Available online at www.sciencedirect.com

ScienceDirect

Comparative Biochemistry and Physiology, Part D 3 (2008) 1–11

CBP

www.elsevier.com/locate/cbpd

Phylogeographic genomics of mitochondrial DNA: Highly-resolved patterns of intraspecific evolution and a multi-species, microarray-based DNA sequencing strategy for biodiversity studies ☆

Steven M. Carr *, H. Dawn Marshall, Ana T. Duggan, Sarah M.C. Flynn, Kimberley A. Johnstone, Angela M. Pope, Corinne D. Wilkerson

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

Received 27 March 2006; received in revised form 14 December 2006; accepted 15 December 2006
Available online 24 February 2007

Abstract

Phylogeographic genomics, based on multiple complete mtDNA genome sequences from within individual vertebrate species, provides highly-resolved intraspecific trees for the detailed study of evolutionary biology. We describe new biogeographic and historical insights from our studies of the genomes of codfish, wolffish, and harp seal populations in the Northwest Atlantic, and from the descendants of the founding human population of Newfoundland. Population genomics by conventional sequencing methods remains laborious. A new biotechnology, iterative DNA "re-sequencing", uses a DNA microarray to recover 30–300 kb of contiguous DNA sequence in a single experiment. Experiments with a single-species mtDNA microarray show that the method is accurate and efficient, and sufficiently species-specific to discriminate mtDNA genomes of moderately-divergent taxa. Experiments with a multi-species DNA microarray (the "ArkChip") show that simultaneous sequencing of species in different orders and classes detects SNPs within each taxon with equal accuracy as single-species-specific experiments. Iterative DNA sequencing offers a practical method for high-throughput biodiversity genomics that will enable standardized, coordinated investigation of multiple species of interest to Species at Risk and conservation biologists.

Keywords: Evolutionary genomics; Biodiversity; Phylogeography; Mitochondrial DNA; Microarrays; Iterative sequencing; "ArkChip"

1. Introduction

Genomics, the study of complete gene sets in biological organisms, is a new science that can answer some very old questions of population biology. Whereas "genetics" traditionally considers one or a few genes at a time, "Genomic thinking" is a novel analytical approach that uses massively-parallel, high-throughput biotechnologies to obtain information and ask questions about large numbers of interdependent genes simultaneously.

The nuclear genome is the one we usually think about when we think of "genomics" (International Human Genome Sequencing Consortium, 2001). There is however a second genome, the mitochondrial genome or mtDNA, found in the extranuclear organelles involved in cellular respiration in the cells of all eukaryotes. MtDNA is famously a small, circular genome, about 17 kbp in circumference and comprising 38 genes in vertebrate species (Wilson et al., 1985). These are inherited like a single chromosome through a single parent, the mother. Because of this, mtDNA is a useful molecule for tracing maternal lineages in time and space, and has had wide use over the last 25 years in population biology and evolution. Many of these studies have sought to link population genetics and biogeographic evolution, and the approach of examining genetic relationships in their geographic context has been termed phylogeography (Avise, 2000). A limitation of such studies is the limited resolution possible when only one or a few loci are examined.

☆ This paper is based on a presentation given at the session "Genomics in Aquaculture" during the Annual Main Meeting of the Society for Experimental Biology hosted by the Universitat Autonoma de Barcelona, Barcelona, Spain, 11th–15th July 2005.

* Corresponding author. Tel.: +1 1 709 737 4776; fax: +1 1 709 737 3018.
E-mail address: scarr@mun.ca (S.M. Carr).

1744-117X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cbd.2006.12.005

We present here the results of several mtDNA genome studies underway in our laboratory, to illustrate the power of mitochondrial phylogeographic genomics for biodiversity. These investigations include marine species found in the western North Atlantic and elsewhere, several of which are included on Canada's list of Species At Risk of extinction, as well as descendants of the founding human population of the island of Newfoundland. They provide highly-resolved insights into previously-unsuspected phylogeographic patterns, including details of clade structure and relationships, as well as indications of historical population origins and movements. These studies employed conventional methods of PCR, dideoxy sequencing, and contig assembly, methods that remain laborious. We have therefore applied a new biotechnology, iterative sequencing on DNA microarrays ("re-sequencing"), that is able to recover a complete mtDNA genome sequence in a single experiment. We present evidence that the method is accurate for SNP identification within a single species, and show how initial results from a multi-species microarray (the "ArkChip") provide an efficient, practical strategy for simultaneous, iterative sequencing across species.

2. Genomic phylogeography

2.1. Genomic phylogeography and the "Daughters of Eve" in Newfoundland

As an introduction to mitochondrial phylogeographic genomics, we consider first a familiar species, *Homo sapiens*. The first mitogenomic study of humans was that of Ingman et al. (2000), who examined 53 complete genomes from individuals drawn from a variety of ethnic groups. [It should be emphasized that this was a study of individuals, and their genome-types are not necessarily diagnostic or characteristic of the groups with which they are identified]. Their results reinforced the concepts of an "Out of Africa" origin of modern humans, and of a "Mitochondrial Eve," a common female ancestor to whom all living humans trace their (maternally-inherited) mtDNA genomes (Cann et al., 1987). All non-Africans examined shared a common ancestor at <40 KYBP, and that within this clade (an ancestor–descendant lineage), all Europeans fell into one of the two distinct subclades. Pairwise genomic differences among Europeans ranged from 9 to 41 substitutions, compared with differences of up to 106 between African and non-African pairs. Studies of the hypervariable D-loop Control Region (Brown et al., 1979) had previously established a further refinement of the "Eve" hypothesis, the so-called "Daughters of Eve" (Richards et al., 1998), corresponding to the major, more or less distinctive clades within the human population of western Europe. Seven such lineages have been identified and designated H, J, K, T, U, V, and X, or the daughters respectively of "Helena," "Jasmine," "Katrina," "Tara," "Ursula," "Velma," and "Xenia" (Sykes, 2002).

We have examined the complete mtDNA genome sequence of twenty matrilineal descendants of the founders of the population of Newfoundland, an island province off the Atlantic coast of Canada, as part of a study to identify homogeneous population isolates that would be useful in genetic epidemiological approaches to identifying genes associated with complex disease conditions (Pope, 2004). Newfoundland, the first of England's overseas colonies, was settled initially between 1592 and 1830 by a small number of families, mostly from the West Country of England and southeastern Ireland, with a smaller minority from France (Mannion, 1977). Settlement occurred originally in many small "outport" communities around the coast; limited subsequent immigration, geographic isolation, and religious segregation limited genetic exchange among these settlements until the last few generations. The expected consequence of these "founder events" is a loss of genetic biodiversity within communities, due initially to sampling error, and subsequently because members of a closed community eventually become related and variation is lost more quickly in smaller communities over time, simply by chance. The consequences of such a demographic structure include an increased incidence of certain genetic disease conditions, including Bardet–Beidel Syndrome (Moore et al., 2005) and hereditary colorectal cancer (Woods et al., 2005).

Rather than reduced variability, we have found that every Newfoundlander examined has a unique mtDNA sequence (Fig. 1). In combination with representative genomes from Ingman et al. (2000) and elsewhere deposited in GenBank, it can be seen that the founding population of Newfoundland included six separate lineages, corresponding to five of the seven "Daughters of Eve." Relationships among these "daughters" are more sharply and consistently defined in the whole-genome data than the Control Region data alone (cf. Torroni et al., 2006). Most individuals (including English, Irish, and French descendants) occur in the common western European "Helena" clade, as expected, but distinct lineages of English and Irish Newfoundlanders are daughters of "Jasmine" and "Tara" or "Ursula" and "Katrina," respectively. Proportions of these haplotypes are similar to those reported for other western European populations. Daughters of the relatively-rare "Velma" and "Xenia" clades have not yet been discovered in Newfoundland. One French Newfoundlander occurs in the genetically-distinct A or "Aiyana" clade, which is common in northeastern Eurasian natives and North American First Nations peoples, but is otherwise unknown in western Europeans (Mishmar et al., 2003; Reidla et al., 2003). This individual is likely the descendant of a daughter of a First Nations mother and a French father, who was taken into the French community and whose mtDNA lineage has persisted to the present generation.

2.2. "One stock, two stocks, Red Fish, Blue Fish": fisheries phylogeography of gadid codfish

We next consider some marine species that fall under Canada's Species At Risk Act (SARA). SARA establishes a legal list of species considered to be Endangered, Threatened, or of Special Concern with respect to extinction. The list is determined by a national advisory committee, the Committee on the Status of Endangered Wildlife in Canada (COSEWIC). One of the first decisions to be made in this process is whether a species or population constitutes a recognizable Designatable Unit. For such purposes, genomic data are uniquely valuable.

Analysis of complete mtDNA genomes of codfish and their relatives provides a fully-resolved evolutionary tree that clarifies the phylogenetic and biogeographic relationships with this commercially-important group. Within species, pairwise genome sequence differences between Atlantic cod (*Gadus morhua*) on either side of the Atlantic Ocean are smaller (52 differences) than those between Pacific cod (*G. macrocephalus*) from either side of the Pacific Ocean (73 differences). Alaska or Walleye Pollock (*Theragra chalcogramma*) are more closely related to Atlantic cod than either is to Pacific cod: they represent an independent invasion of the Pacific basin, and should be included in the genus *Gadus* as originally described (Carr et al., 1999; Coulson et al., 2006).

Within the Atlantic cod, understanding of sub-structure among populations remains a pressing scientific and practical issue. Following the collapse of the Northern cod stock (Northwest Atlantic Fisheries Organization (NAFO) Divisions 2J3KL) in the late 1980s and the imposition of a moratorium in 1992, the failure of offshore migratory cod to recover, concomitant with the appearance of aggregations of adult fish in the deep inshore bays around the island, has raised important questions about the affinities of cod in this area. Previous measurements of single-locus mtDNA sequences show that

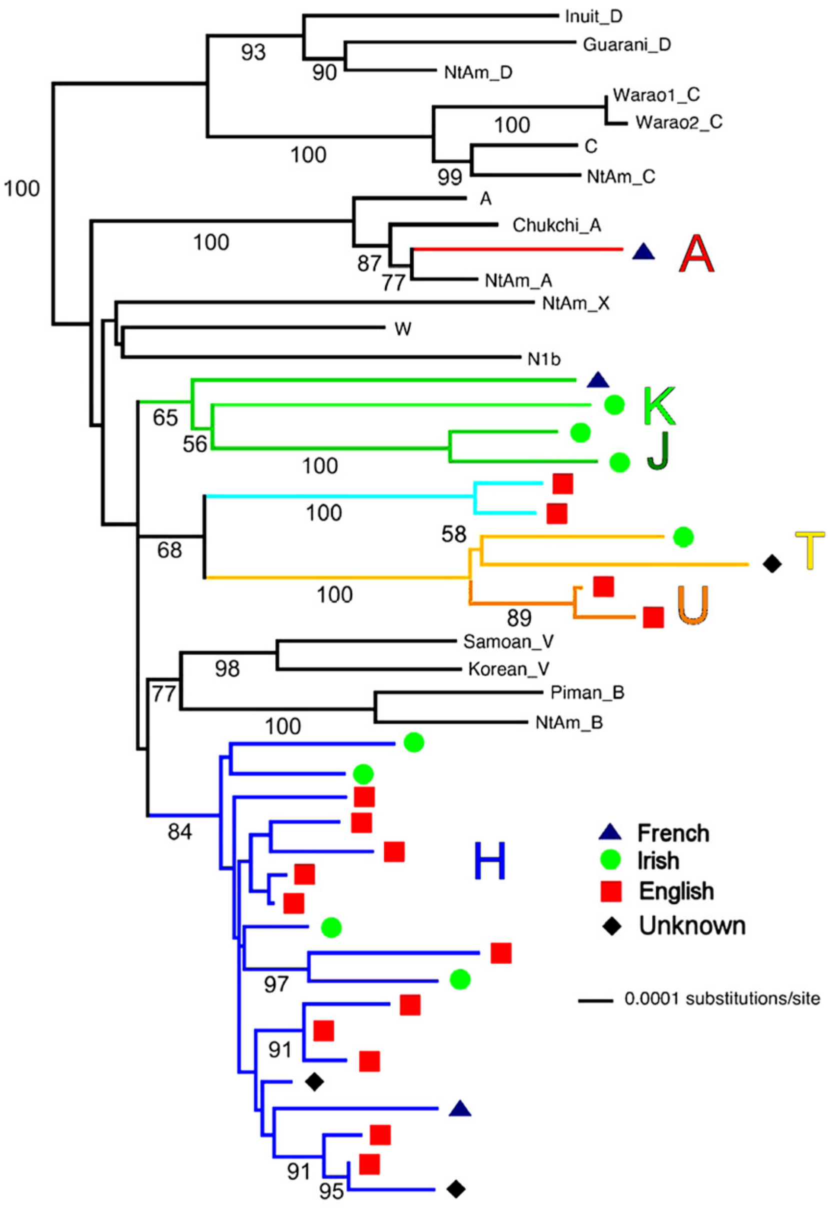


Fig. 1. Genomic phylogeography of the "Daughters of Eve" in Newfoundland. The island of Newfoundland was settled between 1593 and 1830 by a small number of families of English, Irish, and French origin. Subsequent demographic history has kept the original communities small and isolated. Rather than showing loss of genetic variation through founder "effects" and subsequent genetic drift, matrilineal descendants of these settlers are genetically-unique and occur in diverse clades corresponding to five of the seven "Daughters of Eve," the major non-African clades in modern humans.

essentially none of the observed haplotype variance is attributable to among-sample subdivision, consistent with the notion that geographically-disjunct management units in the Northwest Atlantic do not constitute genetically-distinct (or even readily-distinguishable) stocks (Pepin and Carr, 1993). In contrast, population structure in the Northeast Atlantic is markedly different, and a significant component of the genotypic variance is attributable to trans-Atlantic differentiation (Arnason, 2004). Questions about the reality of localized offshore and "bay stocks" remain, and contrasting interpretations from microsatellite variation have been argued (Carr et al., 1995; Carr and Crutcher, 1998).

We have assembled complete genome sequences of fish from two divisions in the Northern cod complex, an offshore seamount at Flemish Cap, and a Norwegian population in the Barents Sea. Fig. 2 shows the genomic "family tree" of 34 individual fish. As with humans, every fish has a unique mitochondrial genome sequence. The tree shows five major clades (A–E), with extensive genetic variation and deep branches across three clades for fish from the Barents Sea (Blue Fish in A, C, and D). In contrast, the majority of Northern cod are closely related within a single lineage (Red Fish in E). Within Northern cod, comparison of fish from Labrador with those from the North Cape of the Grand Banks shows little if any population subdivision. There is a persistent, older clade that shows up a low frequency (B). In contrast, cod at Flemish Cap (Green Fish), an offshore seamount in 3M, show markedly greater genome variation and diversity, and occur in both the western (B and E) and Barents Sea (A) clades. One explanation for these observations is the loss of genome variability in Northern cod as a result of the population crash. Another possibility is the origin of Northern cod through a population "bottleneck" either from the eastern Atlantic or a marine refugium near Flemish Cap, with subsequent migration.

2.3. Genomic differentiation of wolffish Species At Risk

Among the more than 500 species or populations currently on the SARA list, the first marine fish species to be listed as Threatened with extinction under the Canadian Species At Risk Act are spotted and Northern wolffish (*Anarhichas minor* and *A. denticulatus*, respectively); a third species, striped wolffish (*A. lupus*), is listed as of Special Concern. As part of the recovery plan for wolffish, we determined the complete mitochondrial DNA (mtDNA) genome sequences of all three species in order to identify the most variable gene regions for population analysis. The sequencing strategy illustrates our biodiversity strategy. With the known genome sequence of the gadiform *G. morhua* as a reference, aligned to known perciform, pleuronectiform, and salmoniform genomes, we identified conserved DNA sequences across orders that are sufficiently similar to wolffish to serve as entry points into their unknown genome. We used six such regions to design primer pairs for long-range PCR amplification (amplicons >4 kbp), which gives us >95% of the entire genome as three large amplicons (Fig. 3). The sequence of each fragment is read as far as possible, then new sequencing primers are designed to "leap frog" further into the unknown sequence from the known.

Anarhichas genomes each comprise 16,543 bp; 449 SNP sites were identified in the genomes among one individual from each species. Wolffish species differ by 248–286 nucleotide substitutions, about one-half the difference among *Gadus* species. Patterns of intergenic SNP density in *Anarhichas* and *Gadus* genomes are significantly correlated, with some striking exceptions. The Control Region, characterized in many

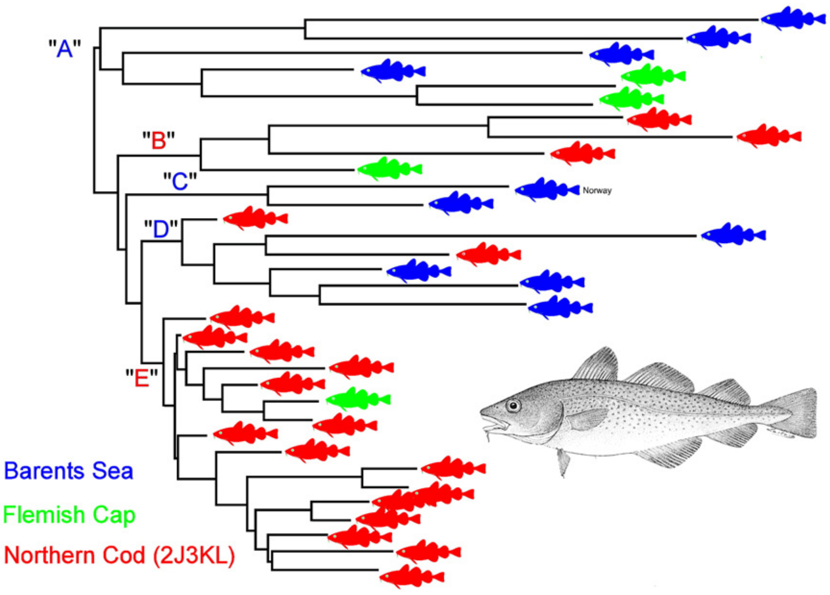


Fig. 2. Genomic phylogeography of Atlantic cod fishing areas. Codfish (*Gadus morhua*) from three different geographic locations fall into five recognizable clades (A–E). Most cod drawn from the Northern cod complex (NAFO 2J3KL) belong to clade E; cod from the Barents Sea are genetically more diverse (clades B, C, and D). Cod from Flemish Cap, an offshore seamount in the west Atlantic, are sometimes more closely related to fish from the Barents Sea than to fish on the adjacent Continental Shelf.

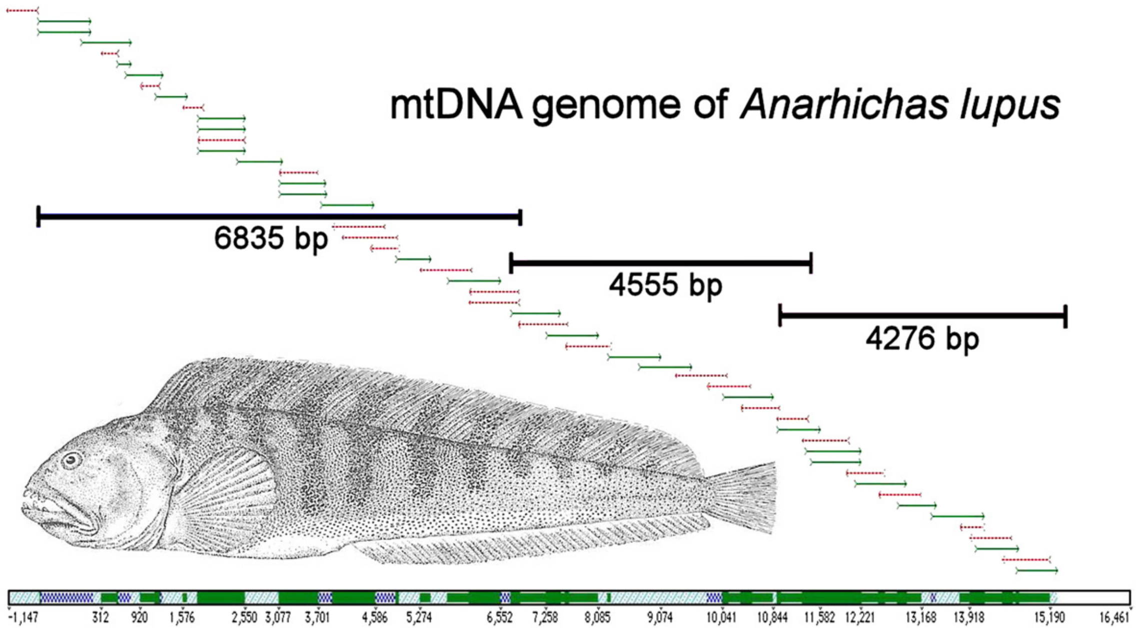


Fig. 3. Long-range PCR amplification and contig sequencing of wolffish mtDNA genomes. Complete mtDNA genomes can be amplified in a small number of overlapping 4–6 kbp segments, and sequenced with internal primers. The diagram shows overlapping forward- and reverse-strand sequences (red and green arrows, respectively) for the Northern wolffish (*Anarhichas lupus*).

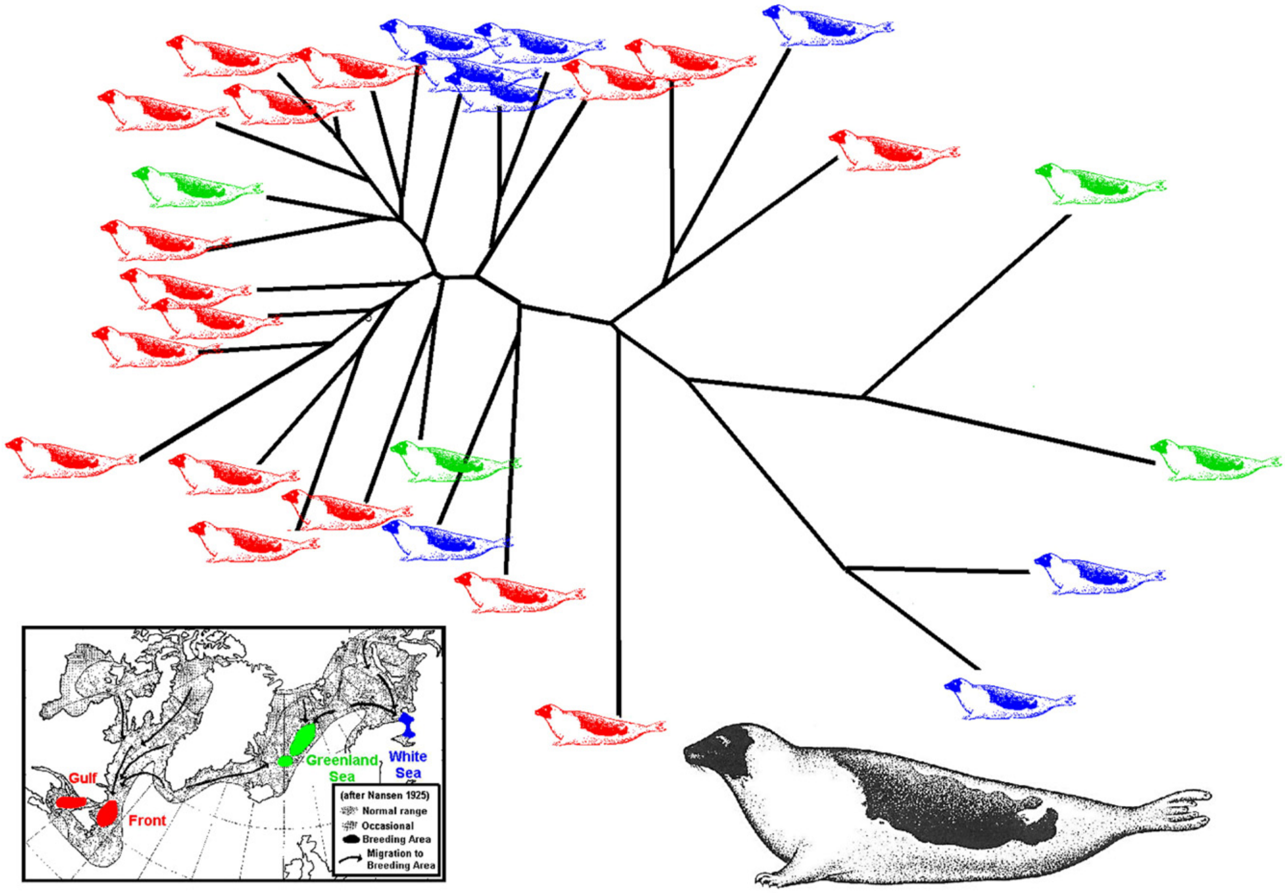


Fig. 4. Genomic phylogeography of harp seal whelping patches. Pupping and mating in harp seals (*Pagophilus groenlandicus*) is confined to three locations in the Northwest Atlantic, Greenland Sea, and White Sea. The oldest lineages are found in the Greenland and White Seas; seals in the western Atlantic have a more recent common ancestor.

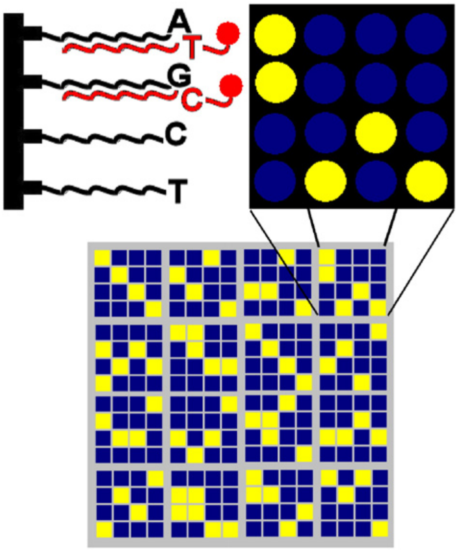


Fig. 5. DNA microarrays as Variant Detector Arrays. The example shows a set of DNA oligonucleotides that differ only at the last position, corresponding to a known SNP site in the genome. Fluorescently-tagged genomic DNA fragments anneal preferentially to those oligos with which they are perfectly complementary: in the example, an allele with a T SNP binds to the A oligo, and an allele with a C SNP binds to the G oligo. A computer reads the position of the two fluorescent tags and identifies the individual as a C/T heterozygote. Similarly, the single spots in the other three columns of the 4 × 4 VDA indicate that the individual is homozygous at the three corresponding SNP positions. The 4 × 4 array fits into one corner of a 256-oligo VDA chip for 64 SNPs (lower right); the current generation of chips includes more than 120,000 oligos (Fig. 7).

species as hypervariable (Faber and Stepien, 1997), was less variable than 10 of 13 protein-coding loci (24.5 SNPs/kbp). For population genetic analyses of wolffish, amplification by long-range PCR and sequence analysis of a contiguous block that spans the ND4–ND5–ND6–CYTB loci (6329 bp) are components of an efficient strategy for evaluating patterns of intra-specific DNA variability (Johnstone et al., 2007).

2.4. Genomic population structure of harp seal whelping patches

Analysis of breeding structure in fish is complicated by the diffuse distribution of spawning over a very wide geographic area. In contrast, breeding and whelping in harp seals (*Pagophilus groenlandicus*) is confined to three population aggregates associated with seasonal pack ice, one off Jan Mayen Island in the Greenland Sea, the second in the "Gorlo" (throat) of the White Sea, and the third in the Northwest Atlantic. The last comprises two sub-populations, one that whelps in the Gulf of St. Lawrence ("Gulf") and one on the southern Labrador/northern Newfoundland coastal shelf ("Ice Front"). Historical and contemporary hunting pressure on the eastern populations, and concerns about increasing population size of the western populations in connection with the decline of Atlantic cod (Stenson et al., 1993), raise questions about genetic intercommunication among populations.

Studies of a 0.4-kbp portion of the mitochondrial Cytochrome *b* locus identified a common mtDNA haplotype shared among all populations. Although the proportions of this haplotype differ significantly between populations in the western and eastern Atlantic (F_{ST} = 0.12), the single-locus data gave no evidence of phylogeographic structure (Perry et al., 2000). In contrast, comparison of coding-region mtDNA genome sequences among seals from each of the four whelping areas has identified several hundred SNPs. As with human and codfish populations, every individual seal has a unique mtDNA sequence (Fig. 4). As is observed in trans-Atlantic cod populations, there is evidence for the existence of deep ancestral clades confined to the eastern populations, and little or no differentiation and relatively close relationships within and between the western populations. The occurrence of individuals taken in the Greenland Sea within this "western" genotype clade suggests contemporary or historical migration from west to east (Marshall, Stenson, and Carr, work in progress).

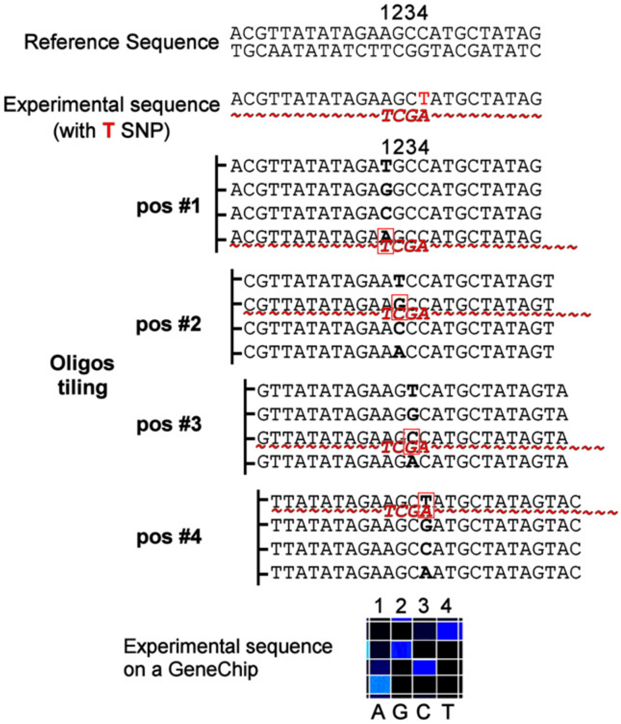

1234
Reference Sequence
ACGTTATATAGAAGCCATGCTATAG
TGCAATATATCTTCGGTACGATATC
Experimental sequence
(with T SNP)
ACGTTATATAGAAGCTATGCTATAG
~~~~~~~~~~~~TCGA~~~~~~~~~~
1234
pos #1
ACGTTATATAGATGCCATGCTATAG
ACGTTATATAGAGGCCATGCTATAG
ACGTTATATAGACGCCATGCTATAG
ACGTTATATAGAAGCCATGCTATAG
TCGA
pos #2
CGTTATATAGAATCCATGCTATAGT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

Fig. 6. Schematic representation of a DNA re-sequencing microarray experiment. A reference DNA sequence is represented in a series of overlapping ("tiled") oligonucleotide probes, each of length 25 bp. For each oligo, three variants are included that vary in the middle (13th) base, one for each of the three alternative code letters. In the example, four successive bases in the reference DNA sequence are **AGCC**: the four alternative oligos tiling the first position are (top to bottom) ***T*GCC**, ***G*GCC**, ***C*GCC**, and ***A*GCC**. The same arrangement occurs for oligos tiling the next three positions; the order of the variant bases in each set of oligos is constant (**T**, **G**, **C**, **A** = 1st, 2nd, 3rd, 4th rows). Consider an experimental DNA sequence with a SNP at the last position: **AGCT**. The sequence of the complementary strand (~~~**TCGA**~~~) is an exact match for only one of the four variant oligos at each tiling position. Mismatch at this position most strongly effects binding: the absolute degree of binding is measured at each oligo, and computer imaging of the microarray shows this as a more or less intense pseudocolour (bottom inset: see Fig. 8). In this case, preferential annealing to the 4th, 3rd, 2nd, and 1st oligos at four successive positions indicates that the original (complementary) experimental sequence is **AGCT**.
~~~~~~~~~~~~

3. New biotechnologies for biodiversity

3.1. Oligonucleotide arrays

The work described thus far was done by current automated methods of fluorescent dideoxy DNA sequencing. Although vastly more efficient than manual methods, it still remains tedious to set up large numbers of separate PCR and sequencing reactions, edit the data for each fragment separately, and finally assemble the separate fragments into contigs for each individual. An alternative approach for large-scale studies is to take advantage of DNA microarray technology. A DNA microarray or "chip" is a small piece of glass with a large number of synthetic oligonucleotides, either glued or grown onto it. A particular set of oligos can be used to interrogate a genome of an individual, for example with regard to its pattern of cDNA expression (Churchill, 2002), or as a Variant Detector Array (VDA) (Wang et al., 1998) to identify allelic variation at known SNP sites within populations (Fig. 5). A recent application extends the idea of a VDA to look at variation in every *potential* SNP site in a reference DNA: that is, the microarray will "re-sequence" complete homologous sequences in new individuals, and identify all SNP differences with respect to the reference DNA (Reider et al., 1998).

3.2. DNA re-sequencing "GeneChips"

The re-sequencing microarray represents a reference sequence of length *n* bases as a series of 4 × *n* overlapping ("tiled") oligonucleotide probes ("oligos") (Fig. 6). For each 25-base oligo, three variant oligos are included that vary in the middle

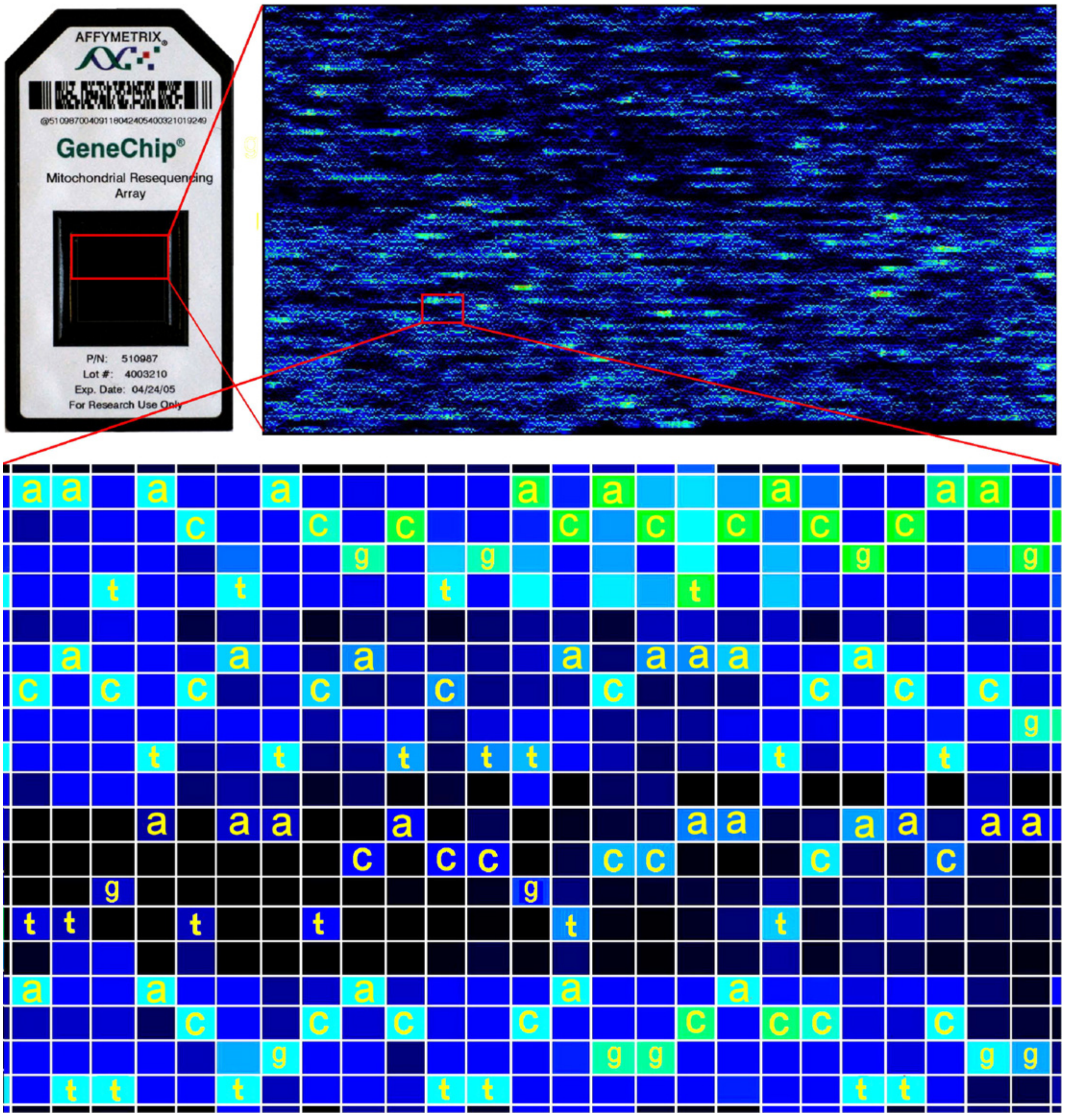


Fig. 7. Human mtDNA re-sequencing microarray. The microarray is a ~1 cm^2 chip set in a cassette that facilitates hybridization. The region shown tiles a reference sequence of 15,452 bases (not including the Control Region) in a 160 row × 488 column array, both for the sense and antisense strands, for a total of > 31 kb and > 123 K oligos. Each nucleotide position is represented in a vertical block of 4 cells in 5 rows (**A**, **C**, **G**, **T**, and a blank). In each block, the cell with the strongest relative intensity of DNA binding identifies the base present at that position. In the magnified view (19 rows × 25 columns), the sequence of bases in each of the four blocks is easily read as the left-to-right order of successive brightest pseudocolour squares. Variation in absolute intensity is influenced primarily by differing [**G**+**C**] ratios among oligos. Accuracy of base calling is determined by an algorithm that compares relative intensities among cells (Fig. 8) (Flynn and Carr, submitted for publication).

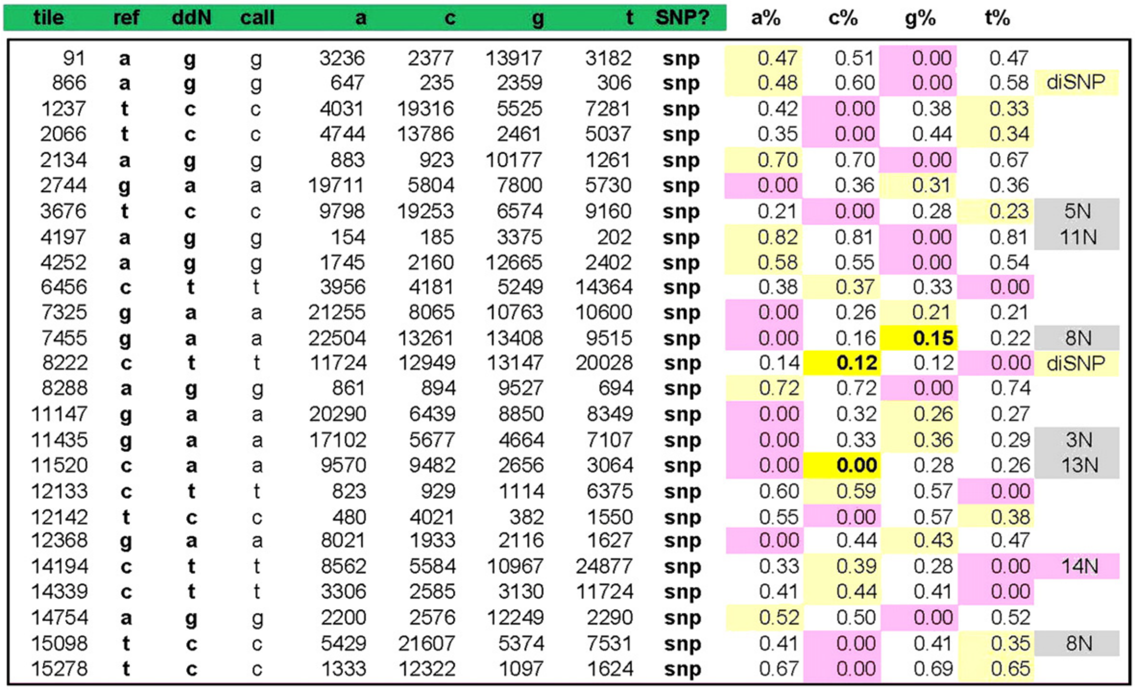
(13th) base, one for each of the three alternative DNA code letters. Mismatch at this position most strongly effects binding, so that an experimental genomic DNA fragment with a SNP variant corresponding to the 13th base will stick preferentially to only one of the four oligos at any tiled position. Fig. 7 shows a “GeneChip” microarray (8 Affymetrix) tiled with the sense and antisense (or heavy and light) strands of a reference human mtDNA sequence (15,452 bases each, not including the Control Region). Each nucleotide position is represented in a vertical block of 4 cells in 5 rows (A, C, G, T and a blank). In the re-sequencing experiment, PCR products that correspond to the complete mtDNA genome sequence are pooled in equimolar proportions, sheared, fluorescently labeled, and hybridized to the chip. Intensity of hybridization is read by a computer: in each block of four, the cell with the strongest relative intensity of DNA binding identifies the base present at that position. In the magnified view, the sequence of bases in each of the four blocks is easily read by the eye as the left-to-right order of successive brightest ‘spots.’ The inclusion of both sense and antisense strands allows each position to be read twice. [The re-sequencing microarray therefore resembles a classical dideoxy autoradiograph, turned on its side, in colour].

We have compared the efficiency and accuracy of re-sequenced human mtDNA genomes with those obtained previously by conventional automated sequencing (Fig. 8). The entire 15,452 bp sequence aligns perfectly with the reference sequence. A quality-control algorithm called correctly 15,211 of 15,452 bases (98.44% efficiency), including all 25 known SNPs (100.00% accuracy); no bases were called incorrectly. Of the remaining 241 positions initially called as ‘N’, 235 were called correctly as the cell with the greatest absolute intensity, where the difference in relative intensity was at least 13% greater than the next most intense cell. Six ‘Ns’ do not satisfy this criterion and remain uncalled (overall 99.96% efficiency).

3.3. Iterative DNA sequencing with a multi-species “ArkChip”

“Re-sequencing” as an approach to population genomics is more aptly termed “iterative sequencing,” to emphasize analysis of homologous genomes from multiple individuals within species. The cost of iterative sequencing of a complete mtDNA genome (US$300–400/microarray) is comparable to that of dideoxy sequencing (20–30 PCR templates per genome, sequenced in both directions at ca. US$5 each, = $200~$300), not counting labor costs. The limiting consideration in the execution of a population genomic study is the design cost for a new species-specific microarray (US$15–20,000), which is prohibitive for individual population biology studies. How can iterative genome sequencing of such species be accomplished?

The first generation of re-sequencing chips accommodated 30,000 nucleotides of reference sequence, enough for a single mtDNA genome. The current generation accommodates



tile	ref	ddN	call	a	c	g	t	SNP?	a%	c%	g%	t%	
91	**a**	**g**	g	3236	2377	13917	3182	**snp**	0.47	0.51	0.00	0.47	
866	**a**	**g**	g	647	235	2359	306	**snp**	0.48	0.60	0.00	0.58	diSNP
1237	**t**	**c**	c	4031	19316	5525	7281	**snp**	0.42	0.00	0.38	0.33	
2066	**t**	**c**	c	4744	13786	2461	5037	**snp**	0.35	0.00	0.44	0.34	
2134	**a**	**g**	g	883	923	10177	1261	**snp**	0.70	0.70	0.00	0.67	
2744	**g**	**a**	a	19711	5804	7800	5730	**snp**	0.00	0.36	0.31	0.36	
3676	**t**	**c**	c	9798	19253	6574	9160	**snp**	0.21	0.00	0.28	0.23	5N
4197	**a**	**g**	g	154	185	3375	202	**snp**	0.82	0.81	0.00	0.81	11N
4252	**a**	**g**	g	1745	2160	12665	2402	**snp**	0.58	0.55	0.00	0.54	
6456	**c**	**t**	t	3956	4181	5249	14364	**snp**	0.38	0.37	0.33	0.00	
7325	**g**	**a**	a	21255	8065	10763	10600	**snp**	0.00	0.26	0.21	0.21	
7455	**g**	**a**	a	22504	13261	13408	9515	**snp**	0.00	0.16	**0.15**	0.22	8N
8222	**c**	**t**	t	11724	12949	13147	20028	**snp**	0.14	**0.12**	0.12	0.00	diSNP
8288	**a**	**g**	g	861	894	9527	694	**snp**	0.72	0.72	0.00	0.74	
11147	**g**	**a**	a	20290	6439	8850	8349	**snp**	0.00	0.32	0.26	0.27	
11435	**g**	**a**	a	17102	5677	4664	7107	**snp**	0.00	0.33	0.36	0.29	3N
11520	**c**	**a**	a	9570	9482	2656	3064	**snp**	0.00	**0.00**	0.28	0.26	13N
12133	**c**	**t**	t	823	929	1114	6375	**snp**	0.60	0.59	0.57	0.00	
12142	**t**	**c**	c	480	4021	382	1550	**snp**	0.55	0.00	0.57	0.38	
12368	**g**	**a**	a	8021	1933	2116	1627	**snp**	0.00	0.44	0.43	0.47	
14194	**c**	**t**	t	8562	5584	10967	24877	**snp**	0.33	0.39	0.28	0.00	14N
14339	**c**	**t**	t	3306	2585	3130	11724	**snp**	0.41	0.44	0.41	0.00	
14754	**a**	**g**	g	2200	2576	12249	2290	**snp**	0.52	0.50	0.00	0.52	
15098	**t**	**c**	c	5429	21607	5374	7531	**snp**	0.41	0.00	0.41	0.35	8N
15278	**t**	**c**	c	1333	12322	1097	1624	**snp**	0.67	0.00	0.69	0.65	

Fig. 8. SNP detection in a human mtDNA re-sequencing experiment. The re-sequencing array result aligns perfectly with the dideoxy sequence, except for a tiling artifact that arises from an error in the published reference sequence. Dideoxy sequencing of the sense strand detects 25 SNPs between the experimental individual (“ddN”) and the reference sequence (“ref”): the re-sequencing chip calls all 25 (“call”) as the cell with the greatest absolute signal intensity (“a”–“t”), as indicated by the red highlight (“Δa”–“Δt”). The relative difference [(‘call’ signal − ‘ref’ signal)/(total signal)] of the SNP with respect to the expected reference base (yellow highlight) averages 43%, with a range of 12–82%. One anomaly (pos 11,520), where the expected SNP signal is <1% greater than that of the reference, is called unambiguously on the complementary strand of the same re-sequencing array (results not shown).

120,000 nucleotides, enough for seven complete homologous mtDNA genomes, from seven separate species. Is it possible to sequence two homologous genomes on the same microarray simultaneously, without interference? The complementary question is whether a microarray designed for the mtDNA genome of one species can accurately sequence the genome of a closely-related species. Flynn and Carr (in review) used the human-specific chip to measure the accuracy of SNP detection and efficiency of re-sequencing of the mtDNA genomes of chimpanzee (*Pan troglodytes*), gorilla (*G. gorilla*), and codfish (*G. morhua*) mtDNA genomes, which differ from that of humans by 8%, 10%, and >30%, respectively. We showed that differential binding of experimental DNAs to the microarray is strongly affected by the number of mismatches in the 25-bp interval spanned by each oligo. Where such intervals contain three or more mismatches, oligo-binding and sequencing efficiency declines log-linearly with respect to sequence divergence, and accuracy of SNP identification drops even more precipitously. Re-sequencing of the codfish genome recovers <4% of the sequence, in short blocks conserved with the human genome.

In the demonstrated absence of interfering cross-hybridization between species-specific oligos and experimental DNA from a distantly-related species, Fig. 9 shows the results of experiments on a 120Kbp multi-species microarray (the "ArkChip") tiled with the complete mtDNA genome sequences (including Control Regions) of seven species, including three fish (Atlantic cod, striped wolffish, and Atlantic salmon), three mammals

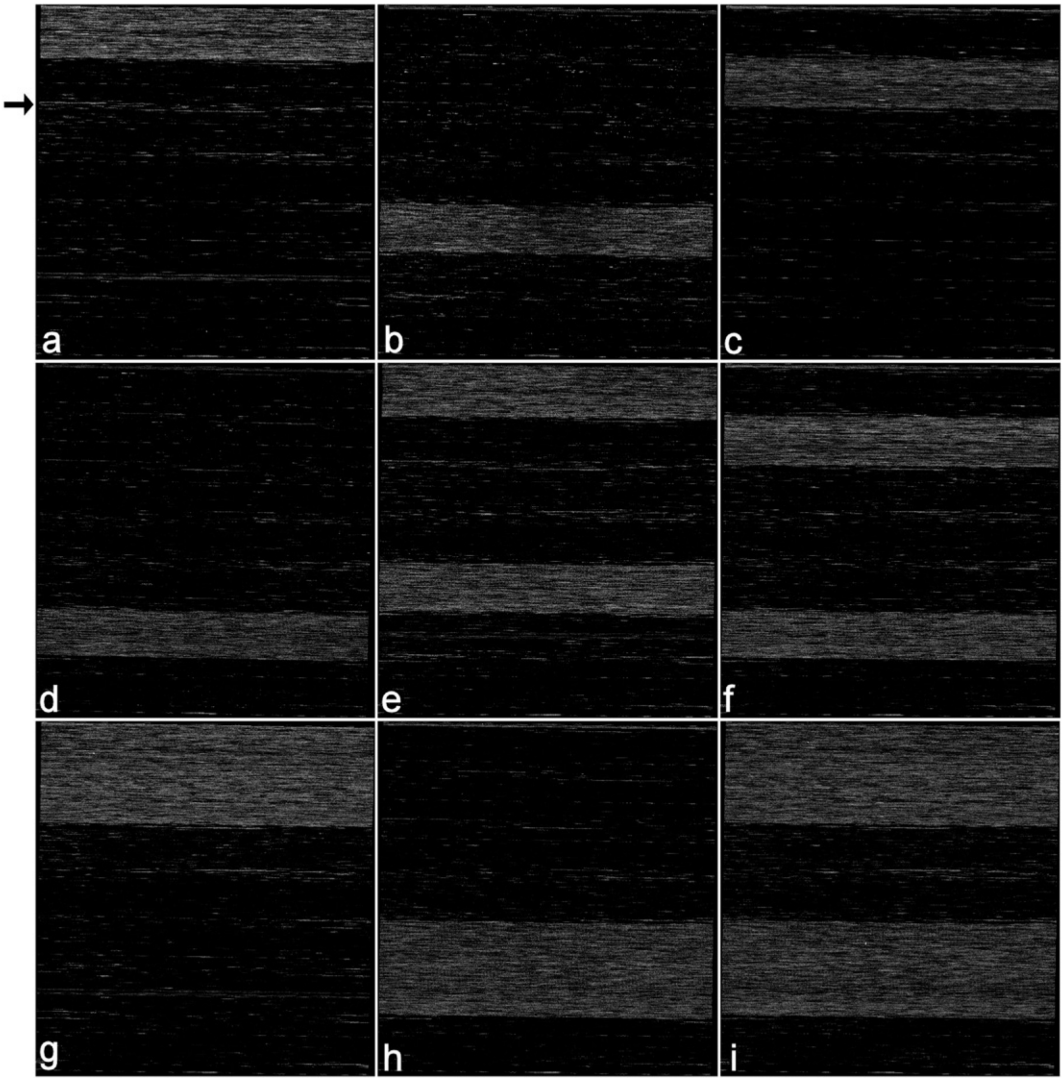


Fig. 9. Design of a multi-species iterative re-sequencing microarray — the "ArkChip". The design for a 120-kbp chip includes the sense and antisense strands of the complete mtDNA genome sequence of three fish, three mammals, and one bird species [Atlantic cod (*Gadus morhua*), Atlantic wolffish (*Anarhichas lupus*), Atlantic salmon (*Salmo salar*), harp seal (*Pagophilus groenlandicus*), Newfoundland caribou (*Rangifer tarandus*), human (*Homo sapiens*), and the blackish oystercatcher (*Haematopus ater*)]. These are tiled on the array in seven successive blocks of oligonucleotides. The nine panels show the results of four separate single-species experiments with mtDNA from cod, caribou, wolffish, and harp seal (blocks 1, 4, 2, & 6, respectively, in panels a–d, respectively), four pairwise expriments with cod / caribou, wolffish / seal, cod / wolffish, caribou / seal (panels e–h), and one with all four species (panel i). Note the species-specificity of mtDNA annealing in each experiment to the appropriate block(s). The arrow in panel (a) indicates a region of intermittent cross-hybridization to a conserved sequence tiled in another species, which occurs in other experiments as well.

(Newfoundland caribou, harp seal, and human), and a bird (blackish oystercatcher). Alignment of the mtDNA genomes of these species shows no blocks of 25 bp or greater that are identical in this region for any interordinal pair. Experiments with four species (two fish and two mammals) show that complete genome sequences are recoverable simultaneously, with efficiency and accuracy equal to those of single-species experiments (A. T. Duggan and S. M. Carr, work in progress) and the human-specific experiments described above.

The next generation of microarrays will accommodate 300 kbp of reference sequence, enough to hold ~20 separate species' mtDNA genomes. Complete reference mtDNA genomes are already available for 28 of COSEWIC's 53 marine Designatable Units (which include not only species but subspecies or geographic populations). By combining multiple species-recovery projects in a single, multiplex "ArkChip", the initial design costs and chip fabrication costs can be reduced as much as 20-fold, rendering the cost of a genomic population analyses comparable to that for a current single-locus project.

Acknowledgments

Development of the concepts and application of the ArkChip have been made possible by the support of numerous individuals and groups. Drs. Carr and Marshall wish in particular to acknowledge the major support for work on fisheries populations provided by the Department of Fisheries & Oceans, and we thank the following staff and scientists at the Northwest Atlantic Fisheries Science Centre for their support during the course of these studies: Mr. Bruce Atkinson, Mr. Julian Goodyear, Dr. Barry McCallum; Drs. Pierre Pepin & John Brattey (cod), Mr. Dave Kulka and Dr. Mark Simpson (wolffish); Dr. Garry Stenson (seals). Facilities for microarray experiments have been provided as part of an ongoing collaboration with Newfound Genomics: Ms. Siobhan Coady, CEO, and Ms. Lynette Peddle. We thank Affymetrix for permission to publish a modified version of one of their figures in Fig. 6. Experiments with caribou are supported by a grant from the Newfoundland Institute for Biodiversity and Ecosystem Science (IBES): Mr. Shane Mahoney. Experiments with the oystercatcher were suggested by Dr. A. J. Baker (Royal Ontario Museum). We thank Ms. Justyna Ciszewska-Carr for editorial assistance with the manuscript. S.M.C dedicates this paper to his father Paul on the occasion of his 82nd birthday.

References

Arnason, E., 2004. Mitochondrial cytochrome *B* DNA variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. Genetics 166, 1871–1885.

Avise, J.C., 2000. Phylogeography: The History and Formation of Species. Harvard University Press, Cambridge, MA.

Brown, W.M., George Jr., M., Wilson, A.C., 1979. Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. U. S. A. 76, 1967–1971.

Cann, R.L., Stoneking, M., Wilson, A.C., 1987. Mitochondrial DNA and human evolution. Nature 325, 31–36.

Carr, S.M., Crutcher, D.C., 1998. Population genetic structure in Atlantic Cod (*Gadus morhua*) from the North Atlantic and Barents Sea: contrasting or concordant patterns in mtDNA sequence and microsatellite data? In: Hunt von Herbing, I., Kornfield, I., Tupper, M., Wilson, J. (Eds.), The Implications of Localized Fishery Stocks. Northeast Regional Agricultural Engineering Service, Ithaca, NY, pp. 91–103.

Carr, S.M., Wroblewski, J.S., Snellen, A.J., Howse, K.A., 1995. Mitochondrial DNA sequence variation and genetic stock structure of Atlantic cod (*Gadus morhua*) from bay and offshore locations on the Newfoundland continental shelf. Mol. Ecol. 4, 79–88.

Carr, S.M., Kivlichan, D.S., Pepin, P., Crutcher, D.C., 1999. Molecular systematics of gadid fishes: implications for the biogeographic origins of Pacific species. Can. J. Zool. 77, 19–26.

Churchill, G.A., 2002. Fundamentals of experimental design for cDNA microarrays. Nat. Genet. 32, 490–495 (Suppl.).

Coulson, M., Marshall, H.D., Pepin, P., Carr, S.M., 2006. Mitochondrial genomics of gadid fish: implications for biogeographic origins and taxonomy. Genome 49, 1115–1130.

Faber, J.E., Stepien, C., 1997. The utility of mitochondrial DNA control region sequence for analyzing phylogenetic relationships among populations, species, and genera of the Percidae. In: Kocher, T.D., Stepien, C.A. (Eds.), Molecular Systematics of Fishes. Academic Press, New York, NY, pp. 129–143.

Flynn, S.M.C., Carr, S.M., in review. Species-specificity of SNP detection on DNA microarrays: efficiency and accuracy of resequencing of chimpanzee and gorilla mtDNA genomes on a human-specific MitoChip. BMC Genomics.

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

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Johnstone, K.A., Marshall, H.D., Carr, S.M., 2007. Biodiversity genomics for Species At Risk: patterns of DNA sequence variation within and among complete mitochondrial DNA genomes of three species of Wolffish (*Anarhichas* spp.). Can. J. Zool. 85, 151–158.

Mannion, J.J., 1977. The Peopling of Newfoundland: Essays in Historical Geography. Institute for Social and Economic Research, St. John's, Newfoundland.

Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A.G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M.D., Sukernik, R.I., Olckers, A., Wallace, D.C., 2003. Natural selection shaped regional mtDNA variation in humans. Proc. Natl. Acad. Sci. U. S. A. 100, 171–176.

Moore, S.J., Green, J.S., Fan, Y., Bhogal, A.K., Dicks, E., Fernandez, B.A., Stefanelli, M., Murphy, C., Cramer, B.C., Dean, J.C., Beales, P.L., Katsanis, N., Bassett, A.S., Davidson, W.S., Parfrey, P.S., 2005. Clinical and genetic epidemiology of Bardet–Biedl syndrome in Newfoundland: a 22-year prospective, population-based, cohort study. Am. J. Med. Genet. 132, 352–360.

Pepin, P., Carr, S.M., 1993. Morphological, meristic, and genetic analysis of stock structure in juvenile Atlantic Cod (*Gadus morhua*) from the Newfoundland Shelf. Can. J. Fish. Aquat. Sci. 50, 1924–1933.

Perry, E.A., Stenson, G.B., Bartlett, S.E., Davidson, W.S., Carr, S.M., 2000. DNA sequence analysis identifies genetically distinguishable populations of harp seals (*Pagophilus groenlandicus*) in the northwest and northeast Atlantic. Mar. Biol. 137, 53–58.

Pope, A.M., 2004. An investigation of the ethnic composition of the Newfoundland population based on whole mitochondrial genomes. B.Sc. (hons) thesis, Memorial University of Newfoundland, St. John's.

Reider, M.J., Taylor, S.L., Tobe, V.O., Nickerson, D.A., 1998. Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. Nucleic Acids Res. 26, 967–973.

Reidla, M., Kivisild, T., Metspalu, E., Kaldma, K., Tambets, K., Tolk, H.V., Parik, J., Loogvali, E.L., Derenko, M., Malyarchuk, B., Bermisheva, M., Zhadanov, S., Pennarun, E., Gubina, M., Golubenko, M., Damba, L., Fedorova, S., Gusar, V., Grechanina, E., Mikerezi, I., Moisan, J.P., Chaventre, A., Khusnutdinova, E., Osipova, L., Stepanov, V., Voevoda, M., Achilli, A., Rengo, C., Rickards, O., De Stefano, G.F., Papiha, S., Beckman, L., Janicijevic, B., Rudan, P., Anagnou, N., Michalodimitrakis, E., Koziel, S., Usanga, E., Geberhiwot, T., Herrnstadt, C., Howell, N., Torroni, A., Villems, R., 2003. Origin and diffusion of mtDNA haplogroup X. Am. J. Hum. Genet. 73, 1178–1190.

Richards, M.B., Macaulay, V.A., Bandelt, H.J., Sykes, B.C., 1998. Phylogeography of mitochondrial DNA in western Europe. Ann. Hum. Genet. 62, 241–260.

Stenson, G., Myers, R., Hammill, M., Ni, I.-H., Warren, W., Kingsley, M., 1993. Pup production of the harp seal, *Phoca groenlandica*, in the Northwest Atlantic. Can. J. Fish. Aquat. Sci. 50, 2429–2439.

Sykes, B.C., 2002. Daughters of Eve. W.W. Norton, New York, NY.

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.

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.

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.

Woods, M.O., Hyde, A.J., Curtis, F.K., Stuckless, S., Green, J.S., Pollett, A.F., Robb, J.D., Green, R.C., Croitoru, M.E., Careen, A., Chaulk, J.A., Jegathesan, J., McLaughlin, J.R., Gallinger, S.S., Younghusband, H.B., Bapat, B.V., Parfrey, P.S., 2005. High frequency of hereditary colorectal cancer in Newfoundland likely involves novel susceptibility genes. Clin. Cancer Res. 11, 6853–6861.