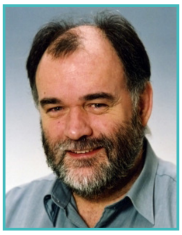
Next generation sequencing and phylogenetic networks



David A. Morrison

Swedish University of Agricultural Sciences, Uppsala, Sweden

Received 20 March 2014; **Published** 26 March 2014

Morrison DA (2014) *EMBnet.journal* **20**, e760. http://dx.doi.org/10.14806/ej.20.0.760

Introduction

Next Generation Sequencing (NGS), or massively parallel sequencing, can potentially provide a fast and cost-effective means of generating multi-locus sequence data for phylogenetics, which is the field that tries to reconstruct the genealogical history of evolutionary change. Unfortunately, the cost for the number of samples typically employed in phylogenetics is currently still beyond the reach of most researchers. This will soon change, and phylogenetics will become phylogenomics.

Phylogeneticists therefore now need to think about the relationship between NGS and their current paradigms, in terms of both data analysis and interpretation. In particular, there has been recent interest among phylogeneticists in using phylogenetic networks rather than phylogenetic trees as the main paradigm for interpretation (Morrison, 2011; Bapteste *et al.*, 2013). Trees are intended only for the study of vertical evolutionary processes, directly from parent to offspring; but networks can accommodate horizontal processes as well, such as recombination, hybridisation, introgression and horizontal gene transfer, all of which are common in one taxonomic group or another. These horizontal processes are represented by reticulations in the network, which do not appear in a tree.

Most of the published discussions about NGS in relation to phylogenetics have focused on trees, rather than networks (Rannala and Yang, 2008; Whelan, 2011; McCormack *et al.*, 2013; Lemmon and Lemmon, 2013). Here, I raise some of the important issues that need to be addressed when using networks.

NGS and phylogenetics

NGS and phylogenetics have so far had only a brief association. McCormack *et al.* (2013) have commented on this:

> *"Despite this obvious potential, NGS has been slow to take root in phylogenetics compared to other fields like metagenomics and disease genetics. We suggest that this lag has been caused by four specific aspects of phylogeographic and phylogenetic research: the predominant focus on non-model organisms, the need for sequencing large numbers of samples per species, the lack of consensus regarding library preparation protocols for particular research questions, and the transitional state of the technology (whole-genome data are still neither cost-effective, nor even desirable for phylogenetics, but are paradoxically easier to collect).*
>
> *Another issue is the historical importance of utilizing gene trees in phylogenetics. Gene trees are most robustly inferred from loci with high information content, for example, a non-recombining locus containing a series of linked SNPs. Individual SNPs, on the other hand, have low information content on a per-locus basis and have been used predominately with classification methods such as Structure and Principal components analysis ... While distance-based genealogies and phylogenies can be built from unlinked SNPs, this ignores models of molecular substitution and probabilistic tree-searching algorithms that have led to more robust phylogenetic inference in the last several decades."*

Furthermore, no-one has yet shown that many of the questions currently being asked by phylogeneticists will actually benefit from genomic data. We may well be able to answer some new questions, but that is quite a different thing from NGS initiating a revolution, as it has done in other fields of biology. The essence here is that, in science, the questions must come first — collecting data for the sake of it is usually unproductive. So, we need a clear demonstration that genomics is actually needed in phylogenetics (as opposed to other disciplines, where it may indeed be very useful). If an increased volume of data will solve a phylogenetic problem, then that is good, but there is no necessary reason to expect that it will happen. Statistically, the extra data can lead to improved precision, but not necessarily improved accuracy. In science, targeted data collection has always been the most productive approach to any clearly stated experimental question.

For example, the estimated relationships among humans, chimpanzees, and gorillas did not change as a result of genome sampling rather than gene sampling (Galtier and Daubin, 2008), nor did those of malaria species (Kuo *et al.*, 2008), nor those of mammal super-orders (Hallström and Janke, 2010) or even the orders of wingless insects (Dell'Ampio *et al.*, 2014). In all four cases, the inferred relationships were just as complex after the genome sequencing as before — the resolution of controversial branches in the phylogenetic trees did not occur as a result of increased access to character data.

In this sense, a small sample of representative gene sequences should reveal just as much of the genealogical truth as will a genome-wide sample. A recent empirical example is presented by O'Neill *et al.* (2013), who found that including less informative loci added so much noise to the phylogenetic signal that the analysis eventually broke down. The issue here is that, as data volume increases, so does the potential occurrence of systematic bias owing to model misspecification.

This sort of problem can easily be visualised using phylogenetic networks. Here, genome-scale data frequently produce unresolved bushes rather than tree-like phylogenies, as shown by Beiko (2011), whose analysis involved 298 completely sequenced bacterial genomes, or Decker *et al.* (2009), who analysed 372 individuals belonging to 48 breeds of cattle. Bush-like phylogenies may represent complex evolutionary histories, but they may also represent a failure of phylogenetic analysis; and it is important to be able to distinguish between these two possibilities.

This all suggests that we will need to think carefully about how to apply phylogenetic networks to genome-scale data. Much of the lack of resolution may very well come from the nature of NGS, rather than from the actual evolutionary history.

NGS and networks

There are a number of potential problems with NGS. These may not matter so much for tree-building algorithms, but it is a different matter for networks. They each need to be thought about to assess whether they are serious problems or only of minor concern.

Increased homoplasy owing to sequencing errors

An error rate of even 0.01% is considered good in NGS (*e.g.*, Roche 454: 1%; Illumina HiSeq: 0.1%; Life SOLiD: 0.01%), but when this is extrapolated to the genome scale, it results in thousands of errors. Networks are sensitive to this magnitude of stochastic error. Indeed, one of the valuable uses of phylogenetic networks is specifically to identify data errors. For example, they have been used for detecting chimeric sequences resulting from laboratory-induced errors (Kong *et al.*, 2008), or detecting possible errors in mitochondrial DNA (miDNA) genomes sequenced to find mutations associated with particular diseases (Bandelt *et al.*, 2009).

Increased homoplasy owing to intra-gene processes

These include substitutions, deletions, duplications (especially tandem repeats), inversions and translocations. These processes can potentially reveal evolutionary history, but we have little idea about how best to process the data in a way that will reveal that history. Currently, we deal with this by lumping most of the processes together in the analysis model as 'indels'. This approach is likely to be inadequate for networks, because these very processes may be involved in horizontal evolution.

Increased homoplasy owing to inter-gene processes

The main processes known to confound attempts to identify reticulate evolution are incomplete lineage sorting and gene duplication–loss. The more genes that are sampled, then the greater will be the effect of these confounding processes. There are several methods available for addressing them in the context of estimating phylogenetic trees (*e.g.*, Knowles and Kubatko, 2010; Blair and Murphy, 2011; Bansal *et al.* 2012), but the applicability of these methods to networks is still being assessed (Kubatko, 2009).

Increased homoplasy in non-coding regions

Sanger sequencing in phylogenetics is usually targeted towards gene-coding regions or their introns, but genome-scale data can include what is currently called 'junk DNA'. The evolution-

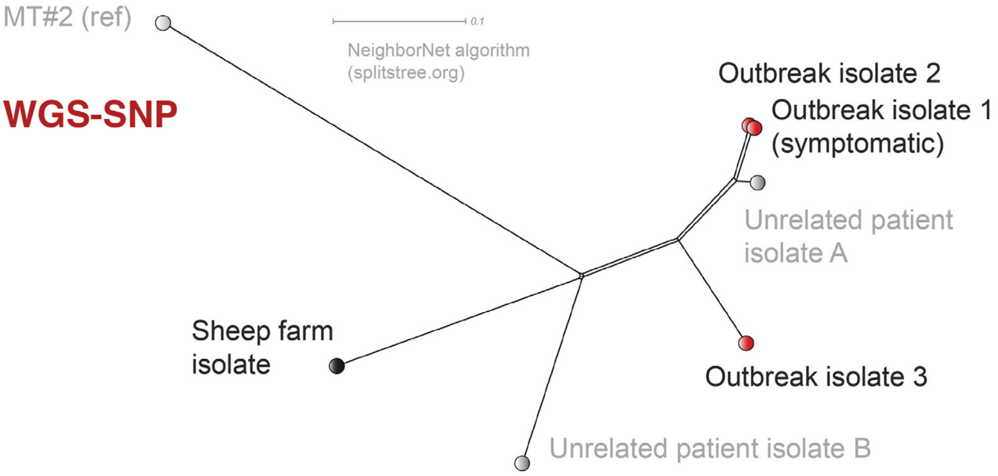


Figure 1. NeighborNet analysis of SNP data from whole-genome shotgun sequencing of seven *E. coli* strains. This network is very tree-like, so that the reticulations are unlikely to represent biologically important processes. The phylogenetic interpretation of the network is thus very straightforward.

ary processes in these regions are currently unknown, so they are difficult to model; and their applicability to phylogenetic analysis has not yet been assessed.

Inadequacies owing to data-processing methods

The analysis of NGS data is often a black art — each published paper seems to provide its own way of processing the data. This has been a cause of concern expressed in the literature (*e.g.*, Check Hayden, 2012; Editorial, 2012a, 2012b; MacArthur, 2012), especially in light of the currently poor documentation and archiving of bioinformatics programs (Cuticchia and Silk, 2004). Perhaps the most talked-about problem is ascertainment bias, especially when SNP (Single Nucleotide Polymorphisms) variants are reported only if they do not match a specified reference genotype. Non-reported variants can just as well be sequencing failures, or coverage gaps, or insufficient evidence for a non-reference variant. Networks generated from such data are likely to consist largely of artefacts.

Network analysis of NGS data

All of this might make the application of networks to phylogenomics problematic in many cases, because we already have enough challenges dealing with the data from Sanger-style sequencing, without having them be orders of magnitude worse. It will therefore be very interesting to see what emerges from the current attempts to apply phylogenetic networks to NGS data. To date, most of the analyses have been ad hoc in nature (Dagan, 2011).

There have been a few applications of EDA (Exploratory Data Analysis) programs, such as SplitsTree[1], mostly involving bacteria and viruses (*e.g.*, Beiko, 2011), and often in the context of detecting recombination. Not all of these studies have produced networks that look bushy, as shown by Figure 1, from Söderlund *et al.* (2013).

SplitsTree is mostly limited by the number of samples, not by the number of characters, so that genomic data are not a particular analysis issue for network algorithms such as NeighborNet. However, it might be necessary to calculate the inter-sample distances outside of this program, unless you want the simple p-distance (popular genome-scale distances include F_{st}).

There have also been programs developed for the study of admixture (or introgression) in human genomes, such as TreeMix[2], AdmixTools[3] and MixMapper[4], and these might repay wider exploration. Essentially, they first construct a phylogenetic tree and then add network reticulations based on various criteria. As is usual with this general approach, there is a problem constructing the initial tree in the presence of reticulation processes. Moreover, there seems to be no clear

[1] www.splitstree.org

[2] https://code.google.com/p/treemix/

[3] genetics.med.harvard.edu/reich/Reich_Lab/Software.html

[4] groups.csail.mit.edu/cb/mixmapper

criterion for when to stop adding reticulations — optimisation criteria always increase as reticulations are added, so that increasingly complex networks will always be preferred mathematically.

Conclusion

We need to make sure that we are getting the most out of NGS that we can in phylogenetics, because the times are changing and we need to move with them. However, when moving, the cart should not be leading the horse, and so the phylogenetic horse needs to think carefully about its relationship to the NGS cart. It should be exciting to see the horse and cart working together well, sometime soon.

References

Bandelt H-J, Yao Y-G, Bravi CM, Salas A, Kivisild T (2009) Median network analysis of defectively sequenced entire mitochondrial genomes from early and contemporary disease studies. *Journal of Human Genetics* **54**, 174-181. http://dx.doi.org/10.1038/jhg.2009.9

Bansal MS, Alm EJ, Kellis M (2012) Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics* **28**, i283-i291. http://dx.doi.org/10.1093/bioinformatics/bts225

Bapteste E, van Iersel L, Janke A, Kelchner S, Kelk S, McInerney JO, Morrison DA, Nakhleh L, Steel M, Stougie L, Whitfield J (2013) Networks: expanding evolutionary thinking. *Trends in Genetics* **29**, 439-441. http://dx.doi.org/10.1016/j.tig.2013.05.007

Beiko RG (2011) Telling the whole story in a 10,000-genome world. *Biology Direct* **6**, 34. http://dx.doi.org/10.1186/1745-6150-6-34

Blair C, Murphy RW (2011) Recent trends in molecular phylogenetic analysis: where to next? *Journal of Heredity* **102**, 130-138. http://dx.doi.org/10.1093/jhered/esq092

Check Hayden E (2012) RNA studies under fire. *Nature* **484**, 428. http://dx.doi.org/10.1038/484428a

Cuticchia J, Silk G (2004) Bioinformatics needs a software archive. *Nature* **429**, 241. http://dx.doi.org/10.1038/429241b

Dagan T (2011) Phylogenomic networks. *Trends in Microbiology* **19**, 483-491. http://dx.doi.org/10.1016/j.tim.2011.07.001

Decker JE, Pires JC, Conant GC, McKay SD, Heaton MP, Chen K, Cooper A, Vilkki J, Seabury CM, Caetano AR, Johnson GS, Brenneman RA, Hanotte O, Eggert LS, Wiener P, Kim J-J, Kim KS, Sonstegard TS, Van Tassell CP, Neibergs HL, McEwan JC, Brauning R, Coutinho LL, Babar ME, Wilson GA, McClure MC, Rolf MM, Kim J, Schnabel RD, Taylor JF (2009) Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proceedings of the National Academy of Sciences of the USA* **106**, 18644-18649. http://dx.doi.org/10.1073/pnas.0904691106

Dell'Ampio E, Meusemann K, Szucsich NU, Peters RS, Meyer B, Borner J, Petersen M, Aberer AJ, Stamatakis A, Walzl MG, Minh BQ, von Haeseler A, Ebersberger I, Pass G, Misof B (2014) Decisive data sets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily wingless insects. *Molecular Biology and Evolution* **31**, 239-249. http://dx.doi.org/10.1093/molbev/mst196

Editorial (2012a) Must try harder. *Nature* **483**, 509. http://dx.doi.org/10.1038/483509a

Editorial (2012b) Error prone. *Nature* **487**, 406. http://dx.doi.org/10.1038/487406a

Galtier N, Daubin V (2008) Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* **363**, 4023-4029. http://dx.doi.org/10.1098/rstb.2008.0144

Hallström BM, Janke A (2010) Mammalian evolution may not be strictly bifurcating. *Molecular Biology & Evolution* **27**, 2804-2816. http://dx.doi.org/10.1093/molbev/msq166

Knowles LL, Kubatko LS (eds) (2010) *Estimating Species Trees: Practical and Theoretical Aspects*. Wiley-Blackwell, Hoboken NJ. http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470526858.html

Kong Q-P, Salas A, Sun C, Fuku N, Tanaka M, Zhong L, Wang C-Y, Yao Y-G, Bandelt H- J (2008) Distilling artificial recombinants from large sets of complete mtDNA genomes. *PLOS One* **3**, e3016. http://dx.doi.org/10.1371/journal.pone.0003016

Kubatko LS (2009) Identifying hybridization events in the presence of coalescence via model selection. *Systematic Biology* **58**, 478-488. http://dx.doi.org/10.1093/sysbio/syp055

Kuo C-H, Wares JP, Kissinger JC (2008) The Apicomplexan whole-genome phylogeny: an analysis of incongruence among gene trees. *Molecular Biology & Evolution* **25**, 2689-2698. http://dx.doi.org/10.1093/molbev/msn213

Lemmon EM, Lemmon AR (2013) High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution & Systematics* **44**, 19.1–19.23. http://dx.doi.org/10.1146/annurev-ecolsys-110512-135822

MacArthur D (2012) Face up to false positives. *Nature* **487**, 427-428. http://dx.doi.org/10.1038/487427a

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* **66**, 526-538. http://dx.doi.org/http://dx.doi.org/10.1016/j.ympev.2011.12.007

Morrison DA (2011) *Introduction to Phylogenetic Networks*. RJR Productions, Uppsala. http://www.rjr-productions.org/Networks/

O'Neill EM, Schwartz R, Bullock CT, Williams JS, Shaffer HB, Aguilar-Miguel X, Parra-Olea G, Weisrock DW (2013) Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Molecular Ecology* **22**, 111-129. http://dx.doi.org/10.1111/mec.12049

Rannala B, Yang Z (2008) Phylogenetic inference using whole genomes. *Annual Review of Genomics and Human Genetics* **9**, 217-231. http://dx.doi.org/10.1146/annurev.genom.9.081307.164407

Söderlund R, Jernberg C, Källman C, Hedenström I, Eriksson E, Bongcam-Rudloff E, Aspán A (2013) Rapid whole genome sequencing investigation of a familial outbreak of *E. coli* O121:H19 with a sheep farm as the suspected source. *Bioinformatics in Action* **19** suppl.A, 89-90. http://journal.embnet.org/index.php/embnetjournal/article/view/657

Whelan N (2011) Species tree inference in the age of genomics. *Trends in Evolutionary Biology* **3**, e5. http://dx.doi.org/10.4081/eb.2011.e5