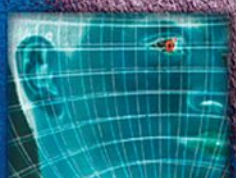
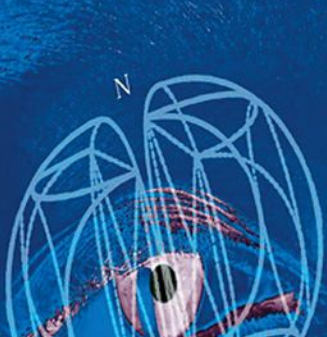


Molecular Photofitting

Predicting Ancestry and Phenotype Using DNA



Anthony Frudakis
with contributions by Mark D. Shriver

AP

MOLECULAR PHOTOFITTING

This page intentionally left blank

MOLECULAR PHOTOFITTING

PREDICTING ANCESTRY AND PHENOTYPE USING DNA

Tony N. Frudakis

With a Chapter 1 Introduction
by Mark D. Shriver

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Acquisitions Editor: Jennifer Soucy
Assoc. Developmental Editor: Kelly Weaver
Project Manager: Christie Jozwiak
Publishing Services Manager: Sarah Hajduk
Cover Designer: Alisa Andreola
Composition: SPi
Printer: China Translation & Printing Services, Ltd.

Academic Press is an imprint of Elsevier
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
84 Theobald's Road, London WC1X 8RR, UK

This book is printed on acid-free paper.

Library of Congress Cataloging-in-Publication Data
Application submitted.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ISBN: 978-0-12-088492-6

For information on all Academic Press publications
visit our Web site at www.books.elsevier.com

Printed in China
07 08 09 10 9 8 7 6 5 4 3 2 1

TABLE OF CONTENTS

FOREWORD

There is a deep desire within us all to find out who we are as individuals by tracing our ancestors through history, asking where they may have come from and what they may have looked like. Only over a relatively short period of time—at best several generations—can most of us follow our personal genealogies, using family trees drawn from oral histories or public records, before we are quickly lost in the depths of inaccessible ancestors. Another approach, only available to us in recent years, is to peer into our genes to examine the DNA record encoded in the human genome. This can provide a wealth of information about our family ties, the level of relatedness within and between populations, and ultimately even the origin of the human lineage. The linking of our shared genetic ancestries with the geographical distribution of prototypical human populations is one of the keys to finding our own affiliations as well as the distribution of physical traits within present day admixed populations. The concept of Biogeographical Ancestry (BGA) is the term comprehensively defined by Tony Frudakis in this landmark reference work, with the motivation being to correlate ancestry and sequence differences within our DNA to an individual's physical appearance. This whole process is referred to as ''molecular photofitting,'' with downstream applications for forensic identification purposes.

The considerable effort expended in characterizing the frequency distribution of single nucleotide changes within human populations has rewarded the DNAPrint Genomics team with a unique set of Ancestry Informative Markers (AIMS). With these tools and a noticeably pragmatic approach, a detailed description is given of the theoretical basis for choosing a model with four main ancestral continental groups (West African, European, East Asian, and Indigenous American). These are the geographical extremes that can be used to plot the admixture of a present day individual. They also allow indirect methods to predict physical traits such as the degree of pigmentation present within the hair, eye, and skin based on the primordial characteristics of these groups. While not definitive, this is a clear improvement on the current inaccurate means of inferring physical appearance from a DNA sample. Although the DNAWitness^{TM} protocol is currently operational as a molecular

photofitting test, the future is in directly correlating physical phenotypes with the genes that are part of the biochemical process determining or modifying these characteristics. The complement of human pigmentation genes is presently being characterized, and this book contains a good description of the polymorphisms that will direct the color traits of hair, eye, and skin. This is a major advancement of a fledgling field; the future will surely be based on an individual's genotype at specific loci. A glimpse of this is seen with correlative genes such as MC1R with hair, OCA2 with eye, and MATP with skin color.

The use of DNA fingerprinting systems such as CODIS are now well accepted in courts of law. This is not yet true for molecular photofitting techniques, and the contentious issue of predicting phenotypes based on ethnic group stereotypes has social connotations beyond forensic analysis. The discussion of the first case reports utilizing DNA for physical profiling shows the need for high levels of accountability with this breakthrough science, testing the limits of understanding of the lay public, police force, and judiciary alike. As a new form of evidence, the ''DNA-witness'' based on accumulating databases of AIMS profiled individuals must be compared with the accepted but not necessarily reliable eye-witness testimony of appearance. The final chapter on the politics and ethics of testing for genetic ancestry, as described in this textbook, is challenging and confrontational to our beliefs about what the idea of race and our own racism represents, and, as such, deserves to be read and considered by a wide audience. *Molecular Photofitting* is a very thoughtful and rigorous treatise on a socially contentious issue, but one that is very likely to help contribute to the policing of our communities.

Richard A. Sturm, PhD
Melanogenix Group
Institute for Molecular Bioscience
University of Queensland
Brisbane, Qld 4072, Australia

PREFACE

This textbook is meant to serve less as an instructive tool for the classroom and more as a reference for the forensic, clinical, and academic scientist. It is my hope that scientists seeking to develop or use methods for the inference of phenotype from DNA will find some of the ideas presented here useful.

Most of the book focuses on data, results, and observations derived from a small number of Ancestry Informative Marker (AIM) panels, which cynics might point out are the same panels my company (DNAPrint Genomics, Inc., Sarasota, FL) sells to the forensics community. This is no accident, and to a point the cynics are right. I founded DNAPrint and invested a substantial portion of its future into developing the data, databases, and tools described herein. I did this because it had not been done before, and believed there existed a corresponding niche that needed to be filled. The fact is that before this text, to my knowledge, there have been no other descriptions of bona-fide AIM panels—describing their expected accuracy and bias, results across populations, comparison of the results with self-held notions, or social constructs and correlation with elements of physical appearance and so forth. Indeed, the AIM panels developed by my colleagues and me at DNAPrint, including Mark Shriver of Penn State University, is among the first ever to be studied in these ways. The statistical methods for estimating individual ancestry, based on what Fisher referred to in 1912 as ''inverse probability,'' were worked out for the first time only in the 1970s. Bayesian methods in the field of phylogenetics and individual ancestry estimation are even more recent, having been developed only within the past 10 years. It has been possible to construct panels of good AIMs—not those found in genes, but bonafide AIMs of neutral evolutionary character—for the first time only in the past decade, as the sequence for the human genome has been released.

No multifactorial human phenotype has ever before been predicted from an appreciation of polymorphisms in human DNA. My colleagues and I were simply among the first to invest in these types of molecular resources and apply them for the purposes described in this book. As the field of molecular photofitting is completely new, having become enabled by the recent completion of

the human genome project, it is just a matter of time before other panels and improved methods will be described. Perhaps future editions can incorporate a more diverse collection of panels and databases.

I have done my best to provide a good theoretical background necessary to appreciate the methods and data discussed. I learned much of the theoretical material in the book (see Chapter 2) during my years in the field from such textbooks as *The Genetics of Human Populations*, by L.L. Cavalli-Sforza, and W.F. Bodmer; *The Handbook of Statistical Genetics*, by D.J. Balding, M. Bishop, and C. Cannings; and the various publications of R. Chakraborty. For the student, the theoretical material discussed in this book may serve as stimulation for further reading from a proper population genetics textbook and papers. For the forensics professional, the lay-description provided for these complex ideas may help in better understanding how the machinery of molecular photofitting works, and possibly obviate the need to go to a proper population genetics textbook (which to a nongeneticist may not be a pleasant experience).

Parts of the book may seem redundant—for example in a later chapter, you might notice an explanation for an idea that was treated fully in an earlier chapter. When dealing with a new topic that requires background information the approach taken here is to recap the background information learned in an earlier chapter rather than to refer you to that earlier chapter. This also serves many readers better—rather than reading the text from cover to cover, most will probably scan quickly for sections of interest, and if the background information was not recapped in that section, the point could be lost.

There are likely to be mistakes discovered in the first edition, which will be corrected in the next edition. We (my colleagues and I)would be grateful for your feedback. As with most books, this one is intended to stimulate thought, discussion, and ultimately activity in the field. It may be that 20 years from now, the field will have advanced very little due to funding limitations and various sources of resistance. On the other hand it may be that 20 years from now molecular photofitting will be standard practice, and we will be doing amazing things, such as using computers to provide most of the information on a person's drivers license from DNA left at a crime scene, even creating ''artist's renderings'' from DNA ''eyewitness testimony.'' Either way, we are honored to have the opportunity to share our work and interpretation of other relevant works with you, and we hope you find this book useful.

ACKNOWLEDGMENTS

A very significant THANKS is given to Mark Shriver, who helped write Chapter 1, provided assistance writing the parts of Chapter 6 that covered his work, and edited the first four chapters. It was in collaboration with Mark that the panel of Ancestry Informative Markers (AIMs) discussed in this book were developed, and he was involved in the collection of many of the samples we used. In addition, he and his colleagues wrote the very first version of the MLE program we used, and we later optimized and expanded this program with his help. Much data from his papers appear within these pages and the book would not have been possible without his efforts.

Matt Thomas managed the laboratory that produced most of the data described in this book, and helped develop many of the ideas, algorithms, and figures that are presented. Without his tireless work over the past six years, most of the data discussed within these pages would not have existed and this book would not have been possible. Zach Gaskin, Shannon Boyd, and Sarah Barrow produced most of the genotype data for the 71AIM and 171 AIM panels and iris color work discussed herein. It was through collaboration with Nick Martin and colleagues at the Queensland Medical Research Institute in Brisbane, Australia that the hair color data and discussion was possible.

Lastly, the most significant THANKS needs to be given to the investors of the company that funded much of the research in this book—DNAPrint Genomics, Inc. Even after countless rejected grant applications, average, every-day citizens invested in the commercial viability of the concepts we describe by buying DNAPrint stock. Although the value of DNAPrint continues to sag, and products based on the ideas and data presented here have not yet sold well in the forensic, academic, or clinical world, the field is a very new one and these investors need only look at the outcome of the Louisiana Multiagency Homicide Task Force Investigation to know that their investment has made a positive difference in the world. This book is partial evidence of the value of their investment, and I thank them for making this work possible.

CHAPTER 1

FORENSIC DNA ANALYSIS: FROM MODEST BEGINNINGS TO MOLECULAR PHOTOFITTING, GENICS, GENETICS, GENOMICS, AND THE PERTINENT POPULATION GENETICS PRINCIPLES

With an Introduction by Mark D. Shriver

PART I: INTRODUCTION: BRIEF HISTORY OF DNA IN FORENSIC SCIENCES

The forensic analysis of DNA is one of the clear successes resulting from our rapidly increasing understanding of human genetics. Perhaps much of this success is because this particular application of the molecular genetic revolution is ultimately pragmatic and because the genetic information required for efforts such as the Combined DNA Index System (CODIS) and The Innocence Project (www.innocenceproject.org) are relatively simple. Although the requirements of DNA in these instances, namely individualization, are indeed, relatively simple, they are somewhat technical, especially for the reader unfamiliar with molecular methods or population genetics. They nonetheless provide an important framework for the bulk of the material presented in this book. Though they are important for the rest of our discussion in the book, in this chapter, we provide only a brief summary of the standard forensic DNA methods, because these are well documented in other recent texts (Budowle et al. 2000; Butler 2001; Rudin & Inman 2002).

Modern forensic DNA analysis began with Variable Number of Tandem Repeats (VNTR), or minisatellite techniques. First discovered in 1985 by Sir Alex Jeffreys, these probes, when hybridized to Southern blot membranes (see Box 1-A), produced highly variable banding patterns that are known as DNA fingerprints (Jeffreys et al. 1985). Underlying these complex multibanded patterns are a number of forms (alleles) of genetic loci that simultaneously appear in a given individual. The particular combinations of alleles in a given individual are highly specific, yet each is visible because they share a common DNA sequence motif that is recognized by the multilocus molecular probe through complementary base pairing. These multilocus probes are

clearly very individualizing, but problematic when it comes to quantifying results. Some statistics can be calculated on multilocus data, but certain critical calculations cannot be made unless individual-locus genotype data are available. In answer to this need, a series of single-locus VNTR probe systems were developed, and these became standard in U.S. forensic labs from the late 1980s through the early 1990s.

Box 1-A

The Southern Blot is named after Edwin Southern, who developed this important first method for the analysis of DNA in 1975. This method takes advantage of several fundamental properties of DNA in order to assay genetic variation, generally called *polymorphism*. The first step is to isolate high molecular weight DNA, a process known as genomic DNA extraction. Next, the DNA is digested with a restriction enzyme, which makes double-stranded cuts in the DNA at every position where there is a particular base pair sequence. For example, the restriction enzyme, *Eco*RI, derived from the bacteria, *Escherichia coli* strain RY13, has the recognition sequence, GAATTC, and will cut the DNA at every position where there is a perfect copy of this sequence. Importantly, sequences that are close to this sequence (e.g., GATTTC) will not be recognized and cut by the enzyme. The restriction digestion functions to reduce the size of the genomic DNA in a systematic fashion, and originally evolved in the bacteria as a defense mechanism as the bacteria's own genomic complement was protected at these sequences through the action of other enzymes.

After DNA extraction the DNA is generally a series of large fragments averaging 25,000 to 50,000 bp in length. Because of the immense size and complexity of the genome, the results of a restriction enzyme digestion are a huge mix of fragments from tens of base pairs to tens of thousands of base pairs. When these fragments are separated by size on agarose gels using the process known as electrophoresis, they form a heavy smear. Although it's hard to tell by looking at these smears since all the fragments are running on top of each other, everyone has basically the same smear since all our DNA sequences are 99.9% identical. Places where the restriction patterns differ because of either changes in the sequence of the restriction sites (e.g., GAATTC → GATTTC) or the amount of DNA between two particular restriction sites are called Restriction Fragment Length Polymorphisms (RFLPs).

The key advancement of the Southern Blot was to facilitate the dissection of these restriction enzyme smears through the ability of DNA to denature (become single-stranded) and renature (go back to the double-stranded configuration), and to do so in a sequence-specific fashion such that only DNA fragments that have complementary sequences will hybridize or renature. The DNA in the gel is denatured using a highly basic solution and then transferred by capillary action, using stacks of paper towels onto a thin membrane, usually charged nylon. After binding the digested DNA permanently to the membrane, we can scan it by annealing short fragments of single-stranded DNA, called probes, which are labeled in such a way that we can detect their presence. The probes will anneal with DNA at locations on the genomic smear to which they have complete, or near complete

complementarity depending on the stringency of the hybridization and wash conditions. Since the probes are radiolabeled or chemiluminescently labeled, the result is a banding pattern where the location of particular sequences on the genome emerge as blobs called bands. The lengths of the bands can be estimated as a function of the position to which they migrated on the gel relative to size standards which are run in adjacent lanes.

The single-locus forensic VNTR systems are highly informative, with each marker having tens to hundreds of alleles. At every locus each person has only two alleles, which together constitute the genotype, one received from the mother and one from the father. Given such a large number of alleles in the population, most genotypes are very rare. A standard analysis with single-locus VNTRs typically included six such single-locus VNTR markers, each run separately on a Southern blot gel. The data from the separate loci would then be combined into a single result expressing one of two outcomes:

- Exclusion—the suspect and evidence samples do not match
- When the genotypes match, a profile or match probability, which is an expression of the likelihood that the two samples matched by chance alone.

Exclusions are pretty intuitive since the lack of genetic match between the samples eliminates any chance that the suspect could have donated the evidence (baring the very rare occurrences of somatic mutation, chimerism and mosaicism, each cell in our bodies has identical DNA). This of course presumes careful lab procedures and an intact and unquestioned chain of evidentiary custody. Given a match, profile probabilities are also quite intuitive, being expressions of the chances or likelihood that a particular genotype exists in a population. Profile probabilities are essentially a means to express the statistical power of a set of makers to demonstrate exclusion. For example, consider that both the suspect and biological evidence have blood type AB, the least common ABO genotype in most populations. There is no exclusion, but does that mean the suspect left the sample? Since about 4% of people have the AB genotype we say that the profile probability is 0.04 and that given no other information, the chances of having a match by chance alone are 1 in 25. Another way to read this profile probability is to say that 4% of the people match the person who left the sample.

Maybe these are good betting odds in the casino, but in both science and in court where the destiny of human lives are at stake, more stringent criteria are required. For one thing, the frequency of 4% in the population does not necessarily mean there is a 1 in 25 chance that the suspect donated the evidence. When tests of such limited power were used, other forms of evidence

that contribute to the prior probability the suspect donated the evidence would have to be taken into account. Generally, genetic markers are not the only evidence against the defendant and other pieces of information can be combined with the genetic data to comprise a preponderance of evidence. With DNA markers commonly used today, profile probabilities are much smaller than 4%, and thus the weight of the evidence is so great that convictions could be and sometimes are made solely on DNA results, without other evidence or prior probabilities taken into account.

Single-locus VNTRs were replaced by newer marker systems that became possible as a result of the Polymerase Chain Reaction (PCR), a process of amplifying DNA *in vitro*, which won a Nobel prize for its inventor, Kary Mullis. These newer markers are most commonly called Short Tandem Repeats (STRs) although they first were referred to as microsatellites since their repeat units are shorter than minisatellites. In many ways STRs are different from VNTRs. For example, STRs generally mutate one or two repeat units at a time and VNTRs mutate in steps of many repeats. There are a number of other differences and similarities in how these markers evolve and how they can be used but these are beyond the scope of this presentation, and interested readers should consult Goldstein and Schlotterer (1999).

Table 1-1 presents a summary of some of the important characteristics of STRs and VNTRs. In terms of how the markers are used in forensic analyses, STRs are quite similar to VNTRs. The most significant difference is that with VNTRs, a process of allelic binning is required to interpret the genotype. Since the range in allele size at forensic VNTRs is large, alleles at a single locus can take the form of both very small and very large fragments, and so a variety of patterns could comprise a given genotype. Gel electrophoresis methods are limited in the resolution of fragments to about 5% of fragment length and as

Table 1-1
Comparison between STRs and VNTRs in forensic DNA analysis.

Characteristic	VNTRs	STRs
Repeat unit size	10–30 base pairs	2–5 base pairs
Repeat array length	10's to 1000's of repeats	10's of repeats
Number of alleles per locus	100's to 1000's	10's
Number of loci in genome	10,000	100,000
Laboratory method	Southern Blot	Polymerase Chain Reaction
Scoring of alleles	Computer-assisted	Automated
Precision of allele size	Binning of alleles	Exact size estimation
Amount of DNA required	1000 ng	0.1 ng