Ciba Foundation Symposium 197

VARIATION IN THE HUMAN GENOME

1996

JOHN WILEY & SONS

Chichester · New York · Brisbane · Toronto · Singapore

VARIATION IN THE HUMAN GENOME

The Ciba Foundation is an international scientific and educational charity (Registered Charity No. 313574). It was established in 1947 by the Swiss chemical and pharmaceutical company of CIBA Limited—now Ciba-Geigy Limited. The Foundation operates independently in London under English trust law.

The Ciba Foundation exists to promote international cooperation in biological, medical and chemical research. It organizes about eight international multidisciplinary symposia each year on topics that seem ready for discussion by a small group of research workers. The papers and discussions are published in the Ciba Foundation symposium series. The Foundation also holds many shorter meetings (not published), organized by the Foundation itself or by outside scientific organizations. The staff always welcome suggestions for future meetings.

The Foundation's house at 41 Portland Place, London W1N 4BN, provides facilities for meetings of all kinds. Its Media Resource Service supplies information to journalists on all scientific and technological topics. The library, open five days a week to any graduate in science or medicine, also provides information on scientific meetings throughout the world and answers general enquiries on biomedical and chemical subjects. Scientists from any part of the world may stay in the house during working visits to London.

Ciba Foundation Symposium 197

VARIATION IN THE HUMAN GENOME

1996

JOHN WILEY & SONS

Chichester · New York · Brisbane · Toronto · Singapore

© Ciba Foundation 1996

Published in 1996 by John Wiley & Sons Ltd
Baffins Lane, Chichester
West Sussex PO19 1UD, England

Telephone National (01243) 779777
International (+44) (1243) 779777

All rights reserved.

No part of this book may be reproduced by any means,
or transmitted, or translated into a machine language
without the written permission of the publisher.

Other Wiley Editorial Offices

John Wiley & Sons, Inc., 605 Third Avenue,
New York, NY 10158-0012, USA

Jacaranda Wiley Ltd, G.P.O. Box 859, Brisbane,
Queensland 4001, Australia

John Wiley & Sons (Canada) Ltd, 22 Worcester Road,
Rexdale, Ontario M9W 1L1, Canada

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01,
Jin Xing Distripark, Singapore 0512

Suggested series entry for library catalogues:
Ciba Foundation Symposia

Ciba Foundation Symposium 197
x + 329 pages, 45 figures, 19 tables

Library of Congress Cataloging-in-Publication Data

Variation in the human genome / [editors, Derek Chadwick and Gail Cardew.
p. cm.—(Ciba Foundation symposium ; 197)
Symposium on Variation in the Human Genome, held at the Ciba Foundation, London, 15 June 1995.
ISBN 0 471 96152 3 (alk. paper)
1. Human population genetics—Congresses. 2. Human genome—Congresses. 3. Human genetics—Variation—Congresses.
I. Chadwick, Derek. II. Cardew, Gail. III. Symposium on Variation in the Human Genome (1995 : Ciba Foundation) IV. Series.
QH455.V37 1996
573.2'1—dc20 95-54159
CIP

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0 471 96152 3

Typeset in 10/12pt Times by Dobbie Typesetting Limited, Tavistock, Devon.
Printed and bound in Great Britain by Biddles Ltd, Guildford.
This book is printed on acid-free paper responsibly manufactured from sustainable forestation, for which at least two trees are planted for each one used for paper production.

Contents

Symposium on Variation in the human genome, held at the Ciba Foundation, London 13–15 June 1995

This symposium is based on a proposal made by K. M. Weiss and R. H. Ward

Editors: Derek Chadwick (Organizer) and Gail Cardew

Participants

J. Armour Department of Genetics, University of Leicester, Adrian Building, University Road, Leicester LE1 7RH, UK

A. Balmain CRC Beatson Laboratories, Department of Medical Oncology, Alexander Stone Building, University of Glasgow, Garscube Estate, Switchback Road, Bearsden, Glasgow G61 1BD, UK

P. Beighton Department of Human Genetics, University of Cape Town Medical School, Observatory 7925, Cape Town, South Africa

J. Bertranpetit Laboratorío d'Antropología, Facultat de Biología, Universitat de Barcelona, Av Diagonal 645, E–08028 Barcelona, Catalonia, Spain

J. Bodmer Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London WC2A 3PX, UK

Sir W. F. Bodmer Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London WC2A 3PX, UK

A. M. Bowcock Department of Pediatrics, University Of Texas, Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75235-8591, USA

A. Cao Istituto di Clinica e Biologia dell'Età Evolutiva, Università degli Studi di Cagliari, Via Jenner s/n, I-09121 Cagliari, Italy

R. Chakraborty Human Genetics Center, School of Public Health, University of Texas, PO Box 20334, Houston, TX 77225, USA

A. Chakravarti Department of Genetics, Case Western Reserve University, School of Medicine, BRB Rm 721, 10900 Euclid Avenue, Cleveland, OH 44106-4955, USA

A. Clark Department of Biology, Penn State University, 208 Mueller Building, University Park, PA 16802, USA

P. Donnelly Departments of Statistics, and Ecology and Evolution, University of Chicago, 5734 University Avenue, Chicago, IL 60637, USA

J. H. Edwards Genetics Laboratory, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK

N. B. Freimer Neurogenetics Laboratory and Center for Neurobiology and Psychiatry, Department of Psychiatry and Programs in Genetics and Biomedical Sciences, University of California, San Francisco, CA 94143–0984, USA

D. L. Hartl Department of Organismic & Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA

N. Hardman Ciba Pharmaceuticals, Wimblehurst Road, Horsham, West Sussex, RH12 4AB, UK

P. S. Harper Institute of Medical Genetics, University of Wales College of Medicine, Heath Park, Cardiff CF4 4XN, UK

H. H. Hobbs Departments of Internal Medicine and Molecular Genetics, University of Texas, Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75235, USA

K. K. Kidd Department of Genetics, Yale University School of Medicine, 333 Cedar Street, PO Box 3333, New Haven, CT 06510–8005, USA

G. M. Lathrop The Wellcome Trust Centre for Human Genetics, University of Oxford, Windmill Road, Oxford OX3 7BN, UK

A. R. Linares *(Bursar)* Departamento de Bioquimica, Facultad de Medicina, Universidad de Antioquia, AA 1226, Medellín, Colombia

C. R. Scriver McGill University–Montreal Children's Hospital Research Institute, 2300 Tupper Street, A-717, Montreal, Quebec H3H 1P3, Canada

C. F. Sing Department of Human Genetics, School of Medicine, University of Michigan, Medical Sciences II M4708, Ann Arbor, MI 48109–0618, USA

G. R. Sutherland Department of Cytogenetics and Molecular Genetics, Centre for Medical Genetics, Women's and Children's Hospital, North Adelaide, SA 5006, Australia

A. R. Templeton Department of Biology, Washington University, St Louis, MO 63130-4899, USA

R. H. Ward Department of Human Genetics, 2100 Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA

Sir D. J. Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK

K. M. Weiss *(Chairman)* Department of Anthropology, The Pennsylvania State University, 409 Carpenter Building, University Park, PA 16802–3404, USA

R. Zechner Institute of Medical Biochemistry, University of Graz, Harrachgasse 21, A-8010 Graz, Austria

Other Ciba Foundation Symposia:

No. 130 **Molecular approaches to human polygenic disease**
Chairman: Sir David Weatherall
1987 ISBN 0 471 91096 1

No. 149 **Human genetic information: Science, law and ethics**
Chairman: Sir Gustav Nossal
1990 ISBN 0 471 92624 8

No. 194 **Genetics of criminal and antisocial behavour**
Chairman: Sir Michael Rutter
1996 ISBN 0 471 95719 4

Introduction

Kenneth M. Weiss

Department of Anthropology, The Pennsylvania State University, 409 Carpenter Building, University Park, PA 16802–3404, USA

London is an appropriate place for a symposium in this field because it's essentially the home of the idea of evolution. During the last century, London was at the centre of a debate about the interpretation of similarities among animal species. Most of the evidence was based on morphology because there was no biochemistry in those days, and there was basically an essentialist or platonic philosophy about structural archetypes among different animals. The problem was to explain the origin and the reality of those archetypes, i.e. whether they were actually present in animals or whether they were just ideals. These issues were addressed by Darwin, who, through the study of variation, proposed the theories of evolution, phylogeny and common descent. These theories did not answer all the questions, but they showed, at least in general terms, how the archetypal concepts could be interpreted in terms of ancestry.

The Human Genome Project, which will create a stereotype of human genetic structure, is in a sense history's greatest exercise in platonic essentialism. It is a stereotype that neither Plato nor Linnaeus would recognize because it is a composite, made up of pieces of chromosomes from different people. It is intended to represent the normal human genome, although there is no guarantee that those who donated their chromosomes for analysis will find out subsequently that they are susceptible to a particular disease. Although this is a stereotype, it is probably one of the most important single projects in biology ever undertaken, and it has already proven to be immensely important.

Data that have been generated in association with the Human Genome Project, together with our understanding of evolution, show that the genome did not arise from a 'great chain of stereotypes'. Genetic variation has been fundamental, not incidental, to the evolutionary generation of our genome. In this introduction I would like to outline a few relevant ideas about evolution and some principles that may be discussed at this symposium.

First, we heavily use the concept of phylogeny, i.e. that present variation in the genome is due to descent with modification. This has now been shown clearly by the results of DNA sequencing.

Second, Victorian biologists were aware of modularity and common body plans in nature. They advanced hypotheses about the central role of

anatomical segmentation in animal evolution (they did not know about modularity in physiological systems). Basic segmental body plans were considered to represent the essence of animal organization, and it was debated whether major advances in evolution occurred by saltations or by large, rapid changes involving these forms. These notions were laughed at for much of this century, but the laughing is now starting to subside because we're realizing that in some ways their ideas about modularity were correct. We now know that the genome itself is modular, from nucleotides and codons to gene families and other higher order structures, and that duplication of whole genes (or clusters of genes) with subsequent modification provides the material for complexity. We now use these concepts routinely in the process of trying to understand complex traits at the level of genetic and molecular physiology.

Third, we are becoming more aware of interactions both between genes, and between genes and their environment. These interactions result in the production of particular somatic phenotypes, but of course only genetic factors are contained within the germline and are, therefore, inherited. The distinction between the 'information' inherited in the germline and the traits that are realized by that information in an individual lifetime is an important distinction that can be traced back to the nineteenth century. Today it takes many forms, including the nature versus nurture debate regarding causation of chronic diseases that involve both genes and environmental exposures. Recent discoveries have shown that more is inherited than just the DNA sequence itself, but the long-term implications of those findings are not yet clear.

Fourth, I would like to mention that evolution has proceeded by a crude form of empirical 'sieving'. That is, natural selection operates on phenotypes, not genotypes, and it accepts any genotype whose phenotype passes the competitive standard. Wallace and Darwin debated whether this standard was, in general, established by the environment (Wallace) or by competition among individuals (Darwin). The former was probably most important for the majority of traits of interest to contemporary biomedical genetics. The critical point is that any genotype that generated an acceptable phenotype was acceptable to selection. There was also undoubtedly a strong component of luck (or genetic drift) in the process. As we shall see in this symposium, one result of evolution by phenotypic rather than genotypic selection is that the genetic basis of simple as well as complex traits is variable: different individuals can have a similar disease for a diverse set of genetic reasons.

A number of generalizations have arisen as a consequence of DNA analysis. All of us here are aware of them, implicitly or explicitly, but they have not always been built explicitly into our models of variation. We do not have a good systematic understanding of the meaning of some of these generalizations, but there are at least some general principles. (1) There are many alleles at any given locus, not just two. In the past we have conceptually thought of most loci as being diallelic but we now know that there are

hundreds of alleles at a given locus. (2) There is a quantitative relationship between genotype and phenotype, even at single loci. Simple concepts such as dominance and recessiveness are becoming obsolete in many ways because we now see that there is a more or less continuous relationship between various alleles at even a single locus and the phenotypes associated with them. This was not fully anticipated until we started looking at DNA. (3) The polygenic model, which originated in the last century, described complex quantitative traits as aggregates of individually unidentifiable genetic components traditionally referred to as polygenes. However, it is only in the last few years that we have been able to identify the quantitative trait loci pertaining to the variation in these traits. We don't yet know how to interpret most of these data. However, at some of these loci what we find are one or two alleles with strong effect and many with minor effect on phenotypic variation. In this sense, modern genetics has unified the previously considered disparate ways in which qualitative and quantitative phenotypes were produced at the gene level.

Most mutations are unique at the haplotype DNA level. These mutations generate cladistic sequence patterns among copies of the given gene in a population, and they retain a strong trace of history. I would say these patterns reflect a type of 'weak law of nature'. There is not the precise relationship between alleles and phenotypes that is generally evoked by ideas of genetic adaptation to environments. Instead, from the point of view of DNA sequences, the genotype to phenotype relationship is rather forgiving or statistically noisy. However, the cladistic structure of DNA sequences themselves retains a reasonably strong trace of population history, and it seems likely that that history, rather than any deterministic force such as adaptive natural selection, is responsible for most of the pattern that we see today. We will see in this symposium how this fact can be used to increase our power to detect the genotype to phenotype signal that exists in any given system.

Genetic identity by state is now usually interpreted as being roughly equivalent to identity by descent, which is different from what I was taught when I was a student. Identical DNA sequences are typically assumed to be descended from a common ancestral chromosome with that sequence (important caveats are needed for some regions of the genome in which real recurrent mutation seems to occur, for various chemical reasons). As a result, when we see different individuals in the same population with a similar disease, we can say that the diseases are clonal, in the sense that they're caused by copies of an allele which can be traced back to a common ancestor.

Alleles with a strong effect on the risk of disease are usually rare. For most alleles, their effect on risk is modest, complex and fairly uncertain. From a public health perspective, the effects of these modest alleles may be more important; however, the alleles with strong effects are the ones that are easily studied by standard scientific methods.

We had not anticipated that the germline genotype is so dynamic from generation to generation. The situation is more complicated than the simple genetic beads on a string model. This has added a new kind of richness to our understanding of genetics. Somatic genotypes are dynamic during the life of an individual, and the germline is also dynamic across generations in complex ways.

The discovery of regulatory genetic elements has also altered what we know about genes and the original beads on a string model of genetics. Short response elements recognized by transcription factors to switch a gene on or off in appropriate tissues may act together as a separate kind of non-coding gene, which may evolve separately from its neighbouring coding sequence. Phenotypic changes may result from mutations in the response elements or in the coding sequence itself.

In our general model of evolution, genotypes are generated by a random process of mutation. Natural selection provides a sieving mechanism on phenotypes only, as mentioned earlier, and the important point again is that any genotype whose associated phenotype can get through that sieve will be an acceptable genotype. In that sense, the phenotypic variation that we see among individuals was produced by a process that went from phenotype indirectly down to genotype. But in applied biomedicine, we're trying to identify genotypes that predict phenotypes accurately. So are we trying to do something that isn't what Nature did to produce us? Nature was only interested in the fact that you can get here, which represents a very different perspective. In this sense, human genetics turns Nature on its head and tries to make a causal connection, from genotype to phenotype, that was not rigorously built into the system of variation as it arose. We refer to the struggles this entails by using terms such as 'complexity', and we will see what this means to biomedicine in many presentations at this symposium.

A new field that is being called 'evolutionary medicine' is generating a lot of interest, at least in the USA. The idea is that a diversity of human traits, including host–pathogen relationships, allergic reactions, anxiety, menstruation and fever, must have had their origin in adaptive evolution by natural section. There is a highly (I would say 'hyper') deterministic approach to human phenotypes, which views pathological variation as something that should be approached with an understanding of adaptive origins so that therapy does not violate the built-in function of the system. For example, if fever is an adaptive response to infection, it should be interfered with only with caution. Closed explanations are appealing, and no one can deny that dysfunction can be understood best in the context of normal function. There will probably be fervid advocacy of this point of view in the near future.

However, not all of life, nor all of disease, reflects a tightly deterministic natural world. This symposium will consider another, more problematic, aspect of evolutionary medicine. The probabilistic role of population history in generating the pattern of genetic variation associated with disease, and the

statistical relationship between specific genotypes and specific phenotypes, pose challenges for a field that hopes to identify the specific causes of human disorders. Genetic variants, even frequent ones, need have no adaptive meaning or 'explanation'. The contingent nature of biological variation and its phenotypic relationships, the essential product of much of evolution at the gene level, is not always so tidy as the adaptationist perspective would suggest.

This symposium is organized into several topical categories, designed to address the issues I have discussed in a systematic way. The first group of presentations will address molecular variation in human populations and its evolution in general as a topic; the second group will address variation in the Mendelian diseases that should be simplest to understand genetically; the third group will address genetic variation and complex causation for traits that we know involve many genes; the fourth group will consider evolutionary principles and methods for aetiological inference that can take advantage of the historical origin of existing variations; and finally, David Weatherall will give us an overview that illustrates most of the ideas contained in the symposium in terms of an elegant example.

Phylogeographic variability in traditional societies

Ryk H. Ward* and Diana Valencia†

*Department of Human Genetics, 2100 Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA and †Departamento de Biologia, Universidad de Puerto Rico, San Juan, PR 00931, Puerto Rico

Abstract. Our perception of the amount, and distribution, of human genetic diversity is becoming radically altered by the introduction of sophisticated molecular techniques into the field of evolutionary biology. Along with the exponential increase in the number of informative DNA markers, has come an increased precision in estimating the evolutionary relationships between populations. Evaluating DNA variability in terms of the phylogenetic analysis of sequence variability at the population level has been especially informative in illuminating the long-term ancestry of our species. An extension of this strategy, phylogeography, aims to evaluate the evolutionary ancestry of specific genomic regions in terms of the geographic distribution of phylogenetic lineages within and among populations. We have started to apply this approach by studying the distribution of mitochondrial DNA sequence diversity within and among a variety of Amerindian tribes. These data provide an illuminating contrast to regional data on sequence variability, especially when analysed within the theoretical framework of the coalescent. To interpret these results, we have analysed a simple model in which the rate of coalescence between subpopulations varies with respect to the rate of coalescence of individual ancestry. The model indicates that extensively isolated subpopulations will have distinct distributions of ancestry, and hence of sequence variability, compared to subpopulations which arise due to a series of rapid fissioning events. Subpopulations within tribes, such as bands, appear to correspond to the latter category, whereas tribal populations appear intermediate between the two extremes.

1996 Variation in the human genome. Wiley, Chichester (Ciba Foundation Symposium 197) p6–24

Evolutionary biology has experienced two major revolutions in the past 30 years: each sparked by advances in laboratory techniques. The first was the ability to apply gel electrophoresis routinely to identify protein variability, leading to direct estimates of genetic heterozygosity in natural populations (Harris 1966, Lewontin & Hubby 1966). This innovation opened new vistas in population genetics and provided a wealth of empirical data to displace

theoretical speculation. Evolutionary biology benefited from unexpected insights and the development of new research directions. However, some of the presumed benefits of the new technology proved surprisingly elusive and data acquisition was sometimes driven more by rote than by hypothesis testing (Lewontin 1991). The application of molecular techniques to allow direct assessment of genomic variability at the level of DNA represents the second innovation. Although still in its infancy, this new technology promises to dwarf completely the impact of protein electrophoresis. With the ability to resolve genetic variability at the level of a single nucleotide, evolutionary biology is poised to embark on an explosive renaissance. However, Lewontin's (1991) retrospective assessment of protein electrophoresis suggests that it will be difficult to predict the direction in which this new research might develop. Despite some obvious applications, it is likely that many fundamental questions will require a paradigm shift before data can be properly collected and analysed.

With the advent of the Human Genome Project, the rate of data acquisition and the development of new techniques will have a profound impact on the strategies that are developed to assess the evolutionary ancestry of our species (Cavalli-Sforza 1990). As new types of data are described, long-held assumptions about the origin of human diversity will be challenged. This will have relevance for both the individual and society: a comprehensive evolutionary description of genomic variability not only defines ancestry but also illuminates the genetic potential for important phenotypes, such as disease.

A primary theme of this symposium is to examine ways in which molecular strategies can provide insights about the evolutionary heritage of individuals and populations, and how that heritage translates into disease susceptibility and other phenotypes with societal relevance. This paper evaluates the distribution of molecular data at the level of traditional human communities, corresponding to what is loosely called a 'tribe'. Results that are beginning to emerge from the study of small-scale traditional populations promise to provide a perspective that will help interpret data derived from large-scale communities. The distribution of molecular variability in tribal populations helps indicate how methods, such as molecular phylogeography, can be used to interpret variability at a higher level.

Phylogeography and ancestral coalescence

Although molecular techniques allow a huge increase in the number of DNA markers (Cavalli-Sforza 1990), the potential to construct genealogies for specific genomic regions is even more important. Application of phylogenetic analysis provides important clues about evolution in terms of how ancestry coalesces backwards in time. Gene genealogies, singly or collectively, can lead

to powerful inferences about the evolutionary history of populations. Consequently, the analysis of the spatial and community-specific distribution of a set of genomic phylogenies has led to the development of a new paradigm in evolutionary biology: phylogeography (Avise et al 1987).

Following the construction of intraspecific molecular phylogenies, a phylogeographic analysis allows the distribution of gene genealogies to be traced in space and time. Assessing the coalescence of lineage ancestry within and among populations provides information about the origin of contemporary genetic variation. In particular, the distribution of genomic phylogenies can be used to draw conclusions about the relative impact of deterministic forces compared to the influence of drift. Phylogeographic analyses can also be used to investigate the influence of past demographic fluctuations (Slatkin & Hudson 1991, Rogers & Harpending 1992). However, within this relatively novel area, the intersection between theory and observation still needs considerable development. In this context, application of phylogeography to small-scale, traditional human populations can help to identify some of the pressing problems that need resolution.

However, not all types of molecular variability are equally informative for phylogenetic reconstruction. Although biallelic DNA polymorphisms, such as restriction fragment length polymorphisms (RFLPs), represent an obvious extension of classical genetic markers, they suffer from the same constraints: individual loci provide relatively little information about gene genealogies. Consequently, such loci are relevant to modern population geneticists only because of their vast number. Classical genetic markers have already proved exceptionally informative about population affiliations on a global scale (Cavalli-Sforza et al 1994). Hence, the incorporation of many more markers will give a substantial increase in both precision and discrimination, as demonstrated by the use of RFLP data to estimate the relative impact of selection and migration on the genetic composition of major ethnic groups (Bowcock et al 1991). Information content increases with the number of alleles, so that highly variable loci, such as minisatellites and microsatellites, hold even more promise. With only 30 microsatellite loci it is possible to define informative phylogenies of genetic kinship between individuals (Bowcock et al 1994), which would be difficult to duplicate with biallelic loci.

The critical requirement for estimating a phylogeny for a specific genomic region is that mutational events must be distinguishable. Although this is well-nigh impossible for microsatellites, the development of the minisatellite variant repeat mapping strategy (Jeffreys et al 1991) holds promise for minisatellites. This technique could be exceptionally useful for defining phylogenies for otherwise uninformative regions, such as the Y chromosome. The ancestral state of mobile elements (e.g. *Alu* elements) can be defined; therefore, these loci hold considerable promise for developing population phylogenies (Batzer et al 1994). However, individual elements are relatively uninformative for a single

genomic region. The incorporation of multiple loci into haplotypes also holds promise, but more data are needed on the relative rate of recombination versus mutation. Ultimately, the *sine qua non* for phylogenetic reconstruction is sequence data. Although the vast bulk of phylogeographic studies are based on mitochondrial DNA sequences, there is a growing body of sequence data for nuclear regions (Fullerton et al 1994). Consequently, there is hope that population phylogenies will soon be based on a number of independent gene genealogies, which will help overcome the intrinsic problem of trying to infer population history from only a single realization of evolution (Slatkin & Hudson 1991, Majoram & Donnelly 1994).

Mitochondrial DNA variability in Amerindian tribes

The phylogenetic analysis of mitochondrial DNA has had a major impact on contemporary theories of human evolution. High resolution sequence analysis of the mitochondrial DNA control region suggested a relatively recent origin for anatomically modern *Homo sapiens* and a surprisingly rapid dispersal from Africa with little if any matrilineal contribution by the human groups that previously occupied large tracts of Eurasia (Vigilant et al 1991). The unimodal distribution of sequence differences and occurrence of a star-like phylogeny suggested that this early migration was associated with a considerable population expansion (Di Rienzo & Wilson 1991). These observations led to methods for estimating the timing and magnitude of past demographic expansions (Rogers & Harpending 1992), with the conclusion that the major demographic expansion of ancestral human populations occurred 80 000 to 30 000 years ago, depending on the ethnic group (Harpending et al 1994).

Despite the numerous data on regional populations, few studies have evaluated sequence variability within small-scale, traditional societies. This is unfortunate, as these populations approximate the ancestral breeding structure that characterized much of our species' recent evolution. Our initial study of mitochondrial DNA sequence variation in the Nuu-Chah-Nulth, an Amerindian tribe of the Pacific Northwest, identified a high level of intratribal sequence variability (Ward et al 1991). The analysis of 63 maternally unrelated Nuu-Chah-Nulth revealed 26 variable positions in a 360 nucleotide stretch of the mitochondrial DNA control region. These variable positions defined 28 mitochondrial lineages. The average sequence diversity among these 28 Nuu-Chah-Nulth lineages was 80% of the value observed in a sample of 62 Japanese, and 60% of the values observed in a sample of 94 sub-Saharan Africans. These data indicated that, as was true for classical markers (Neel & Ward 1970), an appreciable proportion of human molecular variability is contained within tribal populations.

As indicated in Fig. 1, the majority of Nuu-Chah-Nulth mitochondrial DNA lineages fell into four phylogenetic clusters. The existence of these clusters,

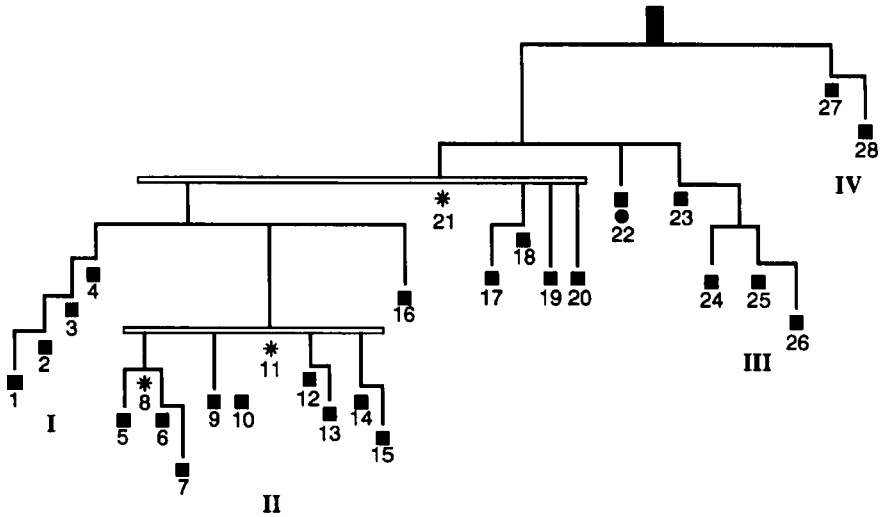

FIG. 1. Phylogeny of 28 mitochondrial DNA lineages found in a sample of 63 Nuu-Chah-Nulth, after Ward et al (1991). The branch lengths in the phylogeny are scaled proportionally to sequence divergence, with the total depth of the tree being 3% sequence divergence. Open boxes denote multifurcations, where the hierarchical branching order cannot be assigned with statistical confidence. Lineages are numbered as in Ward et al (1991), and the three starred lineages are found in tribes throughout the Pacific Northwest. The four major mitochondrial DNA Amerindian clades are indicated with roman numerals.

which appear to predate the colonization of the New World, accounts for a significant fraction of the molecular diversity within the tribal sample. Apart from implying that the ancestral population which entered the Americas contained elements of formerly isolated populations (Ward et al 1991), the existence of these clusters can be used to aid the phylogeographic analysis of mitochondrial DNA variation in Amerindian populations. However, in order to interpret the resulting data from a more intensive analysis, an assessment is needed of the relative importance of the temporal stability of local populations and also of the effect of different levels of genomic variation.

The influence of population coalescence

Like molecular lineages, populations also have an evolutionary history. Population history influences the distribution of coalescent events for the individual genetic lineages; therefore, it will also influence the cladistic structure of molecular phylogenies. This will affect the phylogeographic pattern of molecular lineages within and among extant populations. Demographic change is perhaps the most obvious factor, as phylogenies derived from

populations that had experienced a marked expansion tend to be star shaped with densely bifurcating tips. Such phylogenies result in a strongly unimodal distribution of pairwise sequence differences (Majoram & Donnelly 1994, Rogers & Harpending 1992, Slatkin & Hudson 1991).

Most human populations also display marked extensive population structure. Although characteristic of many natural populations, the cultural dimension of our species tends to magnify the development of population substructure. Moreover, for most tribal populations, the existence of subpopulations is an extremely dynamic process. Local demes (bands, villages etc.) have relatively short histories and tend to evolve through a dynamic process of population interactions that are largely dominated by the sociopolitical relationships between groups. For many tribal populations, this dynamic process can best be described as a 'fission–fusion' process whereby larger populations have a tendency to split into smaller units, with periodic accretion of small subunits into larger ones. From the perspective of a small number of generations, this fission–fusion process can have a marked influence on the magnitude and structure of intratribal genetic differentiation (Ward & Neel 1970, Ward 1972).

The few investigations of the influence of population subdivision on coalescence times (Majoram & Donnelly 1994) have tended to use Wright's paradigm of an island in which subpopulations are fixed, invariant entities. Although instructive, these studies may be less relevant to the human situation than studies which focus on the relationship between coalescence times of individual lineages and population coalescence which involves the coalescence of entire collections of individual lineages. Accretion of small populations into larger ones is a special case of migration that is already incorporated into many existing models (Marjoram & Donnelly 1994, Slatkin & Hudson 1991).

When population coalescence occurs over a much longer time-scale than lineage coalescence, widely divergent clades characterize each subpopulation (Fig. 2). Within each subpopulation, the distribution of coalescence times approximates the standard model and will depend on population size (Kingman 1982, Tavaré 1984). By contrast, the coalescence time between lineages from different subpopulations will be much longer, resulting in phylogeographic patterns that fail to reflect the true population history. Thus, lineages 6–12, represented by the dashed lines in Fig. 2, have been lost, as have lineages 18–20. Hence, population isolation tends to exaggerate the normal process of lineages loss, with consequent inflation of the pairwise sequence differences. The converse situation is obtained when population coalescence occurs much more rapidly than lineage coalescence (Fig. 3). Here, representations of the same lineage tend to be distributed across closely related subpopulations, resulting in extensive lineage sharing. Related populations will only differ in the frequency of their lineages rather than having characteristic lineages.

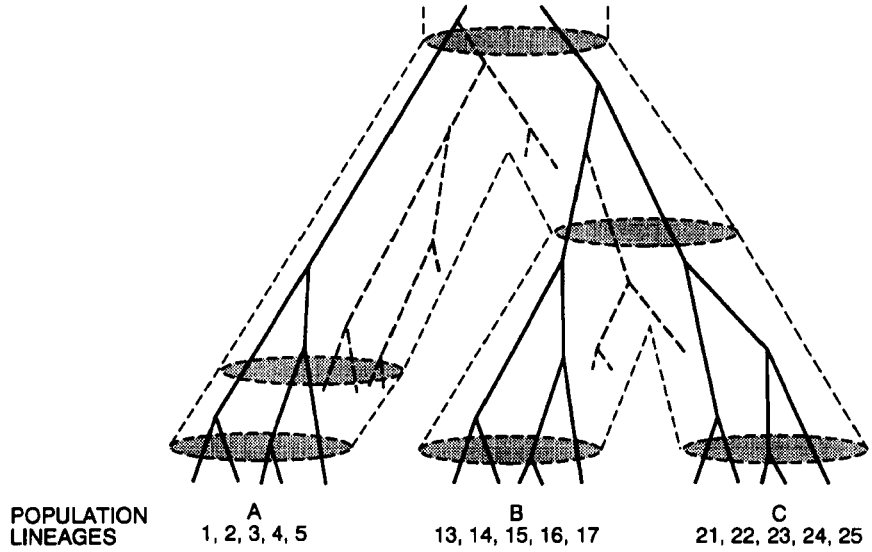

FIG. 2. Representation of 15 molecular lineages distributed in a set of three subpopulations that coalesce to population ancestry at a substantially slower rate than lineage coalescence. This corresponds to populations that have been isolated for long periods of time. Dashed lines indicate the ten molecular lineages that have been lost by random drift.

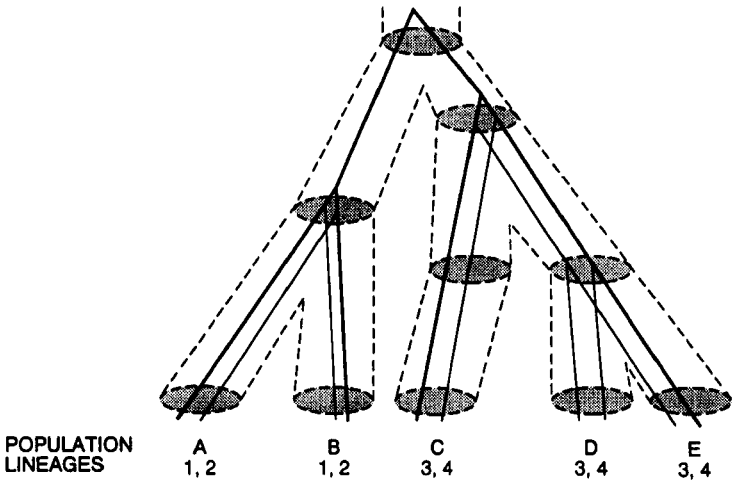

FIG. 3. Representation of four molecular lineages distributed in a set of four subpopulations that coalesce to population ancestry at a much faster rate than lineage coalescence. Dashed lines indicate molecular lineages that have been lost.

TABLE 1 Influence of relative rate of population coalescence on distribution of sequence diversity among subpopulations

Relative rate of population coalescence	*Lineage sharing between populations 1 & 3*	*Lineage sharing between populations 2 & 3*	*Proportion of sequences differing by 0*	*Proportion of sequences differing by 3*	*Number of variable positions*
0.01	0.3	0.1	13.4	5.2	129
0.10	2.3	1.0	13.1	5.7	65
1.00	19.1	10.4	15.6	9.5	24
10.00	63.7	47.3	24.6	10.7	15
100.00	87.0	75.1	29.9	9.4	13

To evaluate the influence of changing rates of population coalescence, we carried out a series of simulations using the same sample parameters that characterized the Nuu-Chah-Nulth data, plus the mutation rates estimated by Lundstrom et al (1992). Relative to the rate of lineage coalescence, the rate of population coalescence ranged over four orders of magnitude; from a 100-fold faster coalescence to a 1% rate of coalescence. The results are summarized in Table 1. They indicate that different rates of population coalescence have a marked impact on the degree of lineage sharing: little lineage sharing occurs among isolated populations (slow coalescence), whereas 10–19% of lineages are shared when the rate of population coalescence equals the rate of lineage coalescence. When coalescence of population ancestry is very short, 75–87% of lineages may be shared between populations.

The distribution of pairwise sequence differences is also influenced by the rate of population coalescence (Table 1). Although the proportion of identical sequences rises monotonically, the proportion of sequences that differ by a specific number of nucleotides has a maximum that depends on the parameters in the model. Thus, the proportion of sequences that differ by three nucleotides rises from 5.2% to a maximum of 11.2%, when the rate of population coalescence is five times the rate of lineage coalescence, then declines. Also, for a given mutation rate, the relative rate of population coalescence has a marked influence on the number of variable sites: isolated populations in this simulation had 129 variable sites (35.8% variability), whereas ephemeral populations had only 13 variable sites (3.6% variability).

An additional issue is the impact of varying mutation rates on the ability to estimate coalescent events from molecular data. Observations from relatively invariant genomic regions will give sparse molecular phylogenies, whereas highly variable regions will yield dense molecular phylogenies. The more complex dense phylogenies will tend to give a more accurate reflection of the

pattern and frequency of coalescent events. An example is afforded by the observation that no sequence variation at the ZFY region was detected in a sample of 38 males (Dorit et al 1995). Although application of the appropriate statistical techniques allows some inference about the time to the most recent common ancestor (Donnelly et al 1995), these data provide much less information about human evolutionary history than an equivalent sample of mitochondrial DNA sequences.

Genomic variability and ancestral inference in the Nuu-Chah-Nulth

To determine whether differing levels of genomic variability influenced estimates of ancestral coalescence, we evaluated sequence variability at three mitochondrial DNA regions in the same set of 60 Nuu-Chah-Nulth. The regions, selected to give a range of variability, were as follows: most variable, 360 nucleotides at the 5′ end of the control region (HVS1 of Vigilant et al 1991); intermediate, 200 nucleotides at the 3′ end of the control region (HVS2); and least variable, 18 RFLPs scattered around the mitochondrial DNA molecule. The restriction sites, chosen for their informativeness in Amerindians (Torroni et al 1992), were equivalent to assaying 510 nucleotides (Valencia 1992). With eight RFLPs being invariant in this sample, the remaining ten sites, representing 2% sequence variability, defined 16 mitochondrial DNA lineages (Table 2). These restriction sites gave low levels of ancestral resolution because 53% of the sample was defined by lineages that occurred eight or more times, and only 10% of the sample was defined by unique lineages.

By contrast, both sets of sequence data gave more resolution. Although the 5′ segment of the control region had slightly lower levels of sequence variability than the 3′ region, it identified almost twice as many mitochondrial DNA lineages, indicating that sequence variability alone may not be the best predictor of phylogeographic informativeness: the resolution of ancestral coalescence is more dependent on the pattern of variable nucleotides than the number. When the 3′ and 5′ data are combined, the number of lineages increases to 40, and the frequency spectrum of lineage distribution changes appreciably (Table 2). Only nine lineages are unique in the 3′ control region sequence data (15% of the sample), whereas two lineages occur nine times each (30% of the sample). When both sets of sequence data are combined, 48% of the sample is defined by 29 unique lineages and the most common lineage occurs only five times (8% of the sample).

Somewhat unexpectedly, the information about ancestral coalescence varies considerably between segments. It seems reasonable that adding more sequence data will mostly increase resolution at the tips of the phylogeny, with little impact on estimates of the coalescence to distant ancestors. This assumption is not borne out. Addition of 5′ sequence data to the two 3′ lineages that each occur nine times, gives seven and four new lineages, respectively, with mean

TABLE 2 Frequency spectrum of mitochondrial DNA lineages in a sample of 60 Nuu-Chah-Nulth as a function of the genomic region sampled. Tabulation of number of lineages in terms of the number of occurrences of each lineage was observed

Genomic region	*Sequence variability*	*Number of occurrences*											*Total lineages*
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*15*	
18 RFLPs[a]	2.0%												
No. of lineages		6	3	0	4	0	0	0	1	1	0	1	16
% of sample		10%	10%	0	27%	0	0	0	13%	15%	0	25%	
3′ (HVS1)	10.5%												
No. of lineages		9	2	2	3	1	1	0	0	2	0	0	17
% of sample		15%	7%	10%	20%	8%	10%	0	0	30%	0	0	
5′ (HVS2)	8.9%												
No. of lineages		15	3	4	2	2	0	0	0	1	0	0	31
% of sample		25%	10%	20%	13%	17%	0	0	0	15%	0	0	
3′ & 5′	9.5%												
No. of lineages		29	5	4	1	1	0	0	0	0	0	0	40
% of sample		48%	17%	20%	7%	8%	0	0	0	0	0	0	

[a]18 restriction fragment polymorphisms scattered around the mitochondrial DNA molecule.

pairwise sequence differences within these ostensibly identical lineages of 3.8 ± 2.5 and 2.6 ± 3.6. This level of sequence divergence (0.7–1.0%) is nearly half the total sequence divergence observed in the entire sample. Further, one of these 3′ lineages falls into three of the four major clades that were defined by the 5′ sequence data (Fig. 1). Conversely, addition of the 3′ data to the lineage that occurs nine times in the 5′ data set results in five new lineages, with an average pairwise difference of 2.3 ± 1.7. Similar results were also obtained by analysing sequence variability within the Chibcha tribes of Central America (Kolman et al 1995, Santos et al 1994). Overall, these results suggest that estimates of ancestral coalescence are likely to have unacceptably large standard errors unless quite large mitochondrial DNA segments are sequenced. Further work will be required to determine whether the approximately 600 nucleotides at the mitochondrial DNA control region is sufficient to give stable estimates of ancestry at the tribal level.

Intratribal phylogeography

A more intensive study of the Nuu-Chah-Nulth analysed sequence data for 119 individuals, sampled from 401 four-generation matrilines, identified by genealogical analysis. These individuals were selected to represent seven bands, with an average sample size of 17 ± 11.7. Sequencing the 360 nucleotides at the 5′ end of the control region identified 36 mitochondrial DNA lineages defined by 35 variable sites, with an average pairwise sequence difference of $1.5\% \pm 0.7\%$. Although the number of lineages observed within each band was relatively small, ranging from six to 13, the mean pairwise sequence difference within bands was virtually identical with the tribal value, ranging from $1.3\% \pm 0.7\%$ to $1.8\% \pm 1.0\%$. Further, the average pairwise sequence difference between different bands was also identical to the tribal average, suggesting that mitochondrial DNA lineages are randomly distributed among the Nuu-Chah-Nulth bands. More detailed analyses failed to identify any association between sequence divergence and geography, language dialect, or sociopolitical grouping (Valencia 1992). This suggests that, analogous to Fig. 3, the coalescence of band ancestry occurs on a much shorter time-scale than coalescence of lineage ancestry. This implies that when using mitochondrial DNA sequences to evaluate ancestral coalescence in Amerindian populations, the tribe, rather than the band (or village), is the more appropriate unit for analysis.

Intertribal phylogeography

Following the analysis of sequence diversity within the Nuu-Chah-Nulth, we evaluated the distribution of mitochondrial DNA lineages within and among Amerindian tribes of the Pacific Northwest (Ward et al 1993) and the circumarctic area (Shields et al 1993). In both cases, the proportion of

mitochondrial DNA lineages shared among tribes (1.1%) is substantially lower than the proportion shared among Nuu-Chah-Nulth bands. However, if the majority of lineages within a tribe tend to be unique, it is relevant to ask whether tribally specific lineages form a distinct clade, similar to the situation depicted in Fig. 1. Analysis of 41 lineages found in 144 individuals sampled from three Amerindian tribes of the Pacific Northwest gave no indication that mitochondrial DNA lineages clustered by tribe (Ward et al 1993). The two Amerind-speaking tribes shared only four lineages, but they had no tribally specific lineage clusters. This was also true for the third (Na-Dene) tribe. Further, the shared lineages occupied a nodal position in the tree (and are marked with an asterisk in Fig. 1). Rather than having been dispersed by admixture, these lineages are likely to be ancestral lineages maintained in all three populations.

A similar result is obtained from evaluating 33 lineages found among 90 circumarctic individuals sampled from a wide geographic range (Greenland to Siberia) and involving representatives of three language phyla (Na-Dene, Eskimo-Aleut, Chukchi-Kamchatka). The phylogeny for these lineages had no evidence of clades that corresponded to geography or language (Shields et al 1993). Hence, the intertribal distribution of mitochondrial DNA lineages is intermediate between the situation depicted in Figs 2 and 3, suggesting that the rate of coalescence of ancestry among tribes may occur at roughly the same rate as the rate of lineage coalescence.

Conclusion

While demonstrating the potential of the phylogeographic approach, these results indicate that the relative rate of population coalescence can exert an important influence on estimates of ancestral coalescence. More detailed models are required to characterize this effect, and analysis of additional sequence data from local populations will provide a guide to the probable magnitude of effect in human populations. The data for Amerindian tribes suggest that bands (or villages) are fairly ephemeral with short coalescence times resulting in a high degree of lineage sharing. Analysis of lineage frequency, rather than phylogeographic structure, is most likely to be informative. However, tribes appear to be more stable entities, with considerable scope for phylogeographic analysis. The degree of sequence variability represents another variable that needs further study at both the theoretical and empirical level. Therefore, the degree of lineage sharing between bands is consistent with the concept that the coalescence times for bands is appreciably shorter than that for lineages.

Acknowledgement

We thank Ken Weiss, not merely for chairing the meeting with infectious enthusiasm, but also for his input in ensuring that the central theme of the symposium would develop along a consistent and productive direction.

References

Advise C, Arnold J, Ball RM et al 1987 Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annu Rev Ecol Syst 18:489–522

Batzer MA, Stoneking M, Alegria-Hartman M et al 1994 African origin of human-specific polymorphic *Alu* insertions. Proc Natl Acad Sci USA 91:12288–12292

Bowcock AM, Kidd JR, Mountain JL et al 1991 Drift, admixture and selection in human evolution: a study with DNA polymorphisms. Proc Natl Acad Sci USA 88:839–843

Bowcock AM, Ruiz-Linares A, Tomforhde J, Minch E, Kidd JR, Cavalli-Sforza LL 1994 High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457

Cavalli-Sforza 1990 How can one study individual variation for three billion nucleotides of the human genome? Am J Hum Genet 46:649–651

Cavalli-Sforza LL, Menozzi P, Piazza A 1994 The history and geography of human genes. Princeton University Press, Princeton

Di Rienzo A, Wilson AC 1991 Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc Natl Acad Sci USA 88:1597–1601

Donnelly P, Tavaré S, Balding DJ, Griffiths RC 1995 On the time since Adam. Science, in press

Dorit RL, Akashi H, Gilbert W 1995 Absence of polymorphism at the ZFY locus on the human Y chromosome. Science 268:1183–1185

Fullerton SM, Harding RM, Boyce AJ, Clegg JB 1994 Molecular and population genetic analysis of allelic sequence diversity at the human β-globin locus. Proc Natl Acad Sci USA 91:1805–1809

Harpending HC, Sherry ST, Rogers AR, Stoneking M 1994 The genetic structure of ancient human populations. Curr Anthrop 34:483–496

Harris H 1966 Enzyme polymorphism in man. Proc R Soc Lond Ser B Biol Sci 164:298–310

Jeffreys AJ, Macleod A, Tamaki K, Neil DL, Monckton DG 1991 Minisatellite repeat coding as a digital approach to DNA typing. Nature 354:204–209

Kingman JFC 1982 On the genealogy of large populations. J Appl Prob 19:27A–43A

Kolman CJ, Bermingham E, Cooke R, Ward RH, Arias TD, Guionneau-Sinclair F 1995 Reduced mtDNA diversity in the Ngöbé Amerinds of Panama. Genetics 140:275–283

Lewontin RC 1991 Twenty-five years ago in genetics: electrophoresis in the development of evolutionary genetics: milestone or millstone? Genetics 128:657–662

Lewontin RC, Hubby JL 1966 A molecular approach to the study of genetic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics 54:595–609

Lundstrom R, Tavaré S, Ward RH 1992 Estimating substitution rates from molecular data using the coalescent. Proc Natl Acad Sci USA 89:5961–5965

Majoram P, Donnelly P 1994 Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics 136:673–683

Neel JV, Ward RH 1970 Village and tribal genetic distances among American Indians and the possible implications for human evolution. Proc Natl Acad Sci USA 65: 323–330