (12) \]

6
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 defined as follows:

$$
idbp : V \times P^2(V) \to [0, 1] \cap R$$

$$
idbp(v, S, L_k) = \frac{gc(v, S)}{\sum_{w \in L_k} gc(w, S)}$$

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

$$
idbp : P^3(V) \to (0, 1] \cap R$$

$$
idbp(F, S, L_k) = \sum_{f \in F} \nidbp(f, S, L_k)$$

$idbp(F, S, L_k)$ 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 $S_B$:

$$
idbp : P^3(V) \to P(R^2)$$

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

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 Specificity 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 defined the concept of specificity of founders as follows:

$$SP : V \times P^3(V) \to [0, 1] \cap R$$

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

$$\begin{align*}
1 - \frac{4}{\pi} \cdot \arctan \left( \frac{idbp(v, S_A, L_k)}{idbp(v, S_B, L_k)} \right) \\
1 & \text{ otherwise}
\end{align*}$$

The specificity of founders is defined 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 specificity of founders indicates the probability that founders share their genes more with one class than the other. The higher the specificity, the higher the probability. Specificity 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 defined in equations 8, 9, 10, and 11. Specificity is not tight to genetic contribution. A founder may be highly specific to a class, while being a small contributor to that class. Conversely, a great genetic contributor to a class may be specific 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, specificity definition emphasizes the differential contribution to two classes, rather than focusing on the absolute level of genetic contribution.

Three classes of founders can be defined based on founder specificity: Specific Layered Founders, Unique Layered Founders, and Ambiguous Layered Founders.

### 6.2 Specific 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 specific to class $S_A$ if their specificity $SP(v, S_A, S_B, L_k)$ is greater than zero. Formally:

$$SLF : P^3(V) \to P(V)$$

$$SLF(S_A, S_B, L_k) =$$

$$\{v \in L_k \mid idbp(v, S_A, L_k) > idbp(v, S_B, L_k)\}$$

(17)
6.3 Unique Layered Founders

\[ ULF(S_A, S_B, L_k) = \{ v \in L_k \mid (\text{idbp}(v, S_A, L_k) > 0) \land (\text{idbp}(v, S_B, L_k) = 0) \} \]

6.4 Ambiguous Layered Founders

\[ ALF(S_A, S_B, L_k) = \{ v \in L_k \mid \text{idbp}(v, S_A, L_k) = \text{idbp}(v, S_B, L_k) \} \]

6.5 Specific, Unique, and Ambiguous IBD Probabilities

\[ S \cdot P_{A,A,k} = \text{idbp}(S \cdot L(S_A, S_B, L_k, L_k, S_A) \]
\[ S \cdot P_{A,B,k} = \text{idbp}(S \cdot L(S_A, S_B, L_k, L_k, S_B) \]
\[ U \cdot P_{A,A,k} = \text{idbp}(U(L(S_A, S_B, L_k, L_k, S_A) \]
\[ A \cdot P_{A,A,k} = \text{idbp}(A(L(S_A, S_B, L_k, L_k, S_A) \]

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

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

\[ S \cdot P_{A,A,k} + S \cdot P_{B,A,k} + A \cdot P_{A,A,k} = 1 \]

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

\[ A \cdot P_{A,A,k} = A \cdot P_{B,B,k} \]

The probabilities are depicted in Figure 2.

6.6 Separability of Layered Founders

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

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

\[ SEP(F_A, F_B, S_A, S_B, L_k) = \frac{(\text{idbp}(F_A, L_k, S_A) - \text{idbp}(F_B, L_k, S_B)) + \text{idbp}(F_B, L_k, S_A))}{2} \]

Separability is commutative:

\[ SEP(F_A, F_B, S_A, S_B, L_k) = SEP(F_B, F_A, S_A, S_B, L_k) \]

Simplifying equation 23, we obtain:

\[ SEP(F_A, F_B, S_A, S_B, L_k) = \frac{\text{idbp}(F_A, L_k, S_A) - \text{idbp}(F_A, L_k, S_B)}{2} \]

Although separability can be computed on any two sets of founders, it is interesting to measure the separability of specific 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 first two are sets of contributing (CF) and anti-contributing (ACF) families for systolic blood pressure (bpd1_glu cosy 2) on chromosome 1 at 195 cM and diastolic blood pressure (aver_bpd1_sup dia) on chromosome 3 at 180 cM as described in [11].

- The last two sets are based on the bi-allelic Scal marker located in the ANP gene. One is the set of CF and ACF families for 24 hours urine sodium excretion (u24hrsna), obtained from a two-point linkage analysis on ANP-Scal and ANP-BstXI, ranked according to their contribution to ANP-Scal LOD score. The other are individuals who have ANP-Scal 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 \textit{ibdp} probabilities to class $S_B$ (first row), the founders at layer $L_k$ (second row) and the \textit{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 first and third rows is 1 (\textit{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 \textit{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 $S_A$ ($UF_A$), specific founders of $S_A$ ($SF_A$), ambiguous founders ($AF$), specific 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 specific to that class: $UF_A \in SF_A$ and $UF_B \in SF_B$. Also, the \textit{ibdp} probability from some class unique founders to that class is included in the \textit{ibdp} probability from the class specific founders to that class. As mentioned before, ancestors unique to a class do not contribute genetically to the other class.
Figure 3. Separability of specific founders. The \( i_bdp \) 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 (\( i_bdp \) probability of \( S_A \)-specific founders to \( S_B \)) minus the length of lower left block (\( i_bdp \) probability of \( S_A \)-specific 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 \( i_bdp \) 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 specific founders.

Figure 4 shows the size of layers computed according to five 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 definitions themselves.

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

\[
JC(A, B) = \frac{|A \cap B|}{|A \cup B|}
\]
Figure 5 reports the Jacquard coefficient for all layers of comparison of the different layer definitions for CF and ACF families on chromosome 1. The Jacquard coefficients 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 coefficient 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 coefficients 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.

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

Specific 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 specific 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 specific founders’ contributions at the first 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.

<table>
<thead>
<tr>
<th>Class</th>
<th>First layer</th>
<th>Last layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome 1 CF</td>
<td>1.0</td>
<td>.20</td>
</tr>
<tr>
<td>Chromosome 1 ACF</td>
<td>1.0</td>
<td>.32</td>
</tr>
<tr>
<td>Chromosome 3 CF</td>
<td>.96</td>
<td>.31</td>
</tr>
<tr>
<td>Chromosome 3 ACF</td>
<td>.98</td>
<td>.29</td>
</tr>
<tr>
<td>ANP-ScaI CF</td>
<td>1.0</td>
<td>.23</td>
</tr>
<tr>
<td>ANP-ScaI ACF</td>
<td>1.0</td>
<td>.34</td>
</tr>
<tr>
<td>ANP-ScaI 2/2</td>
<td>.54</td>
<td>.12</td>
</tr>
<tr>
<td>ANP-ScaI 1/1 or 1/2</td>
<td>.23</td>
<td>.05</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Class</th>
<th>First layer</th>
<th>Last layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome 1</td>
<td>1.0</td>
<td>.41</td>
</tr>
<tr>
<td>Chromosome 3</td>
<td>.98</td>
<td>.42</td>
</tr>
<tr>
<td>ANP-Scal contributors</td>
<td>1.0</td>
<td>.40</td>
</tr>
<tr>
<td>ANP-Scal genotype</td>
<td>.61</td>
<td>.18</td>
</tr>
</tbody>
</table>

genetic contribution of specific 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 first layer, about 40% at the last layer).

To assess the statistical significance of separability of specific 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 significantly far from the one computed on specific 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 genealogy, which was the classical, previously-reported approach. Alternative definitions 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 identified and reported at different distances from current generations; unique founders who uniquely contribute to the genetic pool of a class have been identified too.

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

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

References


14
Table 5. Simulation of separability of specific founders. The letters A, C and d refer respectively to the actual separability, the closest simulated separability and the difference between the two.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Chr 1</th>
<th>Chr 3</th>
<th>ANP-Scal contributors</th>
<th>ANP-Scal genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>C</td>
<td>d</td>
<td>A</td>
</tr>
<tr>
<td>1</td>
<td>1.0</td>
<td>.24</td>
<td>.76</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>.99</td>
<td>.19</td>
<td>.70</td>
<td>.96</td>
</tr>
<tr>
<td>3</td>
<td>.98</td>
<td>.14</td>
<td>.84</td>
<td>.94</td>
</tr>
<tr>
<td>4</td>
<td>.91</td>
<td>.11</td>
<td>.80</td>
<td>.90</td>
</tr>
<tr>
<td>5</td>
<td>.78</td>
<td>.10</td>
<td>.68</td>
<td>.82</td>
</tr>
<tr>
<td>6</td>
<td>.60</td>
<td>.08</td>
<td>.52</td>
<td>.64</td>
</tr>
<tr>
<td>7</td>
<td>.50</td>
<td>.08</td>
<td>.42</td>
<td>.53</td>
</tr>
<tr>
<td>8</td>
<td>.44</td>
<td>.08</td>
<td>.37</td>
<td>.46</td>
</tr>
<tr>
<td>9</td>
<td>.41</td>
<td>.07</td>
<td>.34</td>
<td>.43</td>
</tr>
<tr>
<td>10</td>
<td>.41</td>
<td>.08</td>
<td>.33</td>
<td>.42</td>
</tr>
<tr>
<td>11</td>
<td>.41</td>
<td>.08</td>
<td>.33</td>
<td>.42</td>
</tr>
<tr>
<td>12</td>
<td>.41</td>
<td>.07</td>
<td>.34</td>
<td>.42</td>
</tr>
<tr>
<td>13</td>
<td>.41</td>
<td>.07</td>
<td>.34</td>
<td>.42</td>
</tr>
<tr>
<td>14</td>
<td>.41</td>
<td>.07</td>
<td>.34</td>
<td>.42</td>
</tr>
<tr>
<td>Minimum</td>
<td>.33</td>
<td>.35</td>
<td>.31</td>
<td>.15</td>
</tr>
</tbody>
</table>

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