SMILE: A novel dissimilarity-based procedure for detecting sparse-specific profiles in sparse contingency tables

Mathieu Emily, Christophe Hitte, Alain Mom

SMILE: a novel Dissimilarity-based Procedure for Detecting Sparse-Specific Profiles in Sparse Contingency Tables.

Mathieu Emily[a,d], Christophe Hitte[b], Alain Mom[c,d]

[a] *Agrocampus Ouest, 65, rue de Saint-Brieuc, 35042 Rennes, France*
[b] *Université Rennes 1 - IGDR UMR CNRS 6290, 2 avenue du Professeur Léon Bernard, 35043 Rennes Cedex, France*
[c] *Université Rennes 2, Place du recteur Henri le Moal, 35043 Rennes, France*
[d] *IRMAR UMR CNRS 6625, 263 avenue du Général Leclerc, 35042 Rennes, France*

Abstract

A novel statistical procedure for clustering individuals characterized by *sparse-specific* profiles is introduced in the context of data summarized in sparse contingency tables. The proposed procedure relies on a single-linkage clustering based on a new dissimilarity measure designed to give equal influence to sparsity and specificity of profiles. Theoretical properties of the new dissimilarity are derived by characterizing single-linkage clustering using Minimum Spanning Trees. Such characterization allows the description of situations for which the proposed dissimilarity outperforms competing dissimilarities. Simulation examples are performed to demonstrate the strength of the new dissimilarity compared to 11 other methods. The analysis of a genomic data set dedicated to the study of molecular signatures of selection is used to illustrate the efficiency of the proposed method in a real situation.

Keywords: Dissimilarity, Sparse contingency table, Single-linkage clustering, Conditional profile.

1. Introduction

Let consider a two-way sparse contingency table that displays the number of occurrences of k categories for n individuals. This paper aims at detecting individuals with typical profiles of categories called *sparse-specific* profiles. *Sparse-specific* profiles, formally defined in Definition 2.1, are characterized by two main features. Firstly, the sparse profiles are those profiles for which only very few categories have non-zero counts. Secondly, specific profiles are those profiles presenting specific categories, i.e. categories that are (almost) never observed in the other individuals.

The detection of *sparse-specific* profiles is of interest in various application domains. For example, in genetics, *sparse-specific* profiles are expected to be encountered for breeds (*i.e.* a homogeneous group of domestic animals) under selection [1]. In that context, genetic data can be summarized in a two-way contingency table for which an individual is a breed and a category is a DNA sequence, also called haplotype, of a given chromosomal region. Existing methods for detecting signatures of selection rely on strong assumptions based on population genetic theory that cannot be verified. Therefore, detecting the signatures of selection remains challenging. Alternative statistical methods that are robust to population genetic models are needed to improve the detection of selection [2].

The observation of *sparse-specific* profiles is also expected in other contexts such as text-mining or ecology. In text-mining, collected data are usually stored in a term-document matrix that describes the frequency of terms that occur in a collection of documents [3]. Observing a *sparse-specific* profile for a document in a term-document matrix means that the document has very few different terms that are almost not used in the other documents. In ecology, collected data can be summarized in a site by a species matrix where the abundance of different species is measured in various sites [4]. In that context, a species with a sparse-specific profile is expected to be observed in only very few sites. Those sites are assumed to host very few species, thus characterizing a low species richness. Although *sparse-specific* profiles are likely to be targeted in many applications, their detection raises the issue of detecting a non-symmetric relationship between a set of individuals and a set of categories. Furthermore, the non-symmetric relationship is well-characterized for *sparse-specific* profiles. The main challenge in the detection of *sparse-specific* profiles indeed lies in taking into account sparsity and specificity simultaneously.

In order to group individuals in sparse contingency tables, hierarchical clustering techniques are widely used. As such, the detection of *sparse-specific* profiles can be performed through a similar approach. The quality of the clustering depends on the choice of (1) a dissimilarity measure between individuals and (2) a linkage criterion for the hierarchical clustering. As quoted in [5 - p.506], "Specifying an appropriate dissimilarity measure is far more important in obtaining success with clustering than choice of clustering algorithm". For that reason, attention was first focused on the choice of an appropriate dissimilarity to detect *sparse-specific* profiles. An abundant literature has been dedicated to improving the measure of similarity between individuals in sparse contingency tables, either by applying dimension reduction techniques or by proposing dissimilarity measures [6]. Dimension reduction techniques, such as Latent Semantic Indexing (LSI) [7] and Non-negative Matrix Factorization (NMF) [8] aim at transforming a high dimension space of features to a space of fewer dimensions using linear or non-linear combinations [5]. Applying such techniques can help selecting the most relevant categories, thus improving the quality of the clustering [9]. For instance, LSI and NMF techniques were successfully used prior to the clustering of textual data [6]. However, these techniques does not explicitly account for *sparse-specific* profiles in the reduction of the dimensionality of the feature space. As a consequence, power to detect sparse-specific profiles for dimension reduction techniques is likely to be limited. On the one hand, the similarity between individuals can be measured with many different functions developed to deal with sparse contingency tables. In the text domain, the most well known and commonly used similarity function is the cosine similarity function [10]. In ecology, dedicated dissimilarities are the Bray-Curtis dissimilarity, the Jaccard dissimilarity, the d_1^2 (or Manhattan) distance, the Hellinger distance or the Gower dissimilarity [11].

However, these methods usually work directly with counts which might not be appropriate to detect *sparse-specific* profiles. Indeed, heterogeneity in the marginal counts of individuals gives different weights to individuals, thus leading to inappropriate conclusions. A natural way to control weights given to individuals is to focus on the conditional distribution of the categories, also known as conditional profiles. The analysis of conditional profiles is classically performed by using either the χ^2 distance or the d_2^2 distance (also known as L^2 norm). Nevertheless, capturing *sparse-specific* profile with the χ^2 distance raises some limitations since χ^2 is sensitive to profiles specificities. On the other hand, it will be shown that d_2^2 distance between two profiles is

more influenced by the sparsity than the specificity.

In this paper, we propose a novel dissimilarity called d_s^2 adapted to the detection of *sparse-specific* profiles. d_s^2 is based on the comparison of conditional profiles and gives equal influence to sparsity and specificity of profiles, compared to other dissimilarities. To identify *sparse-specific* profiles, we propose a procedure called SMILE, for Statistical Method to detect sparse-specIfic profiLEs, which consists in a single-linkage hierarchical clustering [12] constructed using the d_s^2 dissimilarity. Selected profiles with the SMILE procedure correspond to the smallest subset of conditional profiles that coalesce at the final step of the hierarchical clustering.

In section 2, we formalize the definition of a *sparse-specific* profile and give details of the SMILE procedure. Furthermore, by considering the parallel between Minimum Spanning Trees and Single-Linkage Cluster Analysis [13], an original characterization of the structure of the individual subset selected by the SMILE procedure is proposed.

In Section 3, we illustrate the performance of the SMILE procedure in a simulation study. For that purpose, a simple simulation algorithm generating contingency tables with respect to sparsity and specificity features was designed. Power for the SMILE procedure is compared to the power of 11 other clustering methods in simulated scenarios highlighting the highest power for the dissimilarity measure d_s^2.

Section 4 is devoted to the application of the SMILE procedure on a real dataset dedicated to the detection of molecular signatures of selection in the domestic dog [14]. Comparing the SMILE procedure to 11 concurrent methods provides illustrative examples of the benefit of using the SMILE procedure for detecting *sparse-specific* profiles in sparse contingency tables.

2. The SMILE procedure

The SMILE procedure aims at detecting *sparse-specific* profiles. To do so, the proposed method selects the smallest subset of conditional profiles that coalesce at the final step of a single-linkage hierarchical clustering constructed with the d_s^2 dissimilarity. In this section, the approach driven by the features characterizing *sparse-specific* profiles is described.

2.1. Description of sparse-specific profiles

Let first consider a two-way contingency table with n individuals and k categories. For example, our illustrative example is $n = 30$ dog breeds for

which a total of k haplotypes have been observed in a given region. Each individual is therefore characterized by a vector of counts $[n_{i1}, \ldots, n_{ik}] \in \mathbb{N}^k$, where n_{ij} is the number of times category j is observed for individual i. Conditional profile for individual i is defined by a $k-$dimensional vector $x_i = [p_1^i, \ldots, p_k^i]'$ where $p_j^i = n_{ij}/n_{i.}$ and $n_{i.} = \sum_{j=1}^{k} n_{ij}$. In the following, E is used to denote the set of those n conditional profiles.

Definition 2.1. *Sparse-specific profiles are characterized by a set of selected individuals, called A, and a set of selected categories, called K. Categories in K are overrepresented for individuals in A and, simultaneously, individuals in A carry categories almost only in K. A sparse-specific profile is then defined by the two following features:*

[Profile sparsity]: for $i \in A$ and $j \notin K$, p_j^i are expected to be very low,
[Profile specificity]: for $i \notin A$ and $j \in K$, p_j^i are expected to be very low.

2.2. The dissimilarity d_s^2

In this section, a novel dissimilarity measure, called d_s^2, that is adapted to the detection of *sparse-specific* profiles is proposed. d_s^2 is designed to equally account for the sparsity and, as well, the specificity of the two profiles. More precisely, d_s^2 is defined by:

Definition 2.2. $\forall x, y \in E$*:*

$$d_s^2(x, y) = \|x\|_2 \, \|y\|_2 \, d_\theta^2(x, y) \tag{1}$$

where

$$d_\theta^2(x, y) = 2(1 - cos(\widehat{xy})) = 2\left(1 - \frac{\langle x, y \rangle_2}{\|x\|_2 \, \|y\|_2}\right) \tag{2}$$

is the square of the angular distance between the lines spanned by x and y, $\langle , \rangle_2$ is the L_2 scalar product and $\| \|_2$ its corresponding norm.

The sparsity of a given profile x is indeed measured by $\|x\|_2$ since the sparser x is, the higher $\|x\|_2$ is. To illustrate this point, let consider two individuals i and i' with profiles x_i and $x_{i'}$. Let assume that x_i and $x_{i'}$ are identical except for two categories j_1 and j_2 such that $n_{i'j_1} + n_{i'j_2} = n_{ij_1}$ and $n_{ij_2} = 0$. x_i is thus sparser than $x_{i'}$ and we have easily $\|x_i\|_2 \geq \|x_{i'}\|_2$. Finally, if individual i has $n_{i.}$ non null coordinates, that are thus all equal to one, the norm of the conditional profile, given $n_{i.}$, is minimum and equals to $\|x_i\|_2 = \sqrt{\frac{1}{n_{i.}}}$.

The specificity between two profiles is measured by the angular distance d_θ. For example, if $\langle x, y \rangle_2 = 0$, x and y have the greatest angular distance $d_\theta^2(x, y) = 2$. If otherwise, x and y are identical, their angular distance is the lowest, $d_\theta(x, y) = 0$.

Thus, from Equation 1, it can be remarked that d_s^2 gives the same importance to the sparsity and the specificity since $0 \leq \|x\|_2 \|y\|_2 \leq 1$, $0 \leq d_\theta^2(x, y)/2 \leq 1$ and multiplying the dissimilarities by the same scalar does not modify the clustering. We can further note that the definition of d_s^2 in Equation 1 can be extended to account for different weights given to sparsity and specificity. To this aim, in Appendix E, we introduced a family of weighted dissimilarity measures, d_α^2, where $\alpha \in [0, 1]$ and $1 - \alpha$ are the weights given to sparsity and specificity, respectively. Since $d_s^2 = d_{\alpha=\frac{1}{2}}^2$, it is straightforward to prove that d_s^2 is a member of the family d_α^2. The results shown in section Appendix E demonstrate that $\alpha = 1/2$ is a reasonable choice to maximize the power of detecting *sparse-specific* profiles. Furthermore, the use of the d_α^2 family in practical situations requires the estimation of α, which goes beyond the scope of this article. For all those reasons, we focus on the dissimilarity d_s^2 in the remainder of the paper.

According to Equation 2, d_s^2 can further be reformulated as:

$$\forall x, y \in E, d_s^2(x, y) = 2(\|x\|_2 \|y\|_2 - \langle x, y \rangle_2). \tag{3}$$

It can easily be remarked that d_s^2 is symmetric since the scalar product is. Moreover, according to the Cauchy-Schwarz inequality, $\forall x, y \in E$ $d_s^2(x, y) \geq 0$, thus proving that d_s^2 is actually a dissimilarity.

2.3. Single-linkage detection

The SMILE procedure is based on single-linkage detection, where single-linkage detection corresponds to the selection of the smaller of the two subsets linked at the final step of a single-linkage hierarchical clustering constructed with some dissimilarity d. In the single-linkage hierarchical clustering, the linkage criterion between two clusters C_i and C_j is defined by $d(C_i, C_j) = min_{x \in C_i, y \in C_j} d(x, y)$ [12]. The clustering is then obtained by iteratively merging the pair of clusters that minimizes the single linkage criterion.

The single-linkage method has a tendency to form long and straggly clusters. This phenomenon, often known as "chaining phenomenon", refers to the gradual growth of a cluster as one element at a time gets added to it [15].

Considered as a potential drawback in some practical situations, the chaining effect is an advantage in our situation by allowing the separation of two highly distinct groups. Another advantage of using the single-linkage criterion is that the dissimilarities between clusters are the original dissimilarities between individuals: dissimilarities remain unchanged during the clustering, preserving their good properties, if any, all through the study. Thus, the single-linkage criterion is adapted to separate individuals with *sparse-specific* profiles from the rest of the population.

2.4. Benefit of the d_s^2 in single-linkage detection

In this section, some properties of d_s^2 in single-linkage detection thanks to the study of Minimum Spanning Tree (MST) are investigated. In graph theory, a spanning tree is a subgraph of a connected undirected graph $\mathcal{G}$ with n vertices that connects all the vertices together with $n-1$ edges. If weights are assigned to each edge of $\mathcal{G}$ then each spanning tree can also be weighted by summing the weights of its edges. A MST is then a spanning tree with weight less than or equal to the weight of every other spanning tree. In our context, we can consider that vertices of $\mathcal{G}$ are the individuals and that the weight between two individuals is the value of the dissimilarity between these two individuals.

In the remainder of this section, an original characterization of the structure of the individual subset selected by the single-linkage detection is first proposed in Theorem 2.1 by considering the parallel between single-linkage clustering trees and MST [13]. Such characterization is valid for any dissimilarity measure and helps in understanding the roles played by A and $\bar{A}$ in single-linkage clustering. The link between d_2^2 and d_s^2 is then stated. Finally, by focusing on the case where A is a singleton, situations for which our dissimilarity, d_s^2, is more powerful than d_2^2 are described in Theorem 2.3.

Impact of the structure of the data on single-linkage detection.

The main goal of this section is to find a necessary and sufficient condition, formulated in Theorem 2.1, for a subset A, with n_A individuals, to be selected at the final step of the single-linkage detection based on a dissimilarity d. Since the set of weights of the edges which leads to the MST is the set of the levels of the hierarchical clustering merger using the single-linkage criterion, the structure of the clusters could be studied by characterizing the MST built on d.

Among the procedures developed to construct the MST, the Kruskal algorithm was chosen [16]. In our context, let consider the complete undirected graph, where the nodes are all individuals in E and edges are weighted according to the dissimilarity between the two corresponding connected nodes. As n individuals belong to E, the complete undirected graph is made by $n(n-1)/2$ edges, or, equivalently, dissimilarities. The Kruskal algorithm starts by ranking the $n(n-1)/2$ dissimilarities in increasing order. At the first step, it chooses the two smallest dissimilarities. Then, it successively selects the smallest dissimilarity, for which the corresponding edge does not create cycle in the current graph. Thus, the way each individual of A is clustered with the other individuals of E is completely defined by the $n-1$ dissimilarities picked up by the Kruskal algorithm. To characterize the clustering of the subset A, we define the quantity $d^{MST}(A)$ as follows:

Definition 2.3. *Let A be a subset of n_A individuals and let consider the $n-1$ edges selected by the Kruskal algorithm in the increasing order. $d^{MST}(A)$ is equal to the length of the $(n_A-1)^{th}$ edge obtained by restricting to edges where at least one node belongs to A.*

Let $\overline{A}$ be the complementary of A. $d^{MST}(\overline{A})$ is defined in the same way by replacing A by $\overline{A}$ and n_A by $n-n_A$ in Definition 2.3. Appendix C provides two illustrations of the calculation of $d^{MST}(A)$ and $d^{MST}(\overline{A})$.

Definition 2.4. *Let A be a subset of E. A is said to be a Kruskal-connected component if there exists a step of the Kruskal algorithm for which A is a connected component of the current graph.*

Lemma 2.1. *A is a Kruskal-connected component if and only if:*

$$d^{MST}(A) < min_{x\in A, y\in\overline{A}} d(x,y) \tag{4}$$

The proof of Lemma 2.1 is given in Appendix B. Furthermore a graphical interpretation of inequality (4) is proposed in Appendix C. Theorem 2.1 that provides a characterization for A to be selected by single-linkage detection can now be set.

Theorem 2.1. *Let A be a subset of E with n_A elements and d be a dissimilarity. A is linked to its complementary $\overline{A}$, at the final step of the hierarchical clustering based on d and using the single-linkage criterion if and only if A and $\overline{A}$ are the two last connected components at the final step of the Kruskal*

algorithm. Considering that $n_A < n - n_A$, A is thus selected by single-linkage detection if and only if:

$$max(d^{MST}(A), d^{MST}(\overline{A})) < min_{x\in A, y\in\overline{A}} d(x,y) \tag{5}$$

Proof of Theorem 2.1 is given in Appendix B. Inequality 5 is also illustrated in Appendix C. According to Theorem 2.1, the structures of A and $\overline{A}$ both play a major role in the selection of A by single-linkage detection. In practice, individuals in the targeted subset A are usually homogeneous sharing all similar *sparse-specific* profiles. However, individuals in $\overline{A}$ are likely to be heterogeneous which might prevent single-linkage detection from correctly selecting A. Thus, to control heterogeneity in $\overline{A}$, the choice of an appropriate dissimilarity is crucial. Theorem 2.1 will also be used, in the next paragraph, to interpret the influence of $\bar{A}$ on the performances of d_s^2 compared to the L_2 norm. Furthermore, the simulated scenarios proposed in Section 3 are based on conclusions drawn by Theorem 2.1.

Link between d_2^2 and d_s^2.

The d_2^2 dissimilarity is defined as the square of the L_2 norm of the difference between two conditional profiles. It can be formulated for each pair of individuals as follows:

$$\forall x, y \in E,\ d_2^2(x,y) = \|x - y\|_2^2 .$$

Thus, according to Equation 3, it can deduced that:

$$d_2^2(x,y) = d_s^2(x,y) + (\|x\|_2 - \|y\|_2)^2. \tag{6}$$

It is noteworthy that $d_2^2(x,y) = d_s^2(x,y)$ if and only if $\|x\|_2 = \|y\|_2$. Moreover, compared to d_s^2, the d_2^2 dissimilarity gives more weight to variation in sparsity between the 2 compared profiles by adding the term $(\|x\|_2 - \|y\|_2)^2$. Thus, compared to d_s^2, d_2^2 is likely to be more sensitive to the heterogeneity, and especially heterogeneity in sparsity, in a given subset.

In the context of *sparse-specific* profile detection, such sensitivity to heterogeneity can be a drawback for d_2^2. Although the targeted subset A is assumed to be homogeneous, meaning that all individuals in A are supposed to have very similar conditional profiles, its complementary, $\overline{A}$, can display very diverse patterns. Assuming that profiles in $\overline{A}$ are highly heterogeneous with some profiles with high sparsity and other with low sparsity, the d_2^2 dissimilarity between two profiles can be high. As a consequence, $\overline{A}$ might

fail at being a Kruskal-connected component with the d_2^2 thus preventing the detection of A (see Theorem 2.1). Therefore, the detection of a targeted subset A with single-linkage detection is less influenced by the structure of $\overline{A}$ when using d_s^2 compared to d_2^2.

The $n_A = 1$ case.

The $n_A = 1$ case is of particular interest since it corresponds to the situation where only one individual has a *sparse-specific* profile. In the illustrative example, such a situation is often encountered. Two main results are given in this section. First, Theorem 2.2 gives sufficient condition in sparsity and specificity for the sparsest profile to be selected by single-linkage detection with d_s^2 or d_2^2. The second main result, given in Theorem 2.3, shows the benefit of using d_s^2 instead of d_2^2 when the targeted individual has the sparsest conditional profile.

It is noteworthy that, when $n_A = 1$, A is a singleton and Inequality (5) becomes:

$$d^{MST}(\overline{A}) < min_{y \in \overline{A}} d(x, y) \text{ where } A = \{x\} \tag{7}$$

To be selected by single-linkage detection, x should hence have the furthest conditional profile according to some dissimilarity d. To understand the role of profile sparsities in d_2^2 and d_s^2, let introduce the sparsest and the least sparse conditional profiles as follows:

Definition 2.5. *The sparsest conditional profile, x_s, and the least sparse conditional profile, x_0, are defined, for all $x \in E$ such that $x \neq x_s$ and $x \neq x_0$, by:*

$$\|x_0\|_2 < \|x\|_2 < \|x_s\|_2 .$$

Let also formalize a hierarchy in the specificity of a profile by defining a totally specific conditional profile (see Definition 2.6) and a nearly totally specific conditional profile (see Definition 2.7).

Definition 2.6. *$x \in E$ is said to be totally specific if and only if: $\forall y \neq x \in E, \langle x, y \rangle_2 = 0$.*

Definition 2.7. *$x \in E$ is said to be nearly totally specific for the dissimilarity d if and only if $\langle x, x_0 \rangle_2 = 0$ and $d(x, x_0) = min_{y \neq x} d(x, y)$.*

It is noteworthy that Definition 2.7 means that the specificity, as defined in Definition 2.1 is not strong enough for modifying the profile ordering imposed by the sparsity.

The hierarchy between totally specific and nearly totally specific is obvious when considering x_s, the sparsest conditional profile, as mentioned in the following lemma.

Lemma 2.2. *If x_s is totally specific, then x_s is nearly totally specific for the dissimilarity $d \in \{d_2^2, d_s^2\}$.*

In the following interest is focused on nearly totally specific profiles for x_s since, according to Lemma 2.2, results are valid for totally specific profiles as long as they are valid for nearly totally specific profiles. Let now formulate two lemmas that are preliminary results needed for Theorems 2.2 and 2.3.

Lemma 2.3. *If x_s is nearly totally specific for $d \in \{d_s^2, d_2^2\}$ then we have $\forall x \neq x_s \in E$:*

$$d(x_0, x) < d(x_0, x_s).$$

Lemma 2.4. *Let consider the k^{th} step of the hierarchical clustering with dissimilarity $d \in \{d_s^2, d_2^2\}$. At the k^{th} step, individuals in E are clustered into $n - k + 1$ Kruskal-connected components. Furthermore, for $x \in E$, $d(x, C_\ell) = min_{y \in C_\ell} d(x, y)$ defines, at this step, the dissimilarity between any individual $x \in E$ and any Kruskal-connected component, called C_ℓ. Let also set C_x as the Kruskal-connected components at the k^{th} step that contains $x \in E$.*

If x_s is nearly totally specific for d, the closest Kruskal-connected component to x_s, according to d, is C_{x_0}:

$$d(x_s, C_{x_0}) = d(x_s, x_0) \leq d(x_s, C_\ell), \forall C_\ell \neq C_{x_s}.$$

Theorem 2.2. *If x_s is nearly totally specific for the dissimilarity $d \in \{d_s^2, d_2^2\}$ then x_s is selected by single-linkage detection.*

Theorem 2.3. *If x_s is nearly totally specific for d_2^2 then x_s is nearly totally specific for d_s^2.*

If x_s is nearly totally specific for d_2^2, then it is selected by single-linkage detection using d_2^2. Theorems 2.2 and 2.3 thus ensure that x_s is selected by single-linkage detection using d_s^2. It is noteworthy that the reciprocity of Theorem 2.3 is false. Conditional profiles selected by d_s^2 might be missed by d_2^2, thus proving the benefit of using d_s^2 instead of d_2^2. This advantage for d_s^2 is further highlighted by the analysis of Region ♯6 in our illustrative example shown in Section 4.2.

3. Simulation study

To evaluate the performance of the SMILE procedure, we compare our novel dissimilarity d_s^2 to 11 other measures in single-linkage detection. We propose an algorithm for the simulation of contingency tables designed to simulate the structure of the set of selected individuals, A, and its complementary $\overline{A}$. Since the structure of $\overline{A}$ plays a major role in the detection of A (see Theorem 2.1), simulated scenarios that focus on the impact of the structure of $\overline{A}$ on single-linkage detection are designed. On the one hand, a simulated scenario, Scenario ♯1, is considered where specific and non-sparse profiles are observed in $\overline{A}$. On the other hand, the impact of heterogeneity in sparsity in analyzed for profiles in $\overline{A}$ thanks to two simulated scenarios, Scenario ♯2 and ♯3.

3.1. Simulation algorithm

In this section, the algorithm used to simulate the different scenarios is set out in details. Let first assume that individuals are clustered into three subsets, A, B and C, with respective sizes of n_A, n_B and n_C. It is also supposed that categories are divided into three groups, K, L_1 and L_2 with sizes given by k_1, ℓ_1 and ℓ_2. Conditional profiles for individuals in A are parameterized by p and p^*, where p is the probability of a category in K and p^* the probability of a category in L_1 and L_2. Regarding individuals in B (resp. C), conditional probabilities for categories in L_1 are given by q_1 (resp. q_2) and for categories in K and L_2 are denoted by q_1^* (resp. q_2^*). The whole set of parameters is summarized in Table 1. To mimic the illustrated example of this paper, n is considered as fixed and equal to $n = 30$. The number of categories is assumed to be equal to $k = 200$ throughout the simulation study. However, since k and n gives the dimensionality of the contingency table, they both may have an impact on the power of detection. To investigate the roles played by n and k in the comparison of the methods, other values of n and k have been tested in Appendix D. The results obtained in Figures D5-D10 show the robustness of our conclusions with respect to n and k.

Simulation of contingency tables is then performed by simulating counts, for each individual, according to a multinomial law with parameters m and the individual probability profile vector, where m corresponds to the number of observations per individual. In order to mimic the mean value observed

Table 1: Summary of the conditional profiles for the n simulated individuals. The cardinal of each set is given in parenthesis such that $n_A + n_B + n_C = n = 30$ and $k_1 + \ell_1 + \ell_2 = k = 200$.

	$K(k_1)$	$L_1(\ell_1)$	$L_2(\ell_2)$
$A(n_A)$	p	p^*	p^*
$B(n_B)$	q_1^*	q_1	q_1^*
$C(n_C)$	q_2^*	q_2	q_2^*

in the real dataset analyzed in section 4, it is assumed to be constant and equal to 30 for all individuals across all simulation scenarios.

Such algorithm has been used to simulate data under three main scenarios, Scenario ♯1, ♯2 and ♯3. Parameter values for each scenario are summarized in Table 2 and explain in detail in section 3.3 for Scenario ♯1 and section 3.4 for Scenarios ♯2 and ♯3.

Table 2: Summary of the values of the parameters used in the three simulated scenarios.

	n_A	n_B	n_C	k_1	ℓ_1	ℓ_2	p	q_1	q_2
Scenario ♯1	1	1	28	1	100	99	1	Not Fixed	0
Scenario ♯2	3	10	17	1	100	99	Not Fixed	0.005	0.005
Scenario ♯3	3	10	17	1	2	197	Not Fixed	0.45	0.1

3.2. Computational details for the compared dissimilarity measures

This section gives some details regarding the computation of the 11 dissimilarities that were compared to d_s^2. Additional information is given in Appendix A.

The R package `ecodist` [17] is used to compute the following dissimilarity metrics based on counts data: Bray-Curtis, d_1^2 (or Manhattan), Jaccard and Gower. Regarding distances based on conditional profiles, either our own functions for χ^2 and d_2^2 distances are proposed or existing R packages such as `vegan` package [11] for the Hellinger dissimilarity and `lsa` package [18] for the cosine distance are used.

d_s^2 is further compared to three reduction dimension techniques dedicated to the analysis of sparse contingency tables. First, `HierarchicalSparseCluster` function from package `sparcl` that implements the method, based on a Lasso-like penalty, developed in [9] was used. Secondly, the `lsa` package [18] to

calculate the latent semantic space, that is based on singular value decomposition of a sparse contingency table was employed. To choose an appropriate number of singular values for the dimensionality reduction in LSA, the `dimcalc_share` function as suggested in `lsa` package was used. Note that other numbers of singular values did not give better results.

Finally, non-negative matrix factorization with the R package `nmf` [19] was done. The quality of the clustering with `nmf` depends on two main choices: the choice of an algorithm for the factorization and the choice of the rank of the factorization. For ease of reading, only results obtained with the standard NMF, based on Euclidean distance [8] are presented. Other algorithms, such as standard NMF based on Kullback-Leibler divergence or Alternating Least Square did not provide better results. Regarding the choice of the rank, several values were tested and the results obtained with the most powerful rank are presented.

3.3. Specific and non-sparse profile in $\overline{A}$ - Scenario ♯1

The goal of this section is to compare the ability of each method to detect *sparse-specific* profiles when specific and non-sparse profiles are observed in $\overline{A}$.

A specific and non-sparse profile displays many categories that are not observed in other individual profiles. To simulate the presence of one such profile in $\overline{A}$, B is first assumed to be a singleton with a non-sparse profile by setting $n_B = 1$ and $\ell_1 = 100$. q_2 is further fixed to 0 so that the specificity of the profile in B is governed by q_1. The specificity of the profile in B indeed increases as q_1 becomes closer to 1. Thus, to evaluate the impact of the specificity q_1 is made to vary from 0 to 1. Furthermore, for ease of reading, A and K are singletons (*i.e.* $n_A = 1$ and $k_1 = 1$). It is also assumed that that $p = 1$ leading to the sparsest possible profile in A. Parameter values are summarized in Table 2.

Results are displayed in Figure 1, where empirical power was estimated from 1,000 Monte-Carlo simulations. It can first be remarked that two methods, Gower and Cosine, have no power in detecting *sparse-specific* profiles. Furthermore, when q_1 is small, *i.e.* when the conditional profile in B is not specific, all other methods show equivalent power to detect A.

However, when q_1 is larger, power for Hellinger, χ^2, Jaccard, d_1^2 and Bray-Curtis drops to zero, thus demonstrating that these methods are very sensitive to the presence of at least one specific and non-sparse profile in $\overline{A}$. Although its power does not vanish to zero, the LSA method also shows a

similar pattern, proving a weakness to distinguish between sparse and non-sparse profiles. On the contrary, the novel dissimilarity d_s^2, as well as d_2^2, NMF and Sparcl keep very high power, regardless of the value of q_1 (*i.e.* the specificity of the profile in B).

It is noteworthy from Figure 1 that d_s^2 and d_2^2 always display a power of 1. With these parameter settings, the conditional profile in A has the highest sparsity and is simultaneously totally specific in the sense of Definition 2.6. Thus these results confirm that, according to Theorem 2.3, single-linkage detection with d_s^2 and d_2^2 always detects A as a *sparse-specific* profile.

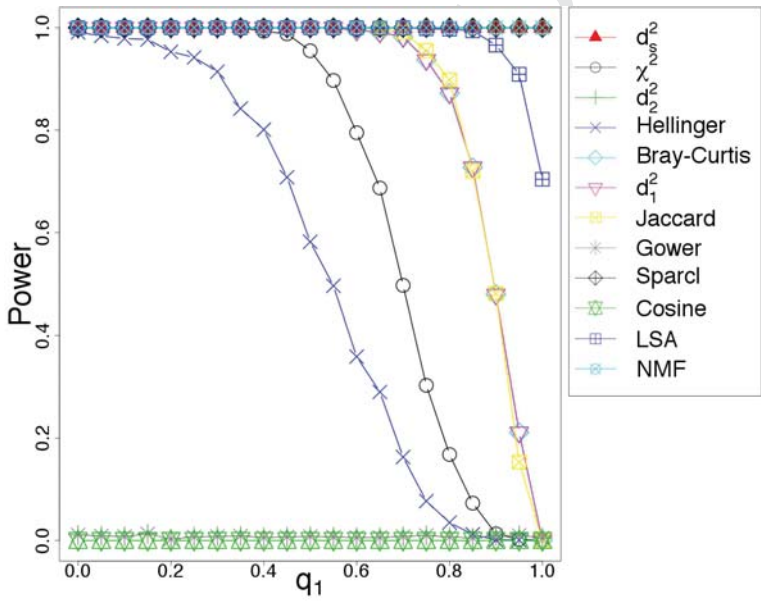


Figure 1: Evaluation of power with respect to q_1 under Scenario ♯1.

3.4. Variation in sparsity within $\overline{A}$ - Scenario ♯2 and ♯3

The aim of this section is to illustrate the benefit of using d_s^2 for selecting *sparse-specific* profiles with single-linkage detection. According to Equation 6, d_2^2 is likely to be more sensitive than d_s^2 to high heterogeneity in sparsity within $\overline{A}$. To highlight such property, power is evaluated in two scenarios: a first scenario with low variation in sparsity within $\overline{A}$ (Scenario ♯2) and a second scenario displaying high variation in sparsity within $\overline{A}$ (Scenario ♯3).

For both scenarios, the structure of A is fixed by setting $n_A = 3$ and $k_1 = 1$. Note that similar conclusions would be obtained with other values for n_A, such as $n_A = 1$. In Scenario ♯2, by setting $q_1 = q_1^* = q_2 = q_2^* = 1/k$, it is assumed that all individuals in $\overline{A}$ have the same conditional profile and without distinction between B and C. As a consequence, variation in sparsity for 2 individuals within $\overline{A}$ is expected to be very small compared to the variation between an individual in A and one in $\overline{A}$. In Scenario ♯3, it is assumed that the profiles of individuals in B and C differ, meaning that $\overline{A}$ is composed of two distinguishable subgroups of individuals. To this end, the size of each subgroup is set to $n_B = 10$ and $n_C = 17$. Further, let set $\ell_1 = 2$, $q_1 = 0.45$ and $q_2 = 0.1$. Therefore, the difference in L_2 norm between individuals in B and individuals in C is likely to be high, thus increasing the d_2^2 dissimilarity between individuals within $\overline{A}$. A summary of the values of the parameters is given in Table 2.

Results displayed in Figure 2 show that d_s^2, d_2^2 and Sparcl are the three most powerful methods with approximatively the same power when all individuals in $\overline{A}$ share the same conditional profile. NMF, followed by Bray-Curtis and d_1^2, are the three next most powerful methods. The five methods, Jaccard, χ^2, LSA, Hellinger and Gower, lack in power when the sparsity in A is not very high (*i.e.* , when $p < 0.8$). Finally, Cosine distance appears to have no power for detecting *sparse-specific* profile.

Results showed in Figure 3 proved that a modification in the structure of $\overline{A}$ does not impact the power for the novel dissimilarity d_s^2. However, power is highly reduced for d_2^2 and even more for Sparcl, thus demonstrating their sensitivity to important differences in sparsity between individuals within $\overline{A}$. Regarding the other methods, it can be remarked that Bray-Curtis, d_1^2, Hellinger, LSA, Gower, have a reasonable power, especially when p is low (see Figure 3). NMF seems to keep roughly the same power regardless of the structure in $\overline{A}$. Finally, Cosine, χ^2 and Jaccard distances, show a very low power. Indeed, because of the sparsity of the simulated tables, few conditional profiles in $\overline{A}$ are very likely to be specific and non-sparse, thus explaining the poor detection power.

Thus, the results demonstrate that the SMILE procedure is adapted to the detection of *sparse-specific* profiles. The novel dissimilarity d_s^2 is indeed the most powerful in the three simulated scenarios. Furthermore, power for d_s^2 is not impacted by the structure in $\overline{A}$, which comes from the fact that d_s^2 gives equal influence to sparsity and specificity of profiles.

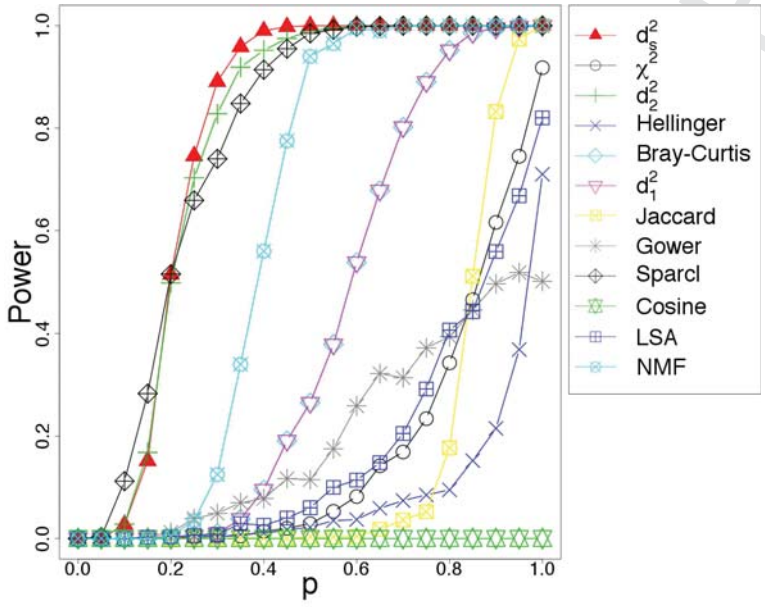


Figure 2: Empirical power with respect to p, the conditional probability of a category in K for an individual in A. Empirical power is estimated assuming that there is no distinction between individuals in $\overline{A}$ ($q_1 = q_2 = q_1^* = q_2^* = 1/k$) as described under Scenario ♯2.

4. Illustrative example

The real dataset chosen to illustrate the SMILE procedure in this paper consists of genomic data from the European consortium LUPA [14]. To study the genetic background of domestic dogs, the LUPA consortium generated molecular data covering the entire genome [20] consisting of 174,000 SNPs (Single Nucleotide Polymorphims). Data from a previous study designed to investigate the genetic background of 30 dog breeds is used [20]. Thus, in the dataset, the set of individuals consists in $n = 30$ individuals so that an individual, at the statistical level, is a dog breed. For each individual (or breed), the number of observations is the number of dogs from that breed in the sample, as shown in Supplementary Table F.4.

Attention is focused to six genomic regions, Regions ♯1, ♯2,...,♯6, defined as small parts of the genome. For each of the six regions of interest, the set of categories are defined as the set of observed DNA sequences, also called the set of haplotypes. The set DNA sequences was obtained using the software FastPHASE [21]. Each region is then characterized by a two-way contingency

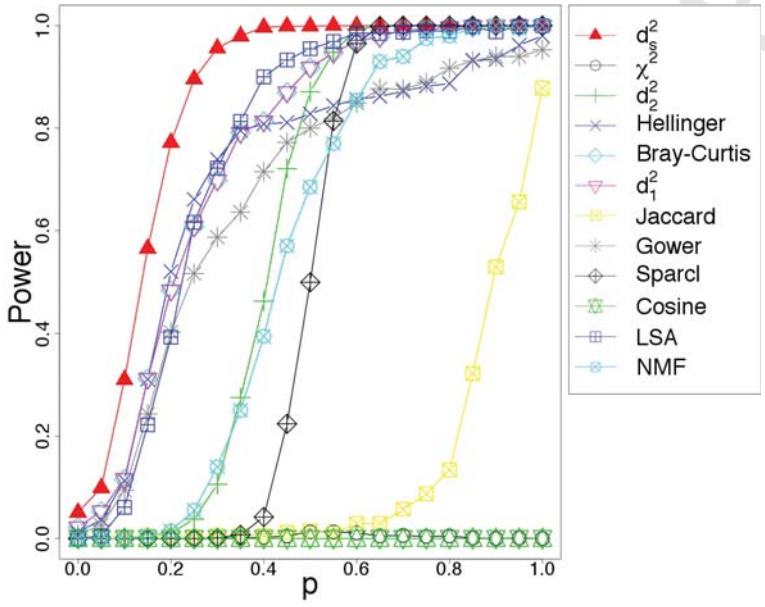


Figure 3: Empirical power with respect to p, the conditional probability of a category in K for an individual in A. $\overline{A}$ is supposed to be divided into two subgroups of individuals ($q_1 = 0.45$ and $q_2 = 0.1$) as described under Scenario ♯3.

table where an individual is a breed and a category is a single DNA sequence. Thus, cell (i, j) gives the number of times the DNA sequence j is observed in dogs from the breed i. It is noteworthy that the number of categories (or DNA sequences), called k throughout this paper, is different from one region to another, as shown in Supplementary Table F.5.

The six regions of interest have been chosen as being previously reported causative for the following morphological traits: brachicephaly, furnishings, wrinkled skin, periodic fever syndrome, chondrodysplasia and curly hair. For each region, genetic studies have shown that the presence or absence of a particular DNA sequence (or category) is associated with the observation of the trait. For example, if we consider Region ♯3 associated with the "wrinkled skin" trait, the only breed with wrinkled skin is the Shar-Pei (ShP) so that the set A is a singleton ($n_A = 1$). Furthermore, genetic studies reveal that the presence of two DNA sequences (or categories) are responsible for the observation of wrinkled skin while the absence of these two DNA sequences leads to the observation of a non wrinkled skin [22]. Thus, from the genetic

study, it can be deduced that k_1, that refers to the number of categories of the *sparse-specific* profile throughout the paper, is equal to $k_1 = 2$. Similar information is known for each region according to various genetic studies dedicated to each trait. Additional biological knowledge, such as the number of the chromosome from which each region is taken and the gene related to that region, can also be obtained from genetic studies. Supplementary Table F.5 provides a summary of what is known for each of the six regions of interest.

These six genomic regions were considered as test regions to compare the ability for the novel dissimilarity d_s^2 and the 11 methods described in Section 3.2 to detect known signals with single-linkage detection. For each of the six regions of interest, the true signal, *i.e.* the breed(s) that is(are) under selection, is(are) known. Thus, for the novel dissimilarity d_s^2 and the 11 competitive methods, we first evaluate the set of breeds detected in each region and then compare it with the known set of breeds that should have been detected according to biological knowledge. The performance of each method, given in Table 3, show that the proposed dissimilarity d_s^2 was the only method able to correctly detect 5 regions. All other methods also failed at finding the only missed region by d_s^2, (Region ♯2), thus demonstrating the strength of d_s^2 in a real situation.

The following sections aim at drawing a parallel between results obtained on the real data set analysis and more general results on the SMILE procedure as detailed in Section 2 and results obtained in the simulation study in Section 3. The low power observed for several dissimilarities is first explained. Regions with only one selected breed, corresponding to $n_A = 1$ case detailed in section 2.4 are then focused on. Finally, results are discussed regarding regions with more than one targeted breed.

4.1. Low power for Jaccard, Bray-Curtis, d_1^2, χ^2, Hellinger, Gower and Cosine

The 7 methods, Jaccard, Bray-Curtis, d_1^2, χ^2, Hellinger, Gower and Cosine, show a very limited power in detecting true signals. These results are in agreement with the simulation study proposed in section 3. Indeed, the analysis of a real dataset confirm that Cosine and Gower have almost no power as displayed in Figure 1. Furthermore, power for Hellinger, χ^2, Jaccard, Bray-Curtis and d_1^2, is expected to vanish to zero in the presence of specific and non-sparse profile in $\overline{A}$ (see Figure 1).

Table 3: Summary of the results obtained for d_s^2 and the 11 compared methods on the six genomic regions used as test regions to validate the method. A x means a correct detection of the signal known from biological experiments. The last column gives the total number of signals correctly found by the corresponding method.

Method	Regions						Total
	♯1	♯2	♯3	♯4	♯5	♯6	
d_s^2	x		x	x	x	x	5
d_2^2	x		x	x	x		4
Sparcl	x		x	x	x		4
NMF	x		x	x			3
LSA	x			x			2
Jaccard			x				1
Bray-Curtis			x				1
d_1^2			x				1
χ^2			x				1
Hellinger			x				1
Gower							0
Cosine							0

For example, low power for χ^2 can be explained by the presence of several breeds showing specific profiles with low sparsity. The observed bias for χ^2 towards specific and non-sparse individual profiles is clearly illustrated by results obtained for Region ♯2. In Region ♯2, χ^2 selected the Beagle (Bgl). In that region, Bgl conditional profile is totally specific, as all categories observed in Bgl are not observed in any other breed. Furthermore, conditional profile for Bgl displays a very low number of zeros leading to a non-sparse profile. Similar arguments can be used for the other regions, thus explaining low power for χ^2, Jaccard, Bray-Curtis, d_1^2, and Hellinger, in our real data example, as shown in the simulated scenario in Section 3.3.

4.2. Genomic regions with one targeted breed

Four regions are characterized by only one breed selected for a phenotypic trait, *i.e.* A is a singleton. These 4 regions (Regions ♯1, ♯3, ♯4 and ♯6), are representative examples of three types of signal theoretically described in the paragraph "The $n_A = 1$ case" in section 2.4.

The first type, observed in Region ♯3, is defined by an ideal *sparse-specific* profile. More precisely, the known selected breed, ShP, is both totally specific

and the sparsest breed. In agreement with Theorem 2.2, d_2^2 norm and d_s^2 correctly select ShP as a *sparse-specific* profile. Furthermore, as expected from the first simulated scenario in Figure 1, Region ♯3 is also correctly detected by Sparcl, NMF and LSA. Finally, it can be remarked that this region is the only region correctly detected by Jaccard, Bray-Curtis, d_1^2, χ^2 and Hellinger. ShP is indeed the only breed for Region ♯3 with a totally specific conditional profile.

The second type of signal, observed in Regions ♯1 and ♯4, illustrates a sparse and nearly totally specific profile for d_2^2 and d_s^2. In Region ♯1, the known selective breed EBD is indeed the sparsest breed. Although EBD shares categories with other breeds, its closest breed is the least sparse breed for d_2^2 and d_s^2. According to Definition 2.7, EBD is nearly totally specific for d_2^2 and d_s^2 and is correctly selected, as a consequence of Theorem 2.2. Similar conclusions are obtained for ShP in Region ♯4. Results obtained for Region ♯1 and Region ♯4 are also illustrative examples of Theorem 2.3 that stipulates that, if the sparsest conditional profile is nearly totally specific for d_2^2 then it is also nearly totally specific for d_s^2. Furthermore, it can be remarked that Sparcl and NMF correctly detected both regions which is in agreement with their high power displayed in Figure 2 in the simulation study.

Region ♯6 displays a third typical situation where the selective signal illustrates the benefit of using d_s^2 compared to d_2^2 in single-linkage detection. On the one hand, the validated breed, StP, is the sparsest breed for Region ♯6. Moreover, StP profile is not totally specific since one category observed in StP is present in other breeds. On the other hand, the least sparse breed TYo is also the closest breed to StP with respect to d_s^2, which means that StP is nearly totally specific for d_s^2. Thus, according to Theorem 2.2, single-linkage detection with d_s^2 is able to correctly detect StP. However, regarding d_2^2 dissimilarity, since the closest breed to StP is not the least sparse breed, StP is not nearly totally specific. Consequently, single-linkage detection with d_2^2 norm is not sure to correctly detect StP and actually wrongly selects GoS. It can further be remarked that $\overline{A}$, where $A = \{StP\}$, is highly heterogeneous in sparsity for Region ♯6. Conditional profile for GoS is indeed highly sparse compared to other profiles in $\overline{A}$. Thus, because of the strong importance of the variation in sparsity in the d_2^2 dissimilarity, GoS is wrongly selected by d_2^2, as illustrated in the second simulated scenario in Section 3.4. Moreover, the fact that d_s^2 is the only method able to correctly detect Region ♯6 is enhanced by the simulation results obtained in Figure 3.

4.3. Genomic regions with more than one targeted breed

In our dataset, Dac and TYo are the only two breeds affected by Chondrodysplasia, characterized by Region ♯5. Results in Table 2 show that single-linkage detection with d_s^2, d_2^2 and Sparcl correctly selects A ={Dac, TYo}. Dac and TYo have very similar conditional profiles and are considered as the two closest breeds for dissimilarities d_s^2, d_2^2 and Sparcl. Heterogeneity in conditional profiles and sparsity in $\overline{A}$ are relatively small, so that $\overline{A}$ is considered as a Kruskal-connected component for d_s^2, d_2^2 and Sparcl. However, single-linkage detection with the other methods fails at correctly selecting A. For example, for the χ^2 distance, although A is considered as a Kruskal-connected component after two steps of the Kruskal algorithm, $\overline{A}$ is not a Kruskal-connected component. Using the χ^2 dissimilarity, EBD is indeed close enough to Dac and TYo to coalesce with A before all breeds in $\overline{A}$ clustered. The results observed for Region ♯5 are illustrated by the results obtained in Figure 2 in the simulation study.

Five breeds are concerned by the signal in the Region ♯2: A ={BoT, IrW, JRT, StP, TYo}. None of the dissimilarities were able to detect A with single-linkage detection. The main reason for such common failure is that A is actually not homogeneous since A can be divided into 3 subsets: {IrW, StP}, {BoT, TYo} and {JRT}. It is noteworthy that subset {IrW,StP} is a Kruskal-connected component for all dissimilarities except Jaccard and Cosine. For χ^2, it can be further remarked that one element in $\overline{A}$, NFd, is close to {IrW,StP} simply because NFd has a low sparsity. Proximity between NFd and {IrW, StP} is not a consequence of similitude in terms of sparsity and specificity, thus proving that χ^2 is not adapted to the detection of molecular signatures of selection. Regarding d_2^2, heterogeneity in sparsity in $\overline{A}$ is high enough to exclude one breed, Dob, thus preventing $\overline{A}$ to be a Kruskal-connected component. Conversely, by putting less weight on the variation in sparsity, d_s^2 is able to detect the subset {IrW,StP}, demonstrating that d_s^2 is more adapted than d_2^2 to detect molecular signatures of selection.

5. Discussion

In this paper, a statistical approach, called SMILE, for detecting *sparse-specific* profiles in contingency tables is proposed. The SMILE procedure selects the smaller of the two subsets linked at the final step of a single-linkage clustering based on a novel dissimilarity measure d_s^2. Compared to

other dissimilarity measures, d_s^2 has been designed to equally account for sparsity and specificity on the targeted profiles.

It is established that the detection of a subset A by single-linkage detection is conditioned by the structure of A and $\overline{A}$. The benefit of using the novel dissimilarity d_s^2 in comparison to one classical dissimilarities d_2^2 is shown. The use of d_s^2 is also compared to 11 other methods in a simulation study, for which three main scenarios are detailed. Some parameters of the simulation setup are chosen to mimic a real dataset that aims at characterizing molecular signatures of selection in the domestic dog. Results obtained for the six genomic regions of interest are consistent with the simulations and general results of single-linkage detection. On the one hand, the simulation study and the analysis of the LUPA dataset demonstrates that Jaccard, Bray-Curtis, d_1^2, χ^2, Hellinger, Gower and Cosine dissimilarities are not adapted to *sparse-specific* profile detection. χ^2, for example, focuses on the specificity of conditional profiles, thus failing at distinguishing between sparse and non-sparse profiles. When the studied table is sparse, specific rare categories are expected to be over-represented. This may fortuitously generate individuals with specific and non-sparse profiles leading to a lack of power of the χ^2 distance when dealing with sparse contingency tables. On the other hand, d_s^2 is outperforming d_2^2, Sparcl, NMF and LSA by being less sensitive to the structure of non targeted individuals. d_s^2 is indeed the only method able to correctly detect 5 regions. All other methods also fail at finding the only missed region by d_s^2, (Region ♯2), thus demonstrating the strength of the novel dissimilarity d_s^2 in a real situation. The good performances of Sparcl and d_2^2 (that both correctly detect the four Regions ♯1, ♯3, ♯4, ♯5), are also in agreement with the simulation study. Region ♯3 and Regions ♯1, ♯4, ♯5, indeed correspond to the simulated scenarios ♯1 and ♯2 shown in Figure 1 and 2, for which Sparcl and d_2^2 have high power. Regarding the furnishing trait, characterized by genomic Region ♯2 in Table 2, none of the methods is able to detect the five selected breeds since more than two clusters have to be detected. From a biological point of view, furnishing trait is a complex phenotype that might be divided into sub phenotypes according to some interactions with other traits such as coat length for example [23].

Results obtained on the real dataset give promising new insights in the detection of selection signatures. First, because of the close proximity between dogs and humans, identifying the targets of selection as well as the genetic variants involved in phenotypic variation of the domestic dog can help the identification of similar variants and novel molecular pathways in

humans [20]. Moreover, the application of our methodology to other species might help in improving the control of genetic variability, thus improving, for example, milk production in agronomy [24].

References

[1] P. Sabeti, S. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. Mikkelsen, D. Altshuler, E. Lander, Positive natural selection in the human lineage, Science 312 (2006) 1614–1620.

[2] M. Wang, X. Huang, R. Li, H. Xu, L. Jin, Y. He, Detecting recent positive selection with high accuracy and reliability by conditional coalescent tree, Molecular Biology and Evolution 31 (2014) 3068–3080.

[3] A. Srivastava, M. Sahami, Text Mining: Classification, Clustering, and Applications, Chapman & Hall/CRC, 1st edition, 2009.

[4] R. H. G. Jongman, C. J. F. Ter Braak, O. F. R. van Tongeren, Data Analysis in Community and Landscape Ecology, Cambridge University Press, 1995.

[5] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference and prediction, Springer, 2 edition, 2009.

[6] C. Aggarwal, C. Zhai, A survey of text clustering algorithms, in: C. C. Aggarwal, C. Zhai (Eds.), Mining Text Data, Springer US, 2012, pp. 77–128.

[7] T. Landauer, P. Foltz, D. Laham, An introduction to latent semantic analysis, Discourse processes 25 (1998) 259–284.

[8] D. D. Lee, H. S. Seung, Algorithms for non-negative matrix factorization, in: In NIPS, MIT Press, 2001, pp. 556–562.

[9] D. M. Witten, R. Tibshirani, A framework for feature selection in clustering, Journal of the American Statistical Association 105 (2010) 713–726.

[10] A. Singhal, Modern Information Retrieval: A Brief Overview, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (2001) 35–42.

[11] J. Oksanen, F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, H. Wagner, vegan: Community Ecology Package, 2015. R package version 2.2-1.

[12] N. Jardine, R. Sibson, Mathematical taxonomy, J. Wiley and Sons, London, 1971.

[13] J. Gower, G. J. S. Ross, Minimum spanning trees and single linkage cluster analysis., Journal of the Royal Statistical Society. Series C (Applied Statistics) 18 (1969) 54–64.

[14] A.-S. Lequarré, L. Andersson, C. André, M. Fredholm, C. Hitte, T. *et. al.*. Leeb, Lupa: A european initiative taking advantage of the canine genome architecture for unravelling complex disorders in both human and dogs, The Veterinary Journal 189 (2011) 155 – 159. Special Issue: Canine Genetics.

[15] T. Calinski, L. C. A. Corsten, Clustering means in anova by simultaneous testing, Biometrics 41 (1985) 39 – 48.

[16] J. B. Kruskal, On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem, in: Proceedings of the American Mathematical Society, 7, pp. 48–50.

[17] S. C. Goslee, D. L. Urban, The ecodist package for dissimilarity-based analysis of ecological data, Journal of Statistical Software 22 (2007) 1–19.

[18] F. Wild, lsa: Latent Semantic Analysis, 2014. R package version 0.73.

[19] R. Gaujoux, C. Seoighe, A flexible R package for nonnegative matrix factorization, BMC Bioinformatics 11 (2010) 367.

[20] A. Vaysse, A. Ratnakumar, T. Derrien, E. Axelsson, G. Rosengren Pielberg, S. *et. al.*. Sigurdsson, Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping, PLoS Genet 7 (2011) e1002316.

[21] P. Scheet, M. Stephens, A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase, American journal of human genetics 78 (2006) 629 – 644.

[22] J. M. Akey, A. L. Ruhe, D. T. Akey, A. K. Wong, C. F. Connelly, J. *et. al.*. Madeoy, Tracking footprints of artificial selection in the dog genome, Proceedings of the National Academy of Sciences 107 (2010) 1160–1165.

[23] E. Cadieu, M. W. Neff, P. Quignon, K. Walsh, K. Chase, H. G. Parker, B. M. Vonholdt, A. Rhue, A. Boyko, A. Byers, A. Wong, D. S. Mosher, A. G. Elkahloun, T. C. Spady, C. André, K. G. Lark, M. Cargill, C. D. Bustamante, R. K. Wayne, E. A. Ostrander, Coat variation in the domestic dog is governed by variants in three genes, Science 326 (2009) 150 – 153.

[24] J. W. Kijas, J. A. Lenstra, B. Hayes, S. Boitard, L. R. Porto Neto, M. *et. al.*. San Cristobal, Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection, PLoS Biol 10 (2012) e1001258.

Acknowledgments

This work has been supported by French CNRS PEPS project MOLoSSE.

Appendix A. Software

Software in the form of R code, together with samples of input data set and documentation is available at
`http://emily.perso.math.cnrs.fr/SMILE/SMILE_CodeR.zip`.

Appendix B. Technical Appendices

Proof of Lemma 2.1.

According to Definition 2.3, it can be remarked that if inequality (4) does not hold, one element of $\overline{A}$ would at least belong to the first connected component that contains the whole subset A. Therefore, the set A, alone, could not be a connected component at any stage of the Kruskal algorithm. □

Proof of Theorem 2.1.

If a subset A is selected by the SMILE procedure, it is connected at the final step of the hierarchical clustering. Furthermore, by Definition 2.4, A and $\overline{A}$ are Kruskal-connected components since they both are connected components at the final step of the Kruskal algorithm. According to Lemma 2.1 we thus have: $d^{MST}(A) < min_{x\in A, y\in\overline{A}} d(x,y)$ and $d^{MST}(\overline{A}) < min_{x\in A, y\in\overline{A}} d(x,y)$, which is equivalent to:

$$max(d^{MST}(A), d^{MST}(\overline{A})) < min_{x\in A, y\in\overline{A}} d(x,y).$$

However, if either A or $\overline{A}$ is not a Kruskal-connected component, one can remark that A would not be selected by the SMILE procedure. In that case, inequality (5) is not satisfied. □

Proof of Lemma 2.2.

Let $x \in E$ such that $x \neq x_0$ and $x \neq x_s$. By Definition 2.6, since x_s is totally specific, $\langle x, x_s\rangle_2 = 0$, leading to:

$$d_s^2(x_0, x_s) = 2\,\|x_0\|_2\,\|x_s\|_2 < 2\,\|x_s\|_2\,\|x\|_2 = d_s^2(x, x_s)$$

and

$$d_2^2(x_0, x_s) = \|x_0\|_2^2 + \|x_s\|_2^2 < \|x_s\|_2^2 + \|x\|_2^2 = d_2^2(x, x_s)$$

□

Proof of Lemma 2.3.

Let $x \in E$ such that $x \neq x_0$ and $x \neq x_s$. Since x_s is nearly totally specific for d_s^2 and d_2^2, we have $\langle x_0, x_s\rangle_2 = 0$. Furthermore, x_s is the sparsest conditional profile, leading to:

$$\begin{aligned} d_s^2(x_0, x) &= 2(\|x_0\|_2\,\|x\|_2 - \langle x_0, x\rangle_2) \leq 2\,\|x_0\|_2\,\|x\|_2 \\ &< 2\,\|x_0\|_2\,\|x_s\|_2 = d_s^2(x_0, x_s), \end{aligned}$$

and:

$$\begin{aligned} d_2^2(x_0, x) &= \|x_0\|_2^2 + \|x\|_2^2 - 2\,\langle x_0, x\rangle_2 \leq \|x_0\|_2^2 + \|x\|_2^2 \\ &< \|x_0\|_2^2 + \|x_s\|_2^2 = d_2^2(x_0, x_s). \end{aligned}$$

□

Proof of Lemma 2.4. Let define the dissimilarity between two Kruskal-connected components, called C and D by:

$$d(C, D) = min_{x \in C, y \in D} d(x, y).$$

Let also introduce, for any Kruskal-connected component C, its closest Kruskal-connected component, called C^*, as follows:

$$d(C, C^*) = min_{C_i \neq C} d(C, C_i) \tag{B.1}$$

Thus, $\forall x \in E$, C^*_x is the closest Kruskal-connected component to its own Kruskal-connected component C_x. By Definition 2.7, since x_s is nearly totally specific, $d(x_0, x_s) = min_{x \neq x_s} d(x, x_s)$. Then:

$$d(x_s, C^*_{x_s}) = d(x_s, x_0)$$

and, given that $x_0 \in C_{x_0}$, $C^*_{x_s} = C_{x_0}$ and thus $\forall C_\ell \neq C_{x_s}$:

$$d(x_s, C_{x_0}) = d(x_s, x_0) \leq min_{x \neq x_s} d(x, x_s) \leq d(x_s, C_\ell)$$

□

Proof of Theorem 2.2. Since the case of $n = 2$ is not of interest, let consider $n > 2$. At the first step of the single-linkage algorithm, all individuals are isolated so that $\forall x \in E, C_x = \{x\}$. Thus we have $C_{x_s} = \{x_s\}$ and $\exists x \notin C_{x_0} = \{x_0\}$.

Let now consider the k^{th} step of the single-linkage algorithm, where $k < n - 1$. Thus, the k^{th} step is not the last step and the number of connected components is $n - k + 1 \geq 3$. Let further assume that x_s is not connected to any other individual, *i.e.* $C_{x_s} = \{x_s\}$. First, let x be an individual not connected to x_0: $C_x \neq C_{x_0}$. Then:

$$min_{C_\ell \neq C_x} d(C_\ell, x) \leq d(C_{x_0}, x) = min_{t \in C_{x_0}} d(t, x) \leq d(x_0, x).$$

Thus, from Lemmas 2.3 and 2.4 we have:

$$min_{C_\ell \neq C_x} d(C_\ell, x) < d(x_s, C_x).$$

Considering the minimum over all elements in C_x in the above inequality and using Equation B.1, we have:

$$d(C_x, C^*_x) < d(C_x, x_s) = d(C_x, C_{x_s}). \tag{B.2}$$

Let now consider the Kruskal-connected component containing x_0. Since we consider the k^{th} step of the single-linkage algorithm with $k < n-1$, there exists a Kruskal-connected component C such as $C \neq C_{x_0}$ and $C \neq C_{x_s}$. Furthermore, according to Lemma 2.3, $\forall x \in C$ we have $d(x_0, x) < d(x_0, x_s)$. Thus by minimizing over all individuals in C and all possible C we obtain that: $min_{C_\ell \neq C_{x_0}} d(x_0, C_\ell) < d(x_0, x_s)$. Since x_s is nearly totally specific and $C_{x_s} = \{x_s\}$, we have:

$$d(C_{x_0}, C^*_{x_0}) \leq min_{C_\ell \neq C_{x_0}} d(x_0, C_\ell) < d(x_0, x_s) = d(C_{x_0}, C_{x_s}) \tag{B.3}$$

Therefore, if $k < n-1$, x_s can neither be linked to any $C_\ell \neq C_{x_0}$, according to inequality B.2, nor to C_{x_0}, according to inequality B.3. Thus at the final step of the single-linkage algorithm, x_s is a singleton and is linked to the other individuals, meaning that x_s would be selected by the SMILE procedure. Furthermore, the level of the last merger is $d(x_0, x_s)$. □

Proof of Theorem 2.3. Since x_s is nearly totally specific for d_2^2, $\langle x_s, x_0 \rangle_2 = 0$. Furthermore, $\forall x \neq x_s,\ d_2^2(x_s, x_0) \leq d_2^2(x_s, x)$, thus:

$$\begin{aligned} d_s^2(x_s, x_0) &= d_2^2(x_s, x_0) - (\|x_s\|_2 - \|x_0\|_2)^2 \\ &\leq d_2^2(x_s, x) - (\|x_s\|_2 - \|x\|_2)^2 = d_s^2(x_s, x) \end{aligned}$$

since $(\|x_s\|_2 - \|x_0\|_2)^2 \geq (\|x_s\|_2 - \|x\|_2)^2$. □

Appendix C. Supplementary interpretation

Appendix C.1. Graphical interpretation of inequality (4) in section 2.4 of the main text

In the first example of Figure C.4, it can be remarked that A and $\overline{A}$ are both Kruskal-connected components at the final step of the algorithm and inequality (1) holds for A and $\overline{A}$. However, in the second example, $\overline{A}$ is not a Kruskal-connected component since the first individual in $\overline{A}$ is linked to A before being linked to any other individual in $\overline{A}$. Inequality (1) is therefore not satisfied as the level of the black dots is smaller than $d^{MST}(\overline{A})$.

Appendix C.2. Graphical interpretation of inequality (5) in section 2.4 of the main text

In the first example of Figure C.4, one can remark that $d^{MST}(A)$ and $d^{MST}(\overline{A})$ are both lower than $min_{x \in A, y \in \overline{A}} d(x, y)$. A and $\overline{A}$ are actually

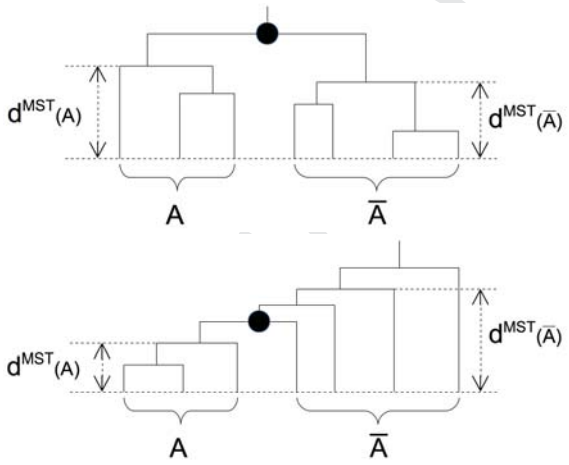


Figure C.4: Two examples of hierarchical clustering with $n = 7$ and $n_A = 3$. The black dots represent the first level of clustering between an individual in A and an individual in $\overline{A}$, *i.e.* $min_{x \in A, y \in \overline{A}} d(x, y)$. The two examples illustrate the calculation of $d^{MST}(A)$ and $d^{MST}(\overline{A})$. In the first example, subset A is selected by the SMILE procedure and inequality (2) in the main text is satisfied. In the second example, subset A is not selected by the SMILE procedure and inequality (2) in the main text does not hold.

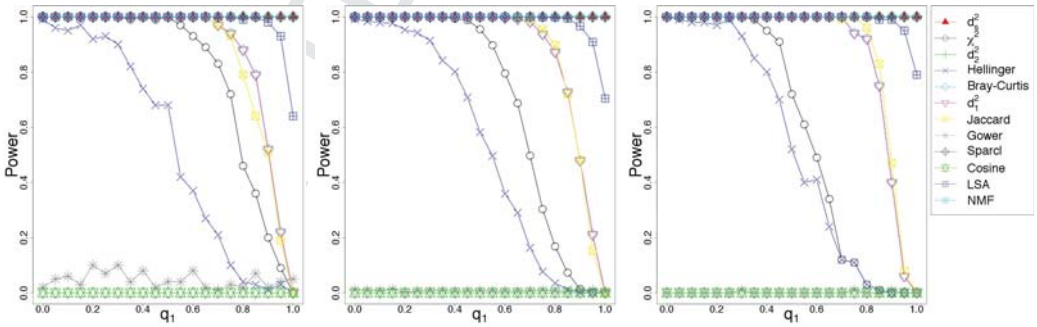
both Kruskal-connected components leading to the selection of A by the SMILE procedure. However in the second example, we have $d^{MST}(A) < min_{x\in A, y\in\overline{A}} d(x,y)$ but $d^{MST}(\overline{A}) > min_{x\in A, y\in\overline{A}} d(x,y)$. Thus, although A is a Kruskal-connected component, A is not selected by the SMILE procedure because of the structure of $\overline{A}$. In other words, the structure of A and $\overline{A}$ both play major roles in the selection of A by the SMILE procedure.

Appendix D. Impact of design parameters on the simulation-based power study

Appendix D.1. Influence of the number of individuals, n.

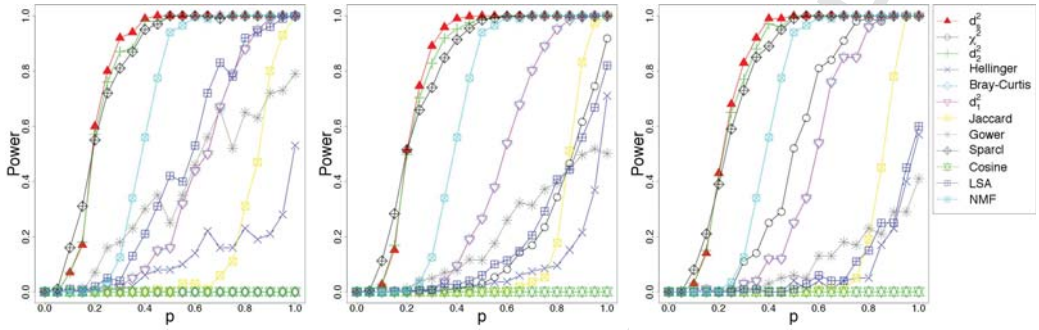
In this section, we study the impact of the number of individuals in the result of our simulation-based power study. For this purpose, simulations under the three scenarios ♯1, ♯2 and ♯3 described in sections 3.3 and 3.4, have been performed with three different values: $n = 15$, $n = 30$ and $n = 50$. The results are shown in Figures D.5, D.6 and D.7 for Scenario ♯1, ♯2 and ♯3, respectively and prove that the impact of n is very limited. Conclusions regarding the good performances of our proposal d_s^2 are still valid for any n. Furthermore, the only methods for which power depends on n are χ^2, LSA, Gower and, to a lesser extent, Hellinger, d_1^2 and Bray-Curtis.

In more detail, we can note In Figure D.5 that n has almost no impact on the power for all methods. Regarding Scenario ♯2, d_s^2, d_2^2 and Sparcl remain the three best methods for all n and their power is not influenced by the number of individuals. We can further remark that power is increasing with n for χ^2, is decreasing with n for LSA and Gower while being constant with n for all other methods. The results in Figure D.7 show that d_s^2 remains the most powerful method with respect to n. We can also note that power for n is not impacted by n. We can further remark that power is increasing with n for χ^2, decreasing with n for Hellinger, d_1^2, Bray-Curtis, Gower and LSA and remains constant for the other methods.



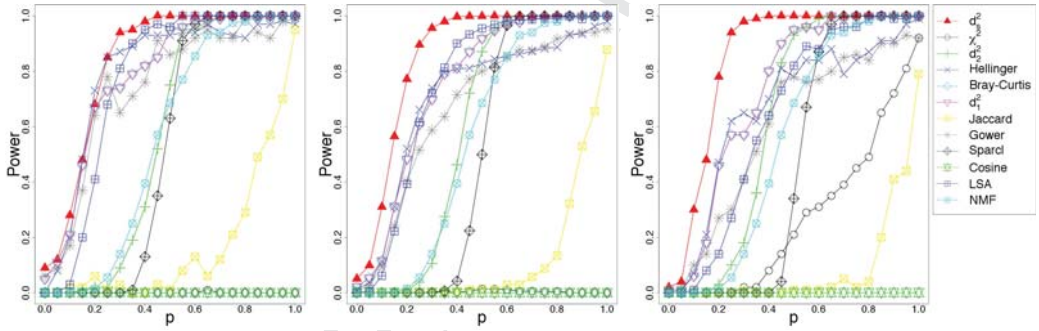
$n = 15$ $n = 30$ $n = 50$

Figure D.5: Impact of n in the simulated scenario with specific and non-sparse profile in $\bar{A}$ as described under Scenario ♯1.



$n = 15$ $n = 30$ $n = 50$

Figure D.6: Impact of n in the simulated scenario with variation in sparsity within $\bar{A}$ as described under Scenario ♯2.



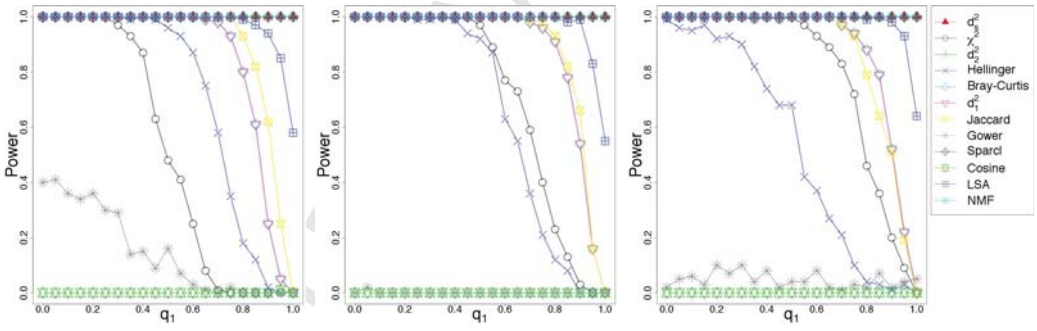
$n = 15$ $n = 30$ $n = 50$

Figure D.7: Impact of n in the simulated scenario with variation in sparsity within $\bar{A}$ as described under Scenario ♯3.

Appendix D.2. Influence of the number of categories, k.

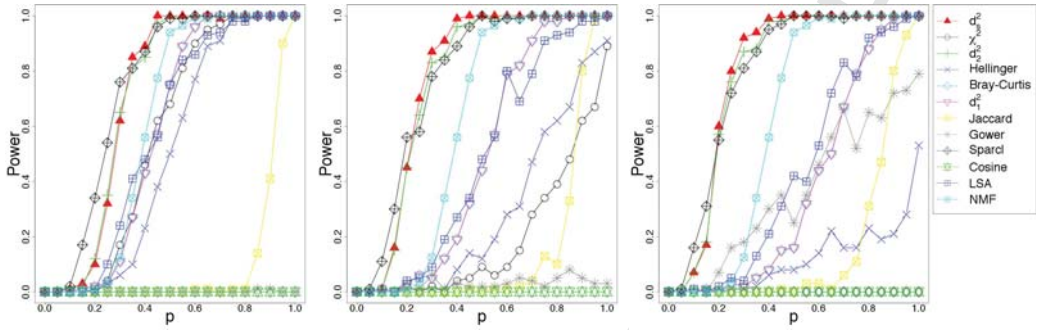
In this section, we study the impact of the number of individuals in the result of our simulation-based power study. For that purpose, simulations under the three scenarios ♯1, ♯2 and ♯3 described in sections 3.3 and 3.4, have been performed with three different values: $k = 40$, $k = 100$ and $k = 200$. The results are shown in Figures D.8, D.9 and D.10 for Scenario ♯1, ♯2 and ♯3, respectively and prove that.

In more detail, we can note in Figure D.8 that k has almost no impact on the power for all methods except for χ^2 for which power slightly increases with k. The results for Scenario ♯2 show that k does not influenced the power for d_s^2, d_2^2 and Sparcl that remain the most powerful methods for all k. Power is decreasing with k for χ^2, LSA, d_1^2, Bray-Curtis and Hellinger, increasing with Gower while being constant for the other methods. Regarding Scenario ♯3, our dissimilarity d_s^2 keeps the most powerful method for any k. Power for LSA and χ^2 tends to decrease with k while power for Hellinger, Bray-Curtis, d_1^2 and Gower increases with k.



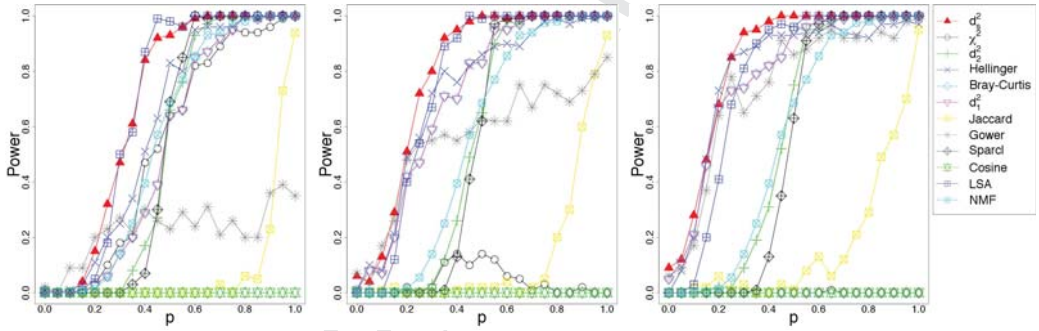
$k = 40$ $k = 100$ $k = 200$

Figure D.8: Impact of k in the simulated scenario with specific and non-sparse profile in $\bar{A}$ as described under Scenario ♯1.



$k = 40$ $k = 100$ $k = 200$

Figure D.9: Impact of k in the simulated scenario with variation in sparsity within $\bar{A}$ as described under Scenario ♯2.



$k = 40$ $k = 100$ $k = 200$

Figure D.10: Impact of k in the simulated scenario with variation in sparsity within $\bar{A}$ as described under Scenario ♯3.

Appendix E. d^2_α: a family of weighted dissimilarity measures

In this section we focus on the importance of the weights given to sparsity and specificity in the dissimilarity. To fit with the features of sparse-specific profiles, our proposed dissimilarity, d^2_s, equally weights sparsity and specificity. Nevertheless, d^2_s can be seen as a member of a more general family of measures, d^2_α, defined as follows:

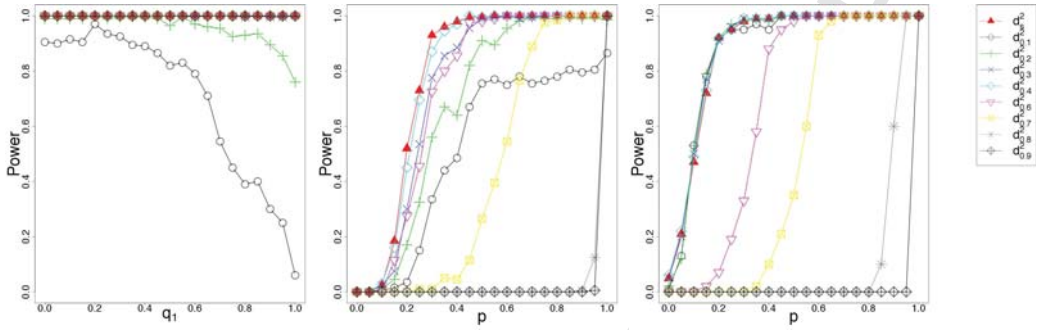
$$\forall x, y \in E : d^2_\alpha(x,y) = \left(\left[\|x\|_2 \|y\|_2 \right]^\alpha \left[d^2_\theta(x,y) \right]^{1-\alpha} \right)^2$$

where

$$d^2_\theta(x,y) = 2\left(1 - \frac{\langle x, y \rangle_2}{\|x\|_2 \|y\|_2}\right)$$

is the square of the angular distance as defined in Equation (2). Since $d^2_s(x,y) = d^2_{\alpha=1/2}(x,y)$, it is straightforward to see that $d^2_\alpha(x,y)$ is a generalization of the definition of d^2_s given in Equation (1). Furthermore, the term $\left[\|x\|_2 \|y\|_2 \right]$ is a measure of the sparsity while $\left[d^2_\theta(x,y) \right]$ is a measure of the specificity. Thus, the measure d_α is a weighted geometric mean of the sparsity and the specificity and our proposal d^2_s is the corresponding geometric means with equal weights to sparsity and specificity.

To evaluate the impact of α, simulations under the three scenarios ♯1, ♯2 and ♯3 described in sections 3.3 and 3.4, have been performed with $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5 \ (i.e.\ d^2_s), 0.6, 0.7, 0.8, 0.9\}$. The results, shown in Figure E.11, confirm that d^2_s is the most powerful method in the three simulated scenario. On one hand, we can remark that choosing a low value for α (for instance $\alpha = 0.1$ or 0.2) tends to decrease power in the Scenario ♯1. On the other hand, using a high value for α (for instance $\alpha \geq 0.6$) leads to a significant loss of power in Scenario ♯3. The results of Scenario ♯2 further illustrates that $\alpha = 0.5$ is an appropriate choice for detecting sparse-specific profiles since the further from 0.5 the value of α, the lower the power.



Scenario ♯1 Scenario ♯2 Scenario ♯3

Figure E.11: Impact of α, the weights given sparsity and specificity, on the power under the three simulated scenarios ♯1, ♯2 and ♯3.

Appendix F. Supplementary tables

Table F.4: Total samples used in the illustrated dataset from the European consortium LUPA.

Breed	Code	Number of dogs
Bernese Mountain Dog	BMD	24
Belgian Tervuren	BeT	20
Beagle	Bgl	24
Border Collie	BoC	32
Border Terrier	BoT	50
Brittany Spaniel	BrS	24
Cocker Spaniel	CoS	28
Dachshund	Dac	24
Doberman Pinscher	Dob	50
English Bulldog	EBD	26
English Setter	ESt	24
Elkhound	Elk	24
Eurasier	Eur	24
Finnish Spitz	FSp	24
Golden Retriever	GRe	50
German Shepherd	GSh	28
Greenland Sledge Dog	GSl	22
Gordon Setter	GoS	24
Greyhound	Gry	24
Irish Wolfhound	IrW	22
Jack Russell Terrier	JRT	24
Labrador Retriever	LRe	28
Newfoundland	NFd	50
Nova Scotia Duck Tolling Retriever	NSD	46
Rottweiler	Rtw	24
Schipperke	Sci	50
Shar-Pei	ShP	22
Standard Poodle	StP	24
Yorkshire Terrier	TYo	24
Weimaraner	Wei	52

Table F.5: Summary of the biological knowledge for the six genomic regions used as test regions to validate our method. The second row, denoted by Chr., refers to the chromosome carrying the region. n_A is the cardinal of the set A and refers to the number of individual(s) (or breed(s)) under selection. k_1 is the cardinal of K and corresponds to the number of category(ies) (or haplotype(s)) involved in the selection. k corresponds to the total number of categories (or haplotypes).

	Region ♯1	Region ♯2	Region ♯3	Region ♯4	Region ♯5	Region ♯6
Chr.	1	13	13	13	18	27
Related gene	*HMGA2*	*RSPO2*	*HSA2*	*HSA2*	*Fgf4*	*KRT71*
Trait	Brachicephaly	Furnished	Skin wrinkling	Periodic Fever Syndrome	Chondro-dysplasia	Curly
Breed(s) under selection	EBD	BoT, IrW, JRT, StP and TYo	ShP	ShP	Dac and TYo	StP
n_A	1	5	1	1	2	1
k_1	1	2	2	2	1	1
k	377	367	124	228	44	55