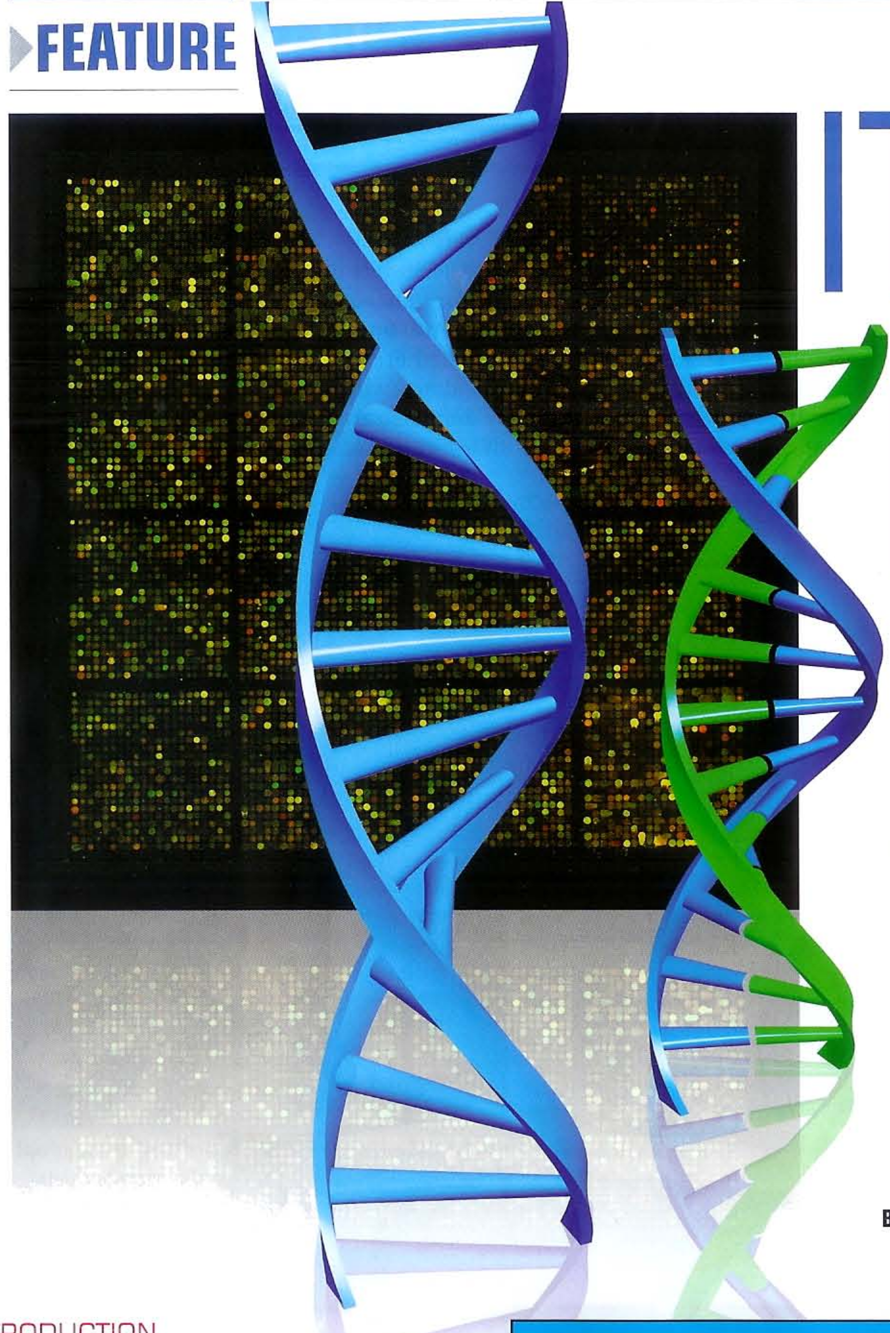
Laboratory Focus 13(2), 8-12

FEATURE



ITERATIVE DNA SEQUENCING ON MICROARRAYS:

a high-throughput NextGen technology for ecological and evolutionary mitogenomics

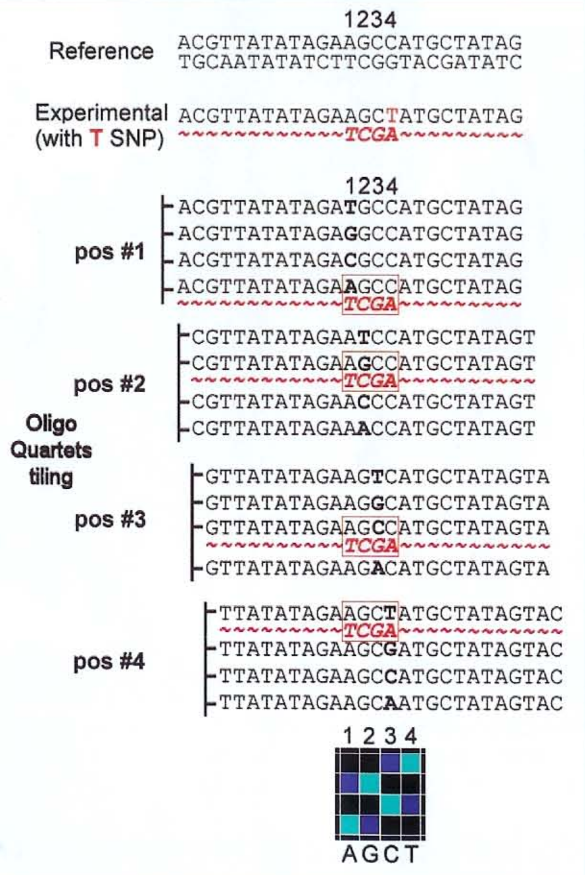
BY: SM CARR*, AT DUGGAN, HD MARSHALL
Dept of Biology, Memorial University of Newfoundland
St John's NL A1B3X9 *author for correspondence: scarr@mun.ca

INTRODUCTION

High-throughput DNA sequencing

Since their introduction in the late 1970s, methods of DNA sequence determination have come to rely on the use of dideoxy terminators[1], in which products of single-stranded DNA replication reactions are terminated base-specifically so as to produce a ladder of products each increasing by the step of a single nucleotide. At its inception, the dideoxy method relied on cloned DNA templates, electrophoretic separation, and autoradiographic detection. Successive technical innovations include PCR[2], use of fluorescent terminators readable by automated laser fluorometry[3] , and the development of capillary-based separation methods[4]. Bioinformatic innovations include algorithmic methods for automated reads[5]. Throughout these developments, the underlying dideoxy method has remained the dependable workhorse technology of the genomic revolution.

Figure 1 **SNP detection on a DNA sequencing microarray** (after Carr et al. [15].


1234
Reference
ACGTTATATAGAAGCCATGCTATAG
TGCAATATATCTTCGGTACGATATC
Experimental (with T SNP)
ACGTTATATAGAAGCTATGCTATAG
TCGA
1234
pos #1
ACGTTATATAGATGCCATGCTATAG
ACGTTATATAGAGGCCATGCTATAG
ACGTTATATAGACGCCATGCTATAG
ACGTTATATAGAAGCCATGCTATAG
TCGA
pos #2
CGTTATATAGAATCCATGCTATAGT
CGTTATATAGAAGCCATGCTATAGT
TCGA
CGTTATATAGAACCCATGCTATAGT
CGTTATATAGAAACCATGCTATAGT
Oligo Quartets tiling
pos #3
GTTATATAGAAGTCATGCTATAGTA
GTTATATAGAAGGCATGCTATAGTA
GTTATATAGAAGCCATGCTATAGTA
TCGA
GTTATATAGAAGACATGCTATAGTA
pos #4
TTATATAGAAGCTATGCTATAGTAC
TCGA
TTATATAGAAGCGATGCTATAGTAC
TTATATAGAAGCCATGCTATAGTAC
TTATATAGAAGCAATGCTATAGTAC
1 2 3 4
A G C T

The reference sequence at Positions 1-4 is AGCC. For Position 1, four alternative oligos are (top to bottom) ~~~TGCC~~~, ~~~GGCC~~~, ~~~CGCC~~~, and ~~~AGCC~~~, where the last corresponds to the reference sequence. Positions 2, 3, and 4 are also tiled with the T, G, C, A variant oligos in the 1st , 2nd, 3rd, 4th rows. Given an experimental sequence with a T SNP at the last position (AGCT), the complementary strand (~~~TCGA~~~) matches only one of the four variant oligos at each tiling position, and binds to it preferentially. Computer imaging of the probe intensity shows binding as a more or less intense pseudocolour (bottom inset). In this case, preferential annealing to the 4th, 3rd, 2nd, and 1st oligos at the four successive positions indicates that the original (complementary) experimental sequence is AGCT. Note that for the first three quartets, the SNP in the experimental sequence necessarily causes mismatch to the reference at Position 4: only in the last quartet is there a perfect match. Reduction of binding specificity in surrounding oligos is a characteristic of SNPs: see text.

Subsequent to completion of the Human Genome Project in 2004, new so-called "Next Generation" sequencing technologies have been introduced, including pyrosequencing (454 Life Sciences), reversible terminator technology (Illumina), and ligation sequencing (ABI SOLiD)[6,7]. A major challenge is to obtain any single genome of interest as a matter of routine[8].

Population and evolutionary geneticists interested in genomic variation within species face a different challenge. Our goal is to harness high-throughput technology to obtain multiple Kbps of sequence from Ns of 100s ~ 1000s of individual organisms, where each N is assayed twice, from either DNA strand. For this goal, another NextGen technology, DNA sequencing on microarrays, holds great promise.

Oligonucleotide Microarray DNA Sequencing

DNA microarrays comprise large sets of synthetic oligonucleotides affixed onto glass "chips"[9,10]. One of many applications is Variant Detector Arrays (VDAs), used to identify allelic variation at known SNP sites within populations[11]. A VDA includes a set of oligos, one of which is perfectly matched to the standard sequence of the SNP region, and a homologue with a single-base variant specific for the alternative base at a known SNP site. Currently available commercial "SNP Chips" can screen for > 100K known SNPs in the human nuclear genome[12].

The VDA concept has been extended to assay variation at every *potential SNP* site in an experimental DNA with respect to a reference DNA[13,14]. In one commercial design, the microarray tiles a reference sequence of length *n* bases as a series of 4 x *n* overlapping 25-b oligonucleotide probes (Fig. 1, after [15]). For each successive 25 base region, the four members of the quartet vary the middle (13th) base through the four possible A C G T SNPs. Mismatch at this position reduces hybridization specificity, such that an experimental DNA with a SNP variant will anneal preferentially to only one of the four oligos. The DNA sequence is thus read as the succession of brightest signals from each quartet (Fig. 2).

Mitochondrial DNA (mtDNA), a highly SNP-dense genome

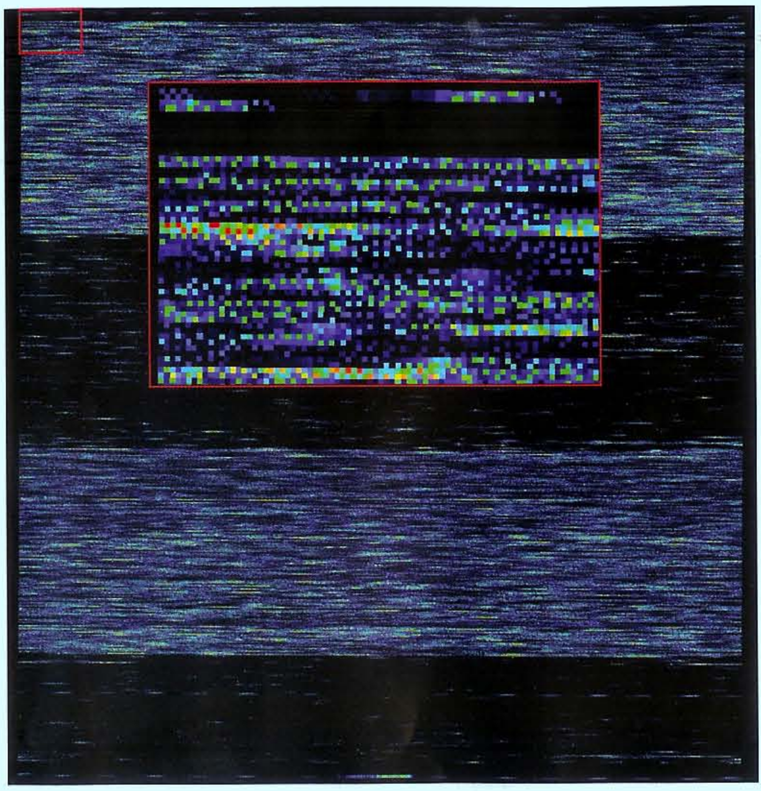
One of the first targets of microarray sequencing has been a particularly SNP-rich portion of the animal genome, the mitochondrial DNA (mtDNA) genome. MtDNA has been the foundation of modern molecular evolutionary genetics[16]. The significance of the mtDNA genome in evolutionary biology rests on an oft-repeated mantra including its small circular size (ca 16-17kbp), maternal inheritance, high mutation rate (and therefore high SNP density) with respect to typical nuclear loci. Demonstrated freedom from adaptive Darwinian natural selection and the absence of recombination mean that mtDNA gene genealogies should be equivalent to (maternal) intraspecific phylogenies, and that diversity should reflect population size[17]. The power of intraspecific whole-genome studies[18] has led to a vast mitogenomic database for genealogical investigation into our own species[19]. The mtDNA genome is also medico-genomically important, as a source of germline mutations that contribute to maternally-inherited diseases, and somatic mutations that contribute to aging and cancer[20].

Efficiency and accuracy of MitoChip sequencing

Experiments with human-mtDNA-specific "MitoChips" show that the technology is both efficient (able to recover sequence accurately) and accurate (able to identify SNPs accurately). Maitra et al.[21] showed that the first-generation oligonucleotide

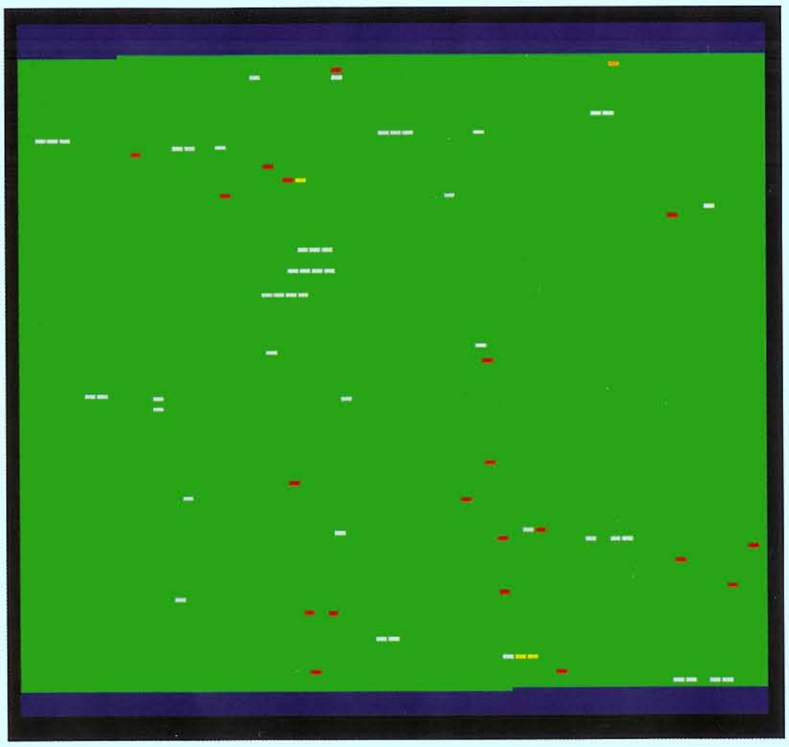
FEATURE

Figure 2 A multispecies ArkChip iterative sequencing microarray



The complete array tiles seven species' mtDNA genomes (115,170 b from both strands). These are three fish (Atlantic Cod [Gadus morhua], Atlantic Wolffish [Anarhichas lupus], and Atlantic Salmon [Salmo]), three mammals (Humans [Homo sapiens], Harp Seals [Pagophilus groenlandicus], and Caribou [Rangifer tarandus]), and one bird (Blackish Oystercatcher [Haematopus ater]). The array shows the hybridization pattern for a four-species experiment (Gadus and Anarhichas above, Pagophilus and Rangifer and below). The inset box shows an enlargement of the left-corner of the Gadus block, including two alignment grids above the main body.

Figure 3 Diagram of information content for a single-species ArkChip experiment



The framed area shows 16,552 cells that represent the complete mtDNA genome sequence of an Atlantic Cod (Gadus morhua) sequenced on a seven-species microarray. Cells in green (16,483) are high-confidence calls that are identical to the reference sequence. Cells in red (20) are high-confidence calls (dS/N > 0.13) that differ from the reference, i.e., SNPs. The one cell in orange [upper right] is a lower-confidence SNP call (dS/N > 0.09). Cells in yellow (3) are lower-confidence ambiguous calls. Cells in grey (45) are lower-confidence calls that do not differ from the reference. Note that grey cells typically occur in short runs, and that the ambiguous calls are all adjacent to SNPs [upper left] or grey cells [bottom right].

tiling array (MitoChip v.1.0: Affymetrix, Santa Clara, CA), which included 15,451 bases of the mtDNA coding region, assigned base calls for an average of 96% of positions, with reproducibility of 99.99%. Flynn and Carr[22] recovered 99.67% of the human sequence with 100.0% accuracy of SNP detection. Zhou et al.[23] developed a MitoChip v.2.0 that tiles the complete 16,569bp revised Cambridge Reference Sequence, including the hypervariable regions. Their average call rates across 33 arrays was 94.6%. Hartmann et al.[24] used the same chip to compare 93 mitochondrial genomes sequences and measured an average call rate of 99.48%, with an accuracy of ≥99.98% for the MitoChip with respect to dideoxy methods. They observed that inaccurate calls were most commonly associated with runs of four or more C bases [which affects G+C content], or within the region ±12 bases of authentic SNPs. The latter phenomenon occurs because the presence of a SNP means that other oligo sets tiling the surrounding bases necessarily have two mismatches to the reference (cf. Fig. 1). "Flynn's Rule" states that, where SNPs are spaced $13 < n \leq 25$b apart, anomalies may be expected among [(2)(25 n) + 1] oligo quartets tiling the intermediate positions[22].

A further consequence of Flynn's Rule is that heterospecific experimental DNAs bind inefficiently to species-specific microarrays, when their sequences differ by more than a few percent. We challenged the human MitoChip v.1.0 with chimpanzee, gorilla, and codfish mtDNA genomes[22], whose sequences differ from the reference by by eight, 10, and >30%, respectively. For the gorilla genome, in which 46% of all 25b regions contain three or more SNP differences from the reference, only 88% of the sequence was recoverable, and one in four SNP identifications were in error. In the codfish genome, less than 4% of the sequence was recoverable. These results indicate that intraspecific cross hybridization should not interfere with the accurate recovery of data from multiple reference sequences tiled on the same microarray, provided that the competing DNA sequences differ on average by >5 mismatches per 25b oligo.

Accordingly, we designed and constructed a multi-species iterative sequencing microarray (the "ArkChip") tiled with a series of reference mtDNA genomes from different vertebrate species whose sequences differ from each other by >20%[15]. Here, we examine the efficiency of sequence recovery and accuracy of intra-specific SNP detection as affected by the presence of multiple homologous mtDNAs from different taxonomic orders and classes.

Methods

Design of the microarray

The Arkchip includes reference mitochondrial DNA genomes (16 ~ 17K bp each) sequences from three fish, thee mammal, and one birds species, for a total of 115,170 bases per strand, from both the light and heavy strands (Fig. 2).

Preparation and microarray sequencing

PCR amplification of complete mitochondrial DNA genomes was done in 3 ~ 24 overlapping fragments, according to species. Each species' amplicons were pooled in equimolar quantities, and the pooled amplicons were fragmented by DNAse treatment to ca. 20 ~ 200 bp. Fragmented DNA was sent to The Centre for Applied Genomics at the Hospital for Sick Children, Toronto, where they were labeled and fluorescently stained according to the Affymetrix GeneChip CustomSeq protocol, v. 2 (2003). Arrays were scanned with an Affymetrix GeneChip Scanner 3000 and data extracted with the GeneChip DNA Analysis Software.

Algorithmic analysis of probe intensities

Output from each array experiment includes eight columns of probe intensity values, corresponding to the A, C, G, and T variants for each position, for both the heavy and light strands. We identify for each position on each strand the strongest of the probe intensities as the presumptive base call, and calculate as a confidence score the differential signal-to-noise score (dS/N), defined as the difference between the strongest and next-strongest intensities, divided by the sum of all intensities at that position. For each position, where the calls from the two strands agree and agree with the reference, the position is called as invariant. Where the calls from the two strands disagree, but the call with the higher dS/N agrees with the reference, the position is called as a weak invariant. Where the calls from the two strands agree, and differ from the reference, we accept as high-confidence SNPs those calls that are made on both strands at $dS/N \geq 0.13$, and as potential (lower-confidence) SNPs those made on both strands at $0.09 \leq dS/N < 0.013$. All calls that cannot be called by these criteria are assigned appropriate ambiguity codes. Irrespective of the above criteria, where total probe intensity at any position is small, the position is scored as N. A rule-of-thumb is to score as N, any position in which the sum of probe intensities is less than that of any position called by the above criteria.

Results

We summarize here results obtained from a series of experiments with one species on the ArkChip, the Atlantic Cod; comparative results on all species[24] will be reported elsewhere (Duggan, Marshall, and Carr, in prep).The tiled *Gadus* genome reference comprises 16,552 positions for each strand.

Fig. 3 shows the information content of a single-species experiment sequenced on the multispecies array. Of 21 dideoxy-detectable SNPs, 20 were detected by at high confidence and one at lower confidence, for an accuracy of 100%. Exclusion of three ambiguous and 45 N sites gives a mean efficiency (fraction of bases called correctly) of the single-species experiment of 99.58%.

In a four-species experiment with the same *Gadus* individual, together with another fish and two mammal species (*Anarhichas, Pagophilus, and Rangifer*) (Fig. 4), there are 19 positions at which anomalous SNP or ambiguous calls are made. In other four-species experiments, these sites are typically identified as ambiguities or weak or strong SNPs. Among 27 dideoxy sequences, none of these sites has been shown to vary intraspecifically[25]. In duplicate seven-species experiments, there are a further 30 positions at which anomalous SNP or ambiguous calls occur in one or the other of both replicates. In only one case does an ambiguity occurs at a position previously known to vary intraspecifically, in this case a variant unique to one fish. The final efficiency excluding these positions is 99.29%.

In almost all cases, these ambiguous positions can be shown to occur at homologous positions in an alternative species that differs from *Gadus* by a single substitution within the 25-base interval spanned by the oligo quartet (Fig. 5). That is, these "pseudo-SNPs" are explainable as inter-specific polymorphisms detected as intra-specific SNPs in conserved in regions, typically the 12S and 16S rRNA regions. Mean efficiency excluding these positions is 99.47%.

Discussion and Conclusions

Where reference species' DNA sequences differ on average >20%, interspecific cross-hybridization does not interfere with the accurate recovery of species-specific data from multispecies microarrays. Authentic SNPs generally occur in intraspecifically variable regions that are highly differentiated interspecifically, such that "cross-talk" does not occur. In more conserved regions, identification and exclusion of 'pseudo-SNPs' sites from each species' analysis eliminates noise due to trans-specific cross-hybridization, without loss of intraspecific phylogenetic signal. Potential interference from 'pseudo-SNPs' may be further minimized if more divergent reference genomes are incorporated, e.g., teleost / lissamphibian / reptilian / mammalian / avian combinations, or even invertebrate taxa.

Parallel collection of data from multiple species on a multiplex microarray dramatically reduces the time and cost of obtaining mtDNA genome sequence data for population studies. The complete genome sequence automatically aligns with the reference, eliminating contig assembly from multiple reads. Automation of the base calling algorithm eliminates the subjectivity of manual editing, and promotes consistency of base calling across experiments. More sophisticated Perl- or Python-script-based algorithms could identify and exclude pseudo-SNPs, flag N-rich regions of poor hybridization, and incorporate context-specific information.

Iterative sequencing microarrays lend themselves to collaborative studies and centralization of the chip-reading technology. Our ArkChip microarray manipulations are performed out-of-house at a Genome Canada-sponsored service facility. Because in-house preparation requires only PCR amplification, quantitation, and fragmentation, in principal collaborative projects on the same ArkChip platform can be dispersed among investigators, with the fragmented amplicon pools fed into a single remote pipeline.

The latest generation of chip designs can incorporate nearly 500Kb of reference sequence. It should be possible to incorporate large numbers of species-specific reference sequences, not all of which need be used in any one experiment. By such strategies, the door to high-throughput "Next-Gen" sequencing technology can be opened to individual investigators of non-"genome enabled species" at a small fraction of the cost of conventional genomics.

Acknowledgments

This review is based in part on a poster presentation by ATD at the 2008 Society for Molecular Biology and Evolution meeting in Barcelona, Spain, as part of the Undergraduate Diversity Mentoring Program. ATD gratefully acknowledges receipt of a travel grant from the Society for Molecular Biology and Evolution. These experiments were funded by the Department of Fisheries & Oceans, White Hills, as part of a Grants and Contributions agreement to SMC through Memorial University. We thank Newfound Genomics, St. John's for their assistance with early experiments, and Drs Chao Lu and Lan He of The Centre for Applied Genomics in Toronto for their generous assistance with microarray processing.

References

1. Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chai terminating inhibitors. *Proc Natl Acad Sci USA* 74:5463 5467.
2. Saiki RK, DH Gelflan, S Stoffel, SJ Scharf, R Higuchi, G. T. Horn, KB Mullis, HA Ehrlich. 1988. Primer directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239:487 491
3. Landegren U, R Kaiser, CT Caskey, L Hood (1988). DNA diagnos tics molecular 3. techniques and automation. *Science* 242,229 237
4. Dolnik V. 1999. DNA sequencing by capillary electrophoresis (review). *J Biochem Biophys Methods.* 1999 41:103 19..
5. Ewing B, Hillier L, Wendl M, Green P. 1998. Base-calling of autoamted sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8:175-185.
6. Bentley D. 2006. Whole-genome re-sequencing. *Curr Opin Genet Dev* 16:25-29.
7. Gibson and Muse 2009. A Primer of Genome Science, 3rd ed. Sinauer: Sunderland MA.
8. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC. The diploid genome sequence of an individual human. *PLoS Biol.* 2007 5:e254.

9. D Gresham, MJ Dunham, D Botstein 2008. Comparing whole genomes using DNA microarrays. Nat Rev Genet 9:291-302.

10. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, Altshuler D. Integrated detection and population genetic analysis of SNPs and copy number variation. *Nat Genet*. 2008 40:1166 74.

11. Wang, D.G., Fan, J. B., Sia, C. J, Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T.J., Lipshutz, R., Chee, M., Lander, E. S., 1998. Large scale identification, mapping, and genotyping of single nucleotide polymorphisms in the human genome. *Science* 280:1077 1082.

12. Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, Shen MM, Kulp D, Kennedy GC, Mei R, Jones KW, Cawley S: Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. *Bioinformatics* 2005, 21:1958 1963.

13. Hacia JG, Brody LC, Chee MS, Fodor SP, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two colour fluorescence analysis. *Nat Genet*. 1996. 14:441 447.

14. Reider MJ, Taylor SL, Tobe VO, Nickerson DA. 1998. Automating the identification of DNA variations using quality based fluorescence re sequencing: analysis of the human mitochondrial genome. *Nuc Acids Res* 26, 967 973.

15. Carr SM, Marshall HD, Duggan AT, Flynn SMC, Johnstone KA, Pope AM, Wilkerson CD Wilkerson. Phylogeographic genomics of mitochondrial DNA: patterns of intraspecific evolution and a multi-species, microarray-based DNA sequencing strategy for biodiversity studies *Comp Biochem Physiol D Genomics and Proteomics* 2008; 3:1-11.

16. Wilson, A.C., Cann, R.L., Carr, S.M., George Jr, M., Gyllensten, U.B., Helm Bychowski, K., Higuchi, R.G., Palumbi, S.R., Prager, E.M., Sage, R.D., Stoneking, M., 1985. Mitochondrial DNA and two perspectives on evolutionary genetics. *Biol. J. Linn. Soc.* 26, 375 400.

17. Marshall HD, Coulson MW, Carr SM. 2008. Near neutrality, rate heterogeneity, and linkage govern mitochondrial genome evolution in Atlantic Cod *(Gadus morhua)* and other gadine fish. *Mol Biol Evol* 26:579-589.

18. Ingman, M., Kaessmann, H., Paabo, S., Gyllensten, U., 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408, 708 713.

19. Torroni, A., Schilli, A., Macaulay, V., Richards, M., Bandelt, H. J., 2006. Harvesting the fruit of the human mtDNA tree. *Trends Genet*. 22, 339 345.

20. Wallace DC. 2008 Mitochondria as chi. *Genetics*. 2008. 179:727 735.

21. Maitra A, Cohen Y, Gillespie SE, Mambo E, Fukushima N, Hoque MO, Shah N, Goggins M, Califano J, Sidransky D, Chakravarti A. 2004. The Human MitoChip: a high throughput sequencing microarray for mitochondrial mutation detection. *Genome Res* 14:812–819.

22. Flynn SMC, Carr SM. 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* 2007; 8,33

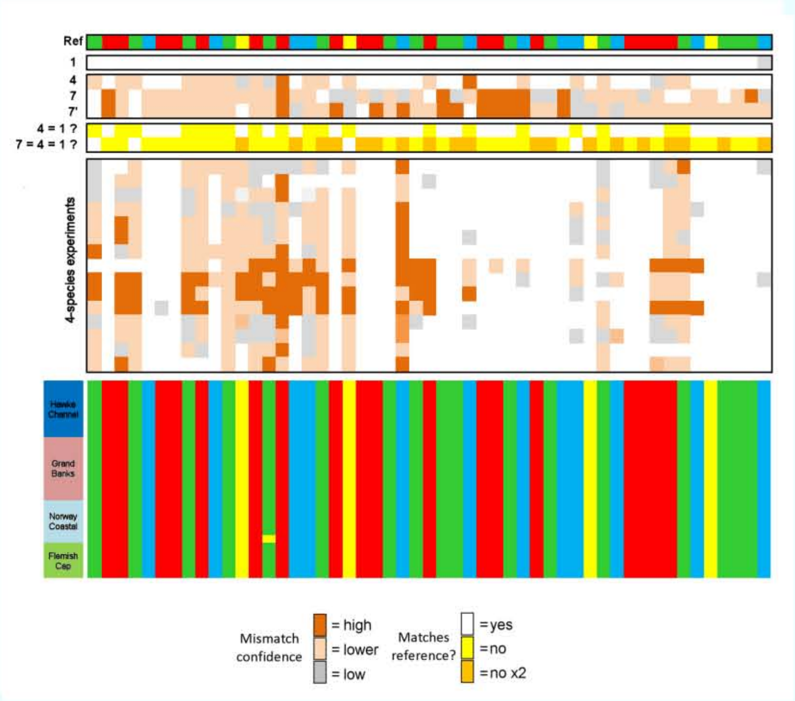
23. Zhou S, Kassauei K, Cutler DJ, Kennedy GC, Sidransky D, Maitra A, Califano J. 2006. An oligonucleotide microarray for high-throughput sequencing of the mitochondrial genome. *J Mol Diagn*. 2006 Sep;8(4):476 82.

24. A Hartmann, M Thieme, LK Nanduri, T Stempfl, C Moehle, T Kivisild, PJ Oefner 2008. Validation of microarray based resequencing of 93 worldwide mitochondrial genomes. *Mutation Res*, Epub ahead of print.

25. Duggan AT. 2007. Efficiency and accuracy of a multi species iterative DNA sequencing microarray. BSc hons thesis, Memorial University of Newfoundland.

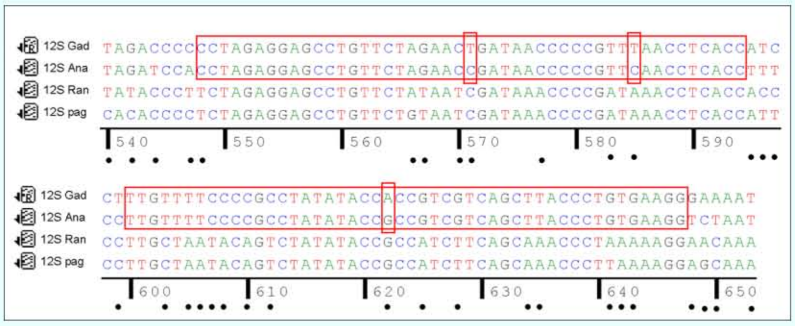
26. Carr SM, Marshall HD. Marshall. Intraspecific phylogeographic genomics from multiple complete mtDNA genomes in Atlantic Cod *(Gadus morhua)*: Origins of the "Codmother," trans-Atlantic vicariance, and mid-glacial population expansion. *Genetics* 2008; 108,381-388.

Figure 4 Schematic diagram of ambiguities in Gadus in four- & seven-species experiments.



ArkChip experiments were done with the same cod mtDNA genome singly ('1'), in combination with three other species ('4'), and in duplicate with six other species ('7' and '7''). The reference sequence [1st row] includes 49 non-consecutive positions, in standard colours A C G T. The single-species sequence is taken as canonical [2nd row], and is invariant at these positions. The four-species replicate [3rd row] differs from this at 19 positions: 17 are ambiguous [light orange] and two are incorrect, high-confidence calls [dark orange]. In two duplicate seven-species experiments [4th & 5th rows], miscalls occur at an additional 30 positions in one [yellow] or both [orange]. Among 15 four-species experiments with different Gadus mtDNA genomes [6th block], most of the first set of 19 positions are called variously as ambiguities, high-confidence SNPs, or correctly but at lower confidence [grey]. None of these positions has been shown to vary among 27 cod genomes sequenced by conventional dideoxy sequencing [7th block], including genomes closely related to those sequenced on the ArkChips; only one position [col 14] called ambiguously in a seven-species experiment is known to vary[25].

Figure 5 "Pseudo-SNPs" in conserved sequence block in the mitochondrial 16S rRNA gene.


12S Gad TAGACCCCCCTAGAGGAGCCTGTTCTAGAACTGATAACCCCCGTTTAACCTCACCATC
12S Ana TAGATCCACCTAGAGGAGCCTGTTCTAGAACCGATAACCCCCGTTCAACCTCACCTTT
12S Ran TATACCCTTCTAGAGGAGCCTGTTCTATAATCGATAAACCCCGATAAACCTCACCACC
12S pag CACACCCCTCTAGAGGAGCCTGTTCTGTAATCGATAAACCCCGATAAACCTCACCATT
540
550
560
570
580
590
12S Gad CTTTGTTTTCCCCGCCTATATACCACCGTCGTCAGCTTACCCTGTGAAGGGAAAAT
12S Ana CCTTGTTTTCCCCGCCTATATACCGCCGTCGTCAGCTTACCCTGTGAAGGTCTAAT
12S Ran CCTTGCTAATACAGTCTATATACCGCCATCTTCAGCAAACCCTAAAAAGGAACAAA
12S pag CCTTGCTAATACAGTCTATATACCGCCATCTTCAGCAAACCCTTAAAAGGAGCAAA
600
610
620
630
640
650

The 197-base region between positions 2361-2558 [Gadus reference] is broadly conserved among fish and mammal species. For example, position 2546 shows an A / G polymorphism between the two fish versus the two mammal species, respectively, bounded by an invariant region >11 b on either side across all species. In the four-species experiments, this position is consistently identifiable in both fish species as a "pseudo-SNP."