Supplement to “Donnelly (1983) and the limits of genetic genealogy” by M.D. Edge and G. Coop

As we were writing our commentary on Donnelly (1983) for the special anniversary edition of *Theoretical Population Biology*, we wrote to Kevin Donnelly, as well as to Elizabeth Thompson, who was his PhD advisor at the time he did the research that resulted in the publication. Both of them generously shared some memories related to the research, and their comments are appended here. -M.D. Edge & G. Coop.

Recollections from Kevin Donnelly on the genesis of Donnelly (1983)

I did a degree in maths and physics at Glasgow University, and followed by an MSc in statistics (trying to become more practical!). I was interested enough in statistics to become interested in doing a PhD in the subject. The reason I was drawn to doing a PhD in Cambridge with Elizabeth Thompson as supervisor was that she had been doing some interesting research on the history of the Icelandic population by studying the statistics of blood groups - genetic loci. As more and more blood groups were becoming available for study, this kind of research seemed to have a great future.

There was a complication, of course. As more and more polymorphic loci became available, more and more of them would be on the same chromosome, and even close together on the same chromosome, and so they would not be inherited independently. As chance would have it, another PhD student at the same college, St John’s in Cambridge, gave me an idea of how to proceed. St John’s had a policy of offering its first year research students accommodation in Merton Hall, a 500 year old house in the college grounds - perhaps to encourage cross-fertilization between subjects. One of the other PhD students---Andrew J.H. Smith, who is now Group Leader, Principal Research Investigator, and Senior Lecturer at the MRC Centre for Regenerative Medicine, University of Edinburgh---was working with Fred Sanger's team at the MRC Laboratory of Molecular Biology on DNA sequencing, which was a new hot topic at the time.Only one or two labs in the world were doing it, and everything was moving so fast that a paper had to go from submission to publication within about two months or else it would be out of date. He told me that complete gene sequencing was eventually going to become commonplace and very cheap.

So that made me look at the whole genome rather than just individual loci. Being a mathematician of sorts, I abstracted things to look at the “chromosome pedigree”, each chromosome being derived from two parent chromosomes (rather than bothering with the pedigree of human individuals). Each chromosome was copied alternately from one or other of its parent chromosomes, with crossovers occurring, rather a big assumption I made, as a Poisson process. Looking simultaneously at all the crossover processes in the pedigree was equivalent to looking at a random walk on the vertices of a hypercube - a continuous-time stochastic process. The “Statistical Laboratory” at Cambridge University was in those days more of a department of

theoretical probability, with a particular interest in continuous-time stochastic processes, so it was natural for me to look at things in that way.

Various interesting genetic events corresponded to the random walk hitting or failing to hit a particular "hitting set" of vertices - events such as an individual sharing some genome with a great×n-grandparent, or sharing some genome with an nth-cousin. I was very interested personally in that kind of thing - detectable relationships - because I had been interested from my teens in my family tree. I had scores and scores of cousins, and had started from about the age of 13 in making a family tree on a huge sheet of paper, trying to sort them all out by talking to elderly uncles and aunts. All my family, all four grandparents, were from Ireland, though, and genealogy is really difficult in Ireland because most of the records, which were never great to begin with because of complications like the Famine, were destroyed in a fire in the troubles of 1922. So anything which DNA could do to help would be wonderful.

Doing calculations involving the 2^n vertices of a hypercube would quickly become impossible, but being a mathematician of sorts, I realised that making use of symmetries could make the problem manageable. The thing to do was to classify the vertices into the orbits under the group of symmetries which left the hitting set invariant, and to just look at the transition probabilities between these orbits. That reduced the computation to order n instead of 2^n, and since Cambridge University had a particularly good computing service, way ahead of its time, that was quite easy.

I remember that one of my conclusions was that proving your descent from William Shakespeare (1564-1616) meant almost nothing as regards the likelihood of you sharing some of his genome, whereas if you could prove your descent from the Scottish poet Robert Burns (1759-1796), you were very likely indeed to be carrying some of his DNA. I was shocked when Elizabeth told me that I should add the words "the Scottish poet" to the paper because people would not know who Robert Burns was. Having been brought up in Ayr in Scotland, I had assumed that Robert Burns was as well known worldwide as William Shakespeare. As far as I know, I am not a descendant of his, but who knows because he had a very large number of illegitimate children. I am very grateful to Elizabeth for encouraging me to submit that TPB paper, as I would not have done it without her encouragement.

Fate has since taken me away from mathematical genetics research. I was keen to return to Scotland and happened to find a job at Glasgow University studying the statistics of solar energy storage, followed by a computing job in Edinburgh with the Forestry Commission research station, and then a job at a tiny college on the island of Skye teaching computing through the medium of Gaelic (so I have since been publishing under the Gaelic version of my name, Caoimhín P. Ó Donnaíle rather than Kevin P. Donnelly). I had become very interested in what computing facilities such as spell checkers could do to assist minority languages. Fate has taken me back a bit to genetics, though. In 2009, I was persuaded by my brother and second cousins to attend a Donnelly family reunion in Carlingford, descendants of our great-

grandparents, John Donnelly and Mary McCann from Glenmore. The reunion was a great success with 120 attending, including 20 from the US. I prepared for it by reviving the family tree which I had started in my teens and putting it into a genealogy program, Webtrees, on the webserver which I ran. That got such interest from students at the college that I have been teaching a genealogy course module through the medium of Gaelic ever since, with 67 students to date. I have been keeping up the research on my own family tree, and have just received delivery of a DNA tester kit from Ancestry - the time has finally come! - and I am looking forward to making contact with more third and fourth cousins.

Other than that, I have not been much involved lately with human genetics, but I have become very interested in family trees of words, the etymology of Gaelic and English and words in every language going back to Proto-Indo-European. Using my computing skills I have developing a project with I have called Bunadas, a network database of cognate words, and have been finding it very exciting, with great potential: http://www.smo.uhi.ac.uk/gaidhlig/faclair/bunadas/

While I was in Cambridge, I only looked at "detectable relationships", whether you could detect at all that two individuals were related. What I had been hoping to go on do, was to look at was whether you could distinguish between two different relationships. You might hope to distinguish fairly well between relationships with different coefficients of kinship, second-cousins and third-cousins, for example, by looking at the proportion of their genome which they share, and that is the kind of thing which Ancestry and 23andMe are now doing. But it might be of interest to look at whether it is possible to distinguish between relationships which have the same coefficient of kinship, to distinguish between double-second-cousins and first-cousins-once-removed, for example, by looking at the distributional pattern of the DNA which they share. I started looking at that to some extent in Glasgow, but I had to leave it to get on with the project I was being paid for!

Recollections from Elizabeth Thompson on the genesis of Donnelly (1983)

I became a graduate student in 1970, and the Cambridge University Statistical Laboratory subscribed to the new TPB journal. It was the journal I followed most closely, awaiting each issue: I recall reading the paper of Sved [TPB, 1971). My advisor, Anthony Edwards, was always a source of knowledge of the literature. Among much else, he loaned me a copy of Fisher's "Theory of Inbreeding" (1949), and I worked carefully through the chapter on the Theory of Junctions. A bit later, other papers such as Franklin [TPB, 1977] also addressed continuous genome models. I had become interested in ideas of identity by descent, and the inference of identity by descent (IBD) from genetic data, the data then being unlinked blood type and enzyme loci. Although intriguing, the relationship of continuous genome models to real data and inference seemed slight.

For me, that changed in 1978, when I visited Mark Skolnick at the University of Utah for 5 months on sabbatical, continuing research connections I had earlier made while

a postdoc at Stanford 1974-5. David Botstein and Ray White were there, and with Mark Skolnick discussed with excitement the prospect that the new RFLP technology would bring genome-wide human genetic linkage maps [Botstein et al. AJHG, 1980]. The relationship between data, inference, and an underlying continuous genome was no longer abstract. In Utah also, we worked on the extended pedigrees extracted from the genealogical data base of the Mormon Church, available to Mark Skolnick and his colleagues through collaborations he had initiated. Would the remote cousins in these pedigrees share any genome IBD?

Kevin Donnelly had become my PhD student in 1977. He was my first PhD student, and I was an inexperienced advisor. However, I discussed ongoing ideas with him, as my advisor had with me, and he became intrigued by the issue of whether, with complete genome information, what would be the limits of finding segments of IBD in more remote relatives. While we discussed these issues, the framework, methodology, and analysis that became the paper of Donnelly [1983, TPB] were Kevin's alone. It was Kevin who translated the IBD question into a random walk on a hypercube. He discussed ideas with me, and also with others working in Applied Probability in Cambridge (for example David Aldous, whose personal communication is cited in the paper). I encouraged the work, which forms the first part of Kevin's 1981 PhD thesis, but in no way can claim any contribution. I encouraged the submission to TPB in February 1981, but again his sole authorship on that paper reflects the realty that was truly his original work in concept and execution.

Kevin returned to Scotland, first to the Forestry Commission in Roslin, and soon after to the Western Isles, where he has been for many years teaching Computer Science and Gaelic at the Gaelic College (Sabal Mor Ostaig) on Skye. It is only with modern-day genomic data and population-based samples of remote relatives that the true relevance and importance of his one statistical genetics publication has become recognized.