

Effects of Purifying and Adaptive Selection on Regional Variation in Human mtDNA

Eduardo Ruiz-Pesini,¹ Dan Mishmar,¹ Martin Brandon,^{1,2}
Vincent Procaccio,¹ Douglas C. Wallace^{1*}

A phylogenetic analysis of 1125 global human mitochondrial DNA (mtDNA) sequences permitted positioning of all nucleotide substitutions according to their order of occurrence. The relative frequency and amino acid conservation of internal branch replacement mutations was found to increase from tropical Africa to temperate Europe and arctic northeastern Siberia. Particularly highly conserved amino acid substitutions were found at the roots of multiple mtDNA lineages from higher latitudes. These same lineages correlate with increased propensity for energy deficiency diseases as well as longevity. Thus, specific mtDNA replacement mutations permitted our ancestors to adapt to more northern climates, and these same variants are influencing our health today.

The human mtDNA exhibits dramatic, region-specific sequence variation in indigenous populations (1) (fig. S1). Although previously this was attributed to genetic drift (2), an analysis of 104 complete human mtDNA sequences (haplotypes) from around the world suggested that the regional distribution of mtDNA haplogroups (specific lineages of related mtDNA haplotypes) has also been influenced by climatic selection (3).

Mitochondrial oxidative phosphorylation (OXPHOS) has two primary physiological functions: adenosine triphosphate (ATP) production and heat generation. The relative allocation of calories between these two energetic functions is determined by OXPHOS coupling: the efficiency of the electron transport chain (ETC) to convert the energy derived from oxidating dietary calories into the mitochondrial inner membrane proton gradient and the efficiency of the ATP synthase to convert the proton gradient into ATP. Tightly coupled OXPHOS would generate more ATP and be advantageous in the tropics, whereas partially uncoupled OXPHOS would produce proportionately more heat and be advantageous in cold climates (4). As expected, the basal metabolic rate of indigenous, circumpolar human populations is greater than that of temperate populations (5).

To further define the role of adaptive mutations in shaping human mtDNA variation, we collected and analyzed 1125 complete mtDNA coding region sequences.

These were assembled into a neighbor-joining phylogenetic tree (fig. S2), and all mtDNA variants were positioned within the tree. Variants common to multiple mtDNA haplotypes were ancient and positioned at internal branches (I), whereas those confined to one mtDNA were recent and relegated to terminal branches (T) (6).

Because deleterious replacement mutations would be eliminated from the population by purifying selection over time, they would be rare in the internal branches but more common at terminal branches of the tree. By contrast, advantageous replacement mutations retained by adaptive selection would be enriched in internal branches relative to terminal branches. Neutral mutations, such as synonymous (S) substitutions, should be uniformly distributed throughout the mtDNA tree. Hence, the frequency of replacement [nonsynonymous (NS)] mutations in different parts of the mtDNA tree can be normalized for varying time intervals by dividing by S [replacement mutations frequency (RF) = NS/S], which permitted the frequency of replacement mutations to be compared for internal branches (RF_I) and terminal branches (RF_T). Thus, lower RF_I values reflect fewer internal missense mutations, implying that most have been removed by purifying selection, whereas higher RF_I values indicate that more missense mutations have become established, indicating the influence of adaptive selection. Hence, a lower RF_I/RF_T (I/T) ratio indicates the predominance of purifying selection, whereas a higher ratio indicates an increasing importance of adaptive selection.

The phenotypic importance of replacement mutations was evaluated by analyzing the inter-specific conservation of the altered amino acids, the conservation indices

(CI). Neutral mutations would have a low CI, whereas deleterious or adaptive mutations would have a high CI. To determine the level of CI that would cause a phenotypic change, we examined the CI of 22 well-characterized human pathogenic mtDNA replacement mutations (www.mitomap.org; supporting online material). This gave a CI of $93 \pm 13\%$ (conserved in an average of 36.4 ± 5.2 species of 39 species compared $\times 100$). Then, we used two standard deviations from this mean (the 95% confidence limits) as the level of replacement mutations likely to have a phenotypic effect (supporting online material).

To determine the overall importance of selection in shaping overall human mtDNA variation, we calculated the RF_I and RF_T and the CI_I and CI_T for all mtDNA variants in the complete (global) mtDNA phylogenetic tree (Table 1). In aggregate, the RF_I was significantly lower than RF_T ($P < 0.0005$), and the CI_I was significantly lower than CI_T ($P < 0.0001$) (Table 1). Therefore, the predominant factor influencing human mtDNA variation throughout human history has been purifying selection, consistent with the essential function of the mtDNA proteins.

We then examined the potential phenotypic consequences of internal branch missense mutations using the 95% confidence limits of the pathogenic mutations as a reference. This revealed that 26% (63/240) of the replacement mutations are likely to be functionally important with a mean CI_I = $85 \pm 9\%$ ($33.2 \pm 3.6 \div 39$), whereas 74% are probably neutral with a mean CI_I = $23 \pm 15\%$ ($9.1 \pm 5.8 \div 39$).

To determine whether the relative effects of purifying versus adaptive selection correlate with climatic zone, we next examined the internal and terminal replacement mutation frequencies of the various regional macro-haplogroups (groups of related haplogroups). The tropic and sub-tropical African macro-haplogroup L mtDNAs had a significantly lower RF_I than did the temperate and arctic Eurasian macro-haplogroups M, N-R, and N-nonR [L versus M ($P = 0.06$), N(R) ($P = 0.013$), or N(NonR) ($P = 0.008$)] (Table 1). Likewise, the I/T ratio of L (0.70) was less than those of M, N(R), or N(nonR) (0.80 to 0.94), and the CI_I of L (36%) was consistently lower than those of M, N(R), or N(nonR) (40 to 44%) (Table 1). Hence, adaptive selection played an increasingly important role as people migrated out of Africa into temperate and arctic Eurasia.

To further clarify the role of arctic selection, we examined only those haplogroups found primarily in the arctic. Haplogroups A, C, D, and G are highly enriched in northeastern Siberia, comprising 75% of arctic versus 14% of temperate Asian mtDNAs. Haplogroups A, C, and D arrived first in northern

¹Center for Molecular and Mitochondrial Medicine and Genetics, University of California, Irvine, CA 92697-3940, USA. ²Information & Computer Science, Institute for Genomics & Bioinformatics, University of California, Irvine, CA 92697-3425, USA.

*To whom correspondence should be addressed. E-mail: dwallace@uci.edu

REPORTS

Table 1. Evolutionary parameters of the global human mtDNA phylogeny and each of the region-specific macro-lineages: Africa [L], Asia [M], and Eurasia [N(R) and N(nonR)]. N, RF_I, RF_T, CI_I, and CI_T indicate sample size, internal and terminal NS/S, and internal and terminal conservation index (supporting

online material), respectively. The *P* values were calculated using the Fisher's exact test (FET) and the Student's *t*-test. L-RF_I, RF_I-RF_T, and CI_I-CI_T indicate comparisons with the African L-RF_I, or within the cluster, respectively. SD, standard deviation.

Haplogroup	N	Internal branches		Terminal branches		<i>P</i> (RF _I -RF _T)	<i>P</i> (CI _I -CI _T)	<i>P</i> (L-RF _I)	I/T
		RF _I (NS/S)	CI _I (SD)	RF _T (NS/S)	CI _T (SD)				
Global	1125	0.41(240/589)	40(31)	0.53(688/1297)	52(34)	<0.0005	<0.0001		0.77
L	106	0.31 (60/191)	36(28)	0.44 (111/254)	54(33)	0.015	0.0002		0.70
M	118	0.42 (28/67)	43(31)	0.52 (90/172)	55(35)	0.071	0.093	0.061	0.80
N(R)	763	0.44(117/265)	40(31)	0.55 (397/725)	51(34)	0.012	0.002	0.013	0.80
N(nonR)	138	0.58 (31/53)	44(34)	0.62 (90/146)	49(33)	0.102	0.489	0.008	0.94

Siberia and thus were in a position to colonize the Americas when the Bering land bridge appeared. Haplogroup G arrived in Siberia after the bridge submerged. Haplogroup B joined A, C, and D in the Americas via an independent migration. Because B is absent in northern Siberia and rare in northwestern North America, it appears to have arrived by a more southerly route. Haplogroup X is found in the Great Lakes region of Central Canada, and is more prevalent in Europe than Asia. Hence, X has also persisted in the northern latitudes for prolonged periods (*I*).

On the basis of this biogeographic analysis, we would expect haplogroups A, C, D, and X to have been subjected to much greater cold stress than haplogroup B or African macro-haplogroup L. Accordingly, the I/T ratios of the arctic haplogroups A, C, D, and X (0.91 to 2.91) were all greater than B (0.75) or L (0.70), and the RF_I values of haplogroups A and X were significantly greater than that of L (*P* < 0.05) (Table 2). Moreover, the CI_I values of haplogroups A (53%) and C (73%) were much higher than B (31%) or L (36%), with the CI_I of C being the highest of any haplogroup (Table 2). Hence, the mtDNA variation of haplogroups A, C, D, and X has been strongly influenced by adaptive selection, whereas that of haplogroup B has not.

To obtain a more comprehensive assessment of the influence of adaptive selection on arctic mtDNA variation, we combined the data for haplogroups A, C, D, and X (designated ACDX) and compared these values with those of all other global mtDNA lineages (designated nonACDX) (Table 2). The RF_I of ACDX was significantly higher than that of nonACDX (*P* < 0.05) or of African L (*P* < 0.01) (Table 2). The CI_I of ACDX was also significantly higher than that of nonACDX or L (*P* ≤ 0.05).

To identify specific adaptive mutations associated with arctic haplogroups A, C, D, and X, we next examined the CI of each of the mutations that lie at the roots of these haplogroups. Haplogroup A was found to have two very well conserved root amino acid variants: one in the mtDNA protein

Table 2. Evolutionary parameters of haplogroups that migrated to the Americas via the arctic (A, C, D, X) and nonarctic (B) routes. L is the African L lineages, ACDX is the combination of the arctic adapted mtDNA haplogroups, and NonACDX is the combination of all other mtDNAs. Abbreviations as in Table 1.

Haplogroup	N	Internal branches		Terminal branches		<i>P</i> (L-RF _I)	I/T
		RF _I (NS/S)	CI _I (SD)	RF _T (NS/S)	CI _T (SD)		
A	33	0.73 (8/11)	53(27)	0.72 (34/47)	45(33)	0.048	1.01
B	31	0.38 (9/24)	31(25)	0.51 (35/69)	54(33)	0.151	0.75
C	22	0.46 (6/13)	73(27)	0.42 (11/26)	76(27)	0.155	1.10
D	63	0.50 (9/18)	42(26)	0.55 (32/58)	64(33)	0.099	0.91
X	19	1.25 (5/4)	36(30)	0.43 (10/23)	46(34)	0.037	2.91
L	106	0.31 (60/191)	36(28)	0.44 (111/254)	54(33)		0.70
ACDX	137	0.61 (28/46)	51(31)	0.56 (87/154)	56(35)	0.008	1.09
NonACDX	988	0.39(212/543)	39(30)	0.53(601/1143)	51(34)		0.74

Table 3. Evolutionary parameters of haplogroups that are indigenous to Europe. V(HV*) encompasses V and non-H HV; I+N1b is a combination of the sister lineages I and N1b. Abbreviations as in Table 1.

Haplogroup	N	Internal branches		Terminal branches		<i>P</i> (L-RF _I)	I/T
		RF _I (NS/S)	CI _I (SD)	RF _T (NS/S)	CI _T (SD)		
H	314	0.48 (29/60)	46(35)	0.61(147/242)	50(34)	0.031	0.79
V(HV*)	55	1.33 (4/3)	32(11)	0.46 (16/35)	60(35)	0.055	2.89
U(Uk)	174	0.36(38/105)	36(29)	0.51 (95/185)	53(34)	0.080	0.71
J	101	0.66 (23/35)	42(31)	0.65 (50/77)	53(35)	0.007	1.02
T	80	0.31 (9/29)	52(42)	0.39 (31/79)	51(35)	0.162	0.79
W	48	0.50 (7/14)	36(36)	0.30 (7/23)	60(40)	0.124	1.67
I+N1b	32	0.63 (10/16)	50(39)	0.72 (28/39)	59(31)	0.050	0.88
X	19	1.25 (5/4)	36(30)	0.43 (10/23)	46(34)	0.037	2.91
L	106	0.31(60/191)	36(28)	0.44(111/254)	50(34)		0.70

ND2 at nucleotide (nt) 4824G, resulting in T119A (CI = 82.1%), and the other in ATP6 at nt 8794T, resulting in H90Y (CI = 72%) (7). The Siberian sub-haplogroup C also contained two conserved root variants: ND4 at nt 11969A (A404T) (CI = 85%) and cytochrome b (cytb) at nt 15204C (I153T) (CI = 85%).

European mtDNAs would also be expected to have been influenced by cold selection because of the episodic periods of cold associated with the repeated continental glaciations (8). Accordingly, RF_I values of European haplogroups H, I + N1b, J, and X were all significantly higher than that of African L [L versus H (*P* < 0.05), I + N1b (*P* = 0.05), J (*P* < 0.01), and X (*P* < 0.05)] (Table 3). The CI_I of haplogroups H, J, and I + N1b

were also higher than African L (Table 3).

European haplogroups J and T provide striking examples of the various effects that adaptive selection has had on European mtDNA variation (Fig. 1). J and T are sister haplogroups that share a common root but diverge early with very different subsequent evolutionary histories. Consequently, they have significantly different RF_I and RF_T values (*P* < 0.05), as well as significantly different overall replacement mutation frequencies (*P* = 0.005) (Table 3).

Haplogroup T has a low RF_I, but this is because it encompasses the most highly conserved internal branch mutation in the ND2 gene observed in the human mtDNA phylogeny. This mutation is in mtDNA protein ND2 at nt 4917G (N150D) (CI = 90%) (Table 3).

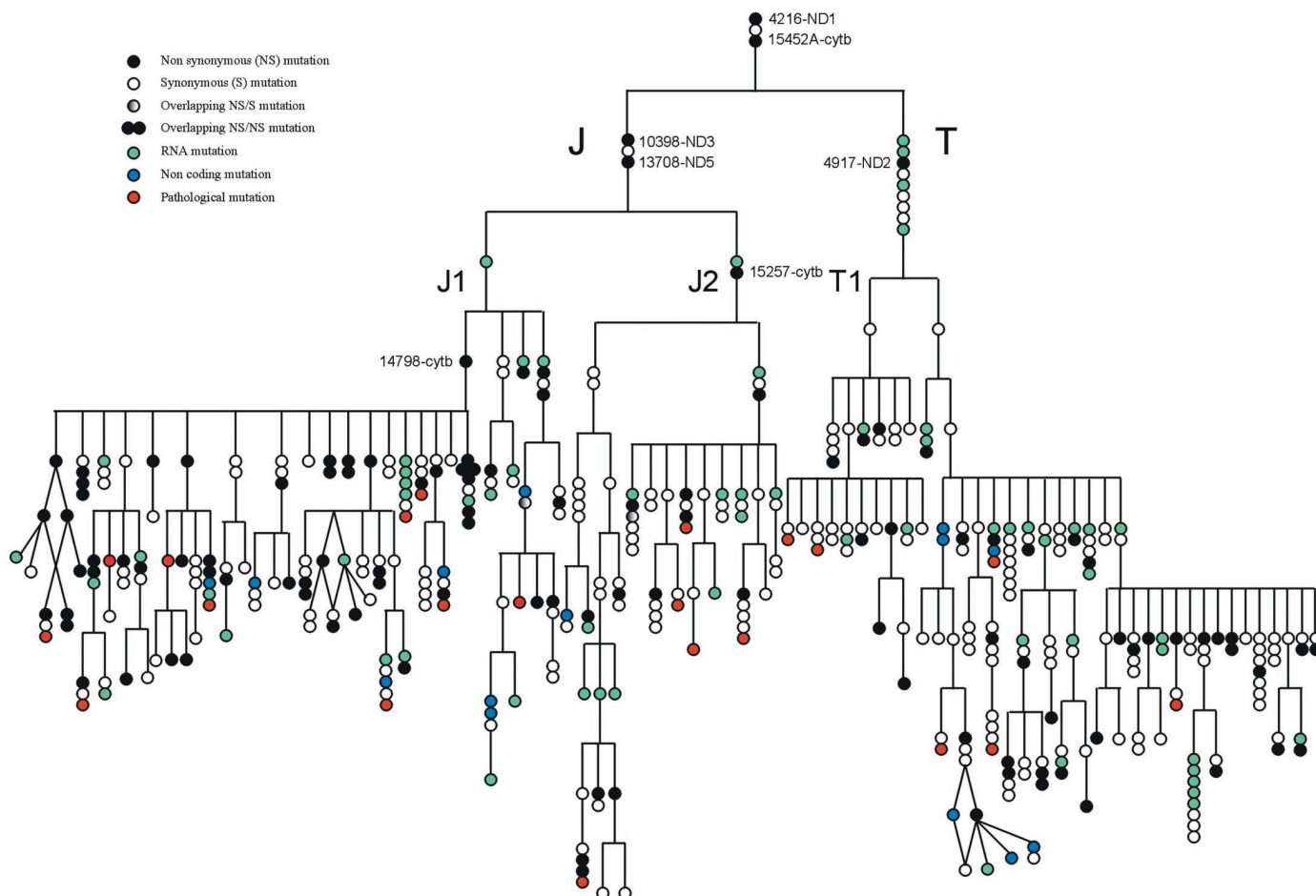


Fig. 1. Phylogeny of the haplogroups J and T. Key internal replacement mutations are designated by the gene name and the nucleotide substitution.

The stem of haplogroup T is additionally unique in encompassing eight other mutations, four that are synonymous and four in RNA genes (Fig. 1). This implies that much of the early diversification of haplogroup T was lost, presumably due to selection, with only the ND2 (N150D) lineage surviving.

In contrast, haplogroup J diverged from T, initially acquiring two additional complex I mutations. It then bifurcated into two sub-haplogroups J1 and J2, each defined by a root cytb replacement mutation (Fig. 1). The J2 cytb mutation is at nt 15257A (D171N) (CI = 95%), whereas the J1 cytb mutation is at nt 14798C (F18L) (CI = 77%). The nt 14798C mutation is also found at the base of haplogroup K (more recently classified Uk).

The nt 15257A (D171N) substitution is located in the outer coenzyme Q_{10} (CoQ)-binding site of cytb (Qo), where it contacts the Rieske iron-sulfur protein (ISP) (9, 10). This variant is likely to be functionally important, because the equivalent mutation (D187N) in *Rhodobacter sphaeroides* alters the photosynthetic apparatus (11). The nt 14798C (F18L) mutation is located at the inner CoQ-binding (Qi) site, within 3.5 Å of the CoQ quinone group (12). Mutations in

this amino acid confer susceptibility to diuron, which blocks electron flow between cytb and cytochrome c1 (13). Because the CoQ binding sites of cytb are thought to be essential for the complex III Q cycle, these mutations could reduce proton pumping and thus coupling efficiency.

The functional relevance of the haplogroup J, T, and K amino acid variants has been shown in human clinical studies. Haplogroup T is associated with moderate asthenozoospermia and a complex I + IV defect (14) as well as with Wolfram syndrome (15). It is also protective against Alzheimer's disease (16). Haplogroup J has been shown to increase the penetrance of the milder complex I gene mutations associated with Leber's hereditary optic neuropathy (LHON) (15, 17, 18), and haplogroups J and K have been linked to increased susceptibility to multiple sclerosis (19). Moreover, haplogroup K is protective against Alzheimer's disease (20), and J and K are protective against Parkinson disease (21) and are associated with increased longevity (22–26).

This combination of the increased predilection to energy deficiency diseases, but protection from neurodegenerative diseases and

aging is consistent with the expectations for mtDNA coupling efficiency mutations. Uncoupling mutations would reduce ATP production, increasing the probability of energetic failure. However, they would also decrease mitochondrial ROS production by increasing the oxidation of the electron transport chain, thus reducing oxidative damage and apoptosis. This could decrease neuronal and other cell loss, thus increasing longevity (27).

Our observations support the hypothesis that certain ancient mtDNA variants permitted humans to adapt to colder climates, resulting in the regional enrichment of specific mtDNA lineages (haplogroups). Today these same variants result in differences in energy metabolism and altered mitochondrial oxidative damage, thus affecting health and longevity. Therefore, to understand individual predisposition to modern diseases, we must also understand our genetic past, the goal of the new discipline of evolutionary medicine.

References and Notes

1. D. C. Wallace, M. D. Brown, M. T. Lott, *Gene* **238**, 211 (1999).
2. M. J. Johnson, D. C. Wallace, S. D. Ferris, M. C. Rattazzi, L. L. Cavalli-Sforza, *J. Mol. Evol.* **19**, 255 (1983).

REPORTS

3. D. Mishmar *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 171 (2003).
4. M. D. Brand, *Exp. Gerontol.* **35**, 811 (2000).
5. W. R. Leonard *et al.*, *Am. J. Hum. Biol.* **14**, 609 (2002).
6. A. R. Templeton, *Genetics* **144**, 1263 (1996).
7. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
8. S. H. Ambrose, *J. Hum. Evol.* **34**, 623 (1998).
9. N. Fisher, B. Meunier, *Eur. J. Biochem.* **268**, 1155 (2001).
10. X. Gao *et al.*, *Biochemistry* **41**, 11692 (2002).
11. V. P. Shinkarev, N. B. Ugulava, E. Takahashi, A. R. Crofts, C. A. Wraight, *Biochemistry* **39**, 14232 (2000).
12. N. Fisher, P. R. Rich, *J. Mol. Biol.* **296**, 1153 (2000).
13. J. P. di Rago, A. M. Colson, *J. Biol. Chem.* **263**, 12564 (1988).
14. E. Ruiz-Pesini *et al.*, *Am. J. Human Genet.* **67**, 682 (2000).
15. S. Hofmann *et al.*, *Genomics* **39**, 8 (1997).
16. P. Chagnon *et al.*, *Am. J. Med. Genet.* **85**, 20 (1999).
17. M. D. Brown, F. Sun, D. C. Wallace, *Am. J. Human Genet.* **60**, 381 (1997).
18. M. D. Brown *et al.*, *Hum. Genet.* **110**, 130 (2002).
19. B. Kalman, F. D. Lublin, H. Alder, *J. Neurol. Sci.* **140**, 75 (1996).
20. G. Carrieri *et al.*, *Hum. Genet.* **108**, 194 (2001).
21. J. M. van der Walt *et al.*, *Am. J. Hum. Genet.* **72**, 804 (2003).
22. R. Ivanova, V. Lepage, D. Charron, F. Schachter, *Gerontology* **44**, 349 (1998).
23. G. De Benedictis *et al.*, *FASEB J.* **13**, 1532 (1999).
24. O. A. Ross *et al.*, *Exp. Gerontol.* **36**, 1161 (2001).
25. A. K. Niemi *et al.*, *Hum. Genet.* **112**, 29 (2003).
26. M. Bonafe *et al.*, *Gene* **286**, 121 (2002).
27. P. E. Coskun, E. Ruiz-Pesini, D. C. Wallace, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 2174 (2003).
28. Supported by NIH grants NS21328, AG13154, NS37167, and HL64017, as well as an Ellison Medical Foundation Senior Investigator Award awarded to D.C.W.

Supporting Online Material

www.sciencemag.org/cgi/content/full/303/5655/223/DC1

Materials and Methods

Figs. S1 and S2

Table S1

References

25 June 2003; accepted 18 November 2003

Global Assessment of Organic Contaminants in Farmed Salmon

Ronald A. Hites,^{1*} Jeffery A. Foran,² David O. Carpenter,³ M. Coreen Hamilton,⁴ Barbara A. Knuth,⁵ Steven J. Schwager⁶

The annual global production of farmed salmon has increased by a factor of 40 during the past two decades. Salmon from farms in northern Europe, North America, and Chile are now available widely year-round at relatively low prices. Salmon farms have been criticized for their ecological effects, but the potential human health risks of farmed salmon consumption have not been examined rigorously. Having analyzed over 2 metric tons of farmed and wild salmon from around the world for organochlorine contaminants, we show that concentrations of these contaminants are significantly higher in farmed salmon than in wild. European-raised salmon have significantly greater contaminant loads than those raised in North and South America, indicating the need for further investigation into the sources of contamination. Risk analysis indicates that consumption of farmed Atlantic salmon may pose health risks that detract from the beneficial effects of fish consumption.

Between 1987 and 1999, salmon consumption increased annually at a rate of 14% in the European Union and 23% in the United States (1). Currently, over half the salmon sold globally is farm-raised in Northern Europe, Chile, Canada, and the United States, and the annual global production of farmed salmon (predominantly Atlantic salmon, *Salmo salar*) has risen from ~24,000 to over 1 million metric tons during the past two decades (2). The health benefits of eating fish such as salmon have been well documented (3, 4). However, salmon are relatively fatty carnivorous fish that feed high in the food web, and as such, they bioaccumulate con-

taminants (5). The potential risks of eating contaminated farmed salmon have not been well evaluated. Three previous studies reporting contaminants in salmon are inconclusive because of their very small sample sizes and narrow geographic representation (6–8). As a result, the extent of this problem and the potential risks to human health remain unclear.

We measured organochlorine contaminants in approximately 700 farmed and wild salmon (totaling ~2 metric tons) collected from around the world. We do not report on other important contaminants, such as methylmercury, because our preliminary study (9) showed no significant difference in methylmercury levels between farmed and wild salmon. Using the data on organochlorine contaminants, we assessed the variation in contaminant loads between farmed and wild salmon and among geographic regions, and we calculated the human health risks of salmon consumption. Farmed Atlantic salmon from eight major producing regions in the Northern and Southern hemispheres were purchased from wholesalers that could obtain fish of the appropriate size within the sam-

pling period; in addition, farmed Atlantic salmon fillets were purchased at supermarkets in 16 large cities in North America and Europe. For comparison, samples of five wild species of Pacific salmon [*chum* (*Oncorhynchus keta*), coho (*O. kisutch*), chinook (*O. tshawytscha*), pink (*O. gorbuscha*), and sockeye (*O. nerka*)] were obtained from three different geographic regions. Wild Atlantic salmon were not studied because few are available commercially; nor did we analyze farmed Pacific salmon because they are not raised in any substantial amounts (2, 10).

A total of 594 individual whole salmon were purchased from wholesalers and filleted; an additional 144 fillets were purchased from retailers in Boston, Chicago, Denver, Edinburgh, Frankfurt, London, Los Angeles, New Orleans, New York, Oslo, Paris, San Francisco, Seattle, Toronto, Vancouver, and Washington, DC. Composites of fillets from whole salmon were made on the basis of the location where they were produced (farmed salmon) or purchased (wild salmon). Composites of fillets from retailers were made on the basis of the retail outlet where they were purchased. Each composite sample consisted of fillets from three salmon per location or three fillets per retail outlet, giving 246 measurable samples. All samples were homogenized and analyzed by gas chromatographic high-resolution mass spectrometry (11). Strict quality assurance and quality control procedures were followed (11). Thirteen samples of salmon feed were purchased from the European, North American, and South American outlets of the two major fish feed companies, which together have ~80% of the global market for fish feed (12), and were analyzed as above.

Contaminant concentrations in farmed and wild salmon were compared by analysis of variance. In comparing wild and farmed salmon, farmed salmon were considered as a single group. In addition, locations at which salmon were farmed were compared by analysis of variance with multiple comparisons of means to test for differences among locations in contaminant levels. In all analyses of vari-

¹School of Public and Environmental Affairs, Indiana University, Bloomington, IN 47405, USA. ²Citizens for a Better Environment, Milwaukee, WI 53202, USA. ³Institute for Health and the Environment, University at Albany, Rensselaer, NY 12144, USA. ⁴AXYS Analytical Services, Post Office Box 2219, 2045 Mills Road, Sidney, British Columbia, Canada V8L 3S8. ⁵Department of Natural Resources, ⁶Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA.

*To whom correspondence should be addressed. E-mail: HitesR@Indiana.edu