GENOMIC FUSION

Gilles Santini - Vintco

Background

In 1983 Fredrich Wendt gave, for the first time ever, a paper on fusion models to the WWRS audience. At that time it was felt as a provocation since the data reported by his AG.MA Model was not anymore the outcome of pure observation.

" We are about to replace a paradigm that ruled over media research for decades by another one." he said.
But no one would have imagined then, the consequences of his ideas, nor the hybrid data world we are in today.

"*This will give rise to a lot of new understanding, but there are many people who will keep their old habit of thinking."* he added.
He was right and the battle went rather fierce for many years.

In 1988, in search for reassurance for those who were not convinced, my question was :
"*Validation of data fusion techniques : what can statistical theory do for us?*".
Frankly speaking, it can do some, but not that much!

In fact, it is experiences, improvements and education which forced over the years, acceptance of the fusion techniques; and, overall, the need to handle a fast changing and more complex reality was the driving force.

Our research community has learned a lot since 1983 but we must continue improving.

Why a new fusion concept ?

In the case of the fusion / ascription of data from two sources, it is usual to look for a proximity between the elements of the donor set and those of the receiving set.

More precisely, for each recipient, one seeks to find a donor who "resembles" him to transfer the data of the latter.

In order to avoid reducing the variance existing in the donor sample, one proceeds in a way that penalize multiple uses of a donor.

Different techniques of searching for good links between donors and recipients exist: some favor stable relations between donors and recipients (i.e. relations where the donor and the recipient are closest to one another) while others try to minimize globally the dissimilarity between donors and recipients.

In all cases, a similarity measure is constructed between donors and recipients. This measure can be frequency based (Chi-2 distance), multivariate (Mahalanobis Distance) or weighted in such a way that if two donors are similar in the donor set, a receiver close to one will necessarily be close to the other (P-weighted distance).

In order to better control the balance of the samples, the search for the donor-recipient links is generally carried out within closed groups (generally sex x age cells).

Nevertheless, conventional techniques do not make it possible to correctly judge the relevance of the choice of these links which are established solely as a function of the distance between donors and recipients.

Aside from cell-based balancing and sometimes additional constraints, it can be said that the classical approach performs information transfers between the donor and recipient groups on the sole fact that the links are established between "resembling" individuals but without consideration to any "compatibility" of them.

The notion of genome aims to overcome this deficiency, the idea being to allow a transfer of information between a donor and a recipient <u>if the two exhibit a close similarity but only if they have compatible genomes.</u>

Sequencing the Genome

The genome of an information unit (in practice an individual) is constructed as a sequence of 4 codes NSML associated with structural characteristics of the studied phenomena: it could be for example, "reader of a magazine" or "consumer of a type of product ".

Each *Character* correspond to a *Gene* which can take 4 *Code values* either N,S,M or L, according to the following concepts:

N - (NONE): Does not present the character.

S - (SMALL): Has a chance to present the character in the lower third of the encountered probabilities set excluding zeros.

M - (MEDIUM): Has a chance to present the character in the middle third the encountered probabilities set excluding zeros.

L - (LARGE): Has a chance to present the character in the upper third of the encountered probabilities set excluding zeros.

The *Genome* of an individual thus looks as a sequence such as: NNSMNLNMSSL ....(Exhibit 1)

Contrary to what exists in biology the place of the characters in the sequence is irrelevant. By convention they are ordered by increasing frequency of N in the dataset (one can say that the sequences are N little endian).

In the subset exhibiting a code different to N for a gene, the number of S,M and L is the same over the donor data set and similarly the recipient data set. However the way the codes values are built may be distinct for the two dataset as long as the observed character is the same (N must mean the same thing in both sets).

The construction of the probabilities on which the sequencing is based can be captured in different ways (e.g. answer to a question, passive measure etc.) and result from different methods of probabilities calculation.

The well-known segmentation / probabilisation method based on an CHAID type algorithm is well suited to create the probabilities; adjustment techniques can be used on the probabilities to calibrate them before the codes valuation.

Building Genome groups

Two information units are said to be compatible if their genome sequences are similar enough.

One can think building groups of genome sequences that exhibit mostly the same values for the genes.

Two information units will be said to be compatible if their genome sequences belong to the same Genome group.

Describing a clustering algorithm where the center of each cluster is equal to the mode of each variable over the cases in the cluster, John Hartigan [2] call the operation Dittoing. It is well fitted for the current goal.

This algorithm however is rather computer resources greedy and cannot be applied in all cases for that reason.

An alternative algorithm called k-modes presented by Huang [3] can be used instead on big datasets.

Both algorithms build for each group a cluster center defined as the sequence of the modes of each gene over the units inside that group excluding the cases when the gene is not active (i.e. with code value = N).

To decide to which group a genome sequence should be allocated one compare the distance of this sequence to clusters centers and pick the closest (Exhibit 2).

To calculate the distance between two genome sequence (g_1, g_2) one uses the following dissimilarity measure :

$$\delta(g_1, g_2) = \frac{\sum_{k}^{K} \left\{ g_1[k] \neq g_2[k] \mid g_1[k] \neq N \;\&\; g_2[k] \neq N \right\}}{K_{12}}$$

Where K_{12} is the number of genes which are actives in at least one of the two genome sequence and $g[k] \in \{N, S, M, L\}$ le code value of the k^{th} gene for the sequence g .

The upper term can simply be read as the number of cases when the two sequences exhibits different code values excluding the case when the codes values are N.

In the present case , it can be proved that the use of the previous distance is equivalent to use of Tanimoto coefficient which is often referred to in molecular fingerprints clustering (see Appendix I).

Since by design, the matching between donors and recipients operates separately within the number of suitable groups will depend on the size of the samples.

A guideline is that each group should be large enough to capture enough variance of the information units to reflect distinct profiles between them.

A rule of thumb is to set a minimum size to 4% of the sample size.

Matching within Genome groups

Within each genome group matching between donors and recipients is performed according to the following well known paradigm (Exhibits 3,4).

Let i be a donor from the donors set D and j be a recipient from the recipients set R ;

Let $d(i,j \mid X_i, Xj)$ be a quantity measuring the difference between i and j profiles according to a set of common known attributes with values in $[0,1]$;

Let r_i bet the number of times i is used as a donor and $\varphi(d,r)$ a penalization monotone increasing function of d and r such that $\varphi(d,1)=d$ and $0<d<\varphi(d,r)<1$;

Let optionally $\varpi(i,j)$ be a weighting value that account for additional imperative or fuzzy constraints such that $\varpi(i,j)=1$ if matching of i and j break no constraint, $\varpi(i,j)=\infty$ if matching of i and j break at least an imperative constraint and $\varpi(i,j)>1$ otherwise where);

The matching algorithm will look for a solution that minimize the cost function :

$$\Delta = \sum_{i \in D} \sum_{j \in R} \varpi(i,j) \varphi\left(d(i,j), r_i\right)$$

The algorithm may try to find a "global" optimum or restrict the search to a subset of solutions that enforce matching when there is a strong "local" proximity (See Appendix II).

The choice of a minimization algorithm is a matter of taste and dataset size since it is a heavy processing task.

On the contrary the choice of the type of distance must be carefully done.

Very many distances derived from the indicator matrix that represents in binary form the dataset are possible : Hamming, Chi-2, Hartigan, MCA-Factorial to name a few.

These distances can be weighted in various ways but it is worth mentioning that a very efficient way to do so is to rely on the weighting scheme described in Appendix III.

A case study

The first real life project where the Genomic Fusion has been used was supported with the aim to enrich the SimmTGI study which is currently widely used in France for consumers data and audience analysis. This data resource is the heir of a respectable line of consumer studies which have learned over the years how to perfectly master print audience data. However, facing the need for Internet sites data and considering the growing trend of analyzing magazine audience as a global brand audience, it had not surprisingly been decided to enhance this study traditionally based on questionnaire data by fusion with a secondary source of meter data.

It was not in the culture of the French TGI Team to take such a route recklessly and a very impressive amount of time have been devoted to quality control.

In order to measure if a tangible improvement could be gained by this method, testing the new Genomic fusion was part of that effort.

The fusion has been run with 2957 donors (meter data) and 10912 recipients (questionnaire respondents data).

The questionnaire data was also available for the donors.

Based on the "*Has visited the web site*" question 361 genes were built using 94 segmentation variables. These genes where clustered into 15 genome groups with sizes ranging from 2614 to 646.

Using those Genome groups to assess compatibility and an additional set of 132 variables chosen in the questionnaire data to build the similarity distances between donors and recipients a cell bey cell (sex x age) fusions has be run.

Cross tabulating , category by category, each of the 132 variables by the other ones and testing for chi-2 discrepancy between values before and after fusion no abnormal levels where found as showed by the following table which exhibits the % of significant chi-2 at level $\alpha = 0.05$ and the number of cross tables with enough counts used to perform the test.

MALE – 15-34	**4.55%**	5669
MALE – 35-54	**3.68%**	7279
MALE – 54 +	**4.93%**	7641
FEMALE – 15-34	**3.06%**	7260
FEMALE – 35-54	**3.43%**	8737
FEMALE – 54 +	**3.79%**	8122

What the TGI team was looking for was a good fit between consumer behavior and brand sites visits when those where mainly exclusive. Here is an example among others which have led the SimmTGI team to use the Genomic fusion method in operations :

BANK Site Visit among BANK Account owners over BANK Site Visit among TOTAL Population	TOTAL Population	BANK A Account	BANK B Account	BANK C Account	BANK D Account
BANK A Web Site	**100**	**953**	**47**	**77**	**108**
BANK B Web Site	**100**	**27**	**377**	**32**	**58**
BANK C Web Site	**100**	**121**	**81**	**409**	**105**
BANK D Web Site	**100**	**66**	**61**	**69**	**688**

This table exhibits as index but the SimmTGI team also calibrate the sites audience levels to the standard levels used by the industry so, although the final dataset do not come from a single source, the Consumer data can be securely used as a target for media planning without changing the audience levels.

Go fast!

Modern data fusion systems must go fast.

Firstly, fusions operations are more numerous:

- Users are expecting a greater value for their data and look for that by integration of several sources.
- Datasets are updated more frequently to follow fast changing markets.

Secondly, data sets used for fusions are bigger :

- More than two sources may be implied or multifold datasets of donors may be required
- Passive, Sites centric, Open or Operators data, generally leads to massive data processing.

Cloud processing is clearly very valuable to adapt the required computer resources to needs. It also offers ways to accelerate the operations using NoSQL blocks oriented storage facilities which are well fitted for the type of calculations required by fusion algorithms.

However the big leap forward will come from the use of machines with massive parallel facilities (gpu). This is already current in other contexts such as IA or cryptanalysis but given the essentially parallel nature of the calculations required by fusion algorithms (or the possibility to organize them in such a way) the gain will be many orders of magnitude and we should be able to see it soon.

Looking backwards

The development, understanding and acceptation of fusion methods applied to media research has been a long road. Since 1981, the WWRS and the PDRF has charted that route. However most of the concepts where already there from the beginning.

The author of this paper was first exposed to a fusion experiment in the mid-seventies, 40 years ago! It was a German-French project team working on the AG.A Model data processing, Dr Fredrich Wendt was the method designer, Lucien Boucharenc was the data scientist, the author was there because the software was coded in APL and he knew from his training in the USA how to use that computer language, well suited to do statistics and which, strangely enough, would have been great on a gpu machine because of its vectorial conception.

The processing time was huge and the budget too…Count the first in days and the second with 5 zeros in Big Mac dollars.

In Fredrich Wendt terms[6] the objective was to create a partnership between the two samples respondents to optimally transfer the topology of the variables relationships. To establish the objective a weighted distance score was minimized by iteration, matching firstly close donors and recipients and secondly looking in more distant halos around.

In today terms the objective is to match the two samples to enhance without distortion the recipient data with the donor data. This is done with a mixed local/global geometric minimization technique applied to data points clouds.

Although the weighting of the distances components was very much handcrafted and the processing a trial and error process this sounds close enough. But, to the best of the author knowledge the notion of compatibility of the donors-recipients partners was not clearly used at that time.

Conclusion

In the light of the above, although it introduces the distinction between compatibility and resemblance, the Genomic Fusion does not imply an in depth change of paradigm but rather sets clarified rules to weave the links between donors and recipients :

A link between a donor and a recipient is acceptable if only if:

i. Their genomes are compatible (**compatibility** criterion)

ii. Their similarity is strong (**resemblance** criterion)

iii. The overall donor-recipient cohesion increases (**coherence** criterion)

iv. Donors are used with parsimony (**distribution** criterion)

A final word.

Even if all four criteria are well satisfied, one cannot guarantee that the fused dataset will be statistically sound and create value. It may looks good, fill a gap and increase data ease of usage but it could lead to spurious decisions unless an educated eye aware of the two sources nature and quality, performs an in depth control, eventually guided by statistical tools, including but not limited to cross-validation, before clearing the fused data set out to the market .

In the author opinion, this **approval** step is key to bring a smooth user experience with the fused data set.

Appendix I

Considering that N play the role of a zero Tanimoto Coefficient can be defined as the number of common characters over the number of existing characters from what one derive a measure of dissimilarity often called Stengel distance:

$$\tau_{12} = \tau\left(g_1[k], g_2[k]\right) = 1 - \frac{n_{12}}{n_1 + n_2 - n_{12}}$$

Where :

$$n_1 = \sum_k^K \left\{ g_1[k] \neq N \right\} \quad (resp\ n_2)$$

$$n_{12} = \sum_k^K \left\{ g_1[k] = g_2[k] \mid g_1[k] \neq N \ \ \& \ \ g_2[k] \neq N \right\}$$

The difference index is :

$$\delta_{12} = \delta\left(g_1, g_2\right) = \frac{d_{12}}{K_{12}}$$

Where :

$$d_{12} = \sum_k^K \left\{ g_1[k] = g_2[k] \mid g_1[k] \neq N \ \ \& \ \ g_2[k] \neq N \right\}$$

Since the values of the genes of the center of a cluster (let's say g_1) are taken by construction within $\{S, M, L\}$, we have $n_1 = K \Rightarrow K_{12} = K$ (where K is the number of genes) and $d_{12} = n_2 - n_{12}$.

From what it finally follows after some simple algebra:

$$\tau_{12} = \frac{K\left(1+2\delta_{12}\right)-n_2}{K\left(1+\delta_{12}\right)}$$ which is a monotone increasing function of δ_{12} since it derivative is

$$\tau'_{12} = \frac{K+n_2}{K\left(1+\delta_{12}\right)^2} > 0$$

.

This prove that the use of τ_{12} is equivalent to the use of δ_{12} for the cluster creation purpose if clusters centers are based on modes.

Appendix II

The search for a global minimum of the cost function Δ may lead to match rather distant donor-recipients couples.

It is said that the matching is not necessary stable in the following sense :

A matching is stable if no donor in an established donor-recipient couple is closest to a recipient which is not its partner and if conversely the same for recipients.

This can formally be written :

$$\left(i,j\right) \Rightarrow \nexists\left(i',j'\right) \mid d\left(i',j\right) < d\left(i,j\right) \wedge d\left(i',j\right) < d\left(i',j'\right)$$

And conversely :

$$\left(i,j\right) \Rightarrow \nexists\left(i',j'\right) \mid d\left(i,j'\right) < d\left(i,j\right) \wedge d\left(i,j'\right) < d\left(i',j'\right)$$

However, stability is difficult to reach.

Some algorithms combine a local/global search in that case, the following rule is a must :

If j^* is the closest neighbor of i and i is the closest neighbor of j then i and j^* must be linked together. This rule can be formalized as :

$$\begin{cases} j^*\left(i\right) = j \\ i^*\left(j\right) = i \end{cases} \Rightarrow Match\left(i,j\right)$$

Additional more complex rules can be designed to approach stability while minimizing Δ (Ref. [5], Chap 11 §. 3.4).

Appendix III

A batch of attributes $X = \left\{X_1, X_2, \boxed{?}, X_n\right\}$ is known for both a donor and a recipient set : $X_{\mathbf{D}}$ and $X_{\mathbf{R}}$; a batch of attributes $Y = \left\{Y_1, Y_2, \boxed{?}, Y_m\right\}$ is known for the donor set only and must be ascribed to the recipient set by a matching algorithm : $Y_{\mathbf{D}}$ and $Y_{\mathbf{R}}$.

A distance d is built on X for donors $i \in \mathbf{D}$ and recipients $j \in \mathbf{R}$:

$d_X^2(i',i'') = \sum_{k=1}^{n} d_{X_k}^2(i',i'')$ where $d_{X_k}^2(i',i'')$ is a distance component based on the k[th] attribute.

Similarly $d_X^2(j',j'') = \sum_{k=1}^{n} d_{X_k}^2(j',j'')$.

A distance δ is also built on Y for the donors only : $d_Y^2(i',i'') = \sum_{k=1}^{n} d_{Y_k}^2(i',i'')$.

One can consider the derived weighted distance based on the X attributes :

$\delta_X^2(i',i'') = \sum_{k=1}^{n} a_k^2 d_{X_k}^2(i',i'')$ where a_k^2 are positive weights such that $\sum_{k=1}^{m} a_k^2 = 1$

$\delta_X^2(j',j'')$ can be calculated likewise since the X attributes are known for the recipient set also.

The idea is to calculate the weights in such a way that : $\delta_X^2(i',i'') = d_Y^2(i',i'') + \varepsilon$.

This can be done as spherical regression problem using an estimation algorithm such as LM (Levenberg-Marquard).

The benefit of this approach is to use the same measure of closeness between a donor and a recipient that between two donors since one can evaluate

$\delta_X^2(i,j)$ for $i \in \mathbf{D}$, $j \in$

So, if j is matched with $i = i(j)$ on the ground of the distance δ_X^2 it should also be in the neighborhood of i when this neighborhood is identified using the Y attributes since $\delta_X^2(i,j)$ provide a close estimate of $d_Y^2(i,j)$. As a consequence it is legitimate to ascribe to j the Y attributes of $i(j)$.

To reach this objective the a_k weights stretch and reduce some of the $d_{X_k}^2$ components just as in the Greek mythology Procrustes was stretching and cutting off the legs of the travelers to fit in his iron bed : this is why such a distances class is sometimes referred as Procrustean but to avoid confusion with a specific distance named Procrustean distance which is used in differential geometry, it is preferable to call them P-Weighted distances.

This weighting method is particularly interesting in combination with a Multiple Correspondence Analysis of the binary indicator matrix derived from the X attributes set because in that case the number n of components is a small number (it is equal to the number of retained factors).

References :

[1] Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*, Academic Press.

[2] Hartigan J. (1975). *Clustering Algorithms*, John Wiley & Sons.

[3] Huang Z. (1997). *A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining*, DMKD, 3(8), 34-39.

[4] Rasler S. (2002). *Statistical Matching*, Springer.

[5] Santini G. (2003). *Mathematical Models & Methods for Media Research*, G.S. IT Services.

[6] Wendt F. (1983). *The AG.MA Model*, WWRS Montreal, Fusion and Modelling 8(3), 393-403.

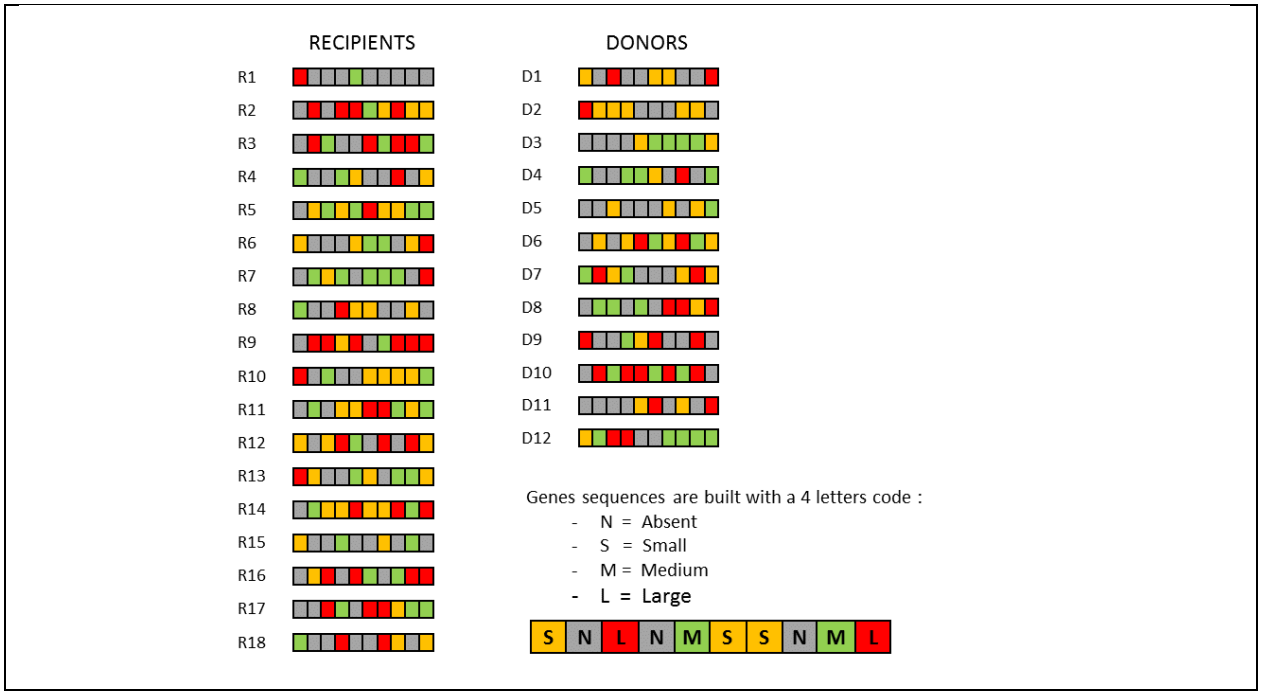

Exhibit 1

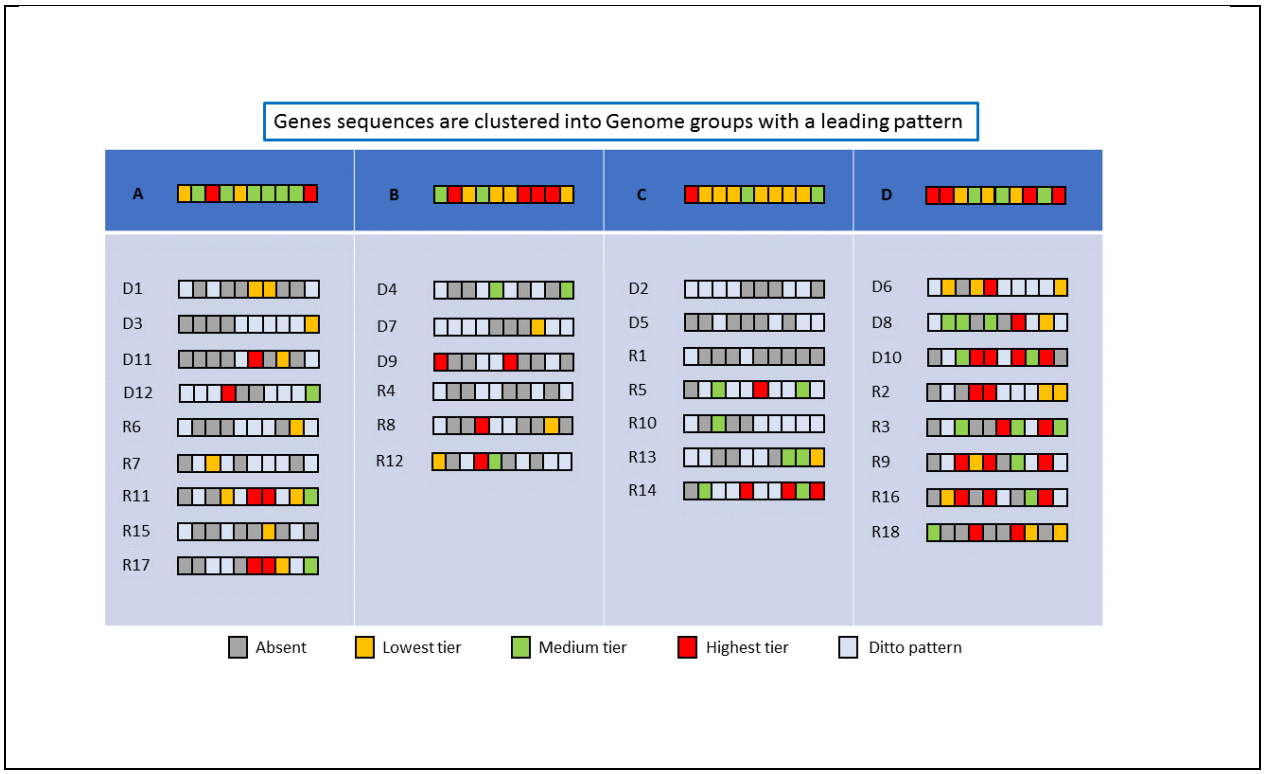

Exhibit 2

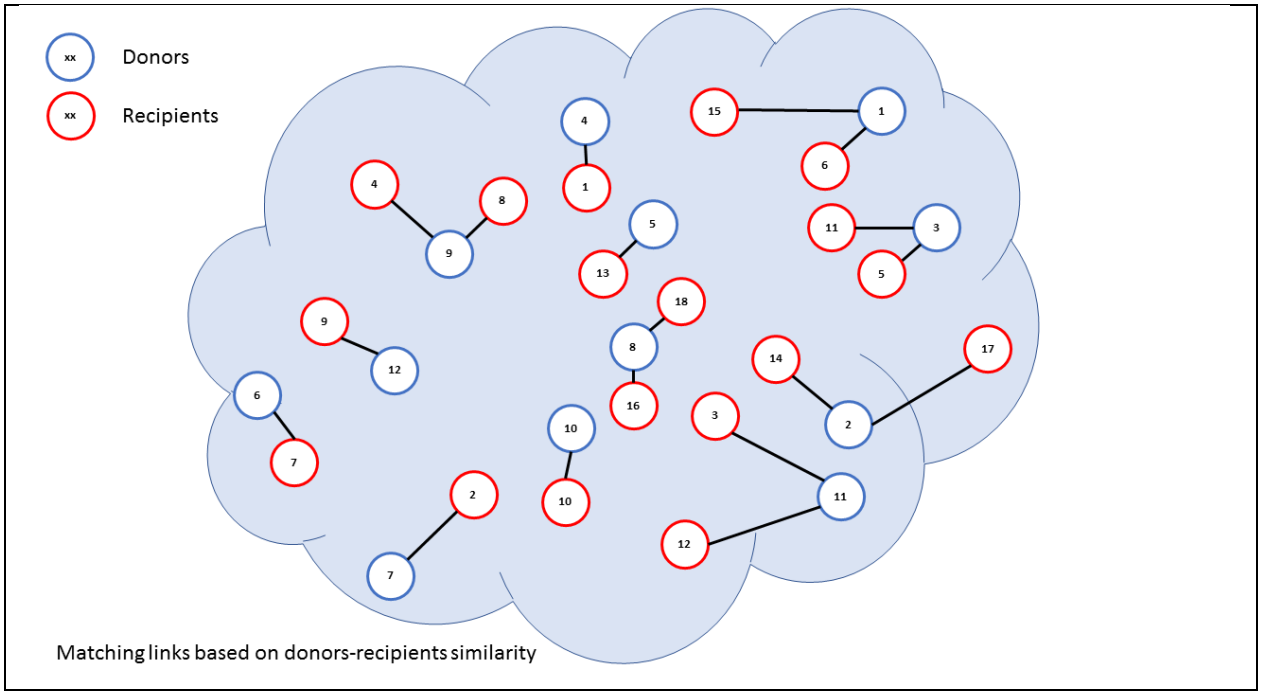

Exhibit 3

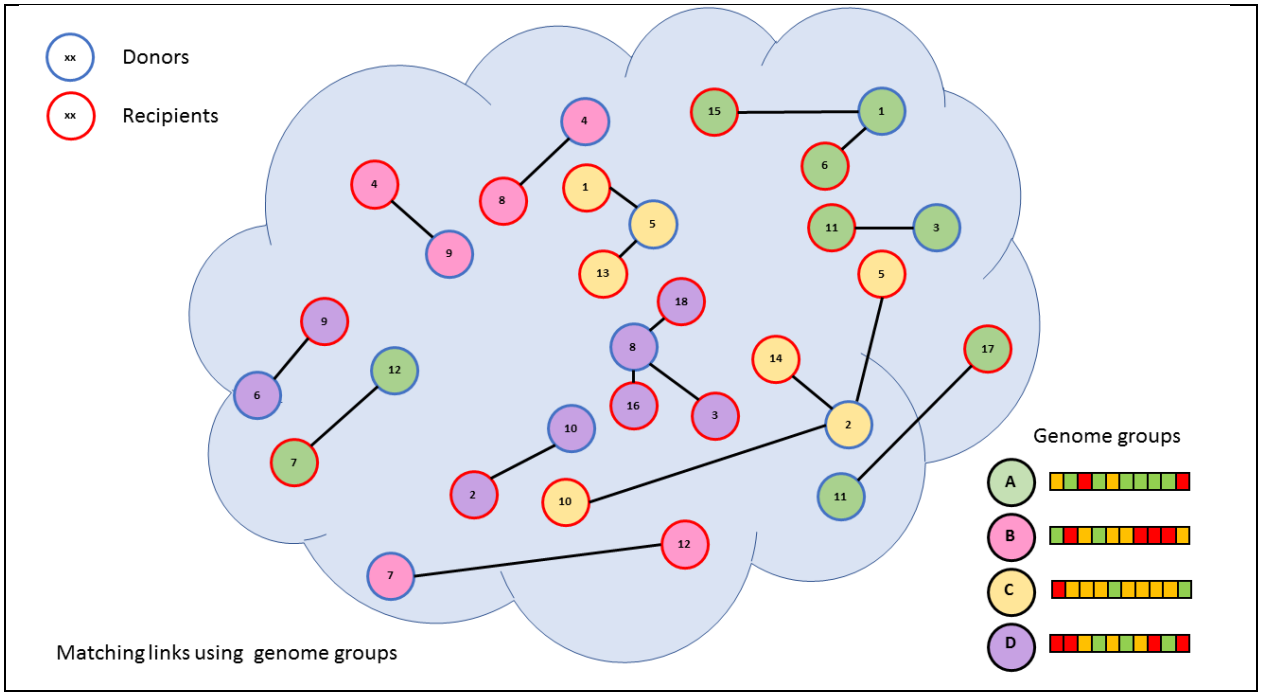

Exhibit 4