Palleschi V, Pagani L, Pagnotta S, Amato G, Tofanelli S. 2015. Application of Graph Theory to the elaboration of personal genomic data for genealogical research. PeerJ Computer Science 1:e27 https://doi.org/10.7717/peerj-cs.27

Application of Graph Theory to the elaboration of personal genomic data for genealogical research

Vincenzo Palleschi, Luca Pagani, Stefano Pagnotta, Giuseppe Amato, Sergio Tofanelli

In this communication a representation of the links between DNA-relatives based on Graph Theory is applied to the analysis of personal genomic data to obtain genealogical information. The method is tested on real data and discussed its applicability to the field of genealogical research. We envisage the proposed approach as a valid tool for a streamlined application to the publicly available data generated by many online personal genomic companies. By this way, anonymized matrices of pairwise genome sharing counts will enable to improve the retrieval of genetic relationship between customers who provided explicit consent to the treatment of their data .

Application of Graph Theory to the Elaboration of Personal Genomic Data for Genealogical Research

Vincenzo Palleschi[1,2], Luca Pagani[3,4*], Stefano Pagnotta[1], Giuseppe Amato[5], Sergio Tofanelli[6]

[1]Institute of Chemistry of Organometallic Compounds
Research Area of National Research Council
Via G. Moruzzi, 1 – 56124 Pisa (ITALY)

[2]Department of Civilizations and Forms of Knowledge
University of Pisa
Via G. Galvani, 1 – 56126 Pisa (ITALY)

[3] Division of Biological Anthropology, University of Cambridge, UK

[4] Department of Biological, Geological and Environmental Sciences,
University of Bologna
Via Selmi 3, 40126 Bologna (ITALY)

[5] Institute of Sciences and Technology of Information
Research Area of National Research Council
Via G. Moruzzi, 1 – 56124 Pisa (ITALY)

[6] Department of Biology
University of Pisa
Via L. Ghini, 13 – 56126 Pisa (ITALY)

ABSTRACT

In this communication a representation of the links between DNA-relatives based on Graph Theory is applied to the analysis of personal genomic data to obtain genealogical information. The method is tested on real data and discussed its applicability to the field of genealogical research. We envisage the proposed approach as a valid tool for a streamlined application to the publicly available data generated by many online personal genomic companies. By this way, anonymized matrices of pairwise genome sharing counts will enable to improve the retrieval of genetic relationship between customers who provided explicit consent to the treatment of their data.

Keywords: DNA Analysis, Personal Genomics, Genealogy, Genetic Genealogy, Statistical methods, Graph Theory, Ancestry reconstruction.

39 *Corresponding Author: Luca Pagani – lp.lucapagani@gmail.com

1. Introduction

In recent years, a number of companies started offering commercial services based on DNA analysis for genealogical research[1-3]. The informatic tools available to interpret such results, usually provided by the same companies or by external services[4], are mainly focused on general population studies (Paternal and Maternal lineages based on Y chromosome and mitochondrial haplogroups, Ancestry Composition/Admixture, etc.). On the other hand, very few tools are provided to investigate the links of one's DNA profile with the relatives made recognizable through personal genomic data. Notably, these pre-compiled tools are often the only way to access the data provided by the DNA testing companies for a panel of hundreds or thousands of individuals. Therefore, the starting point of any downstream analysis based on this kind of data can only rely on the semi-processed input provided by the aforementioned tools. The introduction by the genetic service providers of a wrapped application tool would facilitate users' interpretations and unearth hidden genealogical information. Such tool should enable to implement the mass of data each single DNA test makes available in an easy-to-grasp graphical form. This would be particularly useful to detect the provenience of distant autosomic DNA-relatives from either the paternal or the maternal lineage. In fact this task is often made difficult by the links that might exist between the two parental genealogies due to the custom in closed communities to marry between relatives, especially in the past.

Here we describe and annotate an artificial intelligence tool that helps exploiting the information provided to customers by genealogical genetic services. The original approach of this work is the use of cross-information about the links between the living DNA-relatives of the test user (TU) for obtaining hints about the possible connections with other individuals, in the absence of a-priori genetic or genealogic evidence.

2. Data

The data, consensually provided and anonymously treated, are derived from the results obtained by a test-user (TU) from the personal genomic service 23andMe[2]. Such results typically consist of summary statistics on about one million single nucleotide polymorphisms (SNPs)[5].

A total of 120 anonymized individuals (progressively numbered with an ID from 1 to 120) were considered in the analyses. All of them are 'DNA-relatives' of the TU according to the 23andMe criteria and accepted the invitation to share their DNA information (excluding data related to health conditions). The raw data is available as an Excel online matrix. Since this is a secondary analysis of pre-existing data and the samples are treated in an anonymised version we did not apply for an ethical clearance.

As reference parameter we considered the total amount of autosomal DNA in common between pairs of individuals, calculated as the total length of shared SNP haplotype blocks in mega base-pairs (Mbp) units. This amount, once converted into proportion of shared genome, provides a rough estimate of the number of generations separating any two individuals, under a simple model of "infinite number of ancestors" (Supplementary Table 1). Information either on the relevant chromosomes where the match occurs, or on the number of segments in common was not used. This choice is justified by the fact that only a minimum percentage of the individuals considered shows DNA matches on more than one chromosome. Furthermore, the information about the specific segment of the chromosome where such match occurs is not easily obtainable from the data made available to the users by 23andMe.

Using the Genome-Wide Comparison option in the 23andMe 'Family Traits' feature, the input data were prepared in the form of a symmetric square matrix *C*, whose *C(i,j)* elements correspond to the total length of shared SNP haplotype blocks between the individual *i* and the individual *j*, expressed in Mbp units. Most elements of the matrix are equal to zero, corresponding to the fact that the majority of the individuals does not result genetically related. The sparsity of the matrix *C(i,j)* is visually shown in Figure 1, where the white points indicate a

mutual match of any magnitude between two individuals, and the black correspond to no genetic relation at all.

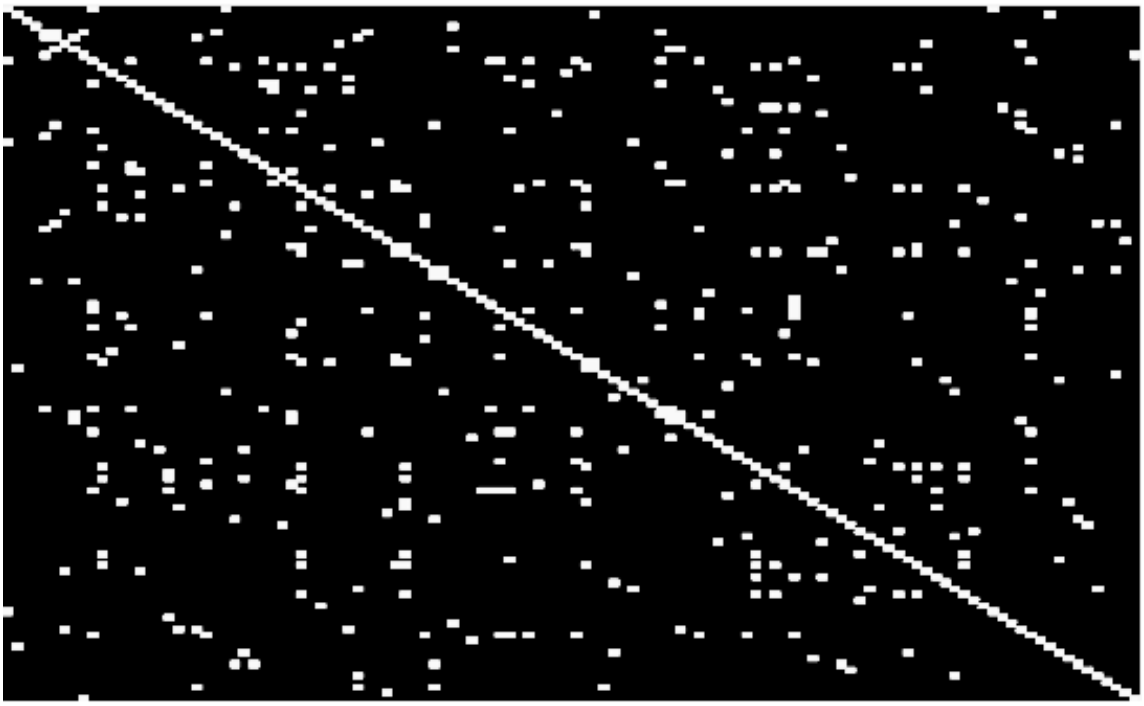

Figure 1 – Visual representation of the correlation between the individuals considered in this work.

3. Classification

The matrix depicted in Figure 1 can be alternatively interpreted as a correlation matrix, a covariance matrix, a similarity matrix[6] or it can be transformed in a distance matrix[7]. Accordingly, the way to elaborate and manipulate the associated information varies depending on the interpretation tasks. Given that the statistical analysis is aimed at simplifying data outputs, a loss of information with respect to the original data has to be expected. The effectiveness of the analysis thus depends on the amount of 'interesting' information unearthed out of the bulk of 'redundant' information. It follows that different methods can be more or less effective according to what is considered, from time to time, interesting or redundant.

To this extent, a number of potential confounders must be considered when dealing with the available genetic similarity matrix. First of all the genetic information on which the analysis is based is intrinsically fuzzy, because of the uncertainty in the data obtained by the service provider (a few 'no-called' SNPs should be routinely expected). Additionally, the presence of identical by state (IBS) other than identical by descent (IBD)[8] SNPs could potentially bias the genealogical interpretation, especially the one associated with distant relationships (Most Recent Common Ancestors distant more than 6/7 generations). Finally, as opposed to uniparental markers, the diploid autosomic data combine information inherited from the paternal and maternal genealogy that should be kept separated when tracing one's ancestry. Therefore, the analysis must be performed using statistical techniques robust enough to sustain these unavoidable uncertainties.

4. Graph theory approach

The ideal framework for studying the complex network of links between the DNA-relatives of a TU is the Graph Theory[9,10]. This approach, widely used in Mathematics, Engineering, and Computer Science, allows the analysis and graphical representation of the links between different entities in a network. In synthesis, the Graph Theory represents the elements in a network as *vertices* (or nodes) connected by *edges*. Edges are often associated with a value representing a *weight*. In our case, the weight of an edge connecting two vertexes is related to the genetic distance between them. A couple of vertexes *a* and *b* can be connected, in principle, by more than one edge. Graphs can be generally oriented, so that the edge from *a* to *b* is different from that linking *b* to *a*. In this way, the distance between the vertexes *a* and *b* can be different from the distance between *b* and *a* (a typical example is driving a car between two points in a city, where the traffic regulations might impose different routes for the direct and return trip, see figure 2).

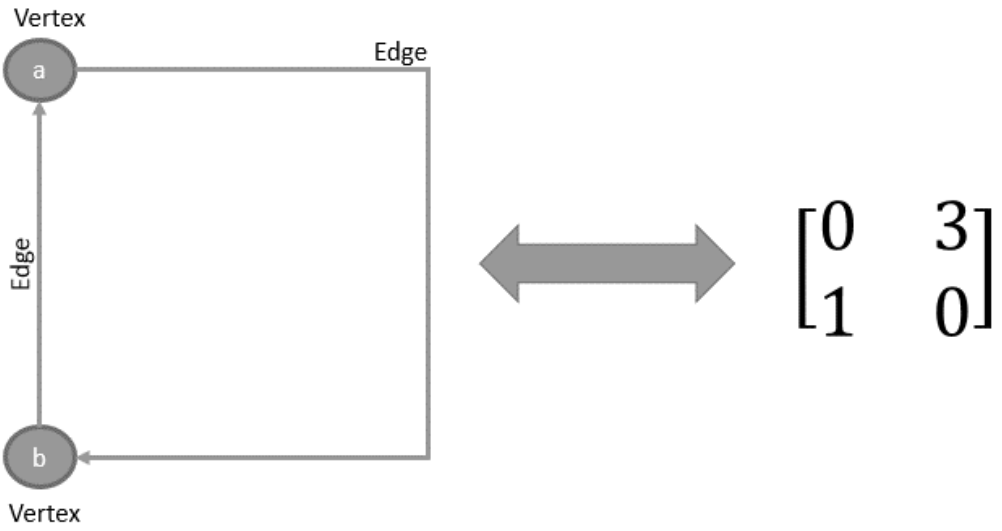

Figure 2 – Graphic representation of a Graph with two vertexes and two edges (oriented Graph). At the right, the corresponding adjacency matrix.

The relation between the vertexes is usually represented in matrix form (adjacency matrix[11]) where the elements out of the diagonal are the weights of the corresponding edges. If the adjacency matrix is symmetric (the distance between two nodes is the same in both the directions) the resulting graph is called *unoriented*.

In our scenario, the correlation matrix *C(i,j)* between the DNA-relatives of the TU is interpreted as a symmetric adjacency matrix. Therefore, we will use unoriented graphs, implemented using the Matlab® code provided in Supplementary Materials.

5. Results

In the dataset analysed here, the adjacency matrix is described by an unweighted Graph with 120 vertexes (individuals) and 196 edges (DNA links between them). The graphical representation of the Graph described by this matrix is shown in Supplementary Figure 1.

The main network connects 100 vertexes (83% of the total) by 190 edges (97% of the total) and sets aside only a few individuals, singularly (10 individuals) or in small groups of two or three persons. A strict interpretation of Supplementary Figure 1 would thus bring to the conclusion that all the individuals belonging to the main group should be considered as somehow related, directly or indirectly, to all the other members of the group. To reduce this connectivity and to assign the various individuals to the TU paternal and maternal ancestries, a further treatment of the input data is thus necessary.

5.1 Pruning

The strength of the DNA cross-links between the individuals can be used to reduce (prune) the connections highlighted in Supplementary Figure 1. Since all the 120 individuals included in this study are, by design, related with the TU, no information can be derived from those that are connected only to the TU. They are represented, in graphical form, as isolated vertexes with no edges associated. Therefore, these individuals can be safely removed from the adjacency matrix without any loss of information. Moreover, as already discussed in Section 3, spurious connections could be introduced by fuzziness of the genetic data and the occurrence of IBS SNPs. These connections can be excluded via the application of an upper threshold on the genetic distances between the individuals. The threshold amount of shared genome for a link to be considered 'real' (i.e., corresponding to IBD SNPs) can be easily converted into expected number of generations, using Supplementary Table 1.

Figure 3 shows the Graph corresponding to the adjacency matrix *C(i,j)* where only the edge weights greater or equal to 24 Mbp (roughly a 8 generations distance between the vertexes/individuals, see Supplementary Table 1) are considered.

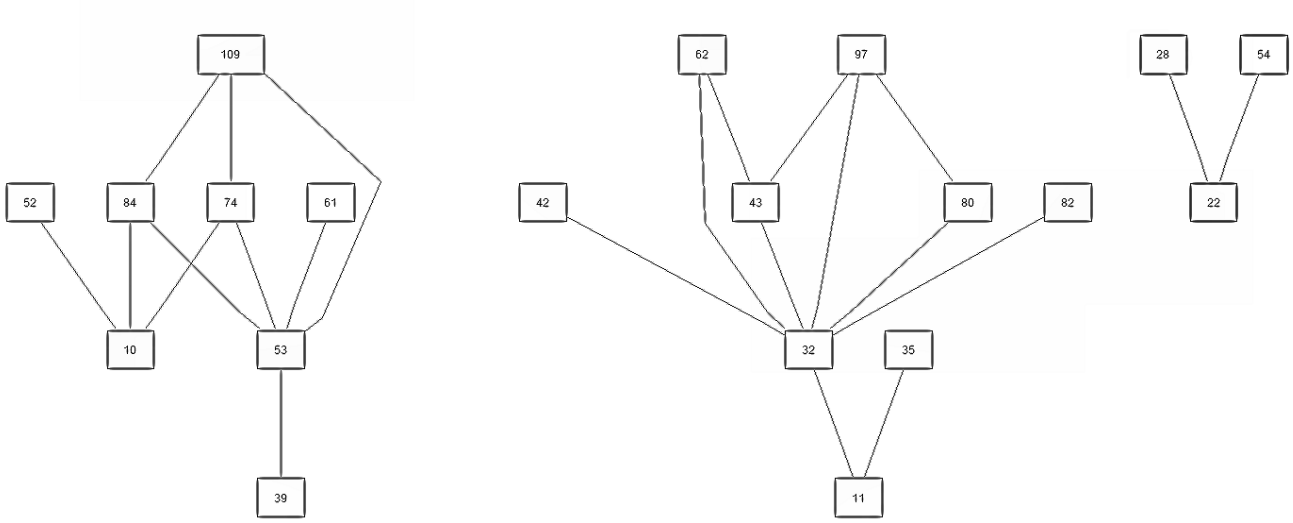

Figure 3 – Graphic representation of the Graph described by the adjacency matrix *C(i,j)* considering only the edges corresponding to DNA-matches greater or equal to 24 Mbp. Isolated individuals and groups of two are not reported in the figure

Figure 3 corresponds to the idea of unconnected graph that we associate with the separation of the different ancestral lines of the TU. Surprisingly enough, when the results of the Graph Theory are compared with the pre-existing genealogical information on some of the matching individuals, it turns out that the two large groups correspond to relatives of the TU related to the maternal grandfather (at the center of the figure) and maternal grandmother (at the left). Another small group of three individuals, at the right in figure 3, shows up, containing an individual associated to the maternal grandmother's lineage of the TU (n.22). The two individuals that can be identified with reasonable certainty as belonging to the paternal grandfather's (n. 118) and grandmother's (n. 96) lineage of the TU, remains unconnected. These results are summarized in Table 1. The individuals underlined and marked in bold are the ones for whom a genealogical evidence exists, and therefore can be assigned with certainty to a given lineage. The ones underlined and marked in italic, on the other hand, cannot be assigned with similar certainty, although there are strong independent clues suggesting that they would actually belong to that lineage.

Table 1 – Classification of the individuals according to their lineage (24 Mbp threshold). Individuals underlined and marked in bold are the ones for whom a genealogical evidence exists, and therefore can be assigned with certainty to a given lineage. The ones underlined and marked in italic, on the other hand, cannot be assigned with similar certainty, although there are strong independent clues suggesting that they would actually belong to that lineage.

Paternal GF	Paternal GM	Maternal GF	Maternal GM	Unclassified
		62	52	28
		97	*109*	54
		42	84	*22*
		43	*74*	
		80	61	
		82	*10*	
		32	*53*	
		35	39	
		11		

The adoption of a conservative threshold (24 Mbp / approx. 8 generations distance / 3rd – 4th cousin range) to define a link between the individuals produced the classification reported in Table I, which is robust and reliable. However, only 17 individuals over a total of 120 (110 with at least one DNA match besides the TU) are attributed to the corresponding ancestral lineage.

Reducing the level of the threshold to 12 Mbp (approx. 9 generations distance) increases the number of individuals that can be associated to the different groups (Supplementary Figure 2). Individual 22 is now correctly associated to the maternal grandmother's group, along with the other members of his/her subgroup. Most importantly the graph now shows an additional group of three individuals (21, 46 and 118) that can be associated to the TU paternal grandfather's lineage, on the basis of independent genealogical information existing for individual 118.

Further lowering the threshold to 6 Mbp (approx. 10 generations distance, i.e. a 4th – 5th cousin range, which is usually considered the lower limit for having a significant DNA match between two individuals) allows to recover important information, graphically represented in figure 4.

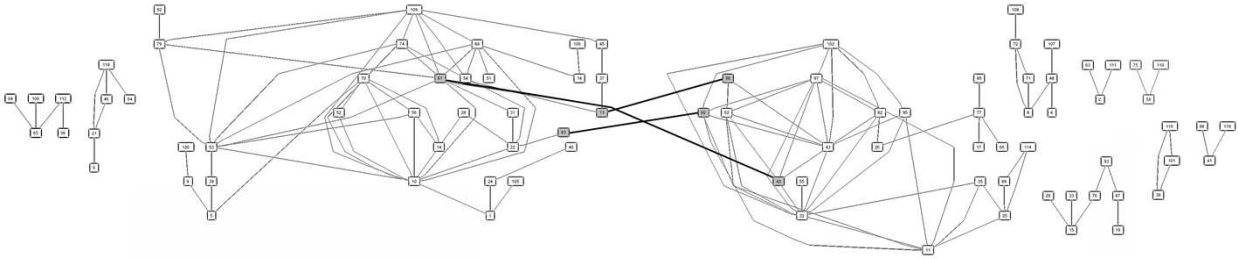

Figure 4 – The same as in figure 3, considering only the edges corresponding to DNA-matches greater or equal to 6 Mbp. The individuals connecting the two main groups and their links are evidenced. Isolated individuals and groups of two are not reported in the figure

From the analysis of figure 4 it is evident that after lowering the threshold to 6 Mbp, a connection appears between the two main groups. The key elements which are linked to both the groups (corresponding to the maternal grandparents of the TU) are individual 83 (initially classified in the maternal GM group) which connects with individual 80 in the maternal GF group), individual 61 of the maternal GM group which connects with individual 42 in the maternal GF group, and individual 86 of the maternal GF group which connects with individual 13 of the maternal GM group.

Lowering the threshold also increased the number of individuals associated to the paternal grandfather of the TU, which at this level formed a group of five persons (118, 46, 21, 6 and 64) connected by the same sub-graph, and recovered a new group of five individuals (96, 112, 65, 100 and 68) that can be associated to the TU paternal grandmother's lineage on the basis of independent genealogical information existing for individual 96.

The main information that can be derived by the comparison of the Graphs obtained using different thresholds on the edge weight is a classification of the individuals according to the different ancestral lineages, with increasing 'levels of confidence'. In that respect, Supplementary Figure 1 would give a minimum level of information, providing classification at the confidence level of the minimum match in the *C(i,j)* matrix, which in our case is 2 Mbp, subsequently refined at higher thresholds of genomic sharing in Figure 3, Supplementary Figure 2 and Figure 4.

The most important results of this paper are shown in Table II, where the classification of the DNA-relatives of the TU is reported according to his maternal and paternal ancestral lineages, with the corresponding confidence level, or 'strength', in brackets. The individuals connecting the groups corresponding to the two maternal grandparents are assigned to both the groups and marked in gray.

Table 2 – Classification of the individuals according to their ancestral lineage. The corresponding level of confidence of the classification is reported in brackets. The individuals connecting the two groups of the maternal grandparents are marked in gray.

Paternal GF	Paternal GM	Maternal GF	Maternal GM	Unclassified
118 (12)	*96* (6)	**97** (24)	52 (24)	116
46 (12)	112 (6)	62 (24)	*109* (24)	101
21 (12)	65 (6)	**42** (24)	84 (24)	38
6 (6)	100 (6)	43 (24)	*74* (24)	----------
64 (6)	68 (6)	80 (24)	61 (24)	29
		82 (24)	*10* (24)	33
		32 (24)	*53* (24)	15
		35 (24)	39 (24)	76
		11 (24)	*70* (12)	93
		102 (12)	83 (12)	87
		95 (12)	54 (12)	19
		25 (6)	51 (12)	----------
		89 (6)	92 (12)	63
		114 (6)	79 (12)	111
		83 (6)	*14* (12)	2
		61 (6)	*22* (12)	----------
		85 (6)	9 (12)	75
		77 (6)	5 (12)	110
		17 (6)	28 (12)	50
		66 (6)	1 (6)	----------
		26 (6)	24 (6)	88
		55 (6)	56 (6)	119
		13 (6)	42 (6)	41
		86 (6)	18 (6)	----------
			106 (6)	108
			31 (6)	72
			13 (6)	71
			86 (6)	8
			45 (6)	48
			37 (6)	4
			120 (6)	107
			105 (6)	

			40 (6)	
			80 (6)	

241

242

243 The Graph Theory method here proposed is capable of reliably classifying 62 individuals at
244 strength 6 (Mbp) over a total of 110 DNA-relatives of the TU (56%). Six other unclassified
245 groups with more than two members can also be determined. Some of them could be
246 connected to the main groups if additional information from new DNA relatives of the TU will
247 become available in the future.

248

Conclusion

The statistical method presented in this work can be usefully exploited for extracting genealogical information from genetic/genomic data. The input data are usually 'fuzzy' and, therefore, the methods used for their analysis should be robust enough for providing useful information. The approach proposed, based on the Graph representation of the adjacency matrix built from the mutual matches between the DNA-relatives of the test user, after the setting of a suitable threshold fulfils this requirement. The method, for which the code is provided at the bottom of this paper, could be easily implementable by the genetic service providers for an easy visualization of the DNA-links existing between the customer and the other users of the service, at different levels of confidence.

References

[1] https://genographic.nationalgeographic.com/

[2] https://www.23andme.com/

[3] https://www.familytreedna.com/

[4] http://www.gedmatch.com/

[5] Nachman, M. W. (2001) Single nucleotide polymorphisms and recombination rate in humans, *Trends in genetics,* 17 (9), 481–485.

[6] Muni S. Srivastava, *Methods of Multivariate Statistics* (2002) Wiley.

[7] Smouse, P. E. and Long, J. C. (1992) Matrix correlation analysis in anthropology and genetics, *American Journal of Physical Anthropology*, 35, 187–213.

[8] Stevens, E.L., Heckenberg, G., Roberson, E.D.O., Baugher, J.D., Downey, T.J., Pevsner, J. (2011) Inference of relationships in population data using identity-by-descent and identity-by-state, *PLoS Genetics*, 7 (9), art. no. e1002287.

[9] Bondy, A., Murty, U.S.R., *Graph Theory* (2008) Springer-Verlag London.

[10] Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R., Bagos, P. G. (2011) Using graph theory to analyze biological networks, *BioData Mining*, 4:10.

[11] Gehlenborg, N., Wong, B. (2012) Points of View: Networks, *Nature Methods*, 9, 115.