Supplementary Material for

Assessment of kinship detection using RNA-seq data

Authors

Natalia Blay[1,2,3], Eduard Casas[1,2,4], Iván Galván-Femenía[1,5], Jan Graffelman[6,7], Rafael de Cid[1,5], Tanya Vavouri[1,2]

Affiliations

1. Program for Predictive and Personalized Medicine of Cancer, Germans Trias i Pujol Research Institute (PMPPC-IGTP), Badalona, Spain
2. Josep Carreras Leukaemia Research Institute (IJC), Campus ICO-Germans Trias i Pujol, Universitat Autònoma de Barcelona, Badalona, Spain
3. Masters Programme in Bioinformatics and Biostatistics, Universitat Oberta de Catalunya (UOC), Barcelona, Spain
4. Doctoral Programme in Biomedicine, Universitat de Barcelona, Barcelona, Spain
5. Genomes for Life - GCAT lab Group - Germans Trias i Pujol Research Institute, Can Ruti Campus, Ctra de Can Ruti, Camí de les Escoles s/n, Badalona, Barcelona, Spain.
6. Department of Statistics and Operations Research Universitat Politècnica de Catalunya, Barcelona, Spain.
7. Department of Biostatistics, University of Washington, Seattle, USA.

Corresponding author: Tanya Vavouri tvavouri@carrerasresearch.org

This PDF file includes:

Supplementary Table 1. Source and identifiers of publicly available datasets used in the manuscript.

Data type	*Identifier / accession #*	*Reference*	*Usage in the manuscript*
RNA-seq	SRR1258217 SRR1258218 SRR1258219 SRR1258220 SRR1258221 SRR1258222 SRR1258223 SRR1258224 SRR1258225 SRR1258226 SRR1258227 SRR1258228 SRR1258229 SRR1258230 SRR1258231 SRR1258232 SRR1258233	Li *et al.*, Am J Hum Genet, 2014. Data retrieved from Gene Expression Omnibus record GSE56961.	Kinship detection using RNA-seq data from the 17-member CEPH-UTAH family 1463.
RNA-seq	SRR9070085 SRR9070086 SRR9070087 SRR9070088 SRR9070089 SRR9070090 SRR9070091 SRR9070092 SRR9070093	Gifford *et al.* Science, 2019. Data retrieved from Gene Expression Omnibus record GSE131322.	Kinship detection using RNA-seq data from iPSC-derived cardiomyocytes from a trio. Three replicates per sample.
RNA-seq	SRR7739176 SRR7739177 SRR7739178 SRR7739179 SRR7739180 SRR7739181 SRR7739182 SRR7739183 SRR7739184 SRR7739185 SRR7739186	Oda *et al.* Front Immunol 2019. Data retrieved from Gene Expression Omnibus record GSE118901	Kinship detection using RNA-seq data from whole blood and peripheral blood mononuclear cells from 1 parent, 1 offspring and 3 unrelated individual with 1-2 replicates per sample.
RNA-seq	SRR6926729 SRR6926730 SRR6926731 SRR6926732 SRR6926733 SRR6926734 SRR6926735 SRR6926736 SRR6926737 SRR6926738 SRR6926739	Boutboul *et al* J Clin Invest 2018. Data retrieved from Gene Expression Omnibus records GSE112706 and GSE112707	Kinship detection using AmpliSeq (targeted RNA-seq) data from Naive CD4+ cells and LPS-stimulated monocytes from 8 unrelated individuals and two siblings with 2 replicates per sample.

	SRR6926740 SRR6926741 SRR6926742 SRR6926743 SRR6926744 SRR6926281 SRR6926282 SRR6926283 SRR6926284 SRR6926285 SRR6926286 SRR6926287 SRR6926288 SRR6926289 SRR6926290 SRR6926291 SRR6926292 SRR6926293 SRR6926294 SRR6926295 SRR6926296 SRR6926297 SRR6926298		
Genotypes	NA12877 NA12878	Eberle, MA *et al.* Genome Res, 2017	Estimation of genotype prediction accuracy.
Genotypes	HG02363 HG02372 HG02377 HG02381 HG02387 HG02388 HG02250 HG02353 HG02371 HG02373 HG02375 HG02386	The 1000 Genomes Project Consortium, Nature, 2015.	Kinship detection using simulated RNA-seq data from genotypes from **related** individuals (**CDX** population).
Genotypes	HG00096 HG00097 HG00099 HG00100 HG00101 HG00102 HG00231 HG00232 HG00233 HG00234 HG00235	The 1000 Genomes Project Consortium, Nature, 2015.	Kinship detection using simulated RNA-seq data from genotypes from **unrelated** individuals (**GBR** population).
Genotypes	HG02133 HG02134 HG02136 HG02137	The 1000 Genomes Project Consortium, Nature, 2015.	Kinship detection using simulated RNA-seq data from genotypes from **unrelated** individuals (**KHV** population).

	HG02138 HG02139 HG01840 HG01841 HG01842 HG01843 HG01844		
Genotypes	HG01669 HG01602 HG01670 HG01605 HG01765 HG01766 HG01500 HG01501 HG01503 HG01504 HG01506	The 1000 Genomes Project Consortium, Nature, 2015.	Kinship detection using simulated RNA-seq data from genotypes from **unrelated** individuals (**IBS** population).
Genotypes	NA19017 NA19019 NA19020 NA19025 NA19026 NA19027 NA19307 NA19308 NA19309 NA19310 NA19312	The 1000 Genomes Project Consortium, Nature, 2015.	Kinship detection using simulated RNA-seq data from genotypes from **unrelated** individuals (**LWK** population).
Genotypes	HG01119 HG01121 HG01122 HG01130 HG01131 HG01142 HG01133 HG01134 HG01136 HG01137 HG01139	The 1000 Genomes Project Consortium, Nature, 2015.	Kinship detection using simulated RNA-seq data from genotypes from **unrelated** individuals (**CLM** population).
Genotypes	HG00759 HG00766 HG00844 HG00851 HG00864 HG00867 HG00978 HG00982 HG01028 HG01029 HG01031	The 1000 Genomes Project Consortium, Nature, 2015.	Kinship detection using simulated RNA-seq data from genotypes from **unrelated** individuals (**CLX** population).
Genotypes	HG01961	The 1000 Genomes	Kinship detection using

	HG01965 HG02006 HG02102 HG02150 HG02252 HG02262 HG02265 HG02266 HG02274 HG02275	Project Consortium, Nature, 2015.	simulated RNA-seq data from genotypes from **unrelated** individuals (**PEL** population).
Genotypes	HG01985 HG02051 HG02052 HG02095 HG02111 HG02283 HG02330 HG02332 HG02334 HG02337 HG02339	The 1000 Genomes Project Consortium, Nature, 2015.	Kinship detection using simulated RNA-seq data from genotypes from **unrelated** individuals (**ACB** population).
Genotypes	HG01500 HG01501 HG01503 HG01504 HG01506 HG01767 HG01768 HG01785 HG01786 HG02219 HG01507 HG01509 HG01510 HG01512 HG01513 HG01515 HG01516 HG01518 HG01519 HG01521	The 1000 Genomes Project Consortium, Nature, 2015.	Identification of a pair of RNA-seq samples from related individuals in a dataset containing 20 unrelated individuals (**IBS** population).

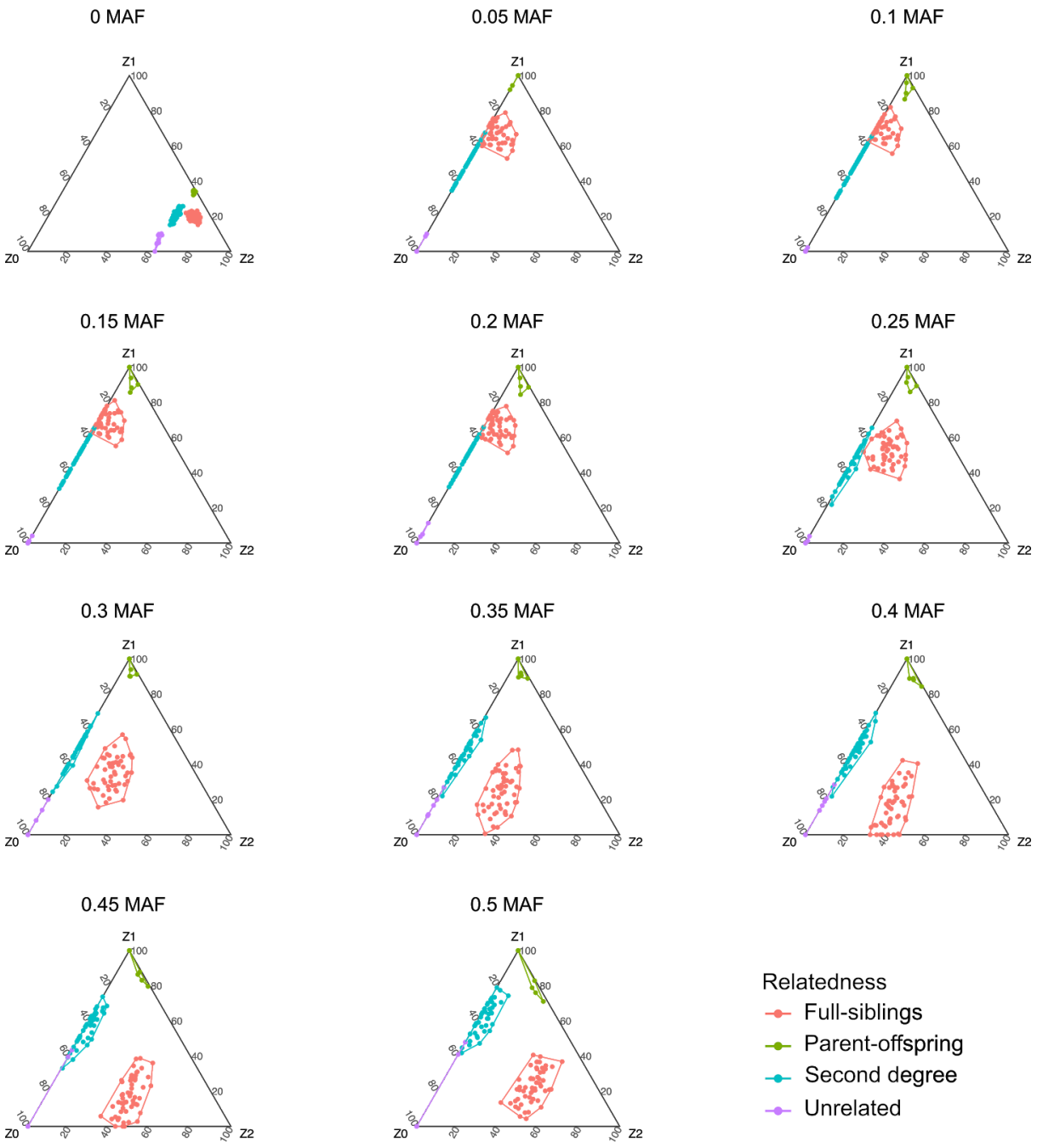
Supplementary Table 2. The effect of SNP filtering on the range of number of SNPs used for IBD estimation, the true positive rate of genotype calls for RNA-seq samples from individuals NA12877 and NA12878 from the CEPH/UTAH family 1463, the clustering of samples in ternary diagrams and pedigree reconstruction of family 1463 using PRIMUS.

Filter	*minimum number of SNPs used for IBD estimation*	*maximum number of SNPs used for IBD estimation*	*% correctly predicted genotypes (NA12878)*	*% correctly predicted genotypes (NA12877)*	*Incorrect clustering of samples in ternary diagram*	*Pedigree predicted correctly*

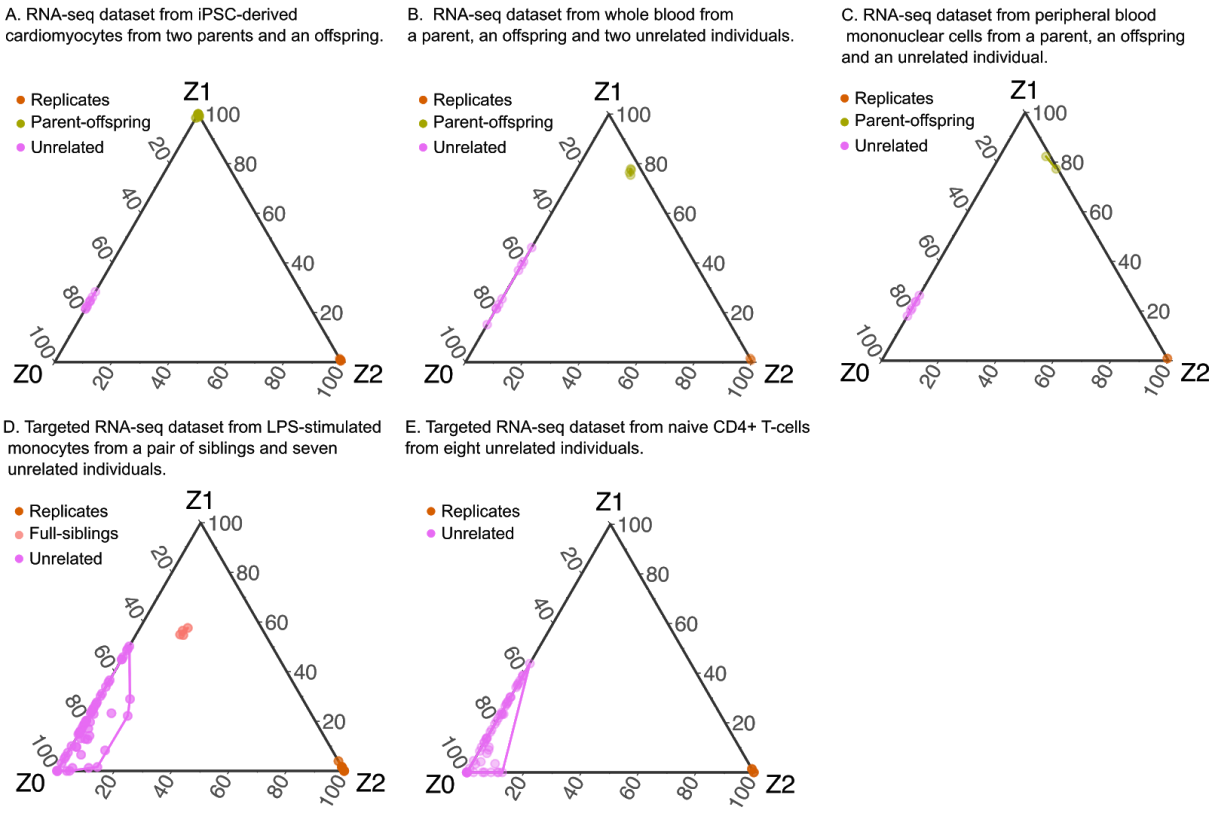
None	6248	9901	99.21	98.66	no	yes
Imprinted genes	6227	9877	99.23	98.74	no	yes
Repeats	5680	8621	99.30	98.72	no	yes
Segmental duplications	5638	8929	99.45	98.95	no	yes

Supplementary Table 3. Expected IBD probabilities of different relationships and degrees of relatedness. Z0, Z1 and Z2 represent the probabilities that a pair of individuals share 0, 1 or 2 identical by descent alleles respectively.

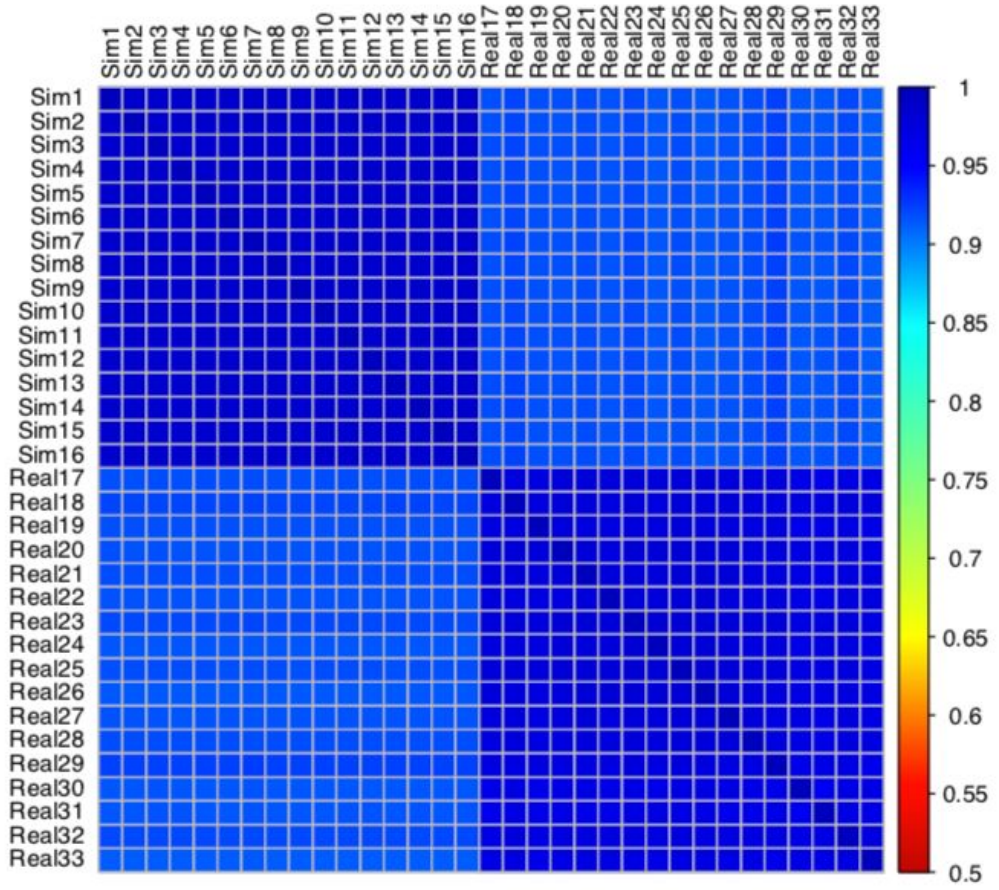
Degree	*Relationship*	*Z0*	*Z1*	*Z2*
1	parent-offspring	0	1	0
	full-siblings	0.25	0.5	0.25
2	grandparental, avuncular, half-siblings	0.5	0.5	0
3	First-cousins, great-grandparental, great-avuncular, half-avuncular	0.75	0.25	0
∞	unrelated	1	0	0



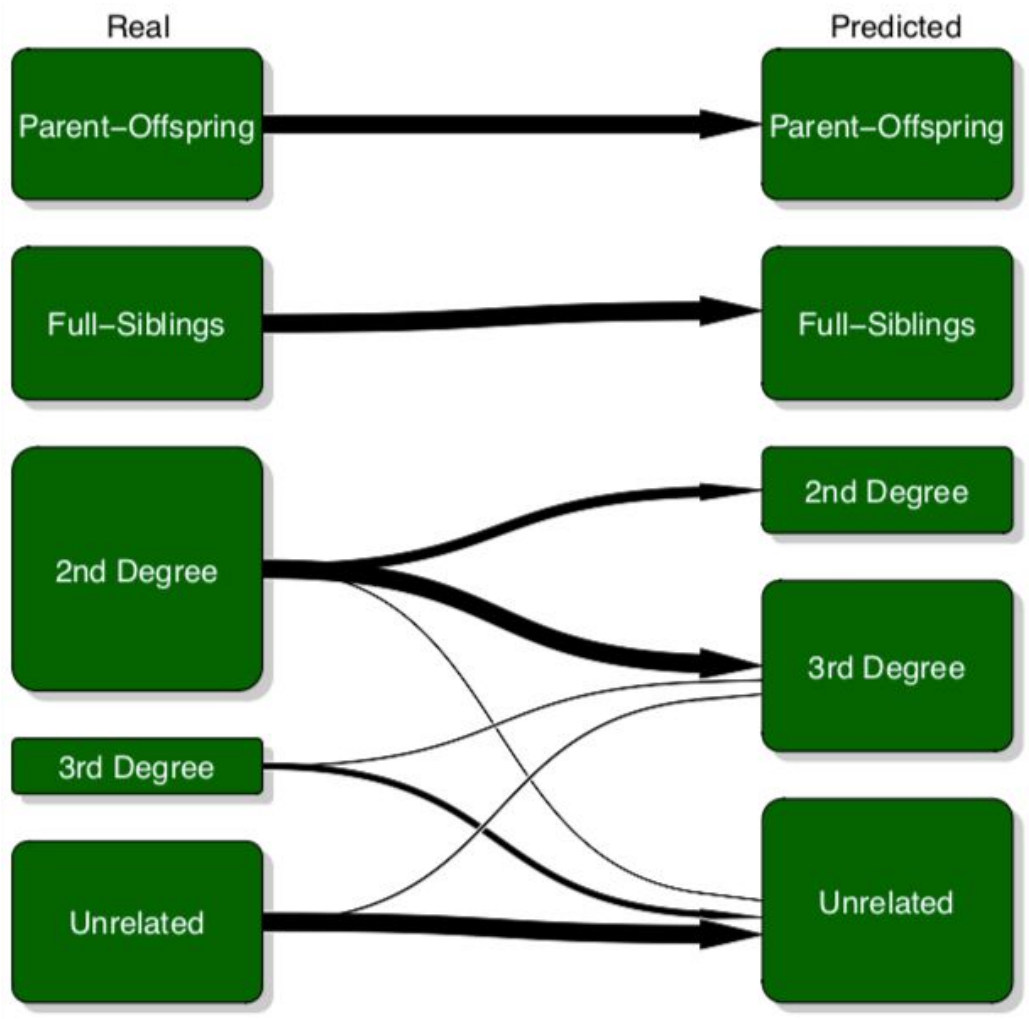
Supplementary Figure 1: Ternary diagrams of IBD estimates for CEPH/UTAH family 1463 using a minor allele frequency threshold ranging from 0 to 0.5.



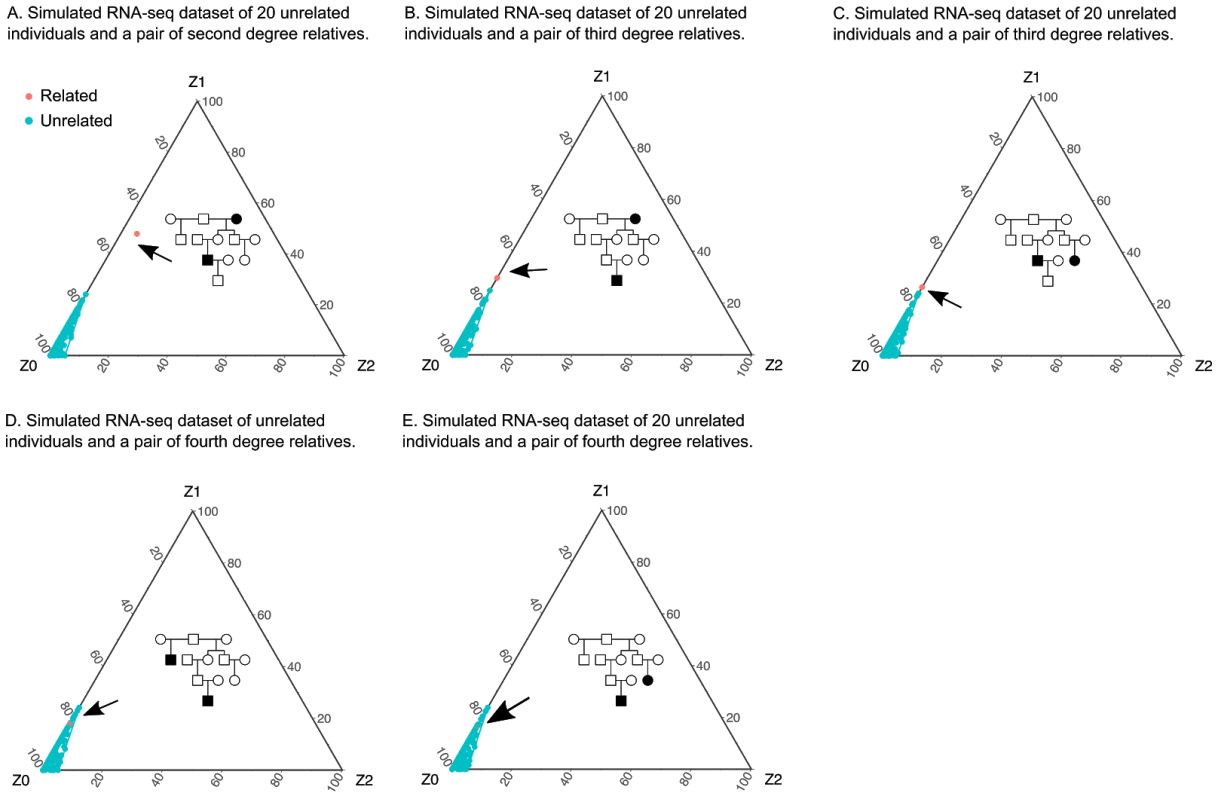
Supplementary Figure 2: Ternary diagrams of IBD estimates from empirical RNA-seq datasets from related and unrelated individuals (A-E). Data from Gifford *et al* Science 2019 (A), Oda *et al* Front Immunol 2019 (B,C) and Boutboul *et al* J Clin Invest 2018 (D,E).



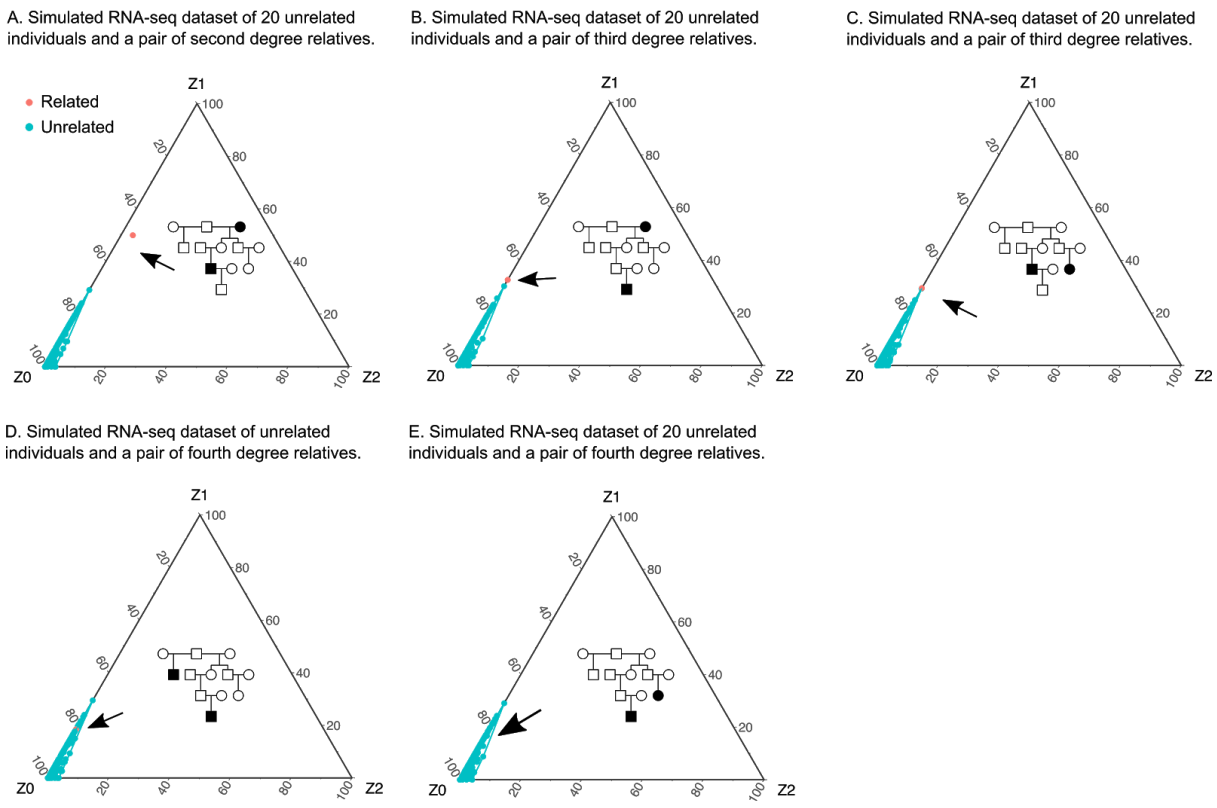
Supplementary Figure 3. Correlation matrix of the number of reads per gene in logarithmic scale between Simulated data (Sim1 - Sim16: individuals 1 - 16 of a type 1 simulated family) and Real data (Real1 - Real17: SRR1258217 - SRR1258233). Mean correlation between Real data: 0.975, between Simulated data: 0.988, between Real data and Simulated data: 0.918.



Supplementary Figure 4: Kinship classification by PRIMUS for simulated families (pairs with higher than third degree relationship are represented here as unrelated).



Supplementary Figure 5: Use of kinship detection for identification of RNA-seq samples from related individuals within a larger datasets of samples from unrelated individuals. Ternary diagrams of IBD estimates for simulated RNA-seq samples from twenty unrelated individuals from the IBS population and a pair of related individuals from a type 4 simulated pedigree (A-E). The pair of related individuals from the pedigree used to spike the dataset of 20 unrelated individuals is shown in black in the pedigree inside each ternary diagram.



Supplementary Figure 6: Ternary diagrams of IBD estimates from simulated RNA-seq datasets consisting of samples from twenty unrelated and a pair of related individuals from the IBS population (genotypes retrieved from the 1000 Genome Project), after removing variants in imprinted genes, repeats and also human segmental duplications from She et al, Nature 2014. The pedigree and five ternary diagrams (A-E) correspond to the same simulated datasets shown in Supplementary Figure 3.