"*Known Knowns, Known Unknowns, & Unknown Unknowns*": Computational Science challenges for analysis of multi-dimensional DNA matrices in Evolutionary & Population Genomics

Steven M. Carr

Genetics, Evolution, and Systematics Laboratory, Department of Biology, and Department of Computer Science, Memorial University of Newfoundland, St John's NL A1B 3X9, Canada

Abstract - *The advent of so-called NextGen DNA sequencing methods has massively increased the rate at which DNA sequence information can be generated, and the volume and complexity of the data matrices that apply to biological questions, including molecular and organismal evolution and population biology. One such approach is the analysis of complete mitochondrial DNA (mtDNA) genomes from multiple species simultaneously, by means of a "sequencing by hybridization" microarray biotechnology, the "ArkChip". I review mitogenomic biology and biotechnology, describe some of the known knowns of bioinformatic information content and its computational challenges, outline new computational strategies for known unknowns of evolutionary trees (phylogeny) and population biology structures in time and space (phylogeography), and speculate on future application of Computational Science to biological unknown unknowns.*

Keywords: Mitogenomics, DNA Microarrays, NextGen Sequencing, Bioinformatics, Evolution, Phylogeography

.

1 Introduction

"*There are known knowns; there are things we know we know. We also know there are known unknowns; ...we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know*" Donald Rumsfeld (2002)

Advances in so-called Next Generation ('NextGen') sequencing methods have created gigantic data sets that test the abilities of computational science both to assemble overlapping primary data as a single robust construct, and then to extract information and detect patterns within that construct, where at least some of the patterns are 'unknown unknowns.'

Where population biologists are interested in multiple individuals per species, a more modest but successful strategy involves the mitochondrial DNA (mtDNA) genome, which has a long history of application in evolutionary and population biology, including resolution of relationships among humans and other Great Apes, and tracing the pre- and post-glacial history of human emergence Out of Africa into Europe, the near and far East, and the Americas.

Figure 1 – Nuclear *versus* vertebrate mitochondrial genomes. The human nuclear genome comprises one set each of chromosomes from the mother and father, for a total of about 3 billion DNA base pairs (bp) encoding just over 20,000 'genes'. In contrast, the human mitochondrial DNA (mtDNA) genome is a small, circular, extra-nuclear molecule inherited solely through the maternal egg cytoplasm. It comprises 38 genes concerned with the cellular 'powerhouse' functions of the mitochondrion [1,2].

2 Mitochondrial Genomics

Unlike genes on separate chromosomes in the nuclear genome that undergo 50% recombination each generation, mtDNA does not undergo genetic recombination, but is passed intact between mother and offspring, and in the next generation passes only through the daughters' cytoplasm, mitochondria in the male sperm making no contribution. This matrilineal inheritance, combined with a higher rate of mutation than typical nuclear genes, makes mtDNA invaluable for tracing patterns of historical migration (vicariance) or descent (evolution) in time and space.

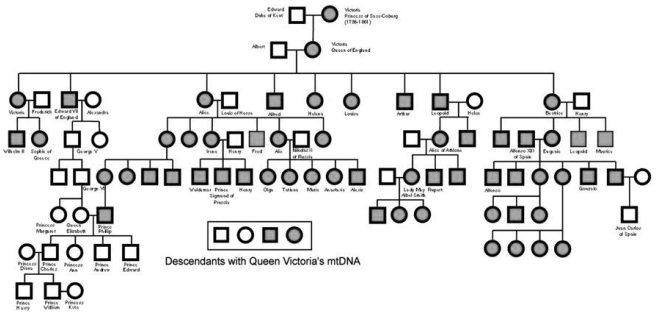

Figure 2 – MtDNA Family Tree of Queen Victoria of England. Victoria is well-known to have carried a nuclear germline mutation for hemophilia, which she passed on as an autosomal recessive allele through her sons and daughters to the royal families of Russia and Spain. She (II-2) is less well-known to have passed her mtDNA genome to all of her children, and via her daughters' daughters' daughters through five generations shown here to her great-great grandson, Prince Phillip (VI-3). Queen Elizabeth II (VI-2) shares her mtDNA with Prince Charles (VII-2), but her grandson and great-grandson Prince William (VIII-2) and Prince George (not shown) have distinct mtDNA genomes inherited from their respective mothers Diana (VII-1) and Kate (VIII-3).

Since the late 1970s, DNA sequence data have been collected by the dideoxy or Sanger method, which involves the use of chemical terminators to produce sets of DNA molecules that differ by plus or minus one base pair, such that the complete sequence is obtained from the nested series. "Pseudo Color"-coding of the terminators and large-scale automation of the separation process culminated in publication of the complete human genome sequence in 2004.

The Sanger method has dominated the field for more than thirty years. Now, "Next Gen" sequencing methods offer increasingly rapid, high-throughput data production that does not rely on linear separation, but rather massively parallel processing of simultaneous reactions. One such method is sequencing by hybridization on a DNA microarray. The method resembles molecular 'velcro', where a known reference sequence is represented on a microarray as a series of short, overlapping oligonucleotide "hooks", and is challenged by an unknown but homologous experimental sequence as a set of "threads". The experimental DNA sticks only to sequence-specific "hooks," which may include single-base variants of the reference sequence. The microarray can return information about widely-separated single nucleotide polymorphisms (SNPs) associated with medical conditions, or where all possible single-base variants are included along with the reference mtDNA sequence, the data are the complete mtDNA sequences of individuals that can differ by from one to hundreds of SNPs [1].

Where a microarray can be designed to accommodate mtDNA reference sequences from several species whose sequences are sufficiently distinct to prevent 'crosstalk', the result is an "ArkChip*"* capable of simultaneous, cost-effective population genomic analysis at the incremental cost of DNA extraction and amplification for each added species [2]. A typical ArkChip experiment generates ca. 1,000,000 features that comprise four A, C, G, and T hybridization signals for the forward and reverse DNA strands of single individuals from each of seven species [4 x 2 x 17,000 x 7] [3]. Projects may include scores or hundreds of individuals per species (Figure 3). *Known knowns* in this process include algorithms that extract individual genome sequences from a 4 x 2 x 17,000 matrix [4]. *Known unknowns* will compare gene patterns along the 17,000 element genome vector within and among species, based on external algorithms applied to exported data [5]. *Unknown unknowns* include creation of algorithms for detection of molecular and evolutionary patterns implicit in fully-annotated higher-order dimensions across genes and species.

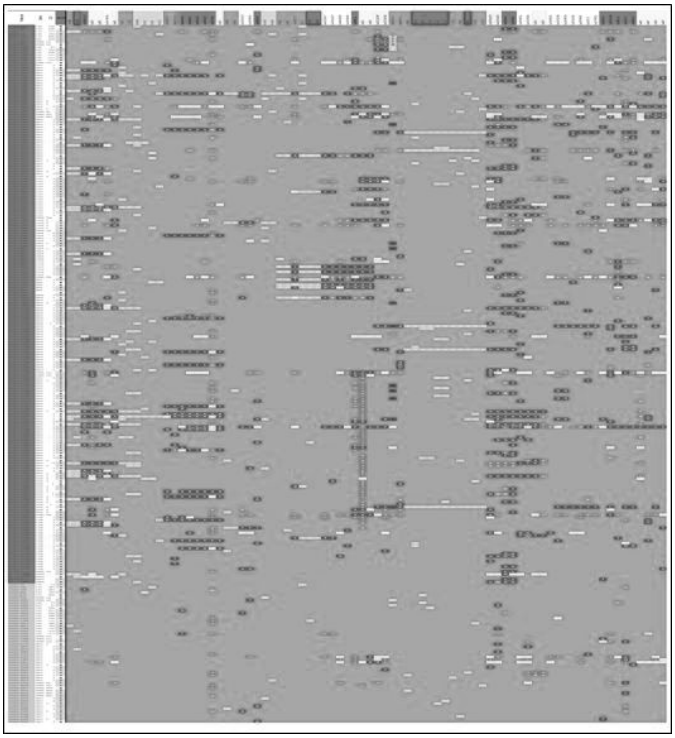

Figure 3 – Schematic of the evolutionary bioinformatic content of the mtDNA genomes from 80 Atlantic Cod (*Gadus morhua*). Each genome comprises 16,553 bp (16.5 Kbp), whose sequence is assembled from a consensus of the forward and reverse DNA strands, so that the complete data set comprises > 1.3 x 10^6 bp (Mbp). Single Nucleotide Polymorphisms (SNPs) have been identified at more the 500 sites. The data have been sorted to highlight more than 200 [dark grey block at left] that are informative as to genetic relationships among fish [2, 3]. Related fish genomes with a common ancestor (clades) have been group by column and are recognizably as bands across columns [4]. Alternative sorting can highlight patterns of molecular evolution by gene or codon position within genes [5]. Color-coding may indicate SNP sites, information content, confidence levels in base calls, patterns of sharing among fish genomes, etc.

Phylogeography is the study of population genetic relationships in space and time. Whereas the field began in the early days of DNA sequencing with short sequences and partially-resolved relationships, the advent of genomic data enables complete resolution of within-species phylogenies and creates new challenges for their interpretation.

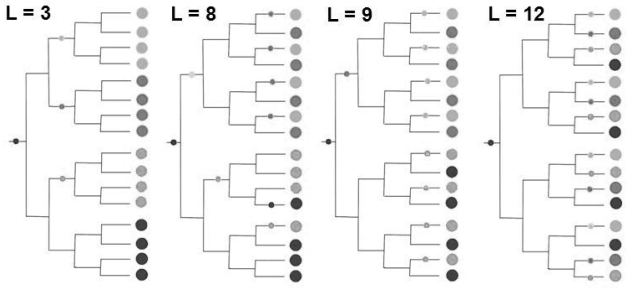

Figure 4 – Thought experiment in phylogeographic genomics: more highly structured family trees are shorter than random trees [after [6]]. Consider 16 individuals found in four distinct breeding locations (four shades of grey), where the darkest shade is considered to be the ancestor of the other three. For an ideal dichotomously-branching phylogeny that shows that individuals in each population are all each other's closest relatives (i.e., none is more closely related to any individual outside the population than to any within) [left], the distribution may be explained by a single vicariance event (historical founding) per descendant population, thus L = 3. Where the phylogeny shows that individuals are uniformly distributed across the tree (i.e., they are no more closely related to other individuals from within the same population than they are to those from outside) [right], the distribution requires the maximum number of steps possible, L=12. Intermediate models require L = 4 ~ 11 events, the more structured models requiring fewer. For example, the trees with L = 8 and L=9 contrast alternative two-population models, in which the shorter has slightly more distinct sub-populations than the latter.

The principles in the idealized model can then be applied to larger data sets with real genomic data. The phylogenetic tree in Figure 5 was derived by one of a variety of well-established "known known" computational algorithms. With genomic data sets, the topological branching order is largely method-independent [7]. Moving backward in time from right to left, the branching order shows successively more inclusive groups of related individuals (clades). The shaded dots are characters attached to each individual, in this case its population of origin. The question is the co-occurrence of clades and populations as an historical biological process.

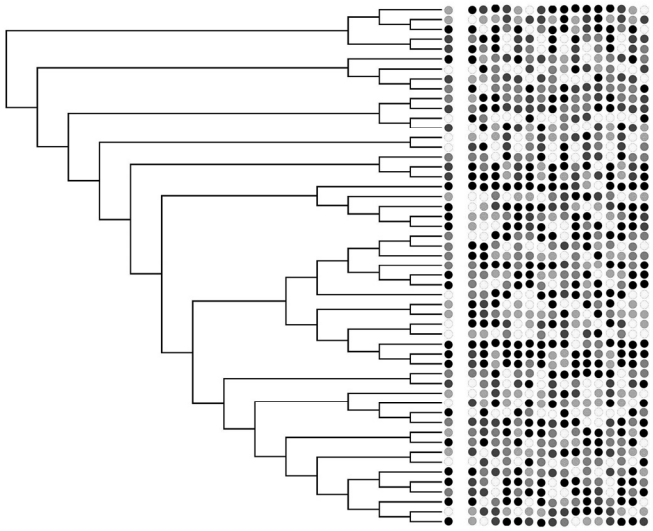

Figure 5 – Monte Carlo randomization of population assignments as a test of phylogeographic structure. For an observed phylogenetic tree [left] that shows the distribution of individuals across populations, the length L of the tree is the minimum number of vicariance events (historical movements) necessary to explain it. By repeatedly randomizing population assignments over the tips of the tree [right] and determining the length of the resultant tree, the observed length may be compared with the random distribution as a test of non-random structure. A set of 10,000 such randomizations gives a stable distribution.

Figure 6 shows the application of the Monte Carlo method to a population genomic data set from Harp Seals (*Pagophilus groenlandicus*) (after [6]). Harp Seals breed in exactly four places in the North Atlantic and adjacent waters, in the White Sea, Greenland Sea, the Newfoundland & Labrador Ice Front, and the southern Gulf of St Lawrence [top]. Whereas the two westernmost breeding sites are known to exchange animals, trans-Atlantic genetic relationships and those among the two eastern populations in particular have been unclear. The well-defined arrangement of populations sets up several *a priori* biogeographic hypotheses, including a linear 'four stepping-stone' model [middle] and a 'two-stone' trans-Atlantic model [bottom].

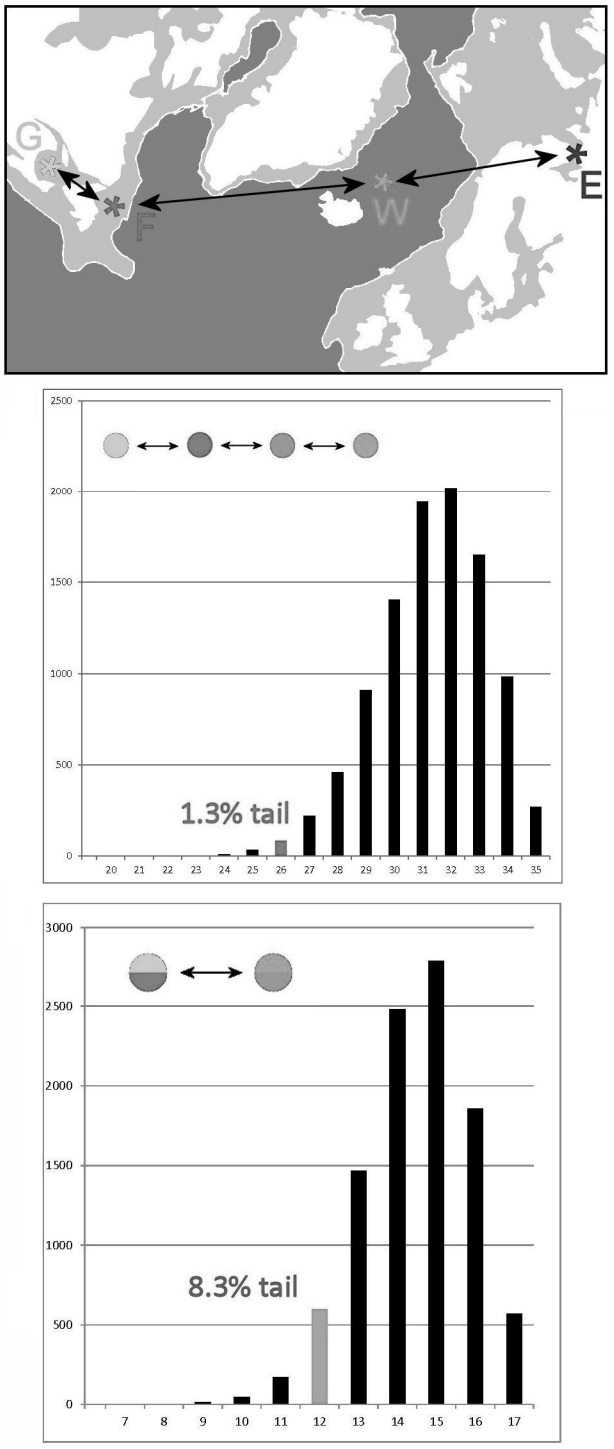

Figure 6 - Results of Monte Carlo simulations of alternative phylogeographic hypotheses for Harp Seals (*Pagophilus groenlandicus*) (after [6]). Phylogeographic models are encoded in a 4x4 matrix, so that it is possible to weight movements among population to reflect hypotheses of random or linear movements, or the likelihood of longer versus shorter movements. Each graph shows the distribution of the length L of 10,000 randomizations by the method in Figure 5, as compared to the observed length [shaded column]. For the linear, four-stone 'stepping stone' model [middle], the observed tree falls within the left-hand 5% tail and thus indicates that the model explains the distribution significantly better than does the random hypothesis of no structure. In contrast, the two-stone model [bottom] that groups the western and eastern population as pairs falls to the right of the 5% tail, such that it is not significantly shorter than random. The four-stone model is a better explanation of the distribution than the two-stone model [6].

Given the Monte Carlo procedure as a means of testing for non-random structure in intra-specific phylogenies as a whole, is it possible to make quantitative distinctions among the component populations of the species? Inspection of the tree may suggest qualitative patterns, for example that two populations seem to differ in their distribution among clades. Traditionally, such comparisons would be quantified by relative frequencies in row-by-column tests. However, when genomic data differentiate every individual, and simple comparison of group frequencies masks the nested nature of those groups as clades, such methods are unproductive. A more productive approach is to derive a numerical proxy for each of the phylogenetic components of the total population.

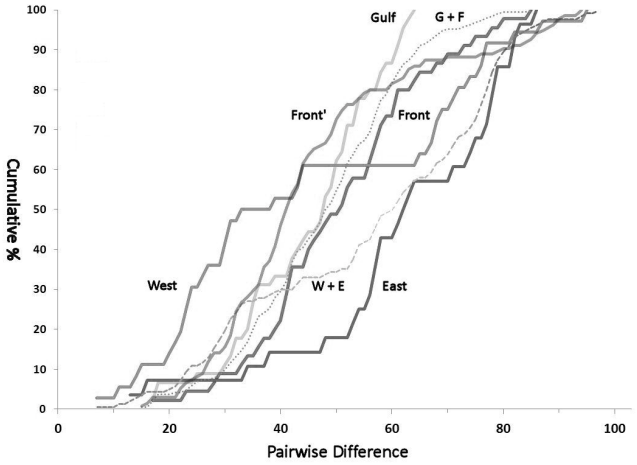

Figure 7 – Cumulative pairwise distance curves for populations of Harp Seals (after [6]). From a matrix of the observed pairwise DNA differences between all individuals in each of five populations, the cumulative curve shows the total fraction of the population differentiated at or below a particular pairwise difference. This curve serves as a quantitative proxy for a time-dependent branching family tree. Compared at 50%, curves displaced to the left indicate relatively 'young' populations in which the majority of animals diverged recently, in contrast to curves displaced to the right that indicate typically 'older' relationships. Differences among curves may be evaluated by a non-parametric Kolmogorov-Smirnov test, which evaluates the single greatest vertical difference between pairs of curves, which in this dimension indicates more or less rapid phylogenetic diversification [6].

3 Conclusions

The advent of NextGen DNA sequencing methods has massively increased the rate at which DNA sequence information can be generated, and the volume and complexity of the data matrices that apply to biological questions, including molecular and evolutionary biology. Questions include known knowns where computational methods can be applied to automated signal processing and ease of comparison among data sets, known unknowns inherent in patterns revealed for the first time by highly-resolved genomic phylogeny and phylogeography, and unknown unknowns lurking in cross-comparisons and pattern-detection among the higher-order dimensions of ordered data matrices. In summary,

- **Biotechnology**
  - **Iterative whole-genome DNA sequencing** on **microarrays: the *ArkChip***
  - ***Known Knowns***: Optimization & Automation of **signal processing algorithm**
  - ***Known Unknowns***: Comparison of data patterns within / between species
- **Phylogenetic Genomics** in *time*
  - ***Known Knowns*:** Reconstruction of intraspecific phylogenies ('family trees')
- **Phylogeographic Genomics** in *space*
  - ***Known Unknowns***: quantitation of phylogeny in space
    - **Monte Carlo** models for testing phylogeographic hypotheses
    - **Non-Parametric** comparison of proxies of phylogenetic topology
- ***Unknown Unknowns* ?**
  - **Higher-order interactions** in microarray data: sequence x species x array
  - **Pattern identification** in multiple dimensions

4 Acknowledgements

The experimental work was supported by research contracts from the Canadian Department of Fisheries and Oceans and a Discovery Grant from the National Science and Engineering Research Council (NSERC). I gratefully acknowledge the contributions of my co-authors and students on the papers referenced below. I am also grateful to my new colleagues in the Department of Computer Science at Memorial University, for stimulus in new research directions and encouragement to attend the BioComp'13 conference. For Justyna, Matilda, and Eowyn, with thanks for their indulgence.

5 References

[1] SMC Flynn & SM Carr. 2007. Interspecies hybridization on DNA resequencing microarrays: efficiency of sequence recovery and accuracy of SNP detection in human, ape, and codfish mitochondrial DNA genomes sequenced on a human-specific MitoChip. BMC Genomics 8, 339.

[2] SM Carr, HD Marshall, AT Duggan, SMC Flynn, KA Johnstone, AM Pope, & CD Wilkerson. 2008. Phylogeographic genomics of mitochondrial DNA: patterns of intraspecific evolution and a multi-species, microarray-based DNA sequencing strategy for biodiversity studies. Comparative Biochemistry and Physiology, D: Genomics and Proteomics 3, 1-11.

[3] SM Carr, AT Duggan, & HD Marshall. 2009. Iterative DNA sequencing on microarrays: a high-throughput NextGen technology for ecological and evolutionary mitogenomics. Laboratory Focus 13, 8-12.

[4] SM Carr & HD Marshall. 2008. Intraspecific phylogeographic genomics from multiple complete mtDNA genomes in Atlantic Cod (*Gadus morhua*): Origins of the "Codmother," trans-Atlantic vicariance, and mid-glacial population expansion. Genetics 108, 381-389.

[5] HD Marshall, MW Coulson, & SM Carr. 2008. Near neutrality, rate heterogeneity, and linkage govern mitochondrial genome evolution in Atlantic Cod (*Gadus morhua*) and other gadine fish. Molecular Biology & Evolution 26, 579-589.

[6] SM Carr, AT Duggan, GB Stenson, & HD Marshall. Quantitative analysis of phylogeographic structure; Whole-mitogenome variation among harp seals (*Pagophilus groenlandicus*) from discrete transatlantic breeding areas, Molecular Ecology, in review.

[7] MW Coulson, HD Marshall, P Pepin & SM Carr. 2006. Mitochondrial phylogeographic genomics of gadine fish: Implications for taxonomy and biogeographic origins. Genome 49, 1115-1130.