Donnelly (1983) and the limits of genetic genealogy

Michael D. Edge & Graham Coop

Center for Population Biology & Department of Evolution and Ecology, University of California, Davis

What do we inherit from our ancestors, and what do we share with our living kin?

There are many ways to answer this question, but with the advent of genetics, biologists realized that genealogical relationships would result in the sharing of genetically identical alleles between pairs of close relatives. Cotterman (1940) formalized the concept of genetic sharing due to a recent common ancestor, which would be advanced by Malécot (1948), and which we now call identity by descent (often abbreviated as IBD, Browning & Browning, 2012; Thompson, 2013).

In the 1970s, Elizabeth Thompson (e.g. Thompson, 1975) applied these ideas to the possibility of inferring genealogical relationships between people using genotypes from several loci. (For recent advances in genealogical inference, see the *TPB* special issue on relatedness estimation, Cussens & Sheehan, 2016.) Because every generation separating a pair of relatives halves the probability of sharing an allele identically by descent at a locus, such methods were limited to identifying close relatives. Still, as the number of markers available increased, the precision of genealogical inferences would increase, eventually allowing them to be applied in many settings, including in conservation biology (Jones & Wang, 2010), quantitative genetics (Pemberton, 2008), and forensics (Bieber et al., 2006; Rohlfs et al., 2012). However, the fundamental limit of genetics to resolve genealogical relationships among individuals was unclear.

Kevin Donnelly, working as a PhD student under Elizabeth Thompson (Cambridge, 1977-1981), studied the sharing of genomic segments identical by descent between related individuals, rather than the sharing of genotypes at specific loci. Donnelly's work was in part inspired by ideas discussed with one of his fellow PhD students—Andrew J.H. Smith, who was working on DNA sequencing with Fred Sanger and is now at the University of Edinburgh. Smith told Donnelly that such sequencing would one day be "commonplace and very cheap" (Supplementary Information). Further inspiration came from Thompson's 1978 sabbatical in Utah with Mark Skolnick, where she talked with David Botstein and Ray White about their ideas for building a linkage map using restriction fragment length polymorphisms (Botstein,White, Skolnick, Davis, 1980).

Donnelly's work inherits from these exchanges of ideas a strikingly modern view of the genome as a continuum, any segment of which might be established to be identically shared between a pair of relatives. Donnelly (1983) noted as motivation that, "The map of the human genome is being filled in increasingly rapidly...and there is the prospect of DNA sequencing becoming commonplace. It may therefore be timely to look tentatively toward the day when measurable informative loci are located densely throughout the genome, so that chromosomes are better represented by line segments, which are broken and respliced by crossovers, than as finite collections of loci." This theoretical choice prefigures the current state of genome-wide inference in genetic genealogy, one that would not obtain for another twenty to thirty years after Donnelly wrote (e.g. Browning & Browning, 2011; Huff et al., 2011).

To analyze shared segments along the linear genome, Donnelly (1983) represented the ancestry along a chromosome as a random walk along the vertices of a hypercube. The vertices of this hypercube encode sets of ancestors from which material at the current genomic location might be inherited, and the transitions between vertices correspond to crossover events that occur as a Poisson process along the chromosome. Donnelly (1983) provides an example of a pair of half-siblings who share a father. If we label the shared father's maternal and paternal chromosomes as 0 and 1, respectively, then we can label the possible states as the vertices of a square. Either both half-siblings inherit the father's maternal chromosome (state 00), they both inherit the father's paternal chromosomes (state 11), or one inherits the father's maternal chromosome and the other inherits the father's paternal chromosome (states 01 and 10). Crossing-over events correspond to changes of a single coordinate on a two-dimensional random walk, and the two half-siblings will have an identical-by-descent segment whenever the walk hits the states 00 or 11. Relationships involving more focal individuals can be represented with higher dimensions. For example, a third half-sibling could be included by adding an additional dimension, and we could consider states in which all three half-siblings are identical by descent (000 and 111). More distant relationships can also be represented by higher-dimensional hypercubes—for example, the process for a pair of half-cousins could be represented by a four-dimensional hypercube, where vertex 0000 might indicate that both half-cousins inherit the maternal copy of the shared grandparent's chromosome at a particular point in the genome. Donnelly's formalism is sufficiently general to allow a variety of questions to be posed about a large range of possible relationships. He also introduced an approximation to the probability that a pair of genealogical relatives share no genetic material, using the idea that the genome is broken into a Poisson number of blocks and that each of these blocks has an independent probability of being shared (an approximation still in use today, e.g. Huff et al., 2011).

Donnelly's computations highlighted an important distinction in genetic genealogy between pairs of genealogical relatives who share vs. do not share any genetic segments (see also Baird, Barton, & Etheridge, 2003; Matsen & Evans, 2008; Gravel & Steel, 2015). Close relatives are virtually certain to share blocks of the genome identical by descent, and thus to be genetically detectable as relatives. But as relationships grow more distant, the probability of genetic sharing decreases rapidly, and a substantial fraction of genealogical relatives will not be "genetic" relatives. Donnelly—who was raised in Ayr, Scotland, childhood home of Robert Burns—gave an example, "This means that someone descended from the Scottish poet Robert Burns (born 1759) [whom Donnelly's assumptions placed 8 generations before the present] probably carries some of his genes, but that someone unilineally descended from the English playwright William Shakespeare (born 1564) is unlikely to have any genes in common with him." Relatively few of one's many ancestors from more than ten generations in the past will have contributed to one's genome.

The distinction between genealogical and genetic relatives emphasized by Donnelly has never been more important. Direct-to-consumer genetic testing is now a large industry, with over 25 million customers (Regalado, 2019), and consumers' eagerness to identify relatives using genetic information is a major driver of demand. As personal genomics databases have grown, many consumers have learned the identities of previously unknown relatives, out to third, fourth, and fifth cousins. These same customers likely have vast numbers of more distant cousins---eighth, ninth, and tenth cousins, say---also in the database, but Donnelly's results imply that the great majority of these genealogical connections have left no genetic trace. The most recent practical application of ideas descended from Donnelly's is long-range forensic searching, in which distant relatives of a person of interest are identified genetically (Erlich et al., 2018; Edge & Coop, 2019). Since 2018, long-range forensic searching has reopened long-cold criminal cases, for example identifying Joseph DeAngelo as the lead suspect in the Golden State Killer case using genetic connections to second, third, and fourth cousins (Jouvenal, 2018). Analysis of the genealogical extent to which genetic relatives can be identified has also been important in studying the privacy concerns raised by long-range searching, as personal decisions about genetic data may expose to one's distant relatives to surveillance by law enforcement. Long-range forensic search is a direct application of the genetic scenario Donnelly (1983) envisioned, in which segments of genomic identity can be readily detected and used to search for genealogical relationships.

Donnelly (Supplementary Information) recently recounted to us his early interest in his own genealogy; as a teenager he sketched a family tree of his many cousins, filling it in by talking with older relatives. As of April 2019, he has just received a personal genomics kit and is "looking forward to making contact with more third and fourth cousins." Donnelly's 1983 paper played a key role in making modern genetic genealogy possible by clarifying the ways in which genetic relationships propagate along our immense family tree, and in which our connections to each other are recorded in our cells. Donnelly's results remind us that genetic connections differ from genealogical connections, a fact that will have growing societal relevance during the coming years.

Acknowledgements

We thank Kevin Donnelly (also known as Caoimhín Ó Donnaíle) and Elizabeth Thompson for sharing their recollections of the genesis of Donnelly ’83 and for commenting on this manuscript. Their comments are available in a supplement to this article. Thanks to Noah Rosenberg for further comments and discussion. We also acknowledge support from NIH (R01-GM108779 and F32-GM130050) and NSF (1262327 and 1353380).

References

Baird, S. J. E., Barton, N. H., & Etheridge, A. M. (2003). The distribution of surviving blocks of an ancestral genome. *Theoretical Population Biology*, *64*:451-471.

Bieber, F. R., Brenner, C. H., & Lazer, D. (2006). Finding criminals through DNA of their relatives. *Science*, *312*:1315-1316.

Browning, B. L., & Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *American Journal of Human Genetics*, *88*: 173-182.

Browning, S. R., & Browning, B. L. (2012). Identity by descent between distant relatives: detection and applications. *Annual Review of Genetics*, *46*: 617-633.

Botstein, D., White, R., Skolnick, M., Davis, R. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*. *32*: 314–331.

Cotterman, C. W. (1940). A calculus for statistico-genetics. PhD Thesis, The Ohio State University, Columbus, Ohio.

Cussens, J., & Sheehan, N. A. (2016). Special issue on New Developments in Relatedness and Relationship Estimation. *Theoretical Population Biology*, *107*: 1-3.

Donnelly, K. P. (1983). The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology*, *23*:34-63.

Edge, M. D. & Coop, G. (2019). How lucky was the genetic investigation in the Golden State Killer case?. *bioRxiv*, 531384.

Erlich, Y., Shor, T., Pe'er, I., & Carmi, S. (2018). Identity inference of genomic data using long-range familial searches. *Science*, *362*:690-694.

Gravel, S. & Steel, M. (2015). The existence and abundance of ghost ancestors in biparental populations. *Theoretical Population Biology*, *101*:47-53.

Huff, C. D., Witherspoon, D. J., Simonson, T. S., Xing, J., Watkins, W. S., Zhang, Y., et al. (2011). Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Research*, *21*:768-774.

Jones, O. R., & Wang, J. (2010). Molecular marker-based pedigrees for animal conservation biologists. *Animal Conservation*, *13*: 26-34.

Jouvenal, J. (2018). To find alleged Golden State Killer, investigators first found his great-great-great-grandparents. *Washington Post*, 30 April 2018.

Malécot, G. (1948). *Mathématiques de l'hérédité*. Paris: Masson et Cie.

Matsen, F. A. & Evans, S. N. (2008). To what extent does genealogical ancestry imply genetic ancestry? *Theoretical Population Biology*, *74*:182-190.

Pemberton, J. M. (2008). Wild pedigrees: the way forward. *Proceedings of the Royal Society B: Biological Sciences*, *275*: 613-621.

Rohlfs, R. V., Fullerton, S. M., and Weir, B. S. (2012) Familial identification: population structure and relationship distinguishability. *PLoS Genetics, 8*:e1002469.

Regalado, A. (2019) More than 26 million people have taken an at-home ancestry test. *MIT Technology Review*, 11 February 2019.

Thompson, E. A. (1975). The estimation of pairwise relationships. *Annals of Human Genetics*, *39*:173-188.

Thompson, E. A. (2013). Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics*, *194*:301-326.