The American Journal of Human Genetics, Volume *103*

Supplemental Data

Inferring Transmission Histories of Rare Alleles in Population-Scale Genealogies

Dominic Nelson, Claudia Moreau, Marianne de Vriendt, Yixiao Zeng, Christoph Preuss, Hélène Vézina, Emmanuel Milot, Gregor Andelfinger, Damian Labuda, and Simon Gravel

Supplemental Material

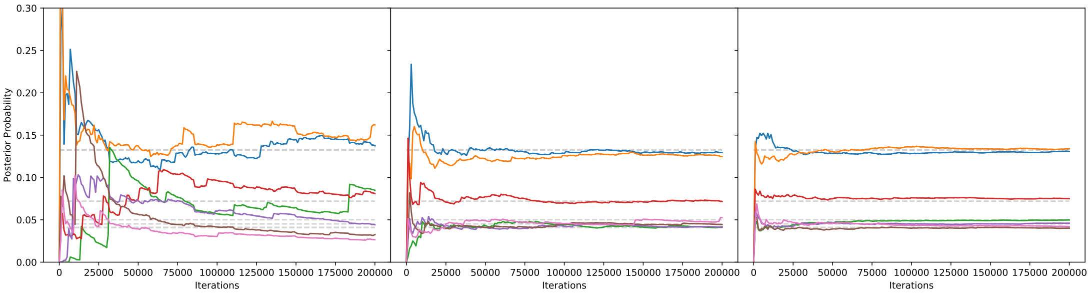
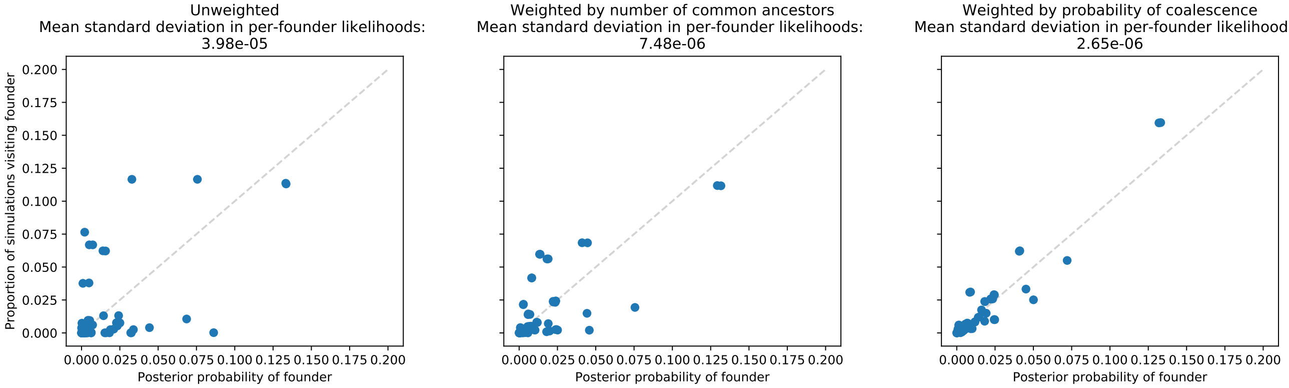


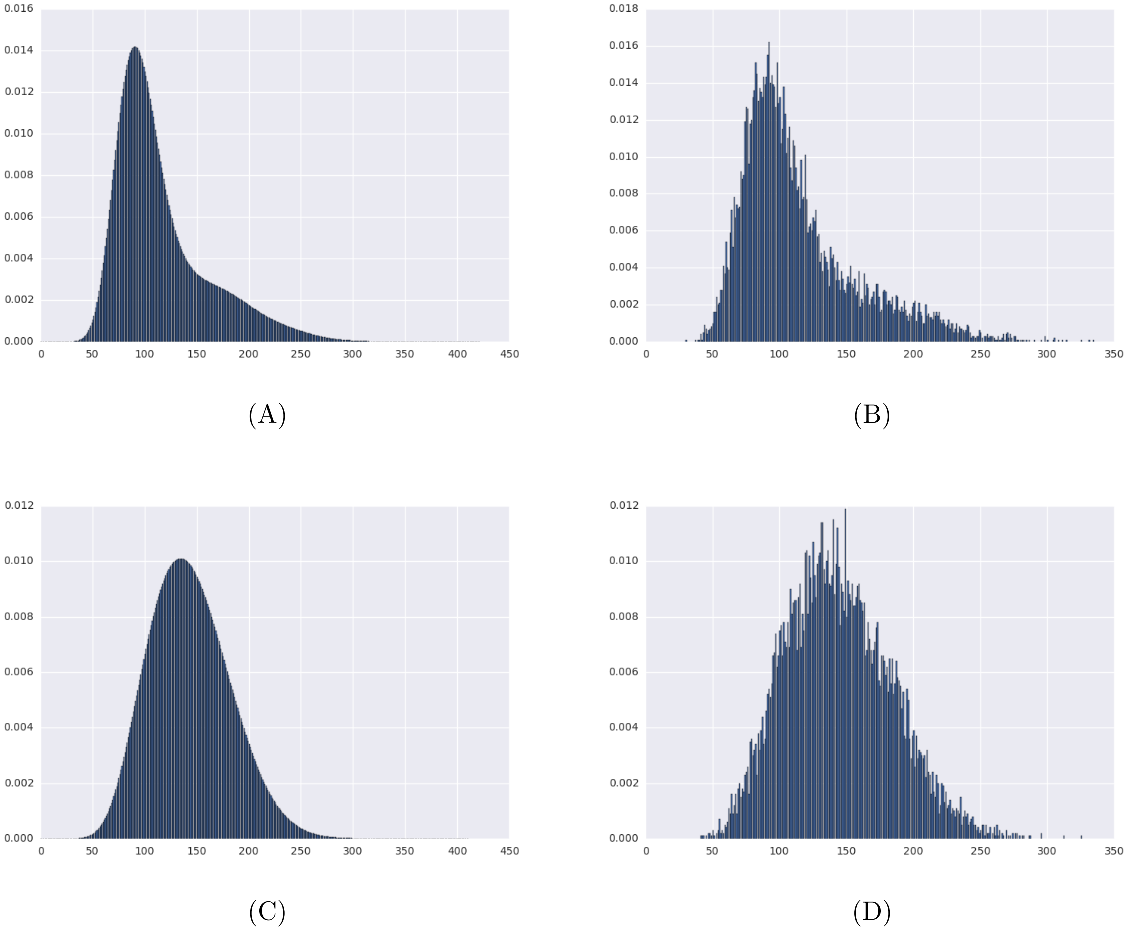
Figure S1: Convergence of likelihood estimates for 7 most-likely ancestors of a minor allele in a single simulated carrier panel. With importance sampling based left-to-right on: a possible path to coalescence only; the number of common ancestors shared with all other simulated carriers of the minor allele; likelihood of coalescing with other lineages.


Unweighted
Mean standard deviation in per-founder likelihoods:
3.98e-05

Weighted by number of common ancestors
Mean standard deviation in per-founder likelihoods:
7.48e-06

Weighted by probability of coalescence
Mean standard deviation in per-founder likelihoods:
2.65e-06

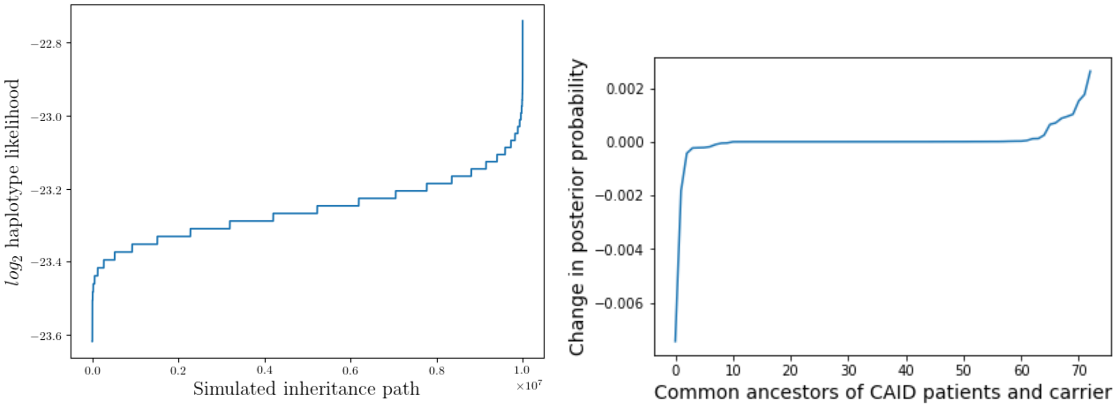
Figure S2: Proportion of simulated inheritance paths which lead to each founder versus converged founder posterior probability. With importance sampling based left-to-right on: a possible path to coalescence only; the number of common ancestors shared with all other simulated carriers of the minor allele; likelihood of coalescing with other lineages. Uses the same simulated carrier panel as Fig. S1. Importance sampling convergence is fastest when outcomes are sampled proportionally to their true probability.[1]



(A) (B)

(C) (D)

Figure S3: Comparison of simulated inheritance path allele frequency distributions (B, D) and their approximation via convolution of the distributions of the tree boundary (A, C) using the method described in Appendix C.



(A) (B)

Figure S4: (A) Log-likelihoods of observing shared 2.9Mb segment in CAID patients and carrier, over all simulated inheritance paths. (B) Impact of incorporating shared haplotype length among CAID patients on estimated posterior probabilities of each common ancestor having been the true origin of the minor allele.

Ind	Father	Mother	Sex
1	11	12	1
2	15	14	2
3	15	14	2
11	102	101	1
12	0	0	2
13	102	101	1
14	0	0	2
15	103	104	1
16	103	104	2
18	105	106	2
19	105	106	2
20	107	108	2
21	107	108	1
101	0	0	2
102	202	201	1
103	0	0	1
104	202	201	2
105	202	201	1
106	0	0	2
107	202	201	1
108	0	0	2
201	0	0	2
202	0	0	1

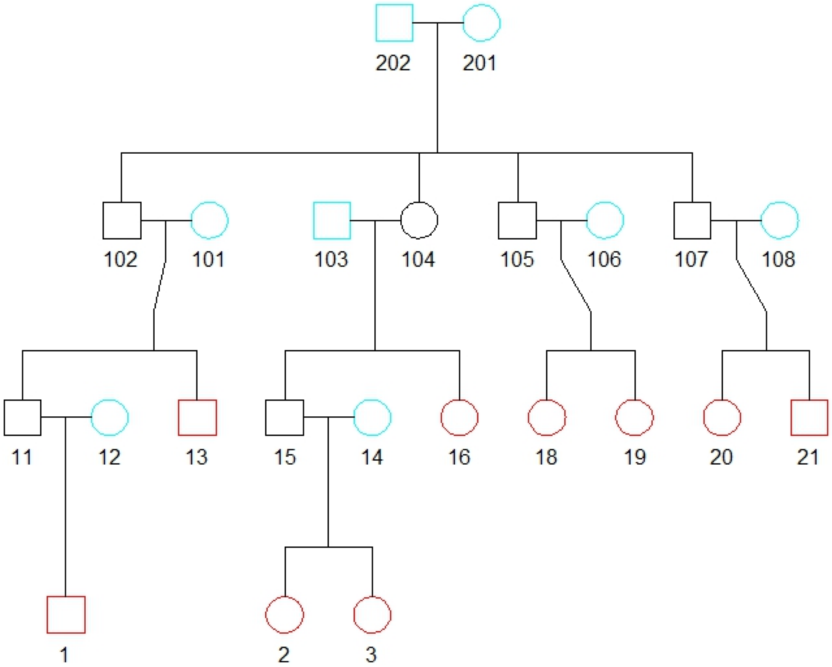
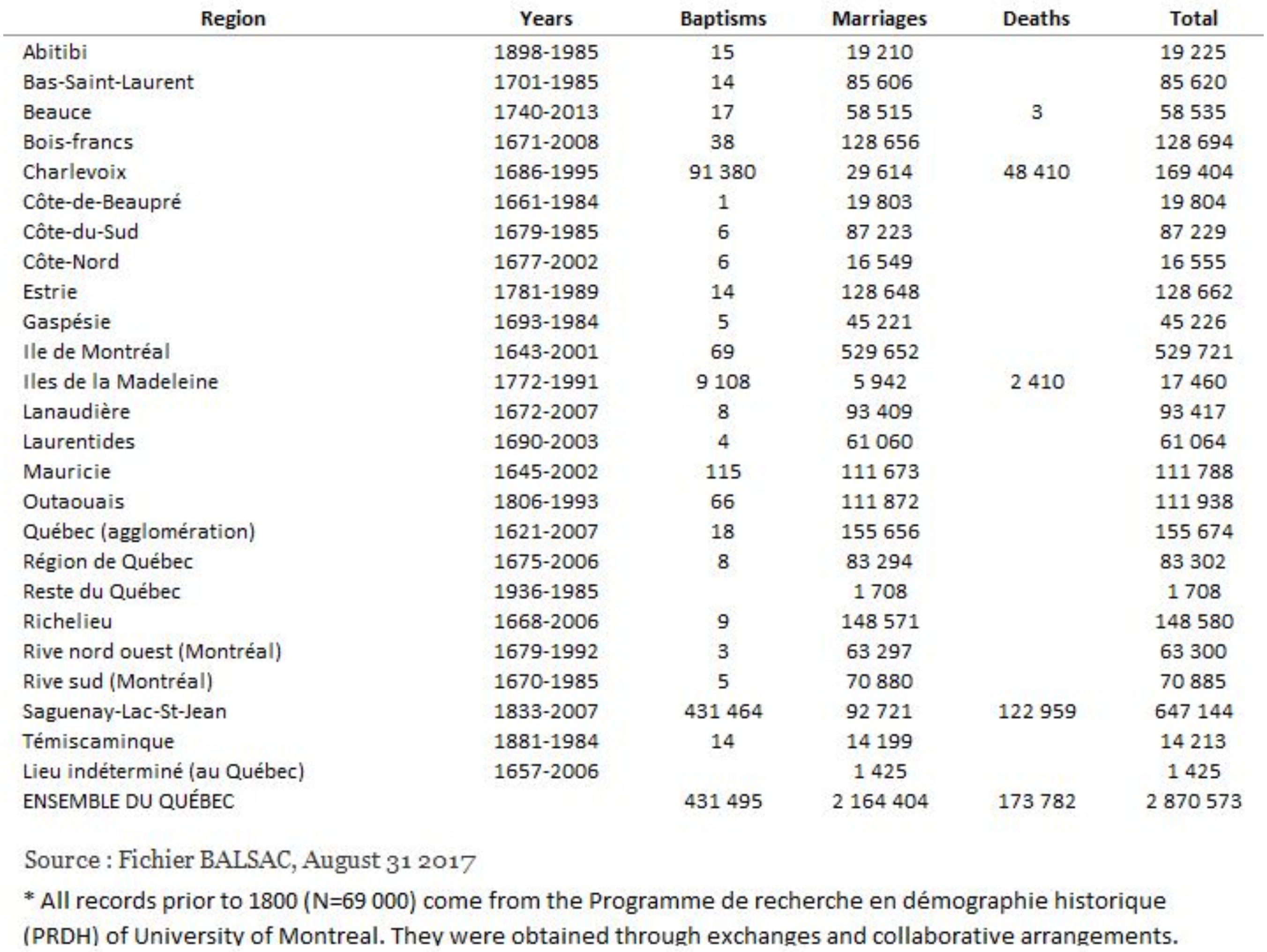


Table S1: Example pedigree and corresponding data format.

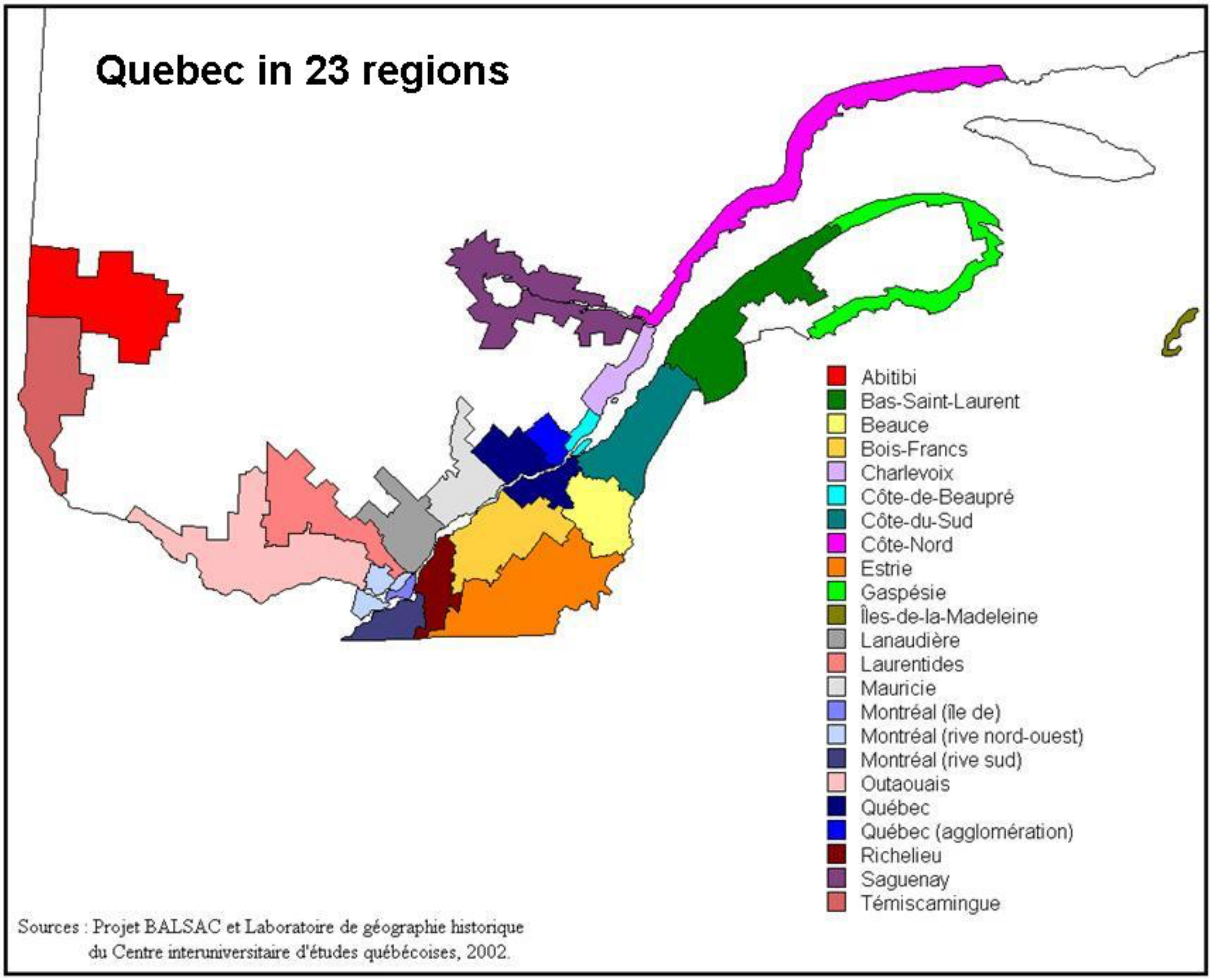


Region	Years	Baptisms	Marriages	Deaths	Total
Abitibi	1898-1985	15	19 210		19 225
Bas-Saint-Laurent	1701-1985	14	85 606		85 620
Beauce	1740-2013	17	58 515	3	58 535
Bois-francs	1671-2008	38	128 656		128 694
Charlevoix	1686-1995	91 380	29 614	48 410	169 404
Côte-de-Beaupré	1661-1984	1	19 803		19 804
Côte-du-Sud	1679-1985	6	87 223		87 229
Côte-Nord	1677-2002	6	16 549		16 555
Estrie	1781-1989	14	128 648		128 662
Gaspésie	1693-1984	5	45 221		45 226
Ile de Montréal	1643-2001	69	529 652		529 721
Iles de la Madeleine	1772-1991	9 108	5 942	2 410	17 460
Lanaudière	1672-2007	8	93 409		93 417
Laurentides	1690-2003	4	61 060		61 064
Mauricie	1645-2002	115	111 673		111 788
Outaouais	1806-1993	66	111 872		111 938
Québec (agglomération)	1621-2007	18	155 656		155 674
Région de Québec	1675-2006	8	83 294		83 302
Reste du Québec	1936-1985		1 708		1 708
Richelieu	1668-2006	9	148 571		148 580
Rive nord ouest (Montréal)	1679-1992	3	63 297		63 300
Rive sud (Montréal)	1670-1985	5	70 880		70 885
Saguenay-Lac-St-Jean	1833-2007	431 464	92 721	122 959	647 144
Témiscaminque	1881-1984	14	14 199		14 213
Lieu indéterminé (au Québec)	1657-2006		1 425		1 425
ENSEMBLE DU QUÉBEC		431 495	2 164 404	173 782	2 870 573

Source : Fichier BALSAC, August 31 2017

* All records prior to 1800 (N=69 000) come from the Programme de recherche en démographie historique (PRDH) of University of Montreal. They were obtained through exchanges and collaborative arrangements.

Figure S5: Number of vital event records per region of Quebec.[2] Table reproduced July 18th, 2018 from http://balsac.uqac.ca/english/balsac-database/overview-of-data/



Sources : Projet BALSAC et Laboratoire de géographie historique du Centre interuniversitaire d'études québécoises, 2002.

Figure S6: Quebec regions used in the BALSAC project.[2] Figure reproduced September 12th, 2018 from `http://balsac.uqac.ca/english/balsac-database/overview-of-data/`

Region	Estimated Allele Frequency	95% Confidence Interval
ABITIBI	0.00128	(0.00127, 0.00129)
BAS SAINT LAURENT	0.00163	(0.00156, 0.00169)
BEAUCE	0.00425	(0.00408, 0.00443)
BOIS FRANCS	0.000882	(0.000858, 0.000908)
CHARLEVOIX	0.00643	(0.00630, 0.00654)
COTE DE BEAUPRE	0.00417	(0.00410, 0.00423)
COTE DU SUD	0.00183	(0.00176, 0.00190)
COTE NORD	0.00253	(0.00249, 0.00258)
ESTRIE	0.00144	(0.00141, 0.00146)
GASPESIE	0.000738	(0.000696, 0.000767)
ILE DE MONTREAL	0.000588	(0.000580, 0.000596)
ILES DE LA MADELEINE	2.61e-05	(2.45e-05, 2.80e-05)
LANAUDIERE	0.000462	(0.000450, 0.000473)
LAURENTIDES	0.000500	(0.000486, 0.000515)
MAURICIE	0.000808	(0.000789, 0.000825)
OUTAOUAIS	0.000349	(0.000340, 0.000356)
QUEBEC (AGGLOMERATION)	0.00183	(0.00179, 0.00187)
REGION DE QUEBEC	0.00118	(0.00113, 0.00124)
RICHELIEU	0.000581	(0.000566, 0.000598)
RIVE NORD OUEST (MTL)	0.000390	(0.000382, 0.000399)
RIVE SUD (MTL)	0.000477	(0.000470, 0.000482)
SAGUENAY (LAC ST JEAN)	0.00520	(0.00512, 0.00527)
TEMISCAMINGUE	0.000794	(0.000786, 0.000802)
All Probands	0.00167	

Table S2: Estimated regional frequencies of the CAID allele within the province of Quebec, among individuals linked to the BALSAC genealogical database. Confidence intervals estimated from bootstrapping over simulated inheritance paths.

Error Measure	Kinship-Based	ISgen	ISgen / Kinship
MAE	0.00105	0.000784	0.74
RMSE	0.00204	0.00169	0.83
MAE (estimated freq < 0.005)	0.000885	0.000591	0.67
RMSE, (estimated freq < 0.005)	0.00146	0.000983	0.67

Table S3: Mean absolute error and root mean squared error in regional allele frequency estimates for ISgen (path-based) and a kinship-based method. We simulated 100 patient panels and corresponding regional allele frequencies. Simulated regional allele frequencies were compared to inference results based on patient panels and estimated global allele frequency.

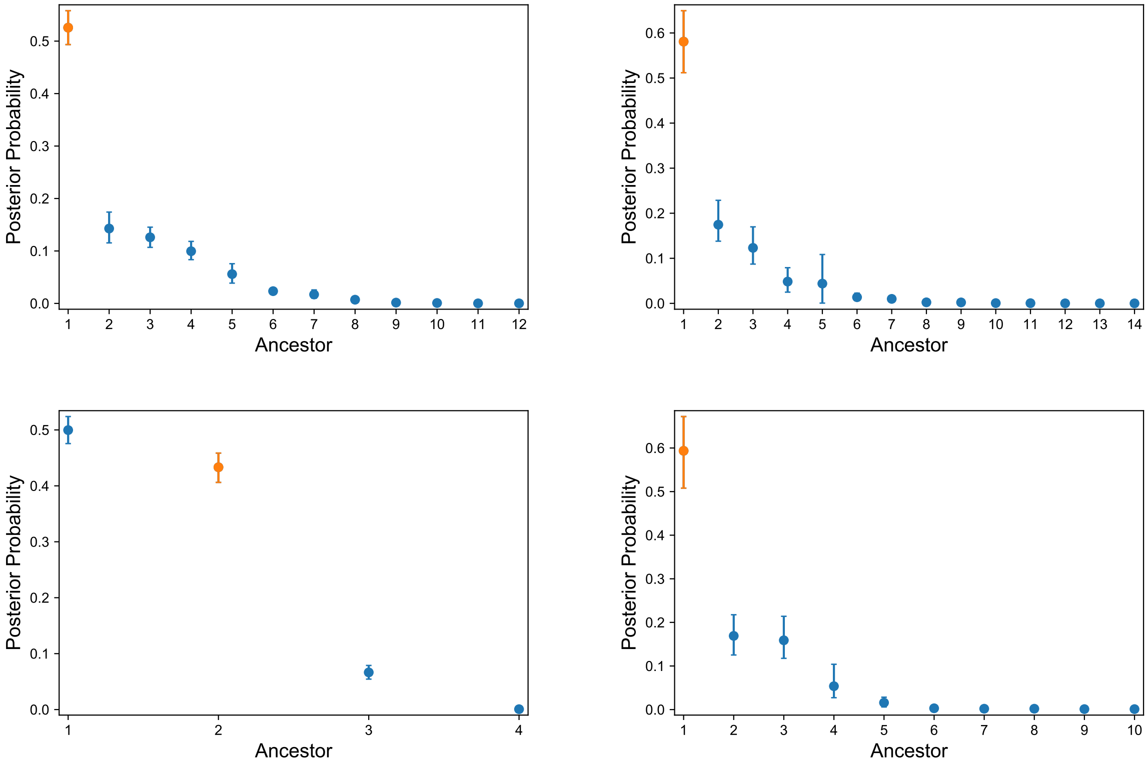


Figure S7: Ancestor posterior probabilities for 4 simulated patient panels, similar to the one displayed in Figure 4. The ancestor generating the panel is shown in orange. Error bars represent uncertainty due to the finite sample size (i.e., the finite number of iterations) in importance sampling. 95% confidence intervals were obtained from bootstrapping over iterations. Only ancestors with nonzero posterior probability are displayed, and ancestor labels represent ordering by posterior probability for a given simulation.

References

[1] Srinivasan, R. (2002). *Importance Sampling: Applications in Communications and Detection.* (Springer-Verlag).

[2] BALSAC. (2018). BALSAC Population Database: 2016-2017 Annual Report. `http://balsac.uqac.ca/english/files/2018/01/BALSAC_RA2017_EN_page_WEB_v2-1.pdf`.