

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

3 **Supplementary Materials**

4 Vincenzo Palleschi^{1,2}, Luca Pagani^{3,4†}, Stefano Pagnotta¹, Giuseppe Amato⁵, Sergio Tofanelli⁶

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

8
9 ²Department of Civilizations and Forms of Knowledge
10 University of Pisa
11 Via G. Galvani, 1 – 56126 Pisa (ITALY)

12
13 ³Division of Biological Anthropology, University of Cambridge, UK

14
15 ⁴Department of Biological, Geological and Environmental Sciences,
16 University of Bologna
17 Via Selmi 3, 40126 Bologna (ITALY)

18
19 ⁵Institute of Sciences and Technology of Information
20 Research Area of National Research Council
21 Via G. Moruzzi, 1 – 56124 Pisa (ITALY)

22
23 ⁶Department of Biology
24 University of Pisa
25 Via L. Ghini, 13 – 56126 Pisa (ITALY)
26

1

2 **Matlab® code used to obtain graphs from the similarity matrix obtained from 23andMe**

```

3 load 'matrix.dat'
4 thr=6;
5 table=tril(matrix);
6 table=table.*(table >= thr);
7 bg=biograph(sparse(table),'','EdgeType','segmented','ShowWeights','off','ShowArrows
8 ','off','LayoutType','hierarchical');
9 bg.view;

```

10

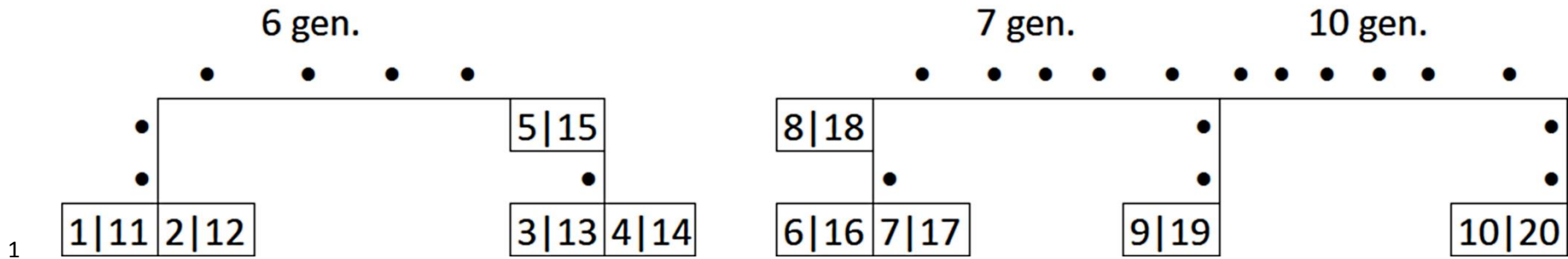
11

12 **Supplementary Table 1:** Expected amount of shared genome between pairs of individuals separated
13 by G number of generations, according to a model of independent sorting among lineages. Using data
14 available from 23andMe we investigated the G range highlighted in gray.

G (number of generations apart)	prop of shared genome	Shared Mbp	N of potential ancestors
1	0.5	3030	2
2	0.25	1515	4
3	0.125	758	8
4	0.0625	379	16
5	0.03125	189	32
6	0.01563	95	64
7	0.00781	47	128
8	0.00391	24	256
9	0.00195	12	512
10	0.00098	6	1024
11	0.00049	3	2048
12	0.00024	1	4096
13	0.00012	1	8192

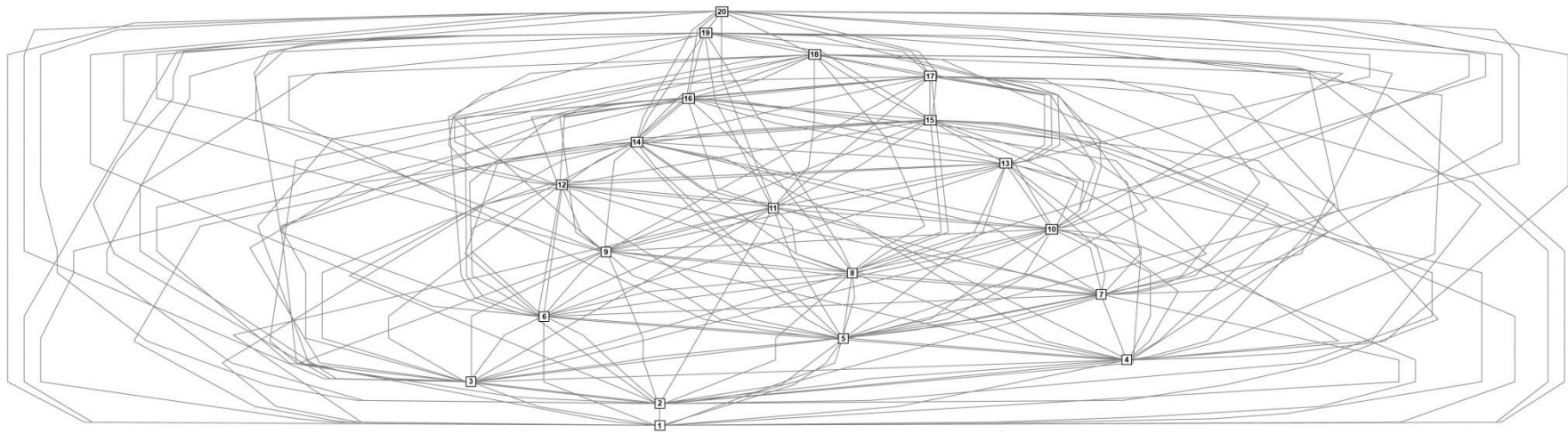
15

16



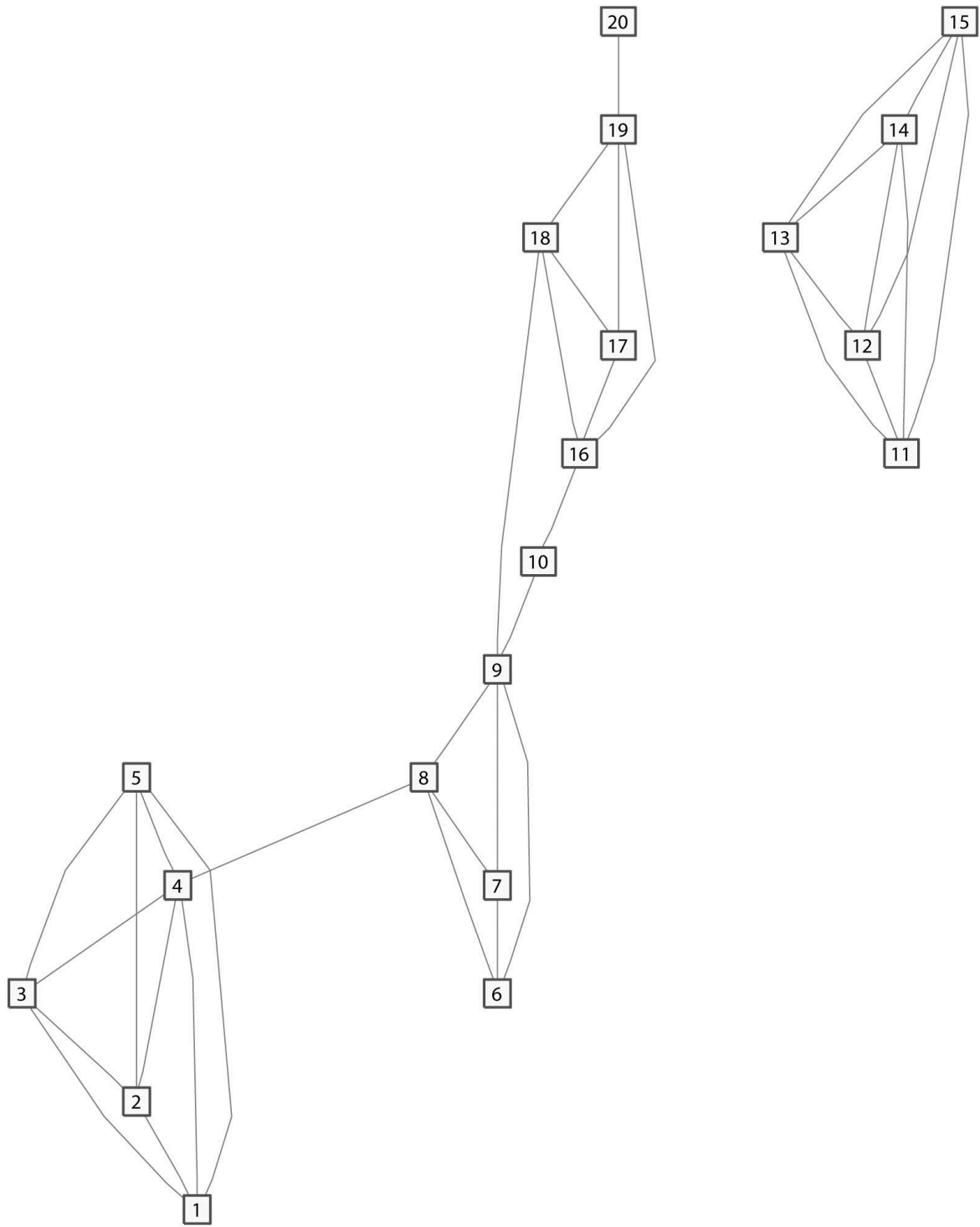
2 **Supplementary Figure 1 – Idealized genealogy.** Two identical genealogies including individuals from 1 to 10 and from 11 to 20, reported
 3 in this scheme, were deployed to infer the theoretical amounts of pairwise genome sharing reported in Supplementary Table 2. Each
 4 dot represents a generation separating any two given individuals within the genealogy.

5



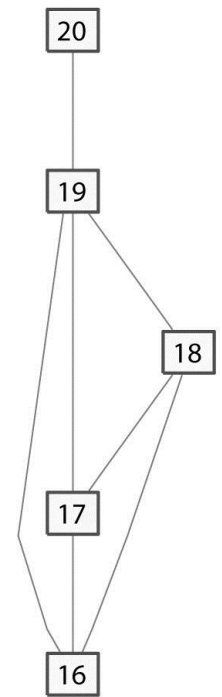
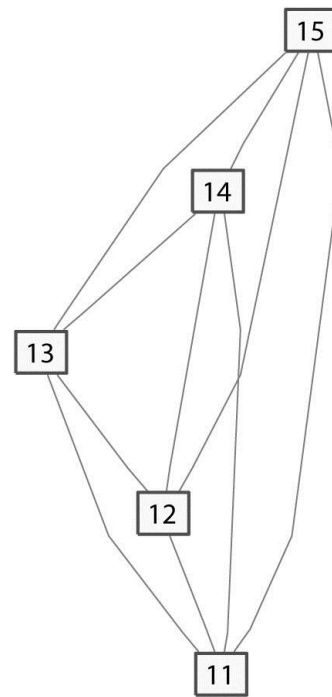
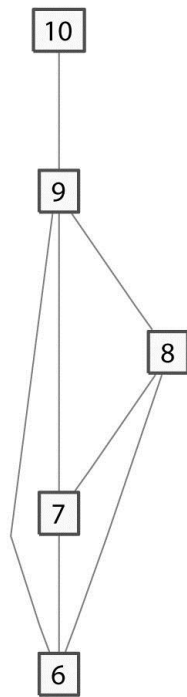
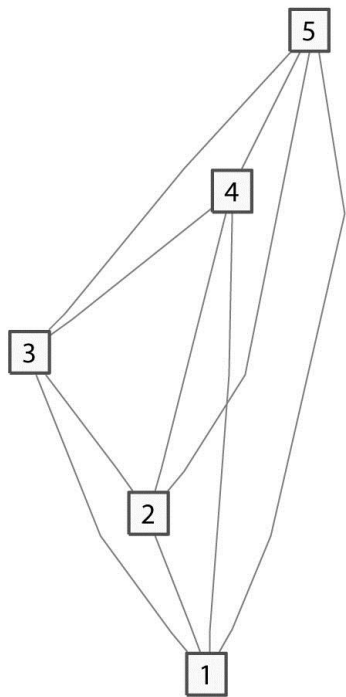
1
2 **Supplementary Figure 2** - Graphic representation of the simulated adjacency matrix $C(i,j)$ reported in Supplementary Table 2. No
3 filtering threshold was applied.

4



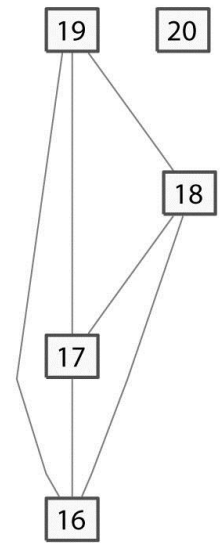
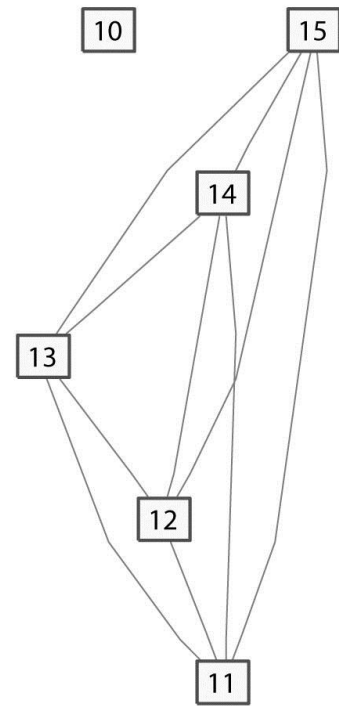
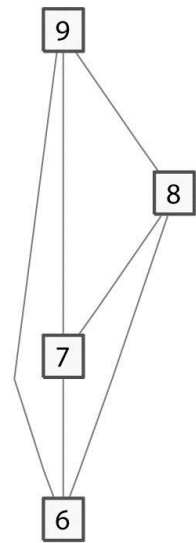
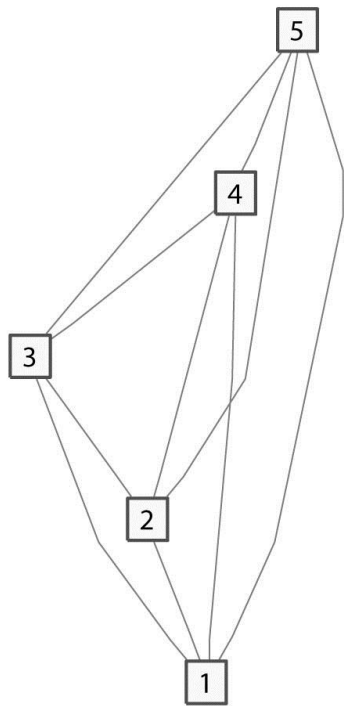
1

2 **Supplementary Figure 3** - Graphic representation of the simulated adjacency matrix $C(i,j)$ reported in
 3 Supplementary Table 2. Only links with more than 2 Mb of genomic sharing are displayed.



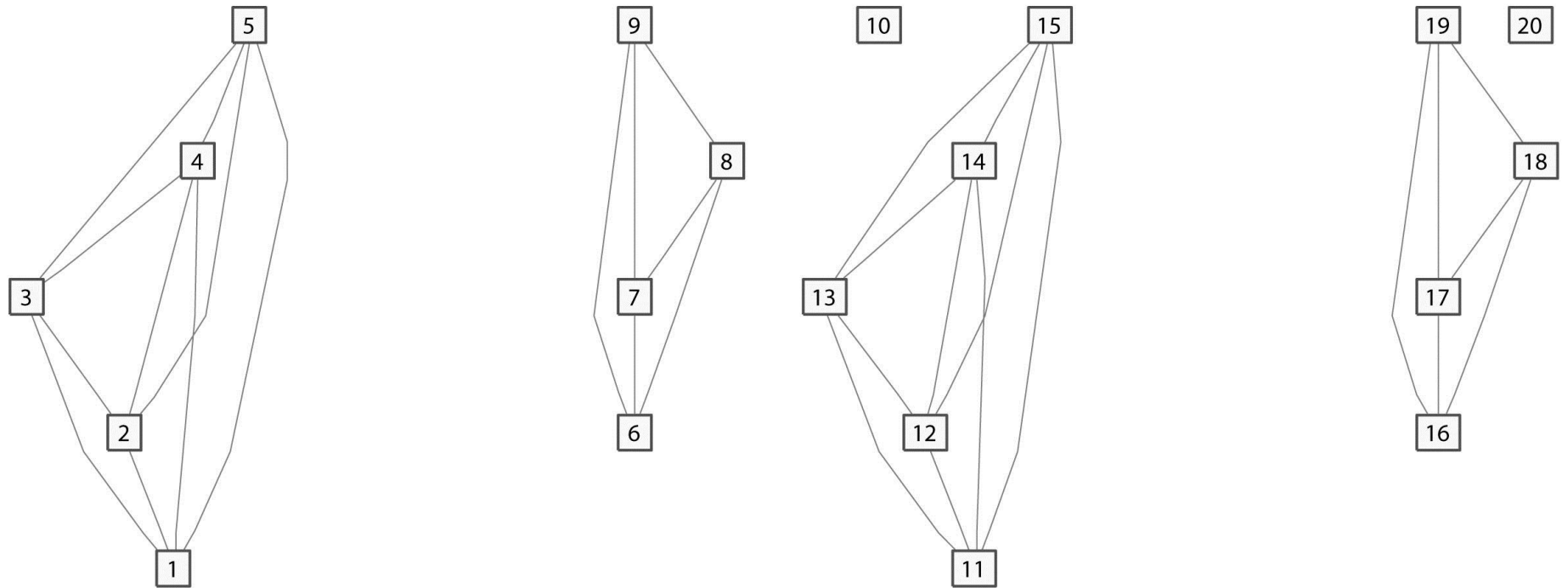
1
 2 **Supplementary Figure 4** - Graphic representation of the simulated adjacency matrix $C(i,j)$ reported in Supplementary Table 2. Only links
 3 with more than 3 Mb of genomic sharing are displayed.

4



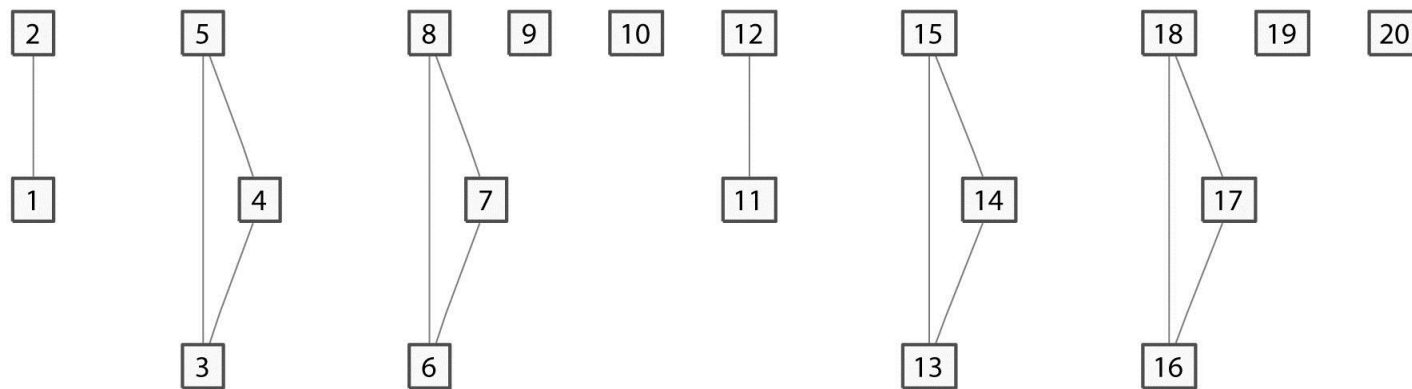
1

2 **Supplementary Figure 5** - Graphic representation of the simulated adjacency matrix $C(i,j)$ reported in Supplementary Table 2. Only links
 3 with more than 7 Mb of genomic sharing are displayed.



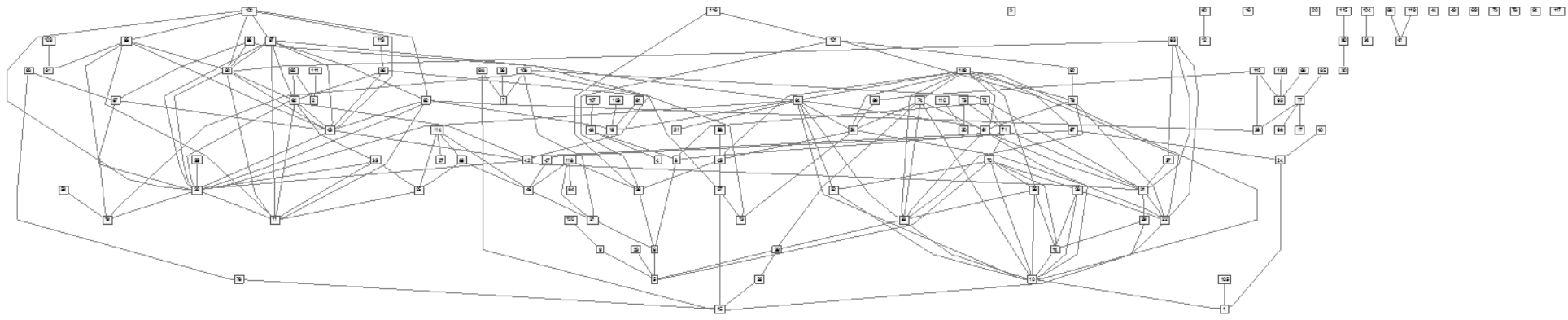
1

2 **Supplementary Figure 6** - Graphic representation of the simulated adjacency matrix $C(i,j)$ reported in Supplementary Table 2. Only links
 3 with more than 12 Mb of genomic sharing are displayed.



1

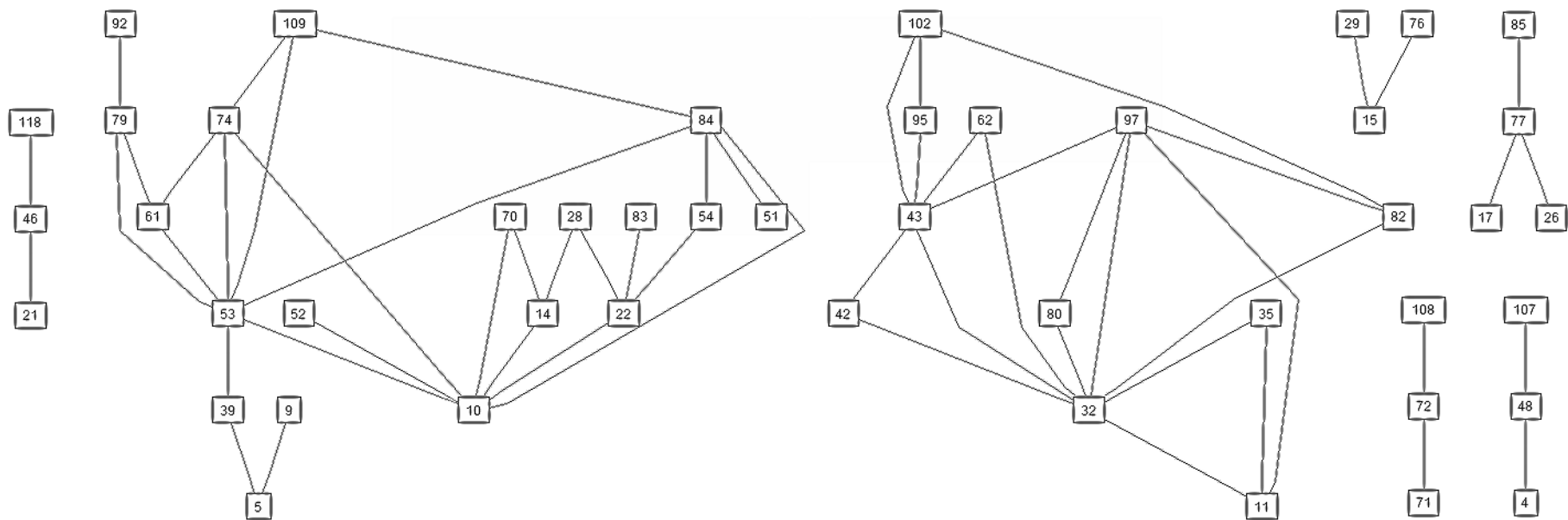
2 **Supplementary Figure 7-** Graphic representation of the simulated adjacency matrix $C(i,j)$ reported in Supplementary Table 2. Only links
 3 with more than 48 Mb of genomic sharing are displayed.



1

2 **Supplementary Figure 8** – Graphic representation of the Graph described by the adjacency matrix $C(i,j)$. The connected subgraph at the
3 left links 100 individuals.

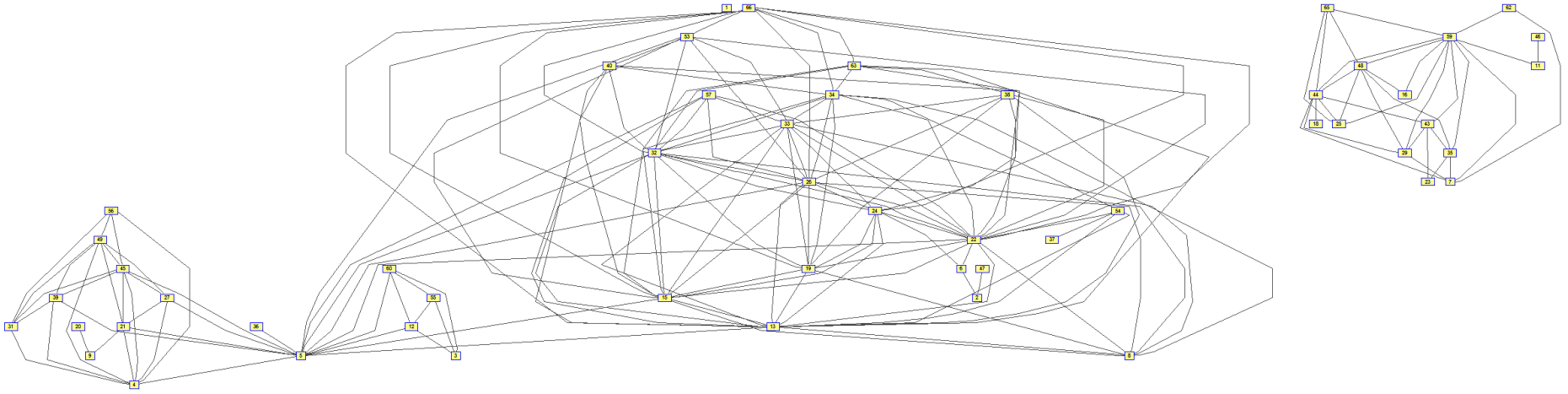
4



1

2 **Supplementary Figure 9** – The same as in figure 3, considering only the edges corresponding to DNA-matches greater or equal to 12 Mbp. Isolated
 3 individuals and groups of two are not reported in the figure.

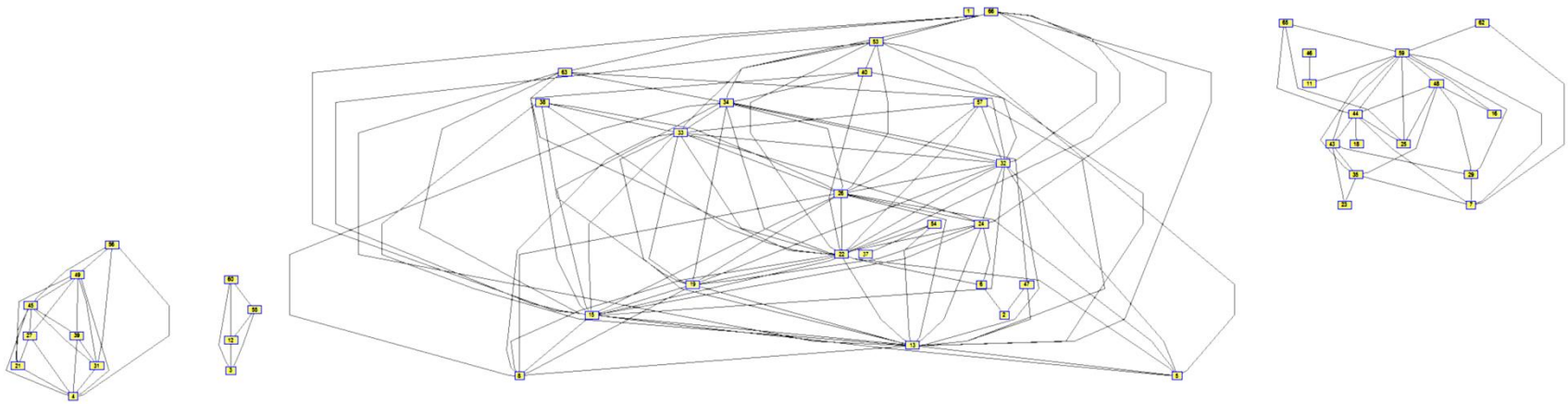
4



1

2 **Supplementary Figure 10** – Graphic representation of the additional adjacency matrix $C(i,j)$ obtained from TU2, reported in Supplementary Table 4.
 3 No filtering threshold was applied. Isolated graphs showing less than three individuals were removed for sake of readability. The subnetwork to the
 4 left hand side of the Graph is the “Mendel family”, a real genealogy made freely available by 23andMe after assigning a mock surname. Given the
 5 lack of known relationship between TU2 and the Mendel family we take the existing link as further support for the need of a 6Mbp threshold when
 6 interpreting the genetic results.

7



1

2 **Supplementary Figure 11** – Graphic representation of the additional adjacency matrix $C(i,j)$ obtained from TU2, reported in Supplementary Table 4.
 3 Only links with more than 6 Mb of genomic sharing are displayed. Isolated graphs showing less than three individuals were removed for sake of
 4 readability. The subnetwork to the left hand side of the Graph is the “Mendel family”, a real genealogy made freely available by 23andMe after
 5 assigning a mock surname, which is show as disconnected from the TU2 graph, as result of the noise reduction achieved with the 6Mb threshold.

6