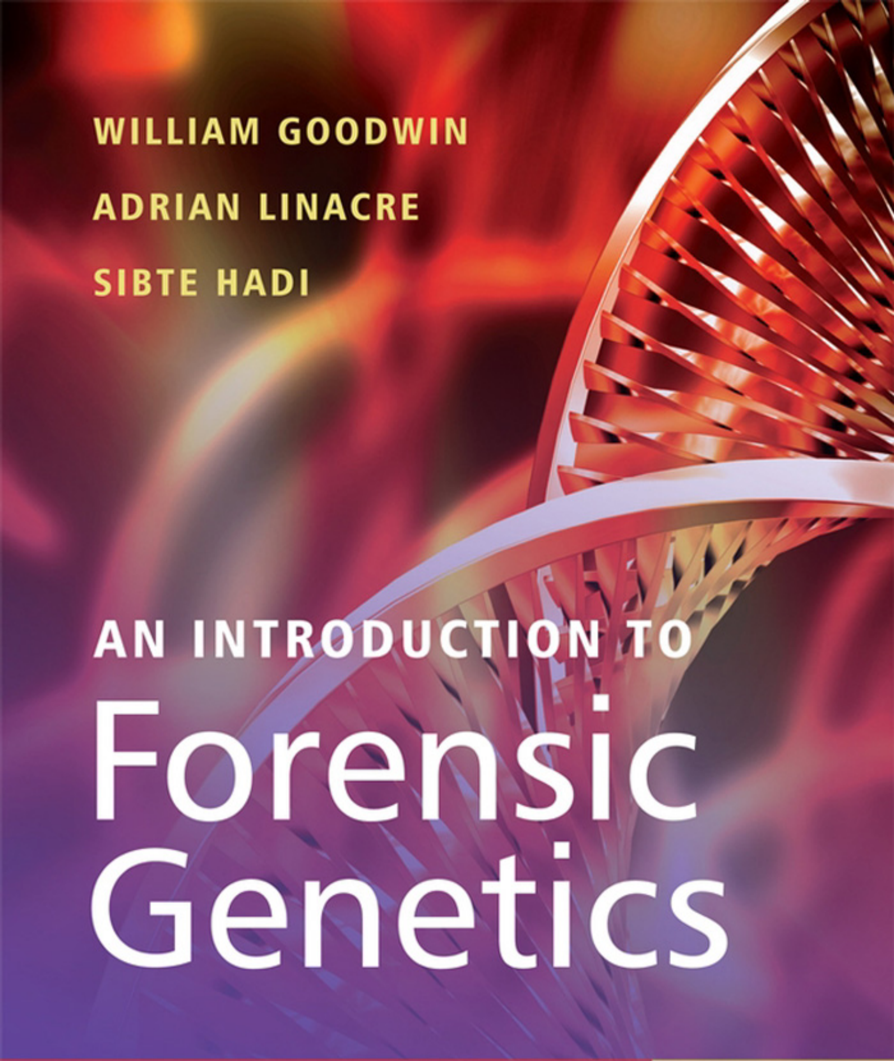


WILLIAM GOODWIN

ADRIAN LINACRE

SIBTE HADI

AN INTRODUCTION TO Forensic Genetics

WILEY

An Introduction to Forensic Genetics

An Introduction to Forensic Genetics

William Goodwin
University of Central Lancashire, UK

Adrian Linacre
University of Strathclyde, UK

Sibte Hadi
University of Central Lancashire, UK

John Wiley & Sons, Ltd

Copyright © 2007 John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher. Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or emailed to permreq@wiley.co.uk, or faxed to (+44) 1243 770620.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The Publisher is not associated with any product or vendor mentioned in this book.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 6045 Freemont Blvd, Mississauga, Ontario, L5R 4J3, Canada

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Anniversary Logo Design: Richard J. Pacifico

Library of Congress Cataloging-in-Publication Data

Goodwin, William, Dr.
An introduction to forensic genetics / William Goodwin, Adrian Linacre, Sibte Hadi.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-470-01025-9 (alk. paper) – ISBN 978-0-470-01026-6 (alk. paper)
1. Forensic genetics. I. Linacre, Adrian. II. Hadi, Sibte. III. Title.
[DNLM: 1. Forensic Genetics–methods. 2. DNA Fingerprinting.
3. Microsatellite Repeats. W 700 G657i 2007]

RA1057.5.G67 2007
614′.1–dc22 2007019041

British Library Cataloguing in Publication Data

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

ISBN (HB) 9780470010259
ISBN (PB) 9780470010266

Typeset in 10.5/12.5pt Times by Aptara, New Delhi, India
Printed and bound in Great Britain by Antony Rowe Ltd. Chippenham, Wiltshire
This book is printed on acid-free paper responsibly manufactured from sustainable forestry
in which at least two trees are planted for each one used for paper production.

Contents

Preface ix
About the Authors x

1 Introduction to forensic genetics **1**
Forensic genetics 1
A brief history of forensic genetics 2
References 5

2 DNA structure and the genome **7**
DNA structure 7
Organization of DNA into chromosomes 7
The structure of the human genome 9
Genetic diversity of modern humans 11
The genome and forensic genetics 11
Tandem repeats 12
Single nucleotide polymorphisms (SNPs) 13
Further reading 14
References 14

3 Biological material – collection, characterization and storage **17**
Sources of biological evidence 17
Collection and handling of material at the crime scene 19
Identification and characterization of biological evidence 19
Evidence collection 20
Sexual and physical assault 21
Presumptive testing 21
Storage of biological material 23
References 24

4 DNA extraction and quantification **27**
DNA extraction 27
General principles of DNA extraction 27
DNA extraction from challenging samples 30
Quantification of DNA 32
DNA IQ^{TM} system 36
References 36

5 The polymerase chain reaction **39**
The evolution of PCR-based profiling in forensic genetics 39
DNA replication – the basis of the PCR 40
The components of PCR 40
The PCR process 42
PCR inhibition 44
Sensitivity and contamination 45
The PCR laboratory 46
Further reading 48
References 48

6 The analysis of short tandem repeats **51**
Structure of STR loci 51
The development of STR multiplexes 51
Detection of STR polymorphisms 54
Interpretation of STR profiles 56
Further reading 61
References 61

7 Assessment of STR profiles **65**
Stutter peaks 65
Split peaks (+/−A) 65
Pull-up 67
Template DNA 68
Overloaded profiles 68
Low copy number DNA 68
Peak balance 70
Mixtures 70
Degraded DNA 71
References 73

8 Statistical interpretation of STR profiles **75**
Population genetics 75
Deviation from the Hardy–Weinberg equilibrium 76
Statistical tests to determine deviation from the Hardy–Weinberg equilibrium 77
Estimating the frequencies of STR profiles 78
Corrections to allele frequency databases 78
Which population frequency database should be used? 83
Conclusions 83
Further reading 84
References 84

9 The evaluation and presentation of DNA evidence 87
Hierarchies of propositions 87
Likelihood ratios 89
Two fallacies 93
Comparison of three approaches 94
Further reading 95
References 95

10 Databases of DNA profiles 97
The UK National DNA database (NDNAD) 97
International situation 102
References 104

11 Kinship testing 105
Paternity testing 105
Identification of human remains 111
Further reading 112
References 112

12 Single nucleotide polymorphisms 115
SNPs – occurrence and structure 115
Detection of SNPs 115
SNP detection for forensic applications 117
Forensic applications of SNPs 119
SNPs compared to STR loci 120
Further reading 121
References 121

13 Lineage markers 125
Mitochondria 125
Applications of mtDNA profiling 127
The Y chromosome 130
Forensic applications of Y chromosome polymorphisms 131
Further reading 132
References 133

Appendix 1 Forensic parameters 137

Appendix 2 Useful web links 139

Glossary 141

Abbreviations 145

Index 147

Preface

It is strange to consider that the use of DNA in forensic science has been with us for just over 20 years and, while a relatively new discipline, it has impacted greatly on the criminal justice system and society as a whole. It is routinely the case that DNA figures in the media, in both real cases and fictional scenarios.

The increased interest in forensic science has led to a burgeoning of university courses with modules in forensic science. This book is aimed at undergraduate students studying courses or modules in Forensic Genetics.

We have attempted to take the reader through the process of DNA profiling from the collection of biological evidence to the evaluation and presentation of genetic evidence. While each chapter can stand alone, the order of chapters is designed to take the reader through the sequential steps in the generation of a DNA profile. The emphasis is on the use of short tandem repeat (STR) loci in human identification as this is currently the preferred technique. Following on from the process of generating a DNA profile, we have attempted to describe in accessible terms how a DNA profile is interpreted and evaluated. Databases of DNA profiles have been developed in many countries and hence there is need to examine their use. While the focus of the book is on STR analysis, chapters on lineage markers and single nucleotide polymorphisms (SNPs) are also provided.

As the field of forensic science and in particular DNA profiling moves onward at a rapid pace, there are few introductory texts that cover the current state of this science. We are aware that there is a range of texts available that cover specific aspects of DNA profiling and where there this is the case, we direct readers to these books, papers or web sites.

We hope that the readers of this book will gain an appreciation of both the underlying principles and application of forensic genetics.

About the Authors

William Goodwin is a Senior Lecture in the Department of Forensic and Investigative Science at the University of Central Lancashire where his main teaching areas are molecular biology and its application to forensic analysis. Prior to this he worked for eight years at the Department of Forensic Medicine and Science in the Human Identification Centre where he was involved in a number of international cases involving the identifications of individuals from air crashes and from clandestine graves. His research has focused on the analysis of DNA from archaeological samples and highly compromised human remains. He has acted as an expert witness and also as a consultant for international humanitarian organisations and forensic service providers.

Adrian Linacre is a Senior Lecturer at the Centre for Forensic Science at the University of Strathclyde where his main areas of teaching are aspects of forensic biology, population genetics and human identification. His research areas include the use of non-human DNA in forensic science and the mechanisms behind the transfer and persistence of DNA at crime scenes. He has published over 50 papers in international journals, has presented at a number of international conferences and is on the editorial board of Forensic Science International: Genetics. Dr Linacre works as an assessor for the Council for the Registration of Forensic Practitioners (CRFP) in the area of human contact traces and is a Registered Practitioner.

Sibte Hadi is a Senior Lecture in the Department of Forensic and Investigative Science at the University of Central Lancashire. His main teaching areas are Forensic Medicine and DNA profiling. He is a physician by training and practised forensic pathology for a number of years in Pakistan before undertaking a PhD in Forensic Genetics. Following this he worked at the Department of Molecular Biology Louisiana State University as a member of the Louisiana Healthy Aging Study group. He has acted as a consultant to forensic service providers in the USA and Pakistan. His current research is focused on population genetics, DNA databases and gene expression studies for different forensic applications.

1 Introduction to forensic genetics

Over the last 20 years the development and application of genetics has revolutionized forensic science. In 1984, the analysis of polymorphic regions of DNA produced what was termed ‘a DNA fingerprint’ [1]. The following year, at the request of the United Kingdom Home Office, DNA profiling was successfully applied to a real case, when it was used to resolve an immigration dispute [2]. Following on from this, in 1986, DNA evidence was used for the first time in a criminal case and identified Colin Pitchfork as the killer of two school girls in Leicestershire, UK. He was convicted in January 1988. The use of genetics was rapidly adopted by the forensic community and plays an important role worldwide in the investigation of crime. Both the scope and scale of DNA analysis in forensic science is set to continue expanding for the foreseeable future.

Forensic genetics

The work of the forensic geneticist will vary widely depending on the laboratory and country that they work in, and can involve the analysis of material recovered from a scene of crime, paternity testing and the identification of human remains. In some cases, it can even be used for the analysis of DNA from plants [3, 4], animals [5, 6] and microorganisms [7]. The focus of this book is the analysis of biological material that is recovered from the scene of crime – this is central to the work of most forensic laboratories. Kinship testing will be dealt with separately in Chapter 11.

Forensic laboratories will receive material that has been recovered from scenes of crime, and reference samples from both suspects and victims. The role of forensic genetics within the investigative process is to compare samples recovered from crime scenes with suspects, resulting in a report that can be presented in court or intelligence that may inform an enquiry (Figure 1.1).

Several stages are involved with the analysis of genetic evidence (Figure 1.2) and each of these is covered in detail in the following chapters.

In some organizations one person will be responsible for collecting the evidence, the biological and genetic analysis of samples, and ultimately presenting the results to a court of law. However, the trend in many larger organizations is for individuals to be

---

An Introduction to Forensic Genetics W. Goodwin, A. Linacre and S. Hadi
© 2007 John Wiley & Sons, Ltd

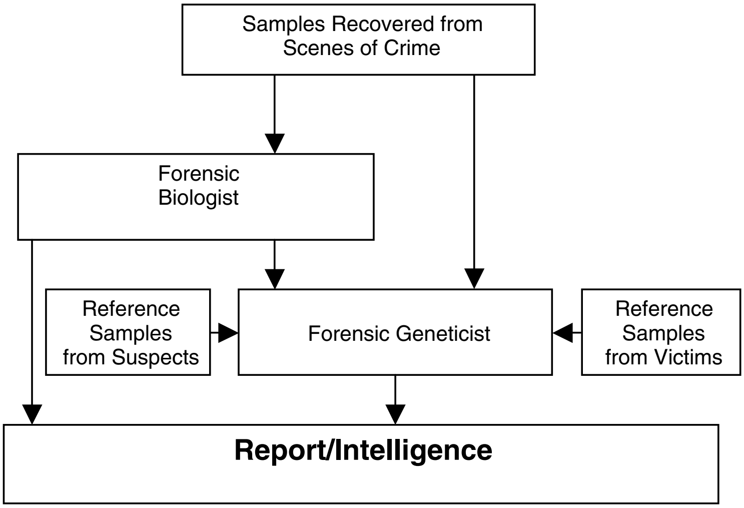

Samples Recovered from Scenes of Crime
Forensic Biologist
Reference Samples from Suspects
Forensic Geneticist
Reference Samples from Victims
Report/Intelligence

Figure 1.1 The role of the forensic geneticist is to assess whether samples recovered from a crime scene match to a suspect. Reference samples are provided from suspects and also victims of crime

responsible for only a very specific task within the process, such as the extraction of DNA from the evidential material or the analysis and interpretation of DNA profiles that have been generated by other scientists.

A brief history of forensic genetics

In 1900 Karl Landsteiner described the ABO blood grouping system and observed that individuals could be placed into different groups based on their blood type. This was the first step in the development of forensic haemogenetics. In 1915 Leone Lattes published a book describing the use of ABO typing to resolve a paternity case and by 1931 the absorption–inhibition ABO typing technique that became standard in forensic laboratories had been developed. Following on from this, numerous blood group markers and soluble blood serum protein markers were characterized and could be analysed in combination to produce highly discriminatory profiles. The serological techniques were a powerful tool but were limited in many forensic cases by the amount of biological material that was required to provide highly discriminating results. Proteins are also prone to degradation on exposure to the environment.

In the 1960s and 1970s, developments in molecular biology, including restriction enzymes, Sanger sequencing [8], and Southern blotting [9], enabled scientists to examine DNA sequences. By 1978, DNA polymorphisms could be detected using Southern blotting [10] and in 1980 the analysis of the first highly polymorphic locus was reported [11]. It was not until September 1984 that Alec Jeffreys realized the potential forensic application of the variable number tandem repeat (VNTR) loci he had been studying [1, 12]. The technique developed by Jeffreys entailed

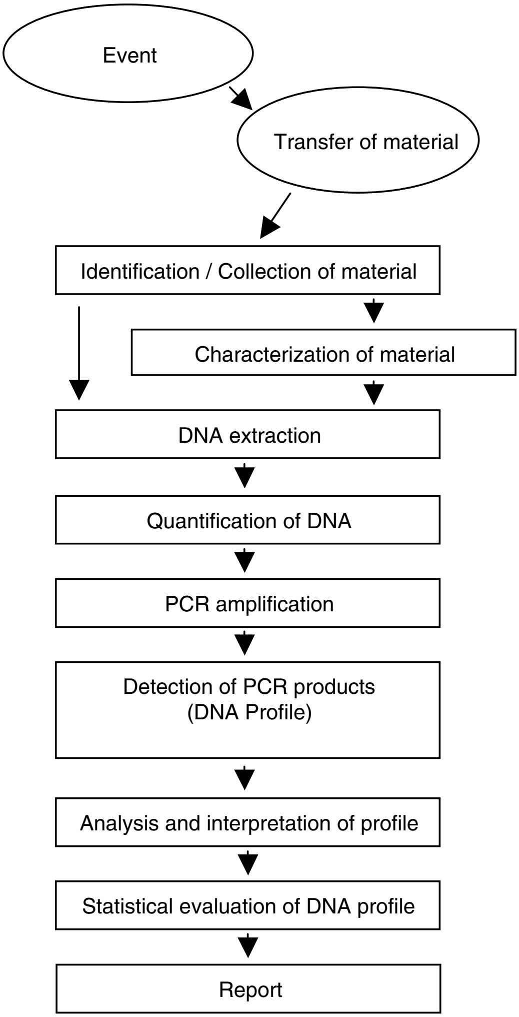

Event
Transfer of material
Identification / Collection of material
Characterization of material
DNA extraction
Quantification of DNA
PCR amplification
Detection of PCR products
(DNA Profile)
Analysis and interpretation of profile
Statistical evaluation of DNA profile
Report

Figure 1.2 Processes involved in generating a DNA profile following a crime. Some types of material, in particular blood and semen, are often characterized before DNA is extracted

extracting DNA and cutting it with a restriction enzyme, before carrying out agarose gel electrophoresis, Southern blotting and probe hybridization to detect the polymorphic loci. The end result was a series of black bands on X-ray film (Figure 1.3). VNTR analysis was a powerful tool but suffered from several limitations: a relatively large amount of DNA was required; it would not work with degraded DNA; comparison between laboratories was difficult; and the analysis was time consuming.

A critical development in the history of forensic genetics came with the advent of a process that can amplify specific regions of DNA – the polymerase chain reaction (PCR) (see Chapter 5). The PCR process was conceptualised in 1983 by Kary Mullis, a chemist

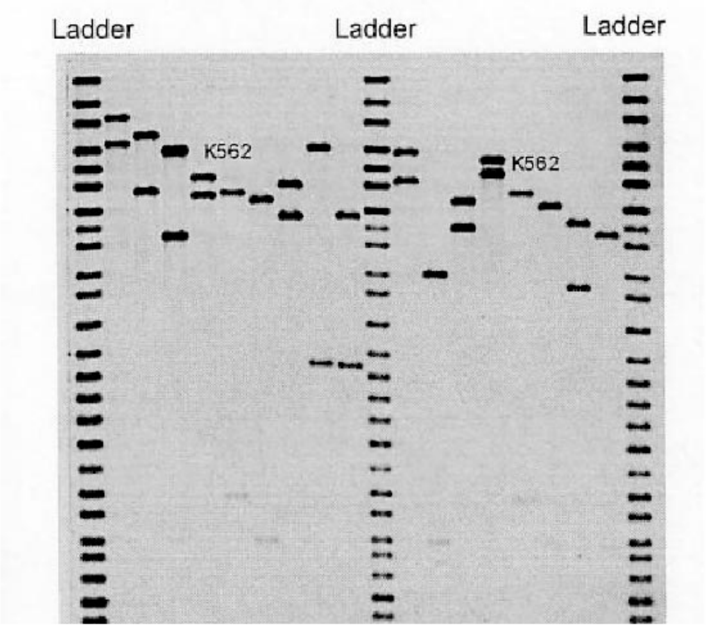


Figure 1.3 VNTR analysis using a single locus probe: ladders were run alongside the tested samples that allowed the size of the DNA fragments to be estimated. A control sample labelled K562 is analysed along with the tested samples

working for the Cetus Corporation in the USA [13]. The development of PCR has had a profound effect on all aspects of molecular biology including forensic genetics, and in recognition of the significance of the development of the PCR, Kary Mullis was awarded the Nobel Prize for Chemistry in 1993. The PCR increased the sensitivity of DNA analysis to the point where DNA profiles could be generated from just a few cells, reduced the time required to produce a profile, could be used with degraded DNA and allowed just about any polymorphism in the genome to be analysed. The first application of PCR in a forensic case involved the analysis of single nucleotide polymorphisms in the DQα locus [14] (see Chapter 12). This was soon followed by the analysis of short tandem repeats (STRs) which are currently the most commonly used genetic markers in forensic science (see Chapters 6 to 8). The rapid development of technology for analysing DNA includes advances in DNA extraction and quantification methodology, the development of commercial PCR based typing kits and equipment for detecting DNA polymorphisms.

In addition to technical advances, another important part of the development of DNA profiling that has had an impact on the whole field of forensic science is quality control. The admissibility of DNA evidence was seriously challenged in the USA in 1987 – '*People* v. *Castro*' [15]; this case and subsequent cases have resulted in increased levels of standardization and quality control in forensic genetics and other areas of forensic science. As a result, the accreditation of both laboratories and individuals is an increasingly important issue in forensic science. The combination of technical advances, high levels of standardization and quality control have led to forensic DNA analysis being recognized as a robust and reliable forensic tool worldwide.

References

1. Jeffreys, A.J. *et al.* (1985) Individual-specific fingerprints of human DNA. *Nature* **316**, 76–79.
2. Jeffreys, A.J. *et al.* (1985) Positive identification of an immigration test-case using human DNA fingerprints. *Nature* **317**, 818–819.
3. Kress, W.J. *et al.* (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 8369–8374.
4. Linacre, A. and Thorpe, J. (1998) Detection and identification of cannabis by DNA. *Forensic Science International* **91**, 71–76.
5. Parson, W. *et al.* (2000) Species identification by means of the cytochrome b gene. *International Journal of Legal Medicine* **114** (1–2), 23–28.
6. Hebert, P.D.N. *et al.* (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London Series B-Biological Sciences* **270**, S96–S99.
7. Hoffmaster, A.R. *et al.* (2002) Molecular subtyping of *Bacillus anthracis* and the 2001 bioterrorism-associated anthrax outbreak, United States. *Emerging Infectious Diseases* **8**, 1111–1116.
8. Sanger, F. *et al.* (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463–5467.
9. Southern, E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology* **98**, 503–517.
10. Kan, Y.W. and Dozy, A.M. (1978) Polymorphism of DNA sequence adjacent to human B-globin structural gene: relationship to sickle mutation. *Proceedings of the National Academy of Sciences of the United States of America* **75**, 5631–5635.
11. Wyman, A.R. and White, R. (1980) A highly polymorphic locus in human DNA. *Proceedings of the National Academy of Sciences of the United States of America* **77**, 6754–6758.
12. Jeffreys, A.J. and Wilson, V. (1985) Hypervariable regions in human DNA. *Genetical Research* **45**, 213–213.
13. Saiki, R.K. *et al.* (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle-cell anemia. *Science* **230**, 1350–1354.
14. Stoneking, M. *et al.* (1991) Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. *American Journal of Human Genetics* **48**, 370–382.
15. Patton, S.M. (1990) DNA fingerprinting: the Castro case. *Harvard Journal of Law and Technology* **3**, 223–240.

2 DNA structure and the genome

Each person's genome contains a large amount of DNA that is a potential target for DNA profiling. The selection of the particular region of polymorphic DNA to analyse can change with the individual case and also the technology that is available. In this chapter a brief description of the primary structure of the DNA molecule is provided along with an overview of the different categories of DNA that make up the human genome. The criteria that the forensic geneticist uses to select which loci to analyse are also discussed.

DNA structure

DNA has often been described as the 'blueprint of life', containing all the information that an organism requires in order to function and reproduce. The DNA molecule that carries out such a fundamental biological role is relatively simple. The basic building block of the DNA molecule is the nucleotide triphosphate (Figure 2.1a). This comprises a triphosphate group, a deoxyribose sugar (Figure 2.1b) and one of four bases (Figure 2.1c).

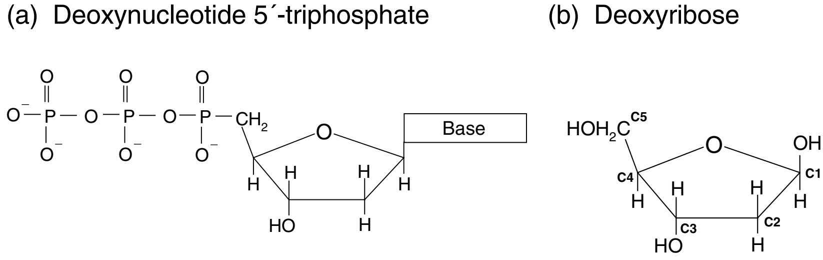
The information within the DNA 'blueprint' is coded by the sequence of the four different nitrogenous bases, adenine, guanine, thymine and cytosine, on the sugar-phosphate backbone (Figure 2.2a).

DNA normally exists as a double stranded molecule which adopts a helical arrangement – first described by Watson and Crick in 1953 [1]. Each base is attracted to its complementary base: adenine always pairs with thymine and cytosine always pairs with guanine (Figure 2.2b).

Organization of DNA into chromosomes

Within each nucleated human cell there are two complete copies of the genome. The genome is 'the haploid genetic complement of a living organism' and in humans contains approximately 3 200 000 000 base pairs (bp) of information, which is organized into 23 chromosomes. Humans contain two sets of chromosomes – one version of each chromosome inherited from each parent giving a total of 46 chromosomes (Figure 2.3). Each chromosome contains one continuous strand of DNA, the largest – chromosome

An Introduction to Forensic Genetics W. Goodwin, A. Linacre and S. Hadi
© 2007 John Wiley & Sons, Ltd


(a) Deoxynucleotide 5´-triphosphate

(b) Deoxyribose

Base
HOH2C
C5
C4
C3
C2
C1
OH
HO

(c) Nitrogenous bases

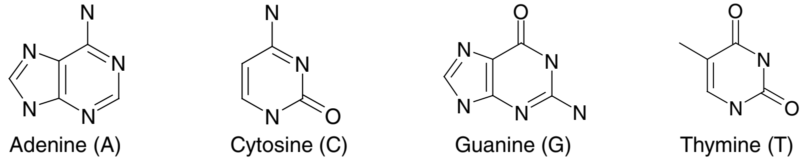

Adenine (A)
Cytosine (C)
Guanine (G)
Thymine (T)

Figure 2.1 The DNA molecule is built up of deoxynucleotide 5′-triphosphates (2.1a). The sugar (2.1b) contains five carbon atoms (labelled C1 to C5); one of four different types of nitrogenous base (2.1c) is attached to the 1 prime (1′) carbon, a hydroxyl group to the 3′ carbon and the phosphate group to the 5′ carbon

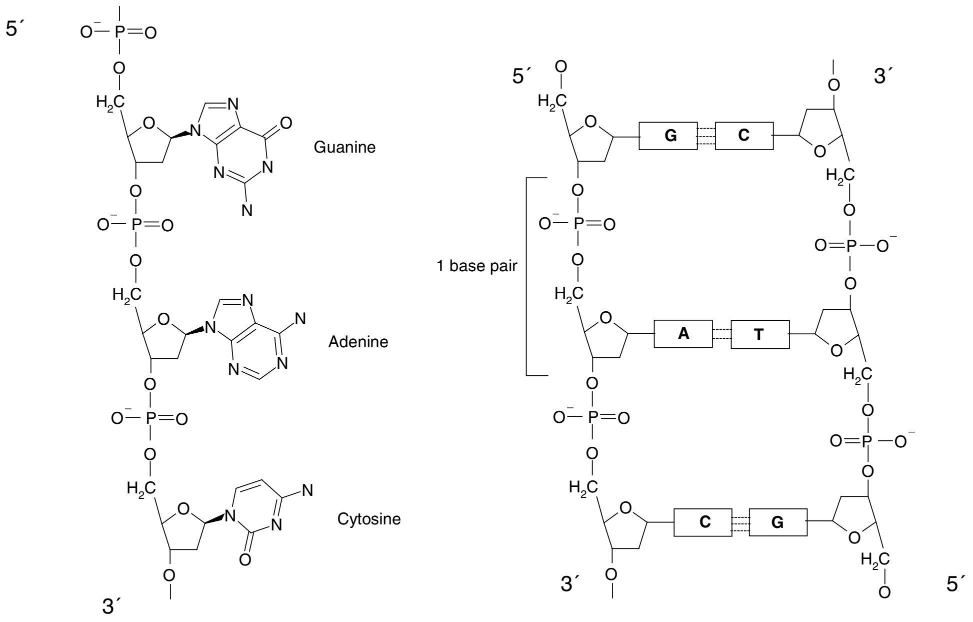

5´
Guanine
Adenine
Cytosine
3´
5´
3´
1 base pair
G
C
A
T
C
G
3´
5´

Figure 2.2 In the DNA molecule the nucleotides are joined together by phosphodiester bonds to form a single stranded molecule (2.2a). The DNA molecule in the cell is double stranded (2.2b) with two complementary single stranded molecules held together by hydrogen bonds. Adenine and thymine form two hydrogen bonds while guanine and cytosine form three bonds

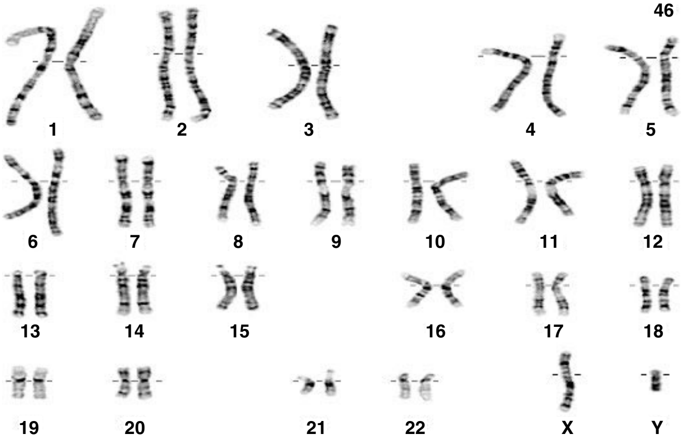


Figure 2.3 The male human karyotype pictured contains 46 chromosomes, 22 autosomes and the X and Y sex chromosomes – the female karyotype has two X chromosomes (picture provided by David McDonald, Fred Hutchinson Cancer Research Center, Seattle and Tim Knight, University of Washington)

1 – is approximately 250 000 000 bp long while the smallest – chromosome 22 – is approximately 50 000 000 bp [2–4].

In physical terms the chromosomes range in length from 73 mm to 14 mm. The chromosomes shown in Figure 2.3 are in the metaphase stage of the cell cycle and are highly condensed – when the cell is not undergoing division the chromosomes are less highly ordered and are more diffuse within the nucleus. To achieve the highly ordered chromosome structure, the DNA molecule is associated with histone proteins, which help the packaging and organization of the DNA into the ordered chromosome structure.

The structure of the human genome

Great advances have been made in our understanding of the human genome in recent years, in particular through the work of the Human Genome Project that was officially started in 1990 with the central aim of decoding the entire genome. It involved a collaborative effort involving 20 centres in China, France, Germany, Great Britain, Japan and the United States. A draft sequence was produced in 2001 that covered 90 % of the euchromatic DNA [3, 4], this was followed by later versions that described the sequence of 99 % of the euchromatic DNA with an accuracy of 99.99 % [2]. The genome can be divided into different categories of DNA based on the structure and function of the sequence (Figure 2.4).

Coding and regulatory sequence

The regions of DNA that encode and regulate the synthesis of proteins are called genes; at the latest estimate the human genome contains only 20 000–25 000 genes and only

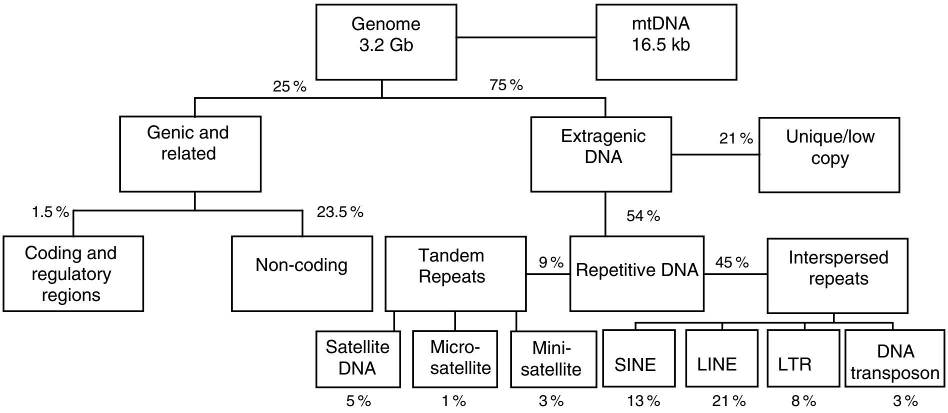

Genome
3.2 Gb
mtDNA
16.5 kb
25 %
75 %
Genic and related
Extragenic DNA
21 %
Unique/low copy
1.5 %
23.5 %
54 %
Coding and regulatory regions
Non-coding
Tandem Repeats
9 %
Repetitive DNA
45 %
Interspersed repeats
Satellite DNA
Micro-satellite
Mini-satellite
SINE
LINE
LTR
DNA transposon
5 %
1 %
3 %
13 %
21 %
8 %
3 %

Figure 2.4 The human genome can be classified into different types of DNA based on its structure and function. Modified with permission from Jasinska, A., and Krzyzosiak, W.J. (2004) Repetitive sequences that shape the human transcriptome. FEBS Letters 567, 136–141).

around 1.5 % of the genome is directly involved in encoding for proteins [2–4]. Gene structure, sequence and activity are a focus of medical genetics due to the interest in genetic defects and the expression of genes within cells. Approximately 23.5 % of the genome is classified as genic sequence, but does not encode proteins. The non-coding genic sequence contains several elements that are involved with the regulation of genes, including promoters, enhancers, repressors and polyadenylation signals; the majority of gene related DNA, around 23 %, is made up of introns, pseudogenes and gene fragments.

Extragenic DNA

Most of the genome, approximately 75 %, is extragenic. Around 20 % of the genome is single copy DNA which in most cases does not have any known function although some regions appear to be under evolutionary pressure and presumably play an important, but as yet unknown, role [6].

The largest portion of the genome – over 50 % – is composed of repetitive DNA; 45 % of the repetitive DNA is interspersed, with the repeat elements dispersed throughout the genome. The four most common types of interspersed repetitive element – short interspersed elements (SINEs), long interspersed elements (LINEs), long terminal repeats (LTRs) and DNA transposons – account for 45 % of the genome [3, 7]. These repeat sequences are all derived through transposition. The most common interspersed repeat element is the *Alu* SINE; with over 1 million copies, the repeat is approximately 300 bp long and comprises around 10 % of the genome. There is a similar number of LINE elements within the genome, the most common is LINE1, which is between 6–8 kb long, and is represented in the genome around 900 000 times; LINEs make up around 20 % of the genome [3, 7]. The other class of repetitive element is tandemly repeated DNA. This can be separated into three different types: satellite DNA, minisatellites, and microsatellites.

Genetic diversity of modern humans

The aim of using genetic analysis for forensic casework is to produce a DNA profile that is highly discriminating – the ideal would be to generate a DNA profile that is unique to each individual. This allows biological evidence from the scene of a crime to be matched to an individual with a high level of confidence and can be very powerful forensic evidence.

The ability to produce highly discriminating profiles is dependent on individuals being different at the genetic level and, with the exception of identical twins, no two individuals have the same DNA. However, individuals, even ones that appear very different, are actually very similar at the genetic level. Indeed, if we compare the human genome to that of our closest animal cousin, the chimpanzee, with whom we share a common ancestor around 6 million years ago, we find that our genomes have diverged by only around 5 %; the DNA sequence has diverged by only 1.2 % [8] and insertions and deletions in both human and chimpanzee genomes account for another 3.5 % divergence [8, 9]. This means that we share 95 % of our DNA with chimps! Modern humans have a much more recent common history, which has been dated using genetic and fossil data to around 150 000 years ago [10, 11]. In this limited time, nucleotide substitutions have led to an average of one difference every 1000–2000 bases between every human chromosome, averaging one difference every 1250 bp [4, 12] – which means that we share around 99.9 % of our genetic code with each other. Some additional variation is caused by insertions, deletions and length polymorphisms, and segmental duplications of the genome. There have been attempts to define populations genetically based on their racial identity or geographical location, and while it has been possible to classify individuals genetically into broad racial/geographic groupings, it has been shown that most genetic variation, around 85 %, can be attributed to differences between individuals within a population [13, 14]. Differences between regions tend to be geographic gradients (clines), with gradual changes in allele frequencies [15, 16].

From a forensic point there is very little point in analysing the 99.9 % of human DNA that is common between individuals. Fortunately, there are well characterized regions within the genome that are variable between individuals and these have become the focus of forensic genetics.

The genome and forensic genetics

With advances in molecular biology techniques it is now possible to analyse any region within the 3.2 billion bases that make up the genome. DNA loci that are to be used for forensic genetics should have some key properties, they should ideally:

- be highly polymorphic (varying widely between individuals);
- be easy and cheap to characterize;
- give profiles that are simple to interpret and easy to compare between laboratories;
- not be under any selective pressure;
- have a low mutation rate.

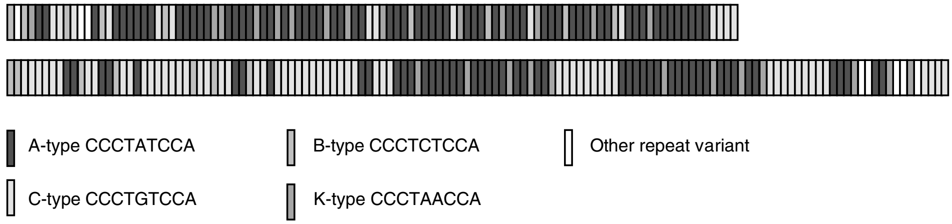


Figure 2.5 The structure of two MS1 (locus *D1S7*) VNTR alleles (Berg et al., 2003) [19]. The alleles are both relatively short containing 104 and 134 repeats – alleles at this locus can contain over 2000 repeats. The alleles are composed of several different variants of the 9 bp core repeat – this is a common feature of VNTR alleles

Tandem repeats

Two important categories of tandem repeat have been used widely in forensic genetics: minisatellites, also referred to as variable number tandem repeats (VNTRs); and microsatellites, also referred to as short tandem repeats (STRs). The general structure of VNTRs and STRs is the same (Figures 2.5 and 2.6). Variation between different alleles is caused by a difference in the number of repeat units that results in alleles that are of different lengths and it is for this reason that tandem repeat polymorphisms are known as length polymorphisms.

Variable number tandem repeats – VNTRs

VNTRs are located predominantly in the subtelomeric regions of chromosomes and have a core repeat sequence that ranges in size from 6 to100 bp [17, 18]. The core repeats are represented in some alleles thousands of times; the variation in repeat number creates alleles that range in size from 500 bp to over 30 kb (Figure 2.5). The number of potential alleles can be very large: the MS1 locus for example, has a relatively short and simple core repeat unit of 9 bp with alleles that range from approximately 1 kb to over 20 kb – which means that there are potentially over 2000 different alleles at this locus [19].

VNTRs were the first polymorphisms used in DNA profiling and they were successfully used in forensic casework for several years [20]. The use of VNTRs was, however, limited by the type of sample that could be successfully analysed because a large amount of high molecular weight DNA was required. Interpreting VNTR profiles could also be problematic. Their use in forensic genetics has now been replaced by short tandem repeats (STRs).

Short tandem repeats – STRs

STRs are currently the most commonly analysed genetic polymorphism in forensic genetics. They were introduced into casework in the mid-1990s and are now the main

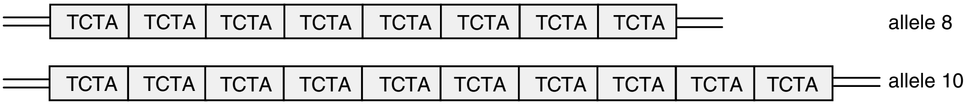

TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA
allele 8
TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA
allele 10

Figure 2.6 The structure of a short tandem repeat. The core repeat can be between 1 and 6 bp. This example shows the structure of two alleles from the locus D8S1179. The DNA either side of the core repeats is called flanking DNA. The alleles are named according to the number of repeats that they contain

tool for just about every forensic laboratory in the world – the vast majority of forensic genetic casework involves the analysis of STR polymorphisms.

There are thousands of STRs that can potentially be used for forensic analysis. STR loci are spread throughout the genome including the 22 autosomal chromosomes and the X and Y sex chromosomes. They have a core unit of between 1 and 6 bp and the repeats typically range from 50 to 300 bp. The majority of the loci that are used in forensic genetics are tetranucleotide repeats, which have a four base pair repeat motif (Figure 2.6).

STRs satisfy all the requirements for a forensic marker: they are robust, leading to successful analysis of a wide range of biological material; the results generated in different laboratories are easily compared; they are highly discriminatory, especially when analysing a large number of loci simultaneously (multiplexing); they are very sensitive, requiring only a few cells for a successful analysis; it is relatively cheap and easy to generate STR profiles; and there is a large number of STRs throughout the genome that do not appear to be under any selective pressure.

Single nucleotide polymorphisms (SNPs)

The simplest type of polymorphism is the SNP; single base differences in the sequence of the DNA. The structure of a typical SNP polymorphism is illustrated in Figure 2.7.

SNPs are formed when errors (mutations) occur as the cell undergoes DNA replication during meiosis. Some regions of the genome are richer in SNPs than others, for example chromosome 1 contains a SNP on average every 1.45 kb compared with chromosome 19, where SNPs occur on average every 2.18 kb [21].

SNPs normally have just two alleles, for example one allele with a guanine and one with an adenine, and therefore are not highly polymorphic and do not fit with the

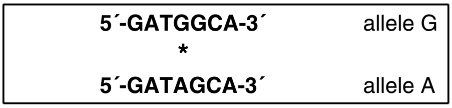


Figure 2.7 A single nucleotide polymorphism (SNP). Two alleles are shown that differ at one position, indicated by the star: the fourth position in allele G is a guanine while in allele A it is an adenine. In most cases, the mutation event at the specific locus which creates a SNP is a unique event and only two different alleles (biallelic) are normally found

ideal properties of DNA polymorphisms for forensic analysis. However, SNPs are so abundant throughout the genome that it is theoretically possible to type hundreds of them. This will make the combined power of discrimination very high. It is estimated that to achieve the same discriminatory power that is achieved using 10 STRs, 50 – 80 SNPs would have to be analysed [22, 23]. With current technology, this is much more difficult than analysing 10 STR loci.

With the exception of the analysis of mitochondrial DNA (see Chapter 13) SNPs have not been used widely in forensic science to date, and the dominance of tandem repeated DNA will continue for the foreseeable future [24]. SNPs are however finding a number of niche applications (see Chapter 12).

Further reading

Brown, T.A. (2007) *Genomes 3*. Garland Science.

WWW resource

The Human Genome Project Information: a website funded by the U.S. Department of Energy which along with and the National Institutes of Health coordinated the project. Contains resources on all aspects of the Human Genome Project. http://www.ornl.gov/sci/techresources/ Human_Genome/home.shtml

References

1. Watson, J., and Crick, F. (1953) A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738.
2. Collins, F.S. *et al.* (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945.
3. Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
4. Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* **291**, 1304–1351.
5. Jasinska, A., and Krzyzosiak, W.J. (2004) Repetitive sequences that shape the human transcriptome. *FEBS Letters* **567**, 136–141.
6. Waterston, R.H. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.
7. Li, W.H. *et al.* (2001) Evolutionary analyses of the human genome. *Nature* **409**, 847–849.
8. Mikkelsen, T.S. et al. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87.
9. Britten, R.J. (2002) Divergence between samples of chimpanzee and human DNA sequences is 5 %, counting indels. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 13633–13635.
10. Cann, R.L. *et al.* (1987) Mitochondrial-DNA and human-evolution. *Nature* **325**, 31–36.
11. Stringer, C.B., and Andrews, P. (1988) Genetic and fossil evidence for the origin of modern humans. *Science* **239**, 1263–1268.
12. Sachidanandam, R. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933.
13. Barbujani, G. *et al.* (1997) An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 4516–4519

14. Lewontin, R.C. (1972) The apportionment of human diversity. *Evolutionary Biology* **6**, 381–398.
15. Cavalli-Sforza, L.L. et al. (1994) *The History and Geography of Human Genes*. Princeton University Press.
16. Serre, D., and Paabo, S.P. (2004) Evidence for gradients of human genetic diversity within and among continents. *Genome Research* **14**, 1679–1685.
17. Jeffreys, A.J. *et al*. (1995) Mutation processes at human minisatellites. *Electrophoresis* **16**, 1577–1585.
18. Jeffreys, A.J. *et al*. (1985) Hypervariable minisatellite regions in human DNA. *Nature* **314**, 67–73.
19. Berg, I. *et al*. (2003) Two modes of germline instability at human minisatellite MS1 (Locus D1S7): Complex rearrangements and paradoxical hyperdeletion. *American Journal of Human Genetics* **72**, 1436–1447.
20. Jeffreys, A.J. *et al*. (1985) Positive identification of an immigration test-case using human DNA fingerprints. *Nature* **317**, 818–819.
21. Thorisson, G.A., and Stein, L.D. (2003) The SNP Consortium website: past, present and future. *Nucleic Acids Research* **31**, 124–127.
22. Gill, P. (2001) An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *International Journal of Legal Medicine* **114**, 204–210.
23. Krawczak, M. (1999) Informativity assessment for biallelic single nucleotide polymorphisms. *Electrophoresis* **20**, 1676–1681.
24. Gill, P. *et al.* (2004) An assessment of whether SNPs will replace STRs in national DNA databases. *Science and Justice* **44**, 51–53.

3 Biological material – collection, characterization and storage

The sensitivity and evidential power of DNA profiling have impacted on the way in which crime scenes are investigated. Because only a few cells are required for DNA profiling, crime scene examiners now have a much wider range of biological evidence to collect and also have a much greater chance of contaminating the scene with their own DNA.

Sources of biological evidence

The human body is composed of trillions of cells and most of these contain a nucleus, red blood cells being a notable exception. A wide variety of cellular material can be recovered from crime scenes (Table 3.1).

Each nucleated cell contains two copies of an individual's genome and can be used, in theory, to generate a DNA profile under optimal conditions [1]. In practice, 15 or more cells are required to generate consistently good quality DNA profile from fresh material. Forensic samples usually show some level of degradation and with higher levels of degradation, more cellular material is required to produce a DNA profile. If the material is very highly degraded then, even with the high sensitivity of DNA profiling, it may not be possible to generate a DNA profile.

The biological material encountered most often at scenes of crime is blood (Figure 3.1). This is mainly because of the violent nature of many crimes and also because it is easier to visualize than other biological fluids such as saliva.

Other frequently encountered samples include seminal fluid, which is of prime importance in sexual assault cases; saliva that may be found on items held in the mouth, such as cigarette butts and drinking vessels, or on bite marks; and epithelial cells, deposited, for example, as dandruff and in faeces. With the increase in the sensitivity of DNA profiling the recovery of DNA from epithelial cells shed on touching has also become possible [2]. Hairs are naturally shed, and can also be pulled out through physical contact and can be recovered from crime scenes. Naturally shed hairs tend to have

© 2007 John Wiley & Sons, Ltd

Table 3.1 Types of biological material that can be recovered from a crime scene. The DNA profiles generated from crime scene material are compared against reference profiles that are provided by suspects, and in some cases, the victims

Scenes of Crime	Reference Samples
Blood	Blood
Semen	Buccal swabs
Hair	Pulled hairs (containing roots)
Epithelial cells – shed skin cells:	
Saliva	
Dandruff	
Clothing	
Cigarette butts	
Drinking vessels/food	
Urine	
Vomit	
Faeces	
Touch DNA	

very little follicle attached and are not a good source of DNA, whereas plucked hairs or hairs removed due to a physical action often have the root attached, which is a rich source of cellular material.

The four most common nucleated cell types that are recovered from scenes of crime are white blood cells, spermatozoa, epithelial cells and hair follicles (Figure 3.2).

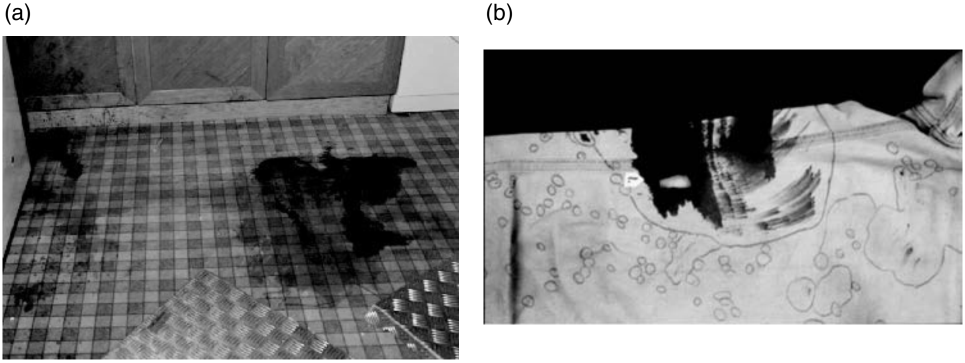


Figure 3.1 Blood is the most common form of biological material that is recovered from crime scenes. (a) Large volumes of blood can be collected using a swab, if the blood is liquid then a syringe or pipette can be used (picture provided by Allan Scott, University of Central Lancashire) (b) Blood on clothing is normally collected by swabbing, or cutting out the stain (picture provided by Elizabeth Wilson) (see plate section for full-colour version of this figure)

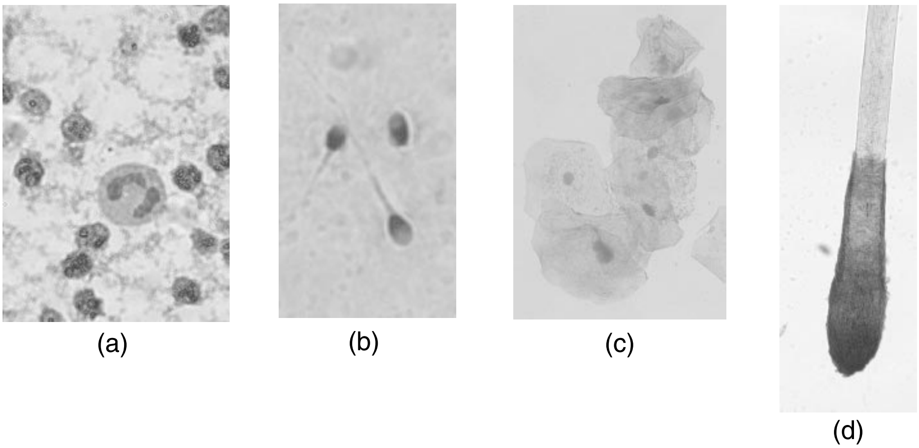


Figure 3.2 Common cell types that are recovered from scenes of crime: (a) white blood cells; (b) spermatozoa; (c) epithelial cells (from saliva); (d) a hair shaft with the follicle attached (the cells have been stained with haematoxylin and eosin)

Collection and handling of material at the crime scene

The high level of sensitivity that makes DNA profiling an invaluable forensic tool can also be a potential disadvantage. Contamination of evidential material with biological material from another source, such as an attending police officer or scene of crime officer, is a very real possibility. It is vital that the appropriate care is taken, such as maintaining the integrity of the scene and wearing full protective suits and face masks during the investigation of the scene [3–5] (Figure 3.3). Improper handling of the evidence can have serious consequences. In the worst cases, it can cause cross contamination, lead to sample degradation, and prevent or confuse the interpretation of evidence.

Identification and characterization of biological evidence

Searching for biological material, both at the crime scene and in the forensic laboratory is performed primarily by eye. In the laboratory, low power search microscopes may help to localize stains and contact marks. Alternative light sources have been found to assist with finding biological material both in the field and in the laboratory. Epithelial cells, saliva and semen stains may fluoresce at different wavelengths of light compared with the background substrate and therefore may become visible [6–8]. A range of light sources is available and these can either operate at fixed wavelengths or a variable number of wavelengths that are suitable for detecting different types of stain.

Searching a crime scene or items recovered from a crime scene for blood can be aided by the use of luminol (3-aminophthalhydrazide). This chemical can be sprayed onto a wide area and will become oxidized and luminescent in the presence of haemoglobin, which is found in red blood cells. It is necessary to be able to darken the area that is being searched in order that the luminescence can be detected. Luminol can also be used

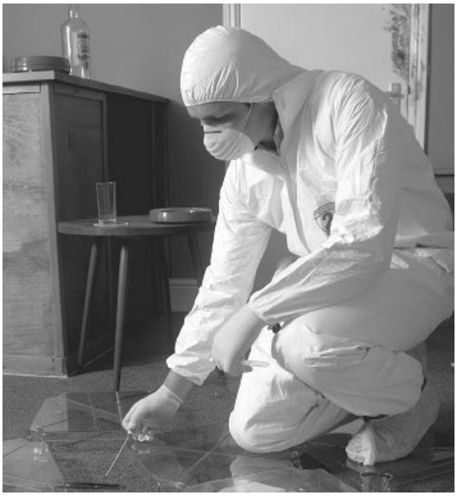


Figure 3.3 It is standard practice for scene of crime officers to wear full overalls, shoe covers, gloves and face masks when collection biological evidence from a scene of crime. Even with these precautions it is possible for crimes to be contaminated by forensic investigators and it is becoming common for the DNA profiles of police officers and scene of crime officers to be stored on a database; any profiles recovered from the scene of crime can be checked against this elimination database to rule out the possibility of a profile coming from an investigating police or scene of crime officer

in the more controlled environment of the forensic laboratory and can be particularly useful when searching clothing for trace amounts of blood.

The success in finding biological material depends upon the search method employed and also on the integrity and state of the scene. In the UK, biological material is found at approximately 12 % of investigated crime scenes, this figure can go up significantly if the crime scene is exhaustively searched.

Evidence collection

The methods used for collection will vary depending on the type of sample. Dry stains and contact marks on large immovable items are normally collected using a sterile swab that has been moistened with distilled water [9, 10]; in other cases, scraping or cutting of material may be more appropriate. Lifting from the surface using high quality adhesive tape is an alternative method for collecting epithelial cells [11]. Liquid blood can be collected using a syringe or pipette and transferred to a clean sterile storage tube that contains anticoagulant (EDTA), or by using a swab or piece of fabric to soak up the stain, which should be air dried to prevent the build up of microbial activity [4]. Liquid blood can also be applied to FTA® paper that is impregnated with chemicals to prevent the action of microbial agents and stabilize the DNA.

Smaller moveable objects, such as weapons, which might contain biological material are packaged at the scene of crime and examined in the controlled environment of the forensic laboratory. The same range of swabbing, scraping and lifting techniques as used in the field can be employed to collect the biological material. Clothing taken from suspects and victims presents an important source of biological evidence. This is also analysed in the forensic biology laboratory where stains and contact areas can be recorded and then cut out or swabbed.

Sexual and physical assault

Following sexual assaults the victim should be examined as soon after the event as possible. Semen is recovered by a trained medical examiner using standard swabs; fingernail scrapings can be collected using a variety of swabs; combings of pubic and head hair are normally stored in paper envelopes. Contact marks, for example bruising caused by gripping or bite marks, can be swabbed for DNA. The same types of evidence (except semen) can be taken after cases of physical assault [4].

Presumptive testing

Identifying a red spot on a wall or a white stain on a bed sheet might indicate the presence of blood or semen. A range of presumptive tests are available that aid the identification of the three main body fluids encountered; blood, semen and saliva. Ideally presumptive tests should be safe, inexpensive, easy to carry out, use a very small amount of the sample, and provide a simple indication of the presence or absence of a body fluid. The presumptive test should have no negative effect on DNA profiling. In addition to helping to locate material for DNA analysis, stain characterization can also provide important probative and circumstantial evidence [12].

Blood

The presumptive tests used to detect the presence of blood take advantage of the peroxidase activity of the haem group which is abundant as part of the haemoglobin molecule within red blood cells – there can be as many as 5 million red blood cells in 1 millilitre of blood. In addition to luminol, two main presumptive tests are available for blood and they work in a similar manner. The haem group can be detected using the colourless reduced dyes Kastle–Meyer (KM) and leuco-malachite green (LMG). If haem is present the colourless substrates are oxidized in the presence of hydrogen peroxide and become coloured. In the case of KM a purple colour develops, and when LMG is used a green colour develops [13, 14]. Any of the tests for blood should be considered as a presumptive test and does not confirm the presence of blood because other naturally occurring compounds, such as plant extracts, coffee and some cleaning fluids, can produce the same colour change or light reaction, thus reducing the specificity of the reaction.

Semen

The positive identification of semen can be extremely important evidence to support an allegation of sexual assault and both presumptive and definitive tests are used. A simple test involves assaying for the presence of the enzyme seminal acid phosphatase that is present in high concentrations in seminal fluid [13]. Other body fluids, such as saliva and vaginal secretions, contain the enzyme albeit in significantly lower concentrations and so can give a positive result [15]. Another marker for the identification of semen is the protein P30 that is a prostate specific antigen (PSA) [16, 17]. The advantage of using PSA compared to the reaction involving acid phosphatase is that PSA is produced independently from the generation of sperm and therefore it can be used for both spermic and azoospermic samples. A definitive test for semen involves treatment with dyes that stain the spermatozoa and allows them to be visualized using a high power microscope; commonly used dyes include haematoxylin-eosin (Figure 3.2b) and Christmas tree stain [18].

Saliva

Saliva is a fluid produced in the mouth to aid in swallowing and the initial stage of digestion. A healthy person produces between 1 and 1.5 litres of saliva every day and can transfer saliva, along with epithelial cells sloughed off from the buccal cavity, in a number of ways. Transfer may be by contact; such as on food products when eating, drinking vessels, cigarette butts, envelopes or in oral sexual assaults. Transfer may also be by aerial deposition of saliva such as on to the front of a mask when worn over the head or onto a telephone when talking into the mouth piece.

Presumptive tests for saliva make use of the enzyme α-amylase which is present at high concentrations and digests starch and complex sugars. The digestion of starch can be measured by the release of dyes that have been covalently linked to insoluble starch molecules [13]. The release of the dye causes a colour change that can easily be detected. Amylases are present in other body fluids such as sweat, vaginal fluid, breast milk and pancreatic secretions; however amylase is present in saliva at concentrations greater than 50 times that in other body fluids.

Epithelial cells

When an object is touched, epithelial cells can be deposited [2]. The amount of cellular material transferred depends upon the amount of time the skin is in contact with the object; the amount of pressure applied; and the presence of fluid such as sweat to mediate the transfer. Some people transfer their skin cells more readily than others; these people are classified as good shedders [5]. This material can be collected from evidential material by swabbing or by tape lifting [11, 19]. Surfaces that the perpetrator(s) of a crime are likely to have had contact with include door handles, the ends of ligatures [20], the handles of weapons and contact marks on victims. These are all potential sources of epithelial cells [21]. In most cases the number of cells is very low and the success rate of DNA profiling is limited. Screening methods, for example using the reagent ninhydrin, which detects the presence of amino acids (and is routinely used to develop latent fingerprints), can be helpful in identifying samples that are likely to contain epithelial cells [22].

Reference samples

In order to identify samples recovered from the scene of crime, reference samples are needed for comparison. Reference samples are provided by a suspect and, in some cases, a victim. Traditionally, blood samples have been taken and these provide an abundant supply of DNA; however, they are invasive and blood samples are a potential health hazard. Buccal swabs that are rubbed on the inner surface of the cheek to collect cellular material have replaced blood samples in many scenarios. In some circumstances plucked hairs may be used but this source of material is not commonly used.

FTA® cards can be used to store both buccal and blood samples (Figure 3.4). The FTA® card is a cellulose based paper which is impregnated with chemicals that cause cellular material to break open – the DNA is released and binds to the card. The chemicals on the card also inhibit any bacterial or fungal growth and DNA can be stably stored on FTA® card for years at room temperature as long as the card remains dry.

Storage of biological material

Biological material collected for DNA analysis should be stored in conditions that will slow the rate of DNA degradation, in particular low temperatures and low humidity. A cool and dry environment limits the action of bacteria and fungi that find biological material a rich source of food and can rapidly degrade biological material.

The exact conditions depend on the nature of the samples and the environment in which the samples are to be stored. Buccal swabs and swabs used to collect material at a crime scene can be stored under refrigeration for short periods and are either frozen directly or dried and then stored at −20 °C for longer term storage. Blood samples will normally be stored at between −20 and −70 °C. Buccal and blood samples collected using FTA® cards can be stored for years at room temperature. Some items of evidence, like clothing, can be stored in a cool dry room; in temperate regions of the world DNA

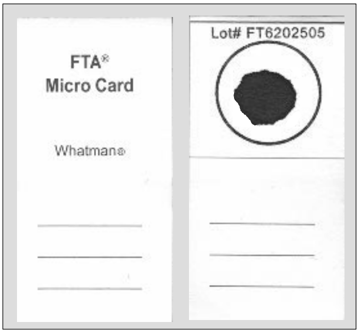


Figure 3.4 FTA cards can be used to store both blood and buccal cells. The cellular material lyses on contact with the card. The DNA binds to the card and is stable for years at room temperature

has been recovered from material stored at room temperature for several years [9]. When samples are not frozen, for example clothing, they are stored in acid-free paper rather than plastic bags, to minimize the build up of any moisture. Once the DNA has been extracted from a sample, the DNA can be stored short term at 4 °C but should be stored at −20 to −70 °C for long term storage.

References

1. Kloosterman, A.D., and Kersbergen, P. (2003) Efficacy and limits of genotyping low copy number DNA samples by multiplex PCR of STR loci. *Progress in Forensic Genetics* **9**, 795–798.
2. van Oorschot, R.A.H., and Jones, M.K. (1997) DNA fingerprints from fingerprints. *Nature* **387**, 767–767.
3. Lee, H.C., and Ladd, C. (2001) Preservation and collection of biological evidence. *Croatian Medical Journal* **42**, 225–228.
4. Lee, H.C. *et al.* (1998) Forensic applications of DNA typing Part 2: Collection and preservation of DNA evidence. *American Journal of Forensic Medicine and Pathology* **19**, 10–18.
5. Rutty, G.N. *et al.* (2003) The effectiveness of protective clothing in the reduction of potential DNA contamination of the scene of crime. *International Journal of Legal Medicine* **117**, 170–174.
6. Soukos, N.S. *et al.* (2000) A rapid method to detect dried saliva stains swabbed from human skin using fluorescence spectroscopy. *Forensic Science International* **114**, 133–138.
7. Stoilovic, M. (1991) Detection of semen and blood stains using polilight as a light-source. *Forensic Science International* **51**, 289–296.
8. Vandenberg, N., and Oorschot, R.A.H. (2006) The use of Polilight® in the detection of seminal fluid, saliva, and bloodstains and comparison with conventional chemical-based screening tests. *Journal of Forensic Sciences* **51**, 361–370.
9. Benecke, M. (2005) Forensic DNA samples–collection and handling. In *Encyclopedia of Medical Genomics and Proteomics* (Fuchs J, and Podda M, Eds), Marcel Dekker, pp. 500–504.
10. Sweet, D. *et al.* (1997) An improved method to recover saliva from human skin: The double swab technique. *Journal of Forensic Sciences* **42**, 320–322.
11. Hall, D., and Fairley, M. (2004) A single approach to the recovery of DNA and firearm discharge residue evidence. *Science and Justice* **44**, 15–19.
12. Juusola, J., and Ballantyne, J. (2003) Messenger RNA profiling: a prototype method to supplant conventional methods for body fluid identification. *Forensic Science International* **135**, 85–96.
13. Ballantyne, J. (2000) Serology. In *Encyclopedia of Forensic Sciences* (Siegel, J.A. *et al.*, Eds), Academic Press, pp.1322–1331.
14. Lee, H. and Pagliaro, E. (2000) Blood identification. In *Encyclopedia of Forensic Sciences* (Siegel, J.A. *et al.*, Eds), Academic Press, pp.13–1338.
15. Steinman, G. (1995) Rapid spot tests for identifying suspected semen specimens. *Forensic Science International* **72**, 191–197.
16. Graves, H.C.B. *et al.* (1985) Postcoital detection of a male-specific semen protein–application to the investigation of rape. *New England Journal of Medicine* **312**, 338–343.
17. Simich, J.P. *et al.* (1999) Validation of the use of a commercially available kit for the identification of prostate specific antigen (PSA) in semen stains. *Journal of Forensic Sciences* **44**, 1229–1231.
18. Allery, J.P. *et al.* (2001) Cytological detection of spermatozoa: Comparison of three staining methods. *Journal of Forensic Sciences* **46**, 349–351.
19. van Oorschot, R.A.H. *et al.* (2003) Are you collecting all the available DNA from touched objects? In *Progress in Forensic Genetics* **9**, 803–807.

20. Bohnert, M. *et al*. (2001) Transfer of biological traces in cases of hanging and ligature strangulation. *Forensic Science International* **116**, 107–115.
21. Wiegand, P., and Kleiber, M. (1997) DNA typing of epithelial cells after strangulation. *International Journal of Legal Medicine* **110**, 181–183.
22. Anslinger, K. *et al*. (2004) Ninhydrin treatment as a screening method for the suitability of swabs taken from contact stains for DNA analysis. *International Journal of Legal Medicine* **118**, 122–124.

4 DNA extraction and quantification

DNA extraction has two main aims: firstly, to be very efficient, extracting enough DNA from a sample to perform the DNA profiling – this is increasingly important as the sample size diminishes – and secondly, to extract DNA that is pure enough for subsequent analysis – the level of difficulty here depends very much on the nature of the sample. Once the DNA has been extracted quantifying the DNA accurately is important for subsequent analysis.

DNA extraction

There are many methods available for extracting DNA. The choice of which method to use depends on a number of factors, including the sample type and quantity; the speed and in some cases ability to automate the extraction procedure [1–4]; the success rate with forensic samples, which is a result of extracting the maximum amount of DNA from a sample and at the same time removing any PCR inhibitors that will prevent successful profiling [1, 5, 6]; the chemicals that are used in the extraction – most laboratories go to great lengths to avoid using hazardous chemicals; and the cost of the procedure. Another important factor is the experience of the laboratory staff.

General principles of DNA extraction

The three stages of DNA extraction can be classified as (i) disruption of the cellular membranes, resulting in cell lysis, (ii) protein denaturation, and finally (iii) the separation of DNA from the denatured protein and other cellular components. Some of the extraction methods commonly used in forensic laboratories are described below.

Chelex® 100 resin

The Chelex® 100 method was one of the first extraction techniques adopted by the forensic community. Chelex® 100 is a resin that is composed of styrene–divinylbenzene

An Introduction to Forensic Genetics W. Goodwin, A. Linacre and S. Hadi
© 2007 John Wiley & Sons, Ltd

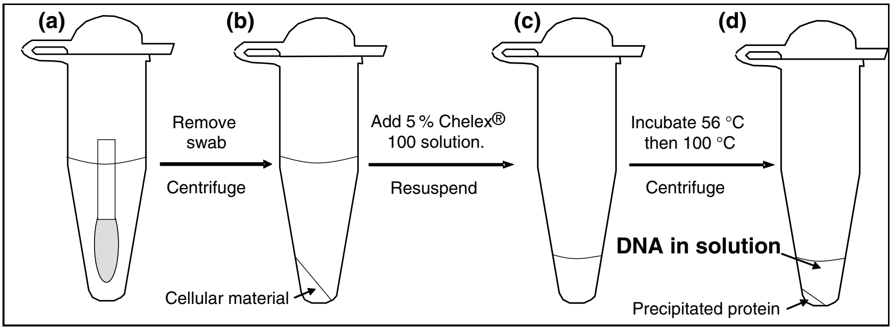


Figure 4.1 The Chelex® 100 extraction is quick and easy to perform. (a) The cellular material is added to 1 ml of TE (1 mM EDTA, 10 mM Tris: pH 8.0) and incubated at room temperature for 10–15 minutes. (b) The tube is centrifuged at high speed to pellet the cellular material and the supernatant is removed. (c) the pellet of cellular material is resuspended in 5 % Chelex®, the tube is incubated at 56 °C for 15–30 minutes and then placed in a boiling water bath for 8 minutes. The tube is centrifuged at high speed for 2–3 minutes to pellet precipitated protein. (d) The supernatant contains the DNA and can be used directly in a PCR

copolymers containing paired iminodiacetate ions [7]. The resin has a very high affinity for polyvalent metal ions, such as magnesium (Mg^{2+}); it chelates the polyvalent metal ions and effectively removes them from solution.

The extraction procedure is very simple, the Chelex® 100 resin, which is supplied as beads, is made into a 5 % suspension using distilled water. The cellular material is incubated with the Chelex® 100 suspension at 56 °C for up to 30 minutes. Proteinase K, which digests most cellular protein, is often added at this point. This incubation is followed by 8–10 minutes at 100 °C to ensure that all the cells have ruptured and that the protein is denatured. The tube is then simply centrifuged to pellet the Chelex® 100 resin and the denatured protein at the bottom of the tube, leaving the aqueous solution containing the DNA to be used in PCR (Figure 4.1). The Chelex® 100 suspension is alkaline, between pH 9.0 and 11.0, and as a result DNA that is isolated using this procedure is single stranded.

The major advantages of this method are: it is quick, taking around 1 hour; it is simple and does not involve the movement of liquid between tubes, thereby reducing the possibility of accidentally mixing samples; the cost is very low; and it avoids the use of harmful chemicals. Importantly, it is amenable to a wide range of forensic samples [7] . The DNA extract produced using this method is relatively crude but sufficiently clean in most cases to generate a DNA profile.

Silica Based DNA Extraction

Within molecular biology generally, the ‘salting out’ procedure has been widely used [8]. The first stage of the extraction involves incubating the cellular material in a lysis buffer that contains a detergent along with proteinase K. The commonly used detergents are sodium dodecyl sulfate (SDS), Tween 20, Triton X-100 and Nonidet P-40. The lysis buffer destabilizes the cell membranes, leading to the breakdown of cellular structure. The addition of a chaotropic salt, for example 6-M guanidine thiocyanate [9] or 6-M sodium chloride, during or after cell lysis, disrupts the protein structure by interfering with hydrogen bonding, Van der Waals interactions, and the hydrophobic interactions. Cellular proteins are largely insoluble in the presence of the chaotropic agent and can be removed by centrifugation or filtration. The reduced solubility of the cellular protein is caused by the excess of ions in the high concentration of salt competing with the proteins for the aqueous solvent, effectively dehydrating the protein. Commonly used commercial kits, for example, the Qiagen kits, exploit the salting-out procedure; the methods to isolate the DNA after the cellular disruption vary widely.

Several DNA extraction methods are based on the binding properties of silica or glass particles. DNA will bind to silica or glass particles with a high affinity in the presence of a chaotropic salt [9, 10]. After the other cellular components have been removed the DNA can be released from the silica/glass particles by suspending them in water. Without the chaotropic salt the DNA no longer binds to the silica/glass and is released into solution. The silica method in particular has been shown to be robust when extracting DNA from forensic samples [1]; it is also amenable to automation [2, 3].

The advantage of the silica based salting-out methods are that they yield high molecular weight DNA that is cleaner than DNA from Chelex® 100 extractions. As with Chelex® 100 extractions, no highly toxic chemicals are involved. The process takes longer than the Chelex® 100 and involves more than one change of tube and so increases the possibility of sample mixing and cross-contamination.

Phenol-chloroform-based DNA extraction

The phenol–chloroform method has been widely used in molecular biology but has been slowly phased out since the mid 1990s, largely because of the toxic nature of phenol. It is still used in some forensic laboratories; in particular it is still widely used for the extraction of DNA from bone samples and soils.

Cell lysis is performed as in the previous method. Phenol–chloroform is added to the cell lysate and mixed and the phenol denatures the protein. The extract is then centrifuged and the precipitated protein forms a pellicle at the interface between the organic phenol–chloroform phase and the aqueous phase; this process is repeated two to three times or until there is no visible pellicle [11]. The DNA is then purified from the aqueous phase by ethanol precipitation or filter centrifugation. The method produces clean DNA but has some drawbacks: in addition to the toxic nature of phenol, multiple tube changes are required and the process is labour intensive.

FTA® paper

In Chapter 3 FTA® paper was described as a method for sample collection and storage, particularly from buccal swabs and fresh blood samples. Once a sample is applied to the FTA® paper it is stable at room temperature for several years. Cellular material lyses on contact with the FTA® paper and the DNA becomes bound to the paper. To analyse the DNA sample, the first step is to take a small region of the card, normally a 2-mm diameter circle, place it into a 1.5-ml tube and the non-DNA components are simply washed off, leaving only DNA on the card. The small circle of FTA® paper is then added directly to a PCR. The FTA paper extractions are very simple to perform and do not require multiple tube changes, thus reducing the possibility of sample mixing.

DNA extraction from challenging samples

The extraction of the many samples encountered in the forensic laboratory, including blood and shed epithelial cells, can be carried out routinely using any of the above techniques. There are however some sample types that require variations on the basic techniques.

Semen

Semen is one of the most commonly encountered types of biological evidence. The extraction of DNA from the spermatozoa is complicated by the structure of the spermatozoa (Figure 4.2). DNA is found within the head of the spermatozoa that is capped by the protective acrosome, which is rich in the amino acid cysteine – a large number of disulphide bridges form between the cysteines in the acrosome. Proteinase K, which is a general proteinase, cannot break the disulphide bonds and this reduces the efficiency of the extraction. The addition of dithiothreitol (DTT), a reducing agent that will break the disulphide bonds, greatly increases the release of spermatozoa DNA [12].

Another complication with semen is that it is often recovered as a mixture of spermatozoa and epithelial cells. The acrosome can be an advantage in these cases as it is possible to perform differential lysis: the epithelial cells are broken down by mild lysis conditions and the spermatozoa can be effectively separated from the lysed epithelial cells [12,13].

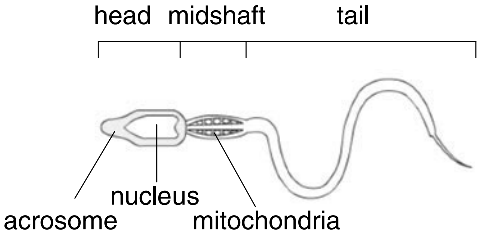


Figure 4.2 The nucleus in the spermatozoa is protected by the acrosome

Hair shafts

Hair shafts that have been pulled out often possess a root that is rich in cellular material and DNA can be extracted using any of the commonly used techniques – plucked roots have been shown to contain as much as 0.5 μg of DNA [14]. Hair that has been shed when it is in the resting telogen phase often contains no cellular material around the root. The hair shafts are composed of keratin, trace metals, air and pigment – cell fragments, including DNA can get trapped in the matrix of the hair and provide enough DNA to produce a profile. However, hair is notoriously difficult to analyse and in many cases it is only possible successfully to profile mitochondrial DNA [14], although nuclear DNA can, in some cases, be recovered [15].

The hair shaft, like the spermatozoa acrosome, is rich in disulphide bridges and requires either mechanical grinding [16] or the addition of a reducing agent such as dithiothreitol [14, 15] that will break the disulphide bonds and allow proteinase K to digest the hair protein and release any trapped nucleic acids. Once released the DNA can be extracted using the salting-out procedure [17] or organic phenol–chloroform based extraction [14–16]. Alternative methods include digestion in a buffer containing proteinase K followed by direct PCR [18,19] or dissolving the hair shaft in sodium hydroxide and, after neutralization, the released DNA is concentrated using filter centrifugation [20].

Because the hair shaft contains very low levels of DNA it is prone to contamination but unlike many other types of biological evidence with low levels of DNA it is possible to clean the hair shaft prior to DNA extraction. Several methods have been used to clean hair including washing in mild detergents, water and ethanol and also using a mild lysis step in the same way as is used in the differential extraction of semen [21].

Hard tissues

Following murders, terrorist attacks, wars and fatal accidents it is desirable to group together body parts from individuals when fragmentation has occurred and ultimately to identify the deceased. If the time between death and recovery of the body is short then muscle tissues provide a rich source of DNA [22], which can be extracted using, for example, any of the Chelex®, salting-out and organic extraction methods. If, however, the soft tissues are displaying an advanced state of decomposition they will not provide any DNA suitable for analysis. When the cellular structure breaks down during decomposition, enzymes that degrade DNA are released and the DNA within the cell is rapidly digested. This process is accelerated by the action of colonizing bacteria and fungi.

Osteocytes are the most common nucleated cells in the bone matrix. In the teeth odontoblasts within the dentine and fibroblasts in the cell rich zone of the pulp cavity provide a source of nucleated cells [23]. The hard tissues of the body, bone and teeth provide a refuge for DNA. In addition to the physical barriers, the hydroxyapatite/apatite mineral, which is a major component of the hard tissues, stabilizes the DNA which becomes closely bound to the positively charged mineral – this interaction limits the action of degrading enzymes [24].

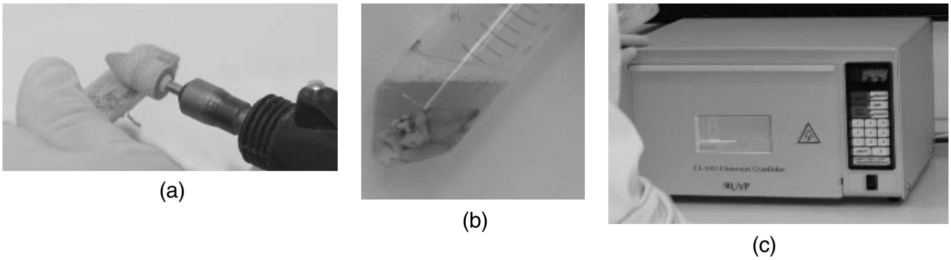


Figure 4.3 Bone and tooth material can be vigorously cleaned using: (a) abrasion to remove the outer surface and (b) washing in detergent and bleach to remove contaminating materials. (c) Exposure to strong UV light introduces thymine dimers into any contaminating exogenous DNA – preventing amplification during PCR

Hard tissues provide an advantage over other forms of biological material because they have a surface that can be cleaned to remove any contaminating DNA by using detergents to remove any soft tissue [25], followed by physical abrasion soaking in sodium hypochlorite (bleach) (Figure 4.3) [26], and exposure to strong ultraviolet light.

After cleaning, the bone/tooth material is normally broken down into a powder by drilling [27] or grinding under liquid nitrogen. The resulting material is decalcified using 0.5-M EDTA either before or at the same time as cell lysis [28]. The organic phenol–chloroform and the silica binding extraction methods are commonly used to extract the DNA [29–34]. The process of extracting DNA from bone samples takes much longer than with any other type of sample.

Quantification of DNA

After extracting DNA an accurate measurement of the amount of DNA and also the quality of the DNA extract is desirable. Adding the correct amount of DNA to a PCR will produce the best quality results in the shortest time; adding too much or not enough DNA will result in a profile that is difficult or even impossible to interpret. This is especially important when profiling forensic samples, when it is very difficult to know the state of preservation of the biological material and also, in many cases, it is difficult to estimate how much cellular material has been collected. It is less important to quantify DNA when using some reference samples – where similar amounts of DNA can be expected to be extracted each time as there are not very many variables. Even so, many laboratories will still quantify the DNA from reference samples as part of their standard analysis. In response to the importance of quantification of samples recovered from the scene of crime, the DNA Advisory Board in the USA adopted rules that made quantification of human DNA mandatory [35].

The quantity of DNA that can be extracted from a sample depends very much on the type of material. Each nucleated cell contains approximately 6 pg of DNA: liquid blood contains 5000–10 000 nucleated blood cells per millilitre; semen contains on average 66 million spermatozoa per millilitre (the average ejaculation produces 2.75 ml of semen) [36]. Biological samples recovered from the scene of crime are not usually in pristine condition and can often consist of a very small number of shed epithelial cells; consequently, the amount of DNA that can be recovered can be extremely low and difficult to quantify.

Visualization on agarose gels

A relatively quick and easy method for assessing both the quantity and quality of extracted DNA is to visualize it on an agarose gel. Agarose is a polymer that can be poured into a variety of gel forms – mini gels approximately 10 cm long are sufficient to visualize DNA. The gel is submerged in electrophoresis buffer and the DNA is loaded into wells that are formed in the gel by a comb; an electric current is applied across the gel and the negatively charged DNA migrates towards the anode. The agarose gel forms a porous matrix and smaller DNA molecules move through the gel more quickly than do larger DNA molecules. Dyes that intercalate with the DNA double helix, such as ethidium bromide [11], can be added to the gel either before or after electrophoresis, the amount of intercalated dye is proportional to the quantity of DNA. An alternative dye, DAPI (4′,6-diamidino-2-phenylindole), can be added directly to the DNA before electrophoresis. This migrates through the gel bound to the minor groove of double stranded DNA [37]. DNA is visualized by placing the gel on a transilluminator that emits UV light at 260 nm. Quantification standards can be run along side the unknown samples to allow the DNA concentrations to be estimated. In addition to showing the presence of DNA, the size of the extracted DNA molecules can also be estimated. High molecular weight DNA can be seen as a single band while degraded DNA or DNA that has been sheared during extraction appears as a smear (Figure 4.4). This makes comparison to the standards difficult as the DNA is spread out over a range rather than in a single band.

The advantages of agarose gel electrophoresis are that it is quick and easy to carry out and also gives an indication of the size of the DNA molecules that have been extracted. The disadvantages are that quantifications are subjective, based on relative

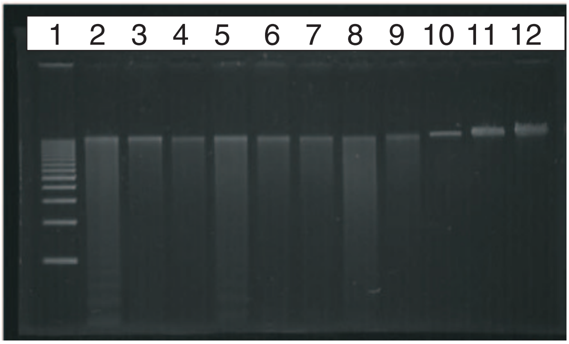


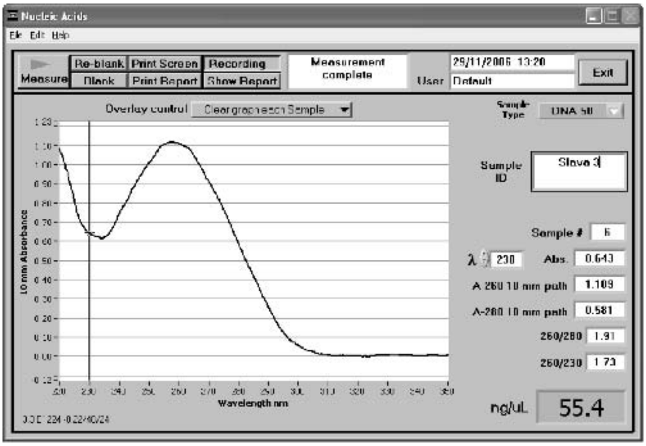
Figure 4.4 DNA run on an agarose gel can be visualized by staining with ethidium bromide and viewing under UV light. The gel shows eight DNA extractions from buccal cells using the QIAamp® Blood DNA extraction kit (Lanes 2–9). Lanes 10–12 contain standards with 25 ng, 50 ng and 75 ng of high molecular weight DNA. A molecular ladder is in lane 1

band intensities; it is difficult to gauge the amounts of degraded DNA as there is no suitable reference standard; total DNA is detected that can be a mixture of human and microbial DNA, and this can lead to over estimates of the DNA concentration; it cannot be used to quantify samples extracted using the Chelex method as this produces single stranded DNA and the fluorescent dyes that intercalate with the double stranded DNA do not bind to the single stranded DNA.

Ultraviolet spectrophotometry

DNA absorbs light maximally at 260 nm. This feature can be used to estimate the amount of DNA in an extract and by measuring a range of wavelengths from 220 nm to 300 nm it is also possible to assess the amount of carbohydrate (maximum absorbance 230 nm) and protein (maximum absorbance 280 nm) that may have co-extracted with the sample. The DNA is placed in a quartz cuvette and light is shone through; the absorbance is measured against a standard. A clean DNA extract will produce a curve as shown in Figure 4.5; if the DNA extract is clean, the ratio of the absorbance at 260 nm and 280 nm should be between 1.8 and 2.0 [11].

Spectrophotometry is commonly used for quantification in molecular biology laboratories but has not been widely adopted by the forensic community. The major disadvantage is that it is difficult to quantify small amounts of DNA accurately using spectrophotometry, it is not human specific and other chemicals, for example, dyes from clothing and humic acids from bone samples, can interfere with the analysis.


Nucleic Acids
File Edit Help
Measure
Re-blank
Blank
Print Screen
Print Report
Recording
Show Report
Exit
10 mm Absorbance
1.10
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
220
230
240
250
260
270
280
290
300
310
320
330
340
350
Wavelength nm

Measurement complete

29/11/2006 13:20

User Default

Overlay control Clear graph each Sample

Sample Type DNA 50

Sample ID Slava 3

Sample # 6

λ 230 Abs. 0.643

A 260 10 mm path 1.109

A-280 10 mm path 0.581

260/280 1.91

260/230 1.73

ng/uL 55.4

Figure 4.5 UV absorbance by a solution containing DNA is maximal at 260 nm. The 260:280 ratio of 1.91 indicates that the extract is not contaminated with proteins

Fluorescence spectrophotometry

Ethidium bromide or DAPI (4′,6-diamidino-2-phenylindole) can be used to visualize DNA in agarose gels – these are both examples of chemicals that fluoresce at much higher levels when they intercalate with DNA. In addition to staining agarose gels, fluorescent dyes can also be used as an alternative to UV spectrophotometry for DNA quantitation. A range of dyes has been developed that can be used with fluorescent microplate readers and these are very sensitive. The PicoGreen® dye is specific for double stranded DNA and can detect as little as 25 pg/ml of DNA [35]. When PicoGreen® binds to DNA the fluorescence of the dye increases over 1000-fold – ethidium bromide in comparison increases in fluorescence 50–100-fold when it intercalates with double stranded DNA [35]. PicoGreen® is very sensitive and is a powerful technique for quantifying total DNA – it does however have the drawback that it is not human specific.

Hybridization

Hybridization based quantification methods have been widely used in forensic genetics since the early 1990s, in particular a commercially available kit Quantiblot® (Applied Biosystems). Extracted DNA is applied to a positively charged nylon membrane using a slot or dot blot process; the DNA is then challenged with a probe that is specific to human DNA. A commonly used target is the D17S1 alpha satellite repeat that is on human chromosome 17 in 500–1000 copies. The probe can be labelled in a number of ways including colorimetric and chemiluminescent.

A series of standards is applied to the membrane, and comparison of the signal from the extracted DNA with the standards allows quantification. The advantage of hybridization-based methods is that the quantification is human specific – agarose gel electrophoresis and spectrophotometry detect total DNA and forensic samples that have been exposed to the environment for any length of time are prone to colonization by bacteria and fungi.

The hybridization systems do suffer from a lack of sensitivity. For samples producing a negative result in many cases it is still possible to generate a profile after PCR. The analysis of the results is also subjective, leading different operators to come to different conclusions. In addition to the limited sensitivity, the process is labour intensive, taking approximately 2 hours to produce the blot. Hybridization-based methods are being widely replaced by real-time PCR systems.

Real-time PCR

When generating a DNA profile, the PCR products are normally analysed at the end point after 28–34 cycles. It is, however, possible to monitor the generation of PCR products as they are generated – real time. This was first developed using ethidium bromide: as PCR products are generated in each cycle, more ethidium bromide intercalates with the double-stranded DNA molecule and fluoresces under UV light. The increase in fluorescence can be detected using a suitable ‘camera’ [38]. Increasingly sensitive

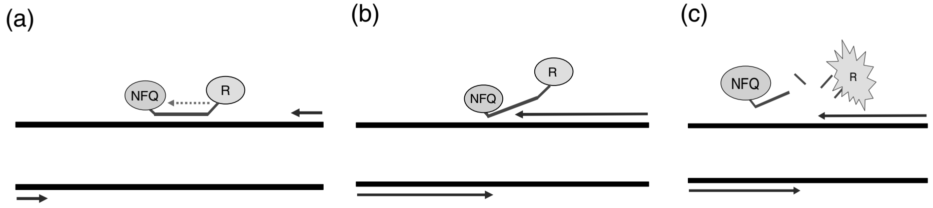


Figure 4.6 (a) The TaqMan® quantification system consists of two PCR primers and an internal probe that hybridizes within the region that is the amplified region; (b) as the primer extends it encounters the probe, the 5′ exonuclease activity of the *Taq* polymerase degrades the probe: (c) the reporter molecule is no longer in proximity to the quencher and fluoresces

assays have been developed, such as SYBR® Green and the TaqMan® system. Using SYBR® Green, as PCR products are generated, the dye binds to the double-stranded product and the fluorescence increases. The TaqMan® system uses a different approach, with two primers and a probe; the probe is within the region defined by the primers and is labelled on the 5′ end with a fluorescent molecule and on the 3′ end with a molecule that quenches the fluorescence. As the primers are extended by the *Taq* polymerase, one of them meets the probe, which is degraded by the polymerase, releasing the probe and the quencher into solution – efficient quenching of the fluorescent molecules only occurs when they are in close proximity on the probe molecule (Figure 4.6).

As more PCR products are generated, more fluorescent molecules are released and the fluorescence from the sample increases. Real-time assays are highly sensitive, human specific and are not labour intensive.

DNA IQ™ system

A novel approach to quantification is used in the commercially available DNA IQ™ Isolation System (Promega Corporation). The isolation method is based on salting-out and binding to silica: a very specific amount of silica coated beads is added to the extraction and these bind a maximum amount of DNA; therefore, when the DNA is eluted from the beads the maximum concentration is known. It has the advantage of combining the extraction and quantification steps but has the disadvantage of not being human specific.

References

1. Castella, V. *et al*. (2006) Forensic evaluation of the QIAshredder/QIAamp DNA extraction procedure. *Forensic Science International* **156**, 70–73.
2. Greenspoon, S.A. *et al*. (2004) Application of the BioMek (R) 2000 laboratory automation workstation and the DNA IQ (TM) system to the extraction of forensic casework samples. *Journal of Forensic Sciences* **49**, 29–39.
3. Montpetit, S.A. *et al*. (2005) A simple automated instrument for DNA extraction in forensic casework. *Journal of Forensic Sciences* **50**, 555–563.
4. Moss, D. *et al*. (2003) An easily automated, closed-tube forensic DNA extraction procedure using a thermostable proteinase. *International Journal of Legal Medicine* **117**, 340–349.

5. Greenspoon, S.A. *et al.* (1998) QIAamp spin columns as a method of DNA isolation for forensic casework. *Journal of Forensic Sciences* **43**, 1024–1030.
6. Vandenberg, N., and van Oorschot, R.A.H. (2002) Extraction of human nuclear DNA from feces samples using the QIAamp DNA stool mini kit. *Journal of Forensic Sciences* **47**, 993–995.
7. Walsh, P.S. *et al.* (1991) Chelex-100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *Biotechniques* **10**, 506–513.
8. Aljanabi, S.M., and Martinez, I. (1997) Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Research* **25**, 4692–4693.
9. Boom, R. *et al.* (1990) Rapid and Simple Method for Purification of Nucleic-Acids. *Journal of Clinical Microbiology* **28**, 495–503.
10. Vogelstein, B., and Gillespie, D. (1979) Preparative and analytical purification of DNA from agarose. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 615–619.
11. Sambrook, J. *et al.* (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press.
12. Gill, P. *et al.* (1985) Forensic application of DNA fingerprints. *Nature* **318**, 577–579.
13. Wiegand, P. *et al.* (1992) DNA extraction from mixtures of body fluid using mild preferential lysis. *International Journal of Legal Medicine* **104**, 359–360.
14. Higuchi, R. *et al.* (1988) DNA typing from single hairs. *Nature* **332**, 543–546.
15. Hellmann, A. *et al.* (2001) STR typing of human telogen hairs – a new approach. *International Journal of Legal Medicine* **114**, 269–273.
16. Wilson, M.R. *et al.* (1995) Extraction, PCR amplification and sequencing of mitochondrial-DNA from human hair shafts. *Biotechniques* **18**, 662–669.
17. Baker, L.E. *et al.* (2001) A silica-based mitochondrial DNA extraction method applied to forensic hair shafts and teeth. *Journal of Forensic Sciences* **46**, 126–130.
18. Allen, M. *et al.* (1998) Mitochondrial DNA sequencing of shed hairs and saliva on robbery caps: Sensitivity and matching probabilities. *Journal of Forensic Sciences* **43**, 453–464.
19. Vigilant, L. (1999) An evaluation of techniques for the extraction and amplification of DNA from naturally shed hairs. *Biological Chemistry* **380**, 1329–1331.
20. Graffy, E.A., and Foran, D.R. (2005) A simplified method for mitochondrial DNA extraction from head hair shafts. *Journal of Forensic Sciences* **50**, 1119–1122.
21. Jehaes, E. *et al.* (1998) Evaluation of a decontamination protocol for hair shafts before mtDNA sequencing. *Forensic Science International* **94**, 65–71.
22. Bar, W. *et al.* (1988) Postmortem stability of DNA. *Forensic Science International* **39**, 59–70.
23. Gaytmenn, R., and Sweet, D. (2003) Quantification of forensic DNA from various regions of human teeth. *Journal of Forensic Sciences* **48**, 622–625.
24. Lindahl, T. (1993) Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715.
25. Arismendi, J.L. *et al.* (2004) Effects of processing techniques on the forensic DNA analysis of human skeletal remains. *Journal of Forensic Sciences* **49**, 930–934.
26. Kemp, B.M., and Smith, D.G. (2005) Use of bleach to eliminate contaminating DNA from the surface of bones and teeth. *Forensic Science International* **154**, 53–61.
27. Ginther, C. *et al.* (1992) Identifying individuals by sequencing mitochondrial-DNA from teeth. *Nature Genetics* **2**, 135–138.
28. Fisher, D.L. *et al.* (1993) Extraction, evaluation, and amplification of DNA from decalcified and undecalcified United States Civil-War bone. *Journal of Forensic Sciences* **38**, 60–68.
29. Bille, T. *et al.* (2004) Novel method of DNA extraction from bones assisted DNA identification of World Trade Center victims. *Progress in Forensic Genetics* **10**, 553–555.
30. Crainic, K. *et al.* (2002) Skeletal remains presumed submerged in water for three years identified using PCR-STR analysis. *Journal of Forensic Sciences* **47**, 1025–1027.
31. Goodwin, W. *et al.* (1999) The use of mitochondrial DNA and short tandem repeat typing in the identification of air crash victims. *Electrophoresis* **20**, 1707–1711.
32. Goodwin, W. *et al.* (2003) The identification of a US serviceman recovered from the Holy Loch, Scotland. *Science and Justice* **43**, 45–47.

33. Holland, M.M. *et al*. (2003) Development of a quality, high throughput DNA analysis procedure for skeletal samples to assist with the identification of victims from the world trade center attacks. *Croatian Medical Journal* **44**, 264–272.
34. Holland, M.M. *et al*. (1993) Mitochondrial-DNA sequence-analysis of human skeletal remains – identification of remains from the vietnam war. *Journal of Forensic Sciences* **38**, 542–553.
35. Nicklas, J.A., and Buel, E. (2003) Quantification of DNA in forensic samples. *Analytical and Bioanalytical Chemistry* **376**, 1160–1167.
36. Carlsen, E. *et al*. (1992) Evidence for decreasing quality of semen during past 50 years. *British Medical Journal* **305**, 609–613.
37. Buel, E. (1995) The use of DAPI as a replacement for ethidium bromide in forensic DNA analysis. *Journal of Forensic Sciences* **40**, 275–278.
38. Higuchi, R. *et al*. (1993) Kinetic PCR analysis – real-time monitoring of DNA amplification reactions. *Bio-Technology* **11**, 1026–1030.

5 The polymerase chain reaction

In 1985, the year of the first DNA 'fingerprint', a new method – the polymerase chain reaction (PCR) was reported [1–3]. The PCR can amplify a specific region of DNA and it has revolutionized all areas of molecular biology, including forensic genetics. The technique allows extremely small quantities of DNA to be amplified. Under optimal conditions, DNA can be amplified from a single cell [4, 5]. The increased sensitivity of DNA profiling using PCR has had a dramatic effect on the types of forensic sample that can be used and it is now possible to analyse trace evidence and highly degraded samples successfully – albeit with less than 100 % success.

The evolution of PCR-based profiling in forensic genetics

PCR technology was rapidly incorporated into forensic analysis. The first PCR based tool for forensic casework amplified the polymorphic HLA DQα locus (the α subunit of the DQ protein is part of the major histocompatibility complex) [6]. It was used for the first time in casework in 1988 to analyse the skeletal remains of a 3-year-old girl [7, 8]. The DQα system's major drawback was that it had a limited power of discrimination.

VNTRs were widely used in casework but required a relatively large amount of DNA. In an attempt to overcome this limitation, PCR technology was applied to the analysis of VNTR loci, and alleles between 5–10 kb could be faithfully amplified from fresh biological material [9]. However, it was of limited value for many forensic samples, which often contained small amounts of DNA that could be highly degraded. To overcome the problems caused by degradation, tandem repeats, called AMP-FLPs (amplified fragment length polymorphisms), that were smaller than 1 kb were selected for PCR based analysis [10–15]. However, as with VNTRs, their use was limited in forensic contexts because of the size of the larger alleles, which were difficult to analyse in degraded samples. By the early 1990s, a large number of short tandem repeats (STRs) had been characterized [16]. The STR loci were simpler and shorter than VNTRs and AMP-FLPs, and were more suitable for the analysis of biological samples recovered from crime scenes [17]. The STR markers were not individually as discriminating as the VNTR and AMP-FLPs but had a major advantage that several of them could be analysed together in a multiplex reaction. The short tandem repeat markers have become

An Introduction to Forensic Genetics W. Goodwin, A. Linacre and S. Hadi
© 2007 John Wiley & Sons, Ltd

the genetic polymorphism of choice in forensic genetics and the PCR is a vital part of the analytical process.

DNA replication – the basis of the PCR

The PCR takes advantage of the enzymatic processes of DNA replication. During every cell cycle the entire DNA content of a cell is duplicated. This copying of DNA can be replicated outside of the cell *in vitro* to amplify specific regions of DNA.

The components of PCR

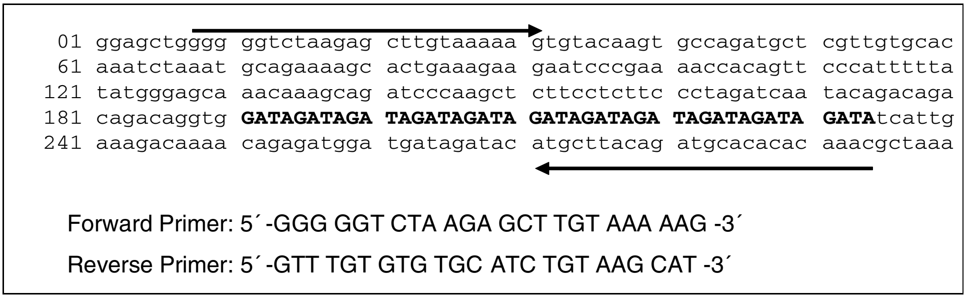
A PCR has the following components: template DNA, at least two primers, a thermostable DNA polymerase such as *Taq* polymerase, magnesium chloride, deoxynucleotide triphosphates and a buffer.

Template DNA

The amount of DNA added to a PCR depends on the sensitivity of the reaction: for most forensic purposes the PCR is highly optimized so that it will work with low levels of template. Most commercial kits require between 0.5 and 2.5 ng of extracted DNA for optimum results. This represents between 166 and 833 copies of the haploid human genome – one copy of the human genome contains approximately 3 pg of DNA. Most forensic profiling can be carried out successfully with fewer templates – even below 100 pg or 33 copies of the genome; however, the interpretation of profiles can become more complex as the amount of template DNA is reduced.

Taq *DNA polymerase*

The first PCRs were carried out using a DNA polymerase that was isolated from *E. coli*; in each cycle of the PCR the enzyme was inactivated by the high temperatures in the denaturation phase and fresh enzyme had to be added [1–3]. Fortunately this is no longer necessary. Scientists were able to isolate the DNA polymerases from the thermophilic bacteria, *Thermus aquaticus* [18], which was discovered in the 1960s in the hot springs of Yellowstone National Park, USA. The *Taq* polymerase enzyme can tolerate the high temperatures that are involved in the PCR and works optimally at 72–80 °C [19]. Using the thermostable enzyme greatly simplifies the PCR procedure and also increases the specificity, sensitivity and yield of the reaction [19]. The *Taq* polymerase enzyme exhibits significant activity at room temperature that can lead to the creation of non-specific PCR products; adding the enzyme to a pre-heated ‘hot start’ reaction reduces the non-specific binding and again improves the specificity and yield of a PCR [20]. Modifications to the commonly used *Taq* polymerase led to the development of the AmpliTaq Gold® polymerase (Applied Biosystems). The enzyme



01 ggagctgggg ggtctaagag cttgtaaaaa gtgtacaagt gccagatgct cgttgtgcac
61 aaatctaaat gcagaaaagc actgaaagaa gaatcccgaa aaccacagtt cccattttta
121 tatgggagca aacaaagcag atcccaagct cttcctcttc cctagatcaa tacagacaga
181 cagacaggtg **GATAGATAGA TAGATAGATA GATAGATAGA TAGATAGATA GATA**tcattg
241 aaagacaaaa cagagatgga tgatagatac atgcttacag atgcacacac aaacgctaaa

Forward Primer: 5′ -GGG GGT CTA AGA GCT TGT AAA AAG -3′

Reverse Primer: 5′ -GTT TGT GTG TGC ATC TGT AAG CAT -3′

Figure 5.1 The forward and reverse primers that are used to amplify the STR locus D16S539 in the Promega Powerplex™ 1.2 kit are shown and the position that they bind to within the sequence is indicated by the arrows. The target sequence is shown in bold – the position of the primers around either side of the target determines the length of the PCR product. The example shown contains eleven repeats of the core GATA sequence

is inactive when it is first added to the PCR – it only becomes active after incubation at 95 °C for approximately 10 minutes [21]. This 'hot start' enzyme allows the PCR to start at an elevated temperature and minimizes the non-specific binding that can occur at lower temperatures [22].

Primers

The primers used in PCR define the region of the genome that will be analysed. Primers are short synthetic pieces of DNA that anneal to the template molecule either side of the target region. The primer sequences are therefore limited to some degree by the DNA sequence that flanks the target sequence. Figure 5.1 illustrates the sequence and positioning of two primers that are used to amplify the D16S539 locus in one of the commercially available kits for analysing STR loci (Promega Corporation).

When designing primers for forensic analysis it is important that they will bind to conserved regions of DNA and therefore effectively amplify human DNA from all populations [23] while at the same time not binding to the DNA of other species. When designing a multiplex PCR, the allelic size ranges are also important considerations for the position of the primer binding sites.

There are a number of basic guidelines for primer design. Primers are normally between 18 and 30 nucleotides long, and have a balanced number of G/C and A/T nucleotides. A primer should not be self complementary or be complementary to any of the other primers that are in the reaction. Self complementary regions will result in the primer pairing with itself to form a loop, whereas primers that are complementary will bind to each other to form primer dimers.

The temperature at which primers anneal to the template DNA depends upon their length and sequence – most primers are designed to anneal between 50 and 65 °C. A basic rule of thumb can be used when designing primers to estimate the melting temperature: for each A or T in the primer 2 °C is added to the melting temperature and for each C or G 4 °C is added – Cs and Gs will bind to the complementary nucleotide with three hydrogen bonds and are therefore more thermodynamically stable. To estimate the annealing temperature 5 °C is subtracted from the melting temperature. PCR primers can be designed manually or with help using software such as Oligo [24] and Primer3 [25].

Magnesium chloride, nucleotide triphosphates and reaction buffer

Magnesium chloride is a critical component of the PCR. The primers bind to the template DNA to form a primer–template duplex: magnesium chloride stabilizes the interaction. The concentration of $MgCl_2$ is typically between 1.5 mM and 2.5 mM; the template–primer stability increases with higher concentrations of $MgCl_2$. The *Taq* polymerase also requires magnesium to be present in order to function.

The building blocks for the PCR are deoxynucleotide triphosphates, which are incorporated into the nascent DNA strand during replication. The four nucleotides are in the PCR in equal concentration, normally 200 μM. The reaction buffer maintains optimal pH and salt conditions for the reaction.

The PCR process

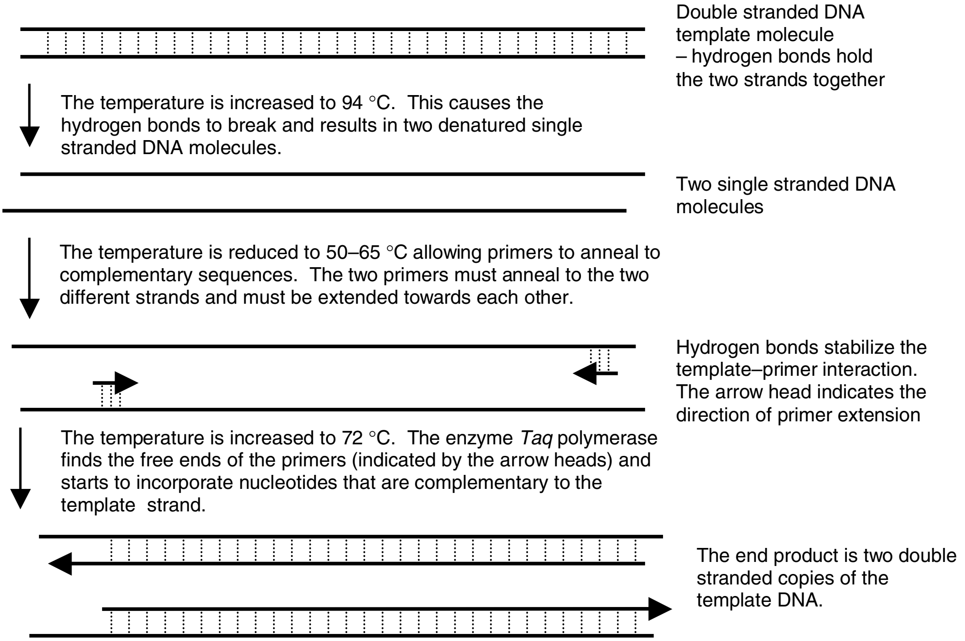
The PCR amplifies specific regions of template DNA. The power of the technique is illustrated in Table 5.1 In theory, a single molecule can be amplified one billion-fold by 30 cycles of amplification; in practice the PCR is not 100 % efficient but does still produce tens of millions of copies of the target sequence [19].

The amplification of DNA occurs in the cycling phase of PCR, which consists of three stages (Figure 5.2): denaturation, annealing and extension. In the denaturation stage the reaction is heated to 94 °C; this causes the double stranded DNA molecule to 'melt' forming two single stranded molecules. DNA melts at this temperature because the hydrogen bonds that hold the two strands of the DNA molecule together are relatively weak. As the temperature is lowered, typically to between 50 and 65 °C, the oligonucleotide primers anneal to the template. The primers are in molar excess to the template strands and bind to the complementary sequences before the template DNA reassociates to form double stranded DNA. After the primers have annealed the temperature is increased to 72 °C, which is in the optimum temperature range for the *Taq* polymerase. Nucleotides are added to the nascent DNA strand at the rate of approximately 40–60 per second [26, 27]. The enzyme catalyses the addition of nucleotides to the $3'$ ends of the primers using the original DNA strand as a template; it has a high processivity, catalysing the addition of approximately 50 nucleotides to the nascent DNA strand before the enzyme dissociates – several *Taq* polymerase enzymes will associate and disassociate during the extension phase of longer PCR products.

The normal range of cycles for a PCR is between 28 and 32. In extreme cases, where the amount of target DNA is very low the cycle number can be increased to up to 34 cycles. It has been demonstrated that going above this cycle number does

Table 5.1 The PCR can theoretically multiply DNA over 1 billion-fold after 32 cycles – in reality it is not 100 % efficient but is still extremely powerful. (The AmpF*l*STR® and PowerPlex® STR kits are described in Chapter 6)

Cycle	Number of PCR products	Comments
1	0	
2	0	
3	2	
4	4	
5	8	
6	16	
7	32	
8	64	
9	128	
10	256	
20	262 144	
28	67 108 864	Standard cycle number using the AmpF*l*STR® SGM Plus® and Identifiler® kits
30	268 435 456	
32	1 073 741 824	Standard cycle number using the PowerPlex® 16 kit
34	4 294 967 296	Maximum number of cycles normally used in forensic analysis



Double stranded DNA template molecule – hydrogen bonds hold the two strands together

The temperature is increased to 94 °C. This causes the hydrogen bonds to break and results in two denatured single stranded DNA molecules.

Two single stranded DNA molecules

The temperature is reduced to 50–65 °C allowing primers to anneal to complementary sequences. The two primers must anneal to the two different strands and must be extended towards each other.

Hydrogen bonds stabilize the template–primer interaction. The arrow head indicates the direction of primer extension

The temperature is increased to 72 °C. The enzyme *Taq* polymerase finds the free ends of the primers (indicated by the arrow heads) and starts to incorporate nucleotides that are complementary to the template strand.

The end product is two double stranded copies of the template DNA.

Figure 5.2 The PCR process – each PCR cycle consists of three phases: denaturing, annealing and extension

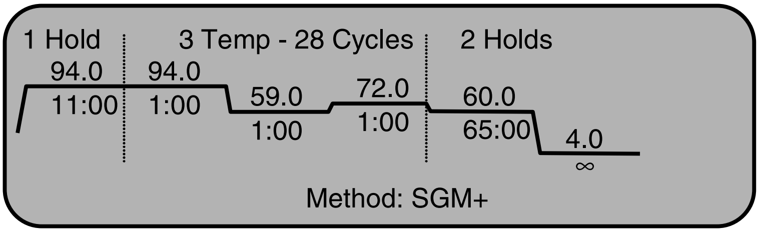


Figure 5.3 A typical PCR programme involves three phases: a pre-incubation at 94 °C, which activates the AmpliTaq Gold® polymerase; the cycling phase; and a terminal incubation that maximizes the non-template addition at the end of the amplification

not increase the likelihood of obtaining a profile but does increase the probability of artefacts forming during the PCR [28]. Using 34 cycles is known as low copy number (LCN) PCR and it is sparingly employed as extreme precautions have to be taken to reduce the chance of contamination – the more cycles the higher the chance of detecting contaminating DNA. The interpretation of the profiles generated using a high cycle number also become more complex.

Following the cycling phase the reaction is incubated between 60 and 72 °C for up to one hour. In addition to the template-dependent synthesis of DNA the *Taq* polymerase also adds an additional residue to the 3′ end of extended DNA molecule, this is non-template dependent; the incubation at the end of the reaction is to ensure that the non-template addition is complete. The conditions of a typical PCR are shown below in Figure 5.3.

The PCR requires tightly controlled thermal conditions and these are achieved by using a thermocycler. This consists of a conducting metal block that contains heating and cooling elements with wells that accommodate the plastic reaction tubes. The temperature of the PCR block is controlled by a small microprocessor. Most thermocyclers also contain a lid that is heated to over 100 °C; this prevents the reaction evaporating and condensing on the cooler lid and thereby maintains the reaction volume, thus keeping the concentration of the reaction components stable throughout the PCR.

After amplification the results of a PCR can be visualized on an agarose gel. In a reaction that amplifies only one locus a single band should be detected (Figure 5.4).

PCR inhibition

When analysing forensic samples a problem that can be encountered is inhibition of the PCR [29]. DNA extraction methods do not produce pure DNA, some chemicals will co-purify and in some cases inhibit the *Taq* polymerase. Potent inhibitors include haem compounds from blood [30–32], bile salts and complex polysaccharides from faeces [33, 34], humic substances from soil [35] and urea from urine [36]. High concentrations of ions, in particular calcium and magnesium, can also act as potent inhibitors of the *Taq* polymerase [37]. EDTA is used in high concentrations for the isolation of DNA

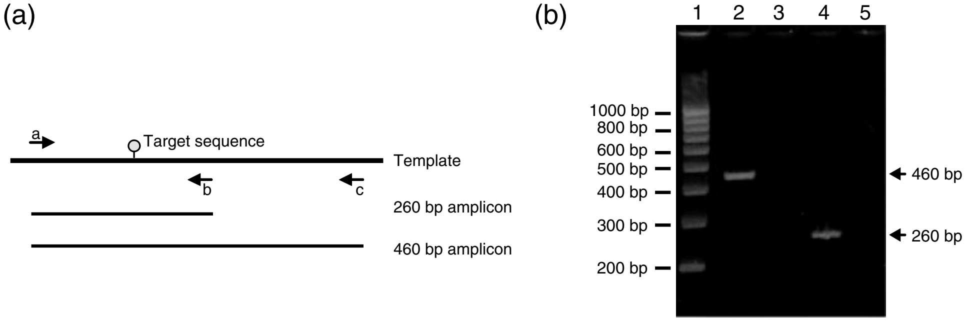


Figure 5.4 (a) The target sequence on the template DNA molecule has been amplified using two different primer pairs: using primers a + c generates a 460 bp product and the a + b primer pair leads to a 260 bp product. (b) Following amplification of template DNA in two separate reactions the products were separated on a 2 % agarose gel and stained with ethidium bromide. The amplified products can be seen in lane 2 (460 bp) and lane 4 (260 bp). Lane 1 contains a 100 bp ladder and lanes 3 and 5 contain negative controls

from bone and will inhibit PCR unless removed as it binds ions such as magnesium ions that are essential for PCR [38]. In a forensic science context, the blue dye in clothing such as denim, called indigo, has an inhibitor effect on PCR [39].

Extraction methods have been developed to remove commonly encountered PCR inhibitors and, for example, the silica binding methods that are commonly used in forensic analysis are effective at removing most inhibitors whereas the methods that produce a cruder extract such as the Chelex® resin are more prone to inhibition. When it is not possible to remove all the potential inhibitors from a DNA extract, the addition of the protein bovine serum albumin (BSA) to the PCR can in many cases prevent or reduce the inhibition of the *Taq* polymerase. The BSA acts as a binding site for some inhibitors and can competitively remove or reduce the concentration of the inhibitor [30, 38]. The action of inhibitors can be detected, for example by spiking a PCR with a known amount of DNA, this alerts the analyst that further purification steps are required [40].

Sensitivity and contamination

The great advantage of PCR is that it will amplify DNA from a template of only a few cells. This high level of sensitivity can also be a potential disadvantage, as DNA from incidental sources can be present and contamination can be introduced. Throughout the handling and analysis of DNA samples extreme care needs to be taken to minimize the chance of introducing this extraneous DNA.

When samples are collected from the scene of an incident, there may be cellular material from persons who had been present at the scene prior to the incident and hence DNA profiles will be generated from people unconnected with the incident. This type of DNA can be termed as incidental as it is not a contamination of the samples. At the time of the incident there is an opportunity for transfer from the perpetrator and it is this cellular material that is pertinent to the investigation. Consider an event such as theft from a house. Prior to the incident there will be cellular material from the owners and from any recent visitors. At the time of the break in there may be transfer from the thief. If the incident is discovered by a neighbour then they will introduce their cellular material after the incident and prior to the scene being secured. When the police are called they have the potential to introduce their cellular material. Once the scene is secured then those entering should be wearing full protection to minimize the opportunity for transfer of their cellular material [41, 42]. If there is any introduction of DNA from those at the crime scene, during collection and transportation, or from laboratory staff, then this is considered as contamination.

The PCR laboratory

Once evidential samples have reached the forensic laboratory there is further potential to introduce contamination. A fundamental feature of PCR laboratories, to reduce the possibility of introducing contamination, is that they are clearly divided into pre- and post-PCR areas (Figure 5.5).

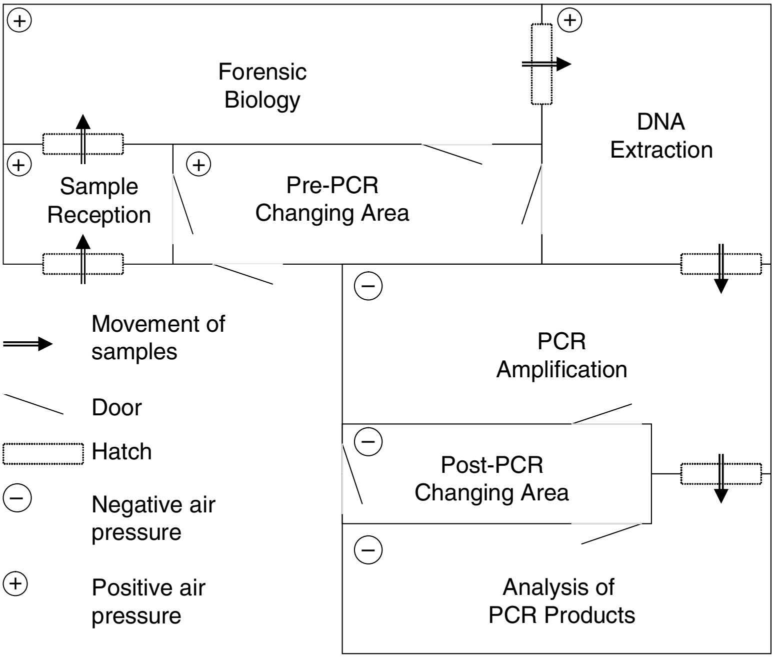
Pre-PCR

Once the samples reach the laboratory, potential contamination comes from the reagents, equipment and the forensic scientists undertaking the analysis. To prevent contamination being introduced from the scientist, protective clothing is worn, including a lab coat, gloves, a face mask, safety glasses/visor and a head cover. Even with these precautions it is still possible to get the scientist's DNA profile showing up – a database of all the people who enter/work in the laboratory can be used to detect when contamination could have been introduced within the laboratory.

When laboratories are engaged in analysing both samples from suspects and from crime scenes it is common to have dedicated areas for the two classes of sample; this prevents any potential cross-contamination of crime scene and suspect DNA. Special dedicated facilities may also be used when dealing with samples that contain very small amounts of DNA, such as hair shafts.

DNA extraction and PCR set-up are commonly carried out in specialized clean hoods that provide a very controlled environment. The hoods have stainless steel surfaces and are easy to keep clean; they have filtered air to prevent any dust or other contaminant getting into the reaction, and they are fitted with a UV-light source that is used to remove any contaminating DNA effectively. The pipetting of any liquids involved in the extraction and PCR set-up is performed using pipette tips with barriers to prevent any DNA carry over.

During the DNA extraction process negative control extractions must always be carried out to monitor for contamination; positive controls that involve extracting material similar to the casework samples, for example buccal swabs or blood stains, can be carried out to monitor that the extraction and amplification procedures are working



Forensic Biology
DNA Extraction
Sample Reception
Pre-PCR Changing Area
PCR Amplification
Post-PCR Changing Area
Analysis of PCR Products
Movement of samples
Door
Hatch
Negative air pressure
Positive air pressure

Figure 5.5 The PCR laboratory is designed so that the work flows through the different processes in one direction starting with sample reception and forensic biology and finishing with the post-PCR analysis. The samples are passed through air-lock hatches to minimize the possibility of any material being transferred from post-PCR to pre-PCR areas. Access to the pre- and post-PCR laboratories is through different changing areas and dedicated staff will work in either pre- or post-PCR areas. Positive air pressure in pre-PCR areas and negative pressure in post-PCR rooms also reduces the possibility of introducing any contamination into the pre-PCR areas

efficiently. The PCR set-up introduces another positive and negative control: the positive control involves setting-up a PCR with DNA of known origin and whose profile is known. Successful analysis demonstrates that the reaction worked. In the negative control PCR, water replaces the DNA to monitor for contamination in the reagents or introduced during the PCR set up.

Post-PCR

The most potent source of contamination is previously amplified PCR products. Following a PCR there are millions of copies of the target sequence that can potentially contaminate subsequent reactions. Each time a PCR tube is opened there is some aerosol spray and a single droplet of aerosol will contain thousands of copies of the amplified target, resulting in transfer of some of the amplified product. The fundamental feature of any laboratory that engages in PCR analysis is that there must be physical separation of the pre-PCR and the post-PCR analysis to minimize the possibility of contaminating

DNA extractions and PCR set-ups with amplified material. In addition to the two physical spaces there should also be dedicated equipment, protective clothing and reagents for each area. There must be a unidirectional work flow through the laboratory – PCR products must never be brought back into the pre-PCR part of the laboratory. There must also be temporal separation of tasks – it is not possible for a scientist who has been working in the post-PCR to then work in the pre-PCR area without the possibility of introducing contamination; an overnight break before returning to the pre-PCR area is normally recommended. Larger laboratories will have scientists who are dedicated to only the pre- or the post-PCR analysis.

Further reading

Dieffenbach C.W., and Dveksler G.S. (2003) *PCR Primer: A Laboratory Manual.* Second Edition. Cold Spring Harbor Laboratory Press.

References

1. Mullis, K., *et al.* (1986) Specific enzymatic amplification of DNA *in vitro* – the polymerase chain-reaction. *Cold Spring Harbor Symposia on Quantitative Biology* **51**, 263–273.
2. Mullis, K.B., and Faloona, F.A. (1987) Specific synthesis of DNA *in vitro* via a polymerase-atalyzed chain-reaction. *Methods in Enzymology* **155**, 335–350.
3. Saiki, R.K., *et al.* (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle-cell anemia. *Science* **230**, 1350–1354.
4. Li, H.H., *et al.* (1990) Direct electrophoretic detection of the allelic state of single DNA-molecules in human sperm by using the polymerase chain-reaction. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 4580–4584.
5. Li, H.H., *et al.* (1988) Amplification and analysis of DNA-sequences in single human-sperm and diploid-cells. *Nature* **335**, 414–417.
6. Saiki, R.K., *et al.* (1986) Analysis of enzymatically amplified beta-globin and HLA-DQ-*ά* DNA with allele-specific oligonucleotide probes. *Nature* **324**, 163–166.
7. Stoneking, M., *et al.* (1991) Population variation of human MtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. *American Journal of Human Genetics* **48**, 370–382.
8. Blake, E., *et al.* (1992) Polymerase chain-reaction (PCR) amplification and human-leukocyte antigen (HLA)-DQ-*ά* oligonucleotide typing on biological evidence samples – casework experience. *Journal of Forensic Sciences* **37**, 700–726.
9. Jeffreys, A.J., *et al.* (1988) Amplification of human minisatellites by the polymerase chain-reaction – Towards DNA Fingerprinting of single cells. *Nucleic Acids Research* **16**, 10953–10971.
10. Boerwinkle, E., *et al.* (1989) Rapid typing of tandemly repeated hypervariable loci by the polymerase chain-reaction–application to the apolipoprotein-B 3′ hypervariable region. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 212–216.
11. Budowle, B., *et al.* (1991) Analysis of the VNTR locus D1S80 by the PCR followed by high-resolution page. *American Journal of Human Genetics* **48**, 137–144.
12. Horn, G.T., *et al.* (1989) Amplification of a highly polymorphic VNTR segment by the polymerase chain-reaction. *Nucleic Acids Research* **17**, 2140–2140.
13. Kasai, K., *et al.* (1990) Amplification of a variable number of tandem repeats (VNTR) locus (PMCT118) by the polymerase chain-reaction (PCR) and its application to forensic-science. *Journal of Forensic Sciences* **35**, 1196–1200.

14. Rand, S., *et al.* (1992) Population-genetics and forensic efficiency data of 4 AmpFLPs. *International Journal of Legal Medicine* **104**, 329–333.
15. Sajantila, A., *et al.* (1991) The polymerase chain-reaction and postmortem forensic identity testing–application of amplified D1S80 and HLA-DQ-$\dot{\alpha}$ loci to the identification of fire victims. *Forensic Science International* **51**, 23–34.
16. Edwards, A., *et al.* (1991) DNA typing and genetic-mapping with trimeric and tetrameric tandem repeats. *American Journal of Human Genetics* **49**, 746–756.
17. Hagelberg, E., *et al.* (1991) Identification of the skeletal remains of a murder victim by DNA analysis. *Nature* **352**, 427–429.
18. Chien, A., *et al.* (1976) Deoxyribonucleic-acid polymerase from extreme thermophile thermus aquaticus. *Journal of Bacteriology* **127**, 1550–1557.
19. Saiki, R.K., *et al.* (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA-polymerase. *Science* **239**, 487–491.
20. Daquila, R.T., *et al.* (1991) Maximizing sensitivity and specificity of PCR by preamplification heating. *Nucleic Acids Research* **19**, 3749–3749.
21. Birch, D.E., *et al.* (1996) Simplified hot start PCR. *Nature* **381**, 445–446.
22. Moretti, T., *et al.* (1998) Enhancement of PCR amplification yield and specificity using AmpliTaq Gold (TM) DNA polymerase. *Biotechniques* **25**, 716–722.
23. Budowle, B., *et al.* (2001) STR primer concordance study. *Forensic Science International* **124**, 47–54.
24. Rychlik, W., and Rhoads, R.E. (1989) A computer-program for choosing optimal oligonucleotides for filter hybridization, sequencing and in-vitro Amplification of DNA. *Nucleic Acids Research* **17**, 8543–8551.
25. Rozen, S., and Skaletsky, H.J. (2000) Primer3 on the WWW for general users and for biologist programmers.. In *Bioinformatics Methods and Protocols: Methods in Molecular Biology* (Krawetz, S., and Misener, S., eds), Humana Press, pp. 365–386
26. Takagi, M., *et al.* (1997) Characterization of DNA polymerase from *Pyrococcus* sp. strain KOD1 and its application to PCR. *Applied and Environmental Microbiology* **63**, 4504–4510
27. Applied Biosystem. A feature guide for PCR enzymes. Avaliable at http://docs.appliedbiosystems.com/pebiodocs/00115226.pdf
28. Gill, P. (2001) Application of low copy number DNA profiling. *Croatian Medical Journal* **42**, 229–232.
29. Wilson, I.G. (1997) Inhibition and facilitation of nucleic acid amplification. *Applied and Environmental Microbiology* **63**, 3741–3751.
30. Akane, A., *et al.* (1994) Identification of the heme compound copurified with deoxyribonucleic-acid (DNA) from bloodstains, a major inhibitor of polymerase chain-reaction (PCR) amplification. *Journal of Forensic Sciences* **39**, 362–372.
31. Akane, A., *et al.* (1993) Purification of forensic specimens for the polymerase chain-reaction (PCR) analysis. *Journal of Forensic Sciences* **38**, 691–701.
32. Defranchis, R., *et al.* (1988) A potent inhibitor of *Taq* polymerase copurifies with human genomic DNA. *Nucleic Acids Research* **16**, 10355–10355.
33. Lantz, P.G., *et al.* (1997) Removal of PCR inhibitors from human faecal samples through the use of an aqueous two-phase system for sample preparation prior to PCR. *Journal of Microbiological Methods* **28**, 159–167.
34. Monteiro, L., *et al.* (1997) Complex polysaccharides as PCR inhibitors in feces: *Helicobacter pylori* model. *Journal of Clinical Microbiology* **35**, 995–998.
35. Tsai, Y.L., and Olson, B.H. (1992) Detection of low numbers of bacterial-cells in soils and sediments by polymerase chain-reaction. *Applied and Environmental Microbiology* **58**, 754–757.
36. Khan, G., *et al.* (1991) Inhibitory effects of urine on the polymerase chain-reaction for cytomegalovirus DNA. *Journal of Clinical Pathology* **44**, 360–365.
37. Abu Al-Soud, W., and Radstrom, P. (1998) Capacity of nine thermostable DNA polymerases to mediate DNA amplification in the presence of PCR-inhibiting samples. *Applied and Environmental Microbiology* **64**, 3748–3753.

38. Kreader, C.A. (1996) Relief of amplification inhibition in PCR with bovine serum albumin or T4 gene 32 protein. *Applied and Environmental Microbiology* **62**, 1102–1106.
39. Larkin, A., and Harbison, S.A. (1999) An improved method for STR analysis of bloodstained denim. *International Journal of Legal Medicine* **112**, 388–390.
40. Kontanis, E.J., and Reed, F.A. (2006) Evaluation of real-time PCR amplification efficiencies to detect PCR inhibitors. *Journal of Forensic Sciences* **51**, 795–804.
41. Rutty, G.N., *et al.* (2003) The effectiveness of protective clothing in the reduction of potential DNA contamination of the scene of crime. *International Journal of Legal Medicine* **117**, 170–174.
42. Port, N.J., *et al.* (2006) How long does it take a static speaking individual to contaminate the immediate environment. *Forensic Science Medicine and Pathology* **2**, 157–164.

6 The analysis of short tandem repeats

Short tandem repeats were first used in forensic casework in the early 1990s [1–3]. By the end of the decade they had become the standard tool for just about every forensic laboratory in the world. Today the vast majority of forensic genetic casework involves the analysis of STR polymorphisms and this situation is unlikely to change in the near future [4].

Structure of STR loci

Short tandem repeats contain a core repeat region between 1 and 6 bp long and have alleles that are generally less than 350 bp long. A large number of STR loci have been characterized [5] but only around 20 are commonly analysed in forensic casework (Table 6.1).

The STRs that are widely used in forensic genetics have either a four or five base-pair core-repeat motif and can be classified as a simple repeat, simple repeat with non-consensus repeats, compound repeat or complex repeat [6] (Figure 6.1).

The development of STR multiplexes

The forensic community has selected STR loci to incorporate into multiplex reactions based on several features including:

- discrete and distinguishable alleles;
- amplification of the locus should be robust;
- a high power of discrimination;
- an absence of genetic linkage with other loci being analysed;
- low levels of artefact formation during the amplification (see Chapter 7);
- the ability to be amplified as part of a multiplex PCR.

An Introduction to Forensic Genetics W. Goodwin, A. Linacre and S. Hadi
© 2007 John Wiley & Sons, Ltd

Table 6.1 The development of STR systems. Two STR systems, the quadraplex (QUAD) and second generation multiplex (SGM) were developed by the Forensic Science Service in the UK. The AmpF*l*STR® SGM Plus® became commercially available in 1999 and has been adopted by a large number of laboratories for routine forensic casework. The AmpF*l*STR® Identifiler® and PowerPlex® 16 both analyse 15 STR including the 13 loci CODIS loci that are required to be analysed for forensic casework in the USA. The two kits are used widely worldwide, particularly for kinship testing

QUAD	SGM	SGM Plus®	Identifiler®	PowerPlex® 16
vWA	Amelogenin	Amelogenin	Amelogenin	Amelogenin
TH01	vWA	D2S1338	D2S1338	D2S1338
F13A1	D8S1179	vWA	vWA	vWA
FES	D21S11	D16S359	D16S359	D16S359
	D18S51	D8S1179	D8S1179	D8S1179
	TH01	D21S11	D21S11	D21S11
	FGA	D18S51	D18S51	D18S51
		TH01	TH01	TH01
		FGA	FGA	FGA
			D13S317	D13S317
			CSF1PO	CSF1PO
			D7S820	D7S820
			TPOX	TPOX
			D5S818	D5S818
		D2S1338	D2S1338	Penta D
		D19S433	D19S433	Penta E

An essential feature of any STR used in forensic analysis is that biological material should give an identical profile regardless of the individual or laboratory that carries out the analysis. Without this standardization it would not be possible to compare results between laboratories and developments like national DNA databases would not be possible [7–11]. All new multiplexes have to be vigorously validated before they are used for the analysis of casework [12–19].

In the UK the Forensic Science Service (FSS) developed the first STR-based typing system that was designed for forensic analysis. Four STR loci were amplified in the same reaction [20–22]. This was replaced by the SGM (second generation multiplex) that was also developed by the FSS [23–25]. Two commercial companies, Applied Biosystems and Promega Corporation, have developed a series of multiplexes that are now used by most laboratories. The AmpF*l*STR® SGM Plus® that is produced by Applied Biosystems replaced the SGM in the UK and has been adopted by many other countries around the world as one of their standard multiplex kits [17]. In the USA, STR technology was adopted into forensic casework following a survey of 17 previously characterized STR loci and in 1997 13 loci were selected as the CODIS (Combined DNA Index System) loci [8, 26]. These loci can be analysed in one PCR using one of two commercially available kits; the AmpF*l*STR® Identifiler® produced by Applied Biosystems [27] and the PowerPlex® 16 produced by Promega Corporation [13]. The STR loci that are incorporated into different multiplexes are shown in Table 6.1.

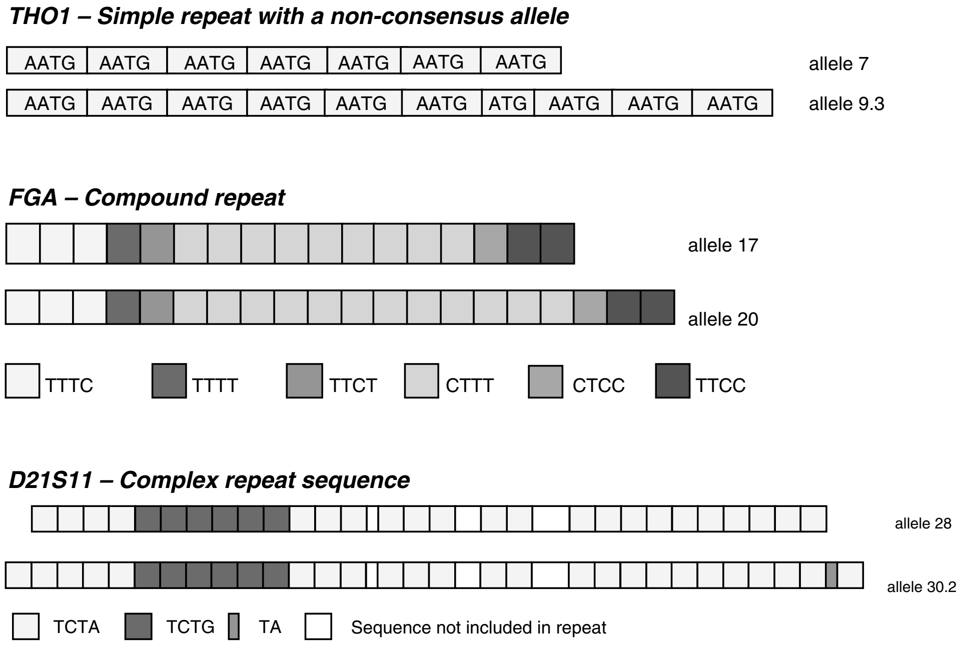

THO1 – Simple repeat with a non-consensus allele
AATG AATG AATG AATG AATG AATG AATG
allele 7
AATG AATG AATG AATG AATG AATG ATG AATG AATG AATG
allele 9.3
FGA – Compound repeat
allele 17
allele 20
TTTC
TTTT
TTCT
CTTT
CTCC
TTCC
D21S11 – Complex repeat sequence
allele 28
allele 30.2
TCTA
TCTG
TA
Sequence not included in repeat

Figure 6.1 The structure of three commonly used STR loci, THO1, FGA and D21S11. The THO1 locus has a simple repeat with a non-consensus allele; in the example the 9.3 allele is missing an A from the seventh repeat*. The FGA locus is a compound repeat composed of several elements. The D21S11 allele is an example of a complex repeat; the three regions not included in the FGA nomenclature are an invariant TA, TCA and TCCATA sequence. *The AATG nomenclature is commonly used but breaks the adopted conventions for STR nomenclature as it represents the bottom strand of the first sequence described in GenBank.

In addition to STR loci, the amelogenin locus which is present on the X and Y chromosomes has been incorporated into all commonly used STR multiplex kits. The amelogenin gene encodes for a protein that is a major component of tooth enamel matrix; there are two versions of the gene, the copy on the X chromosome has a 6 bp deletion and this length polymorphism allows the versions of the gene on the X and Y chromosomes to be differentiated (Figure 6.2) [28].

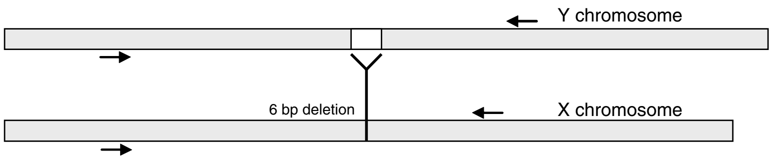

Y chromosome
6 bp deletion
X chromosome

Figure 6.2 The amelogenin locus is present on both the X and Y chromosomes. The gene that is present on the X chromosome has a 6 bp deletion. The primers (schematically shown by the arrowed lines) that were reported by Sullivan *et al.* (1993) [28] lead to products of 106 bp from the X chromosome and 112 bp from the Y chromosome

Detection of STR polymorphisms

After STR polymorphisms have been amplified using PCR, the length of the products must be measured precisely – some STR alleles differ by only one base pair. Gel electrophoresis of the PCR products through denaturing polyacrylamide gels can be used to separate DNA molecules between 20 and 500 nucleotides long with single base pair resolution [29]. Early systems detected the PCR products after electrophoresis on polyacrylamide slab-gels using silver staining [30, 31] but this limited the number of loci that could be incorporated into the multiplexes because the allelic size ranges of the different loci could not overlap. To overcome this limitation, fluorescence labelling of PCR products followed by multicolour detection has been adopted by the forensic community. A series of fluorescent dyes has been developed that can be covalently attached to the 5′ end of one of the PCR primers in each primer pair and detected real-time during electrophoresis. Up to five different dyes can be used in a single analysis which allows for considerable overlap of loci (Figure 6.3). The electrophoresis platforms have evolved from systems based on slab-gels to capillary electrophoresis (CE) that use a narrow glass tube filled with an entangled polymer solution to separate the DNA molecules [32–36]. Applied Biosystems provide the most commonly used

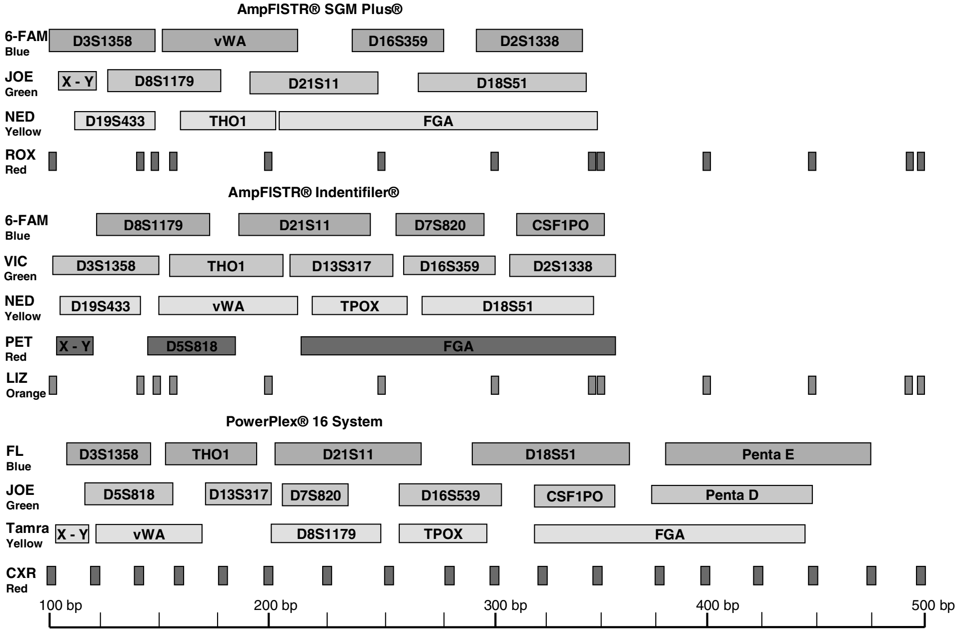

AmpFlSTR® SGM Plus®
6-FAM Blue
D3S1358
vWA
D16S359
D2S1338
JOE Green
X - Y
D8S1179
D21S11
D18S51
NED Yellow
D19S433
THO1
FGA
ROX Red
AmpFlSTR® Indentifiler®
6-FAM Blue
D8S1179
D21S11
D7S820
CSF1PO
VIC Green
D3S1358
THO1
D13S317
D16S359
D2S1338
NED Yellow
D19S433
vWA
TPOX
D18S51
PET Red
X - Y
D5S818
FGA
LIZ Orange
PowerPlex® 16 System
FL Blue
D3S1358
THO1
D21S11
D18S51
Penta E
JOE Green
D5S818
D13S317
D7S820
D16S539
CSF1PO
Penta D
Tamra Yellow
X - Y
vWA
D8S1179
TPOX
FGA
CXR Red
100 bp
200 bp
300 bp
400 bp
500 bp

Figure 6.3 PCR multiplexes use up to five different dyes to label PCR products. The allelic ranges of three commonly used multiplexes, the AmpF*l*STR® SGM Plus®, AmpF*l*STR® Identifiler® and the PowerPlex® 16 are shown. The use of multiple dyes allows the detection of the internal-lane size standard (ROX in SGM Plus®, LIZ in Identifiler® and CXR in PowerPlex® 16) and three to four overlapping STR loci, where the use of different dyes allows the alleles to be assigned to the correct locus

capillary electrophoresis systems and all these have multicolour detection capacity. The ABI PRISM® 310 Genetic Analyzer that has a single capillary and analyses up to 48 samples per day, the ABI PRISM® 3100 and Applied Biosystems 3130*xl* Genetic Analyzers, which have 16 capillaries and can analyse over 1000 samples per day, and the ABI PRISM® 3700 and Applied Biosystems 3730*xl* Genetic Analyzers, which can have up to 96 capillaries that can analyse over 4000 samples per day.

Before electrophoresis, the PCR sample is prepared by mixing approximately 1 μl of the reaction with 10–20 μl of deionized formamide. The internal-lane size standard is also added at this point. The deionized formamide denatures the DNA, heating the samples to 95 °C is routinely done to ensure that the PCR products are single stranded. The samples are transferred into the capillary using electrokinetic injection, a voltage is applied and charged molecules, including the amplified DNA fragments and the internal-lane size standards, migrate into the capillary. After injection, a constant voltage is applied across the capillary and the PCR products migrate towards the positively charged anode, travelling through the polymer, which fills the capillary and acts as the sieving matrix. Urea and 2-pyrrolidinone in the gel polymer and a temperature of 60 °C help to prevent the formation of any secondary structure during electrophoresis [37]. Throughout the period of electrophoresis, an argon ion laser is shone through a small glass window in the capillary and as PCR products labelled with fluorescent dyes travel past the window they are excited by the laser, emit fluorescence that is detected by a charged coupled device camera (CCD), and then are recorded by collection software [38] (Figure 6.4). The electrophoresis of a sample takes up to 30 minutes after which the polymer in the capillary is replaced with fresh polymer and the next sample can be analysed.

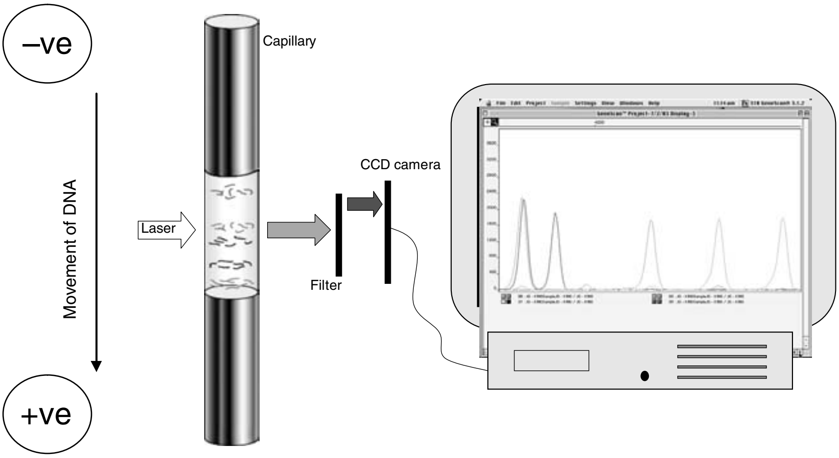


Figure 6.4 During electrophoresis an argon laser is shone through the window in the capillary. As the labelled PCR products migrate through the gel towards the anode they are separated based on their size. When the laser hits the fluorescent label on the PCR products, the lable is excited and emits fluorescent light that passes though a filter to remove any background noise, and then on to a charged coupled device camera that detects the wavelength of the light and sends the information to a computer where software records the profile (see plate section for full-colour version of this figure)

Interpretation of STR profiles

The spectra of the dyes used to label the PCR products overlap and the raw data contains peaks that are composed of more than one dye colour. After data collection the GeneScan® or GeneMapper™ *ID* software removes spectral overlap in the profile and calculates the sizes of the amplified DNA fragments. The software calculates how much spectral overlap there is between each dye and subtracts this from the peaks within the profile (Figure 6.5). A good matrix file, which contains information on the amount of overlap in the spectra, will produce peaks within the profile that are composed of only one colour. The height of the peaks is measured in relative fluorescent units (rfu) – the height is proportional to the amount of PCR product that is detected.

To be able to size the PCR products an internal-lane size standard is used. The internal-lane size standards contain fragments of DNA of known lengths that are labelled

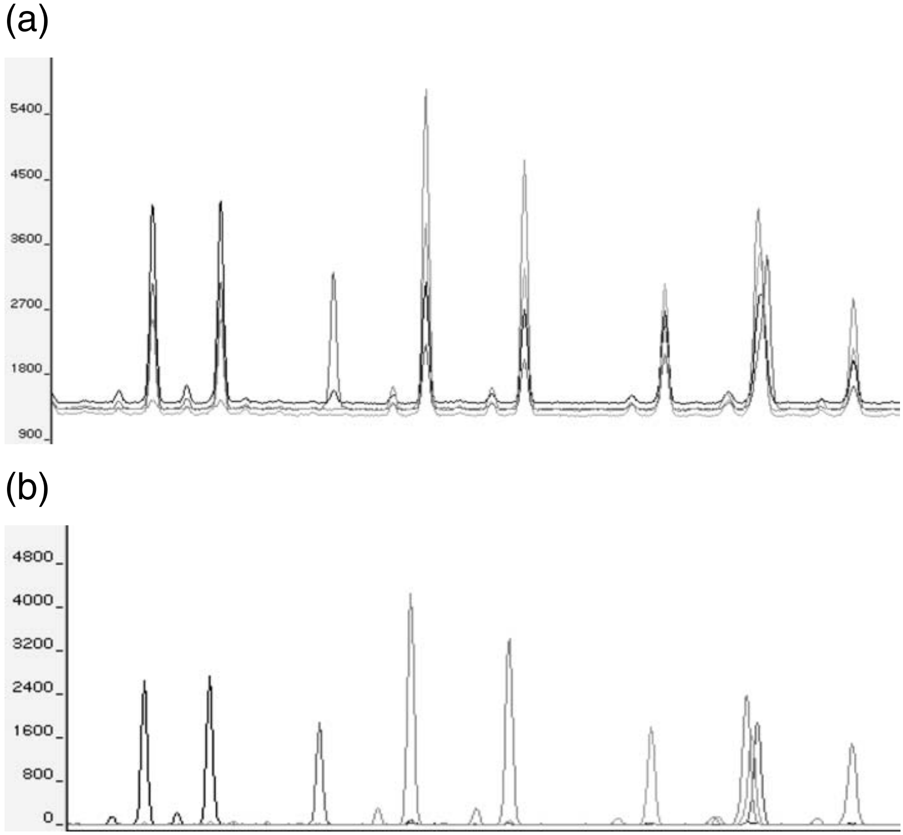


Figure 6.5 The application of a matrix file, using the GeneScan® or GeneMapper™ *ID* software removes the spectral overlap from the raw data (a) to produce peaks within the profile that are composed of only one colour (b) (see plate section for full-colour version of this figure). The scale of the X-axis is relative fluorescent units (rfu)

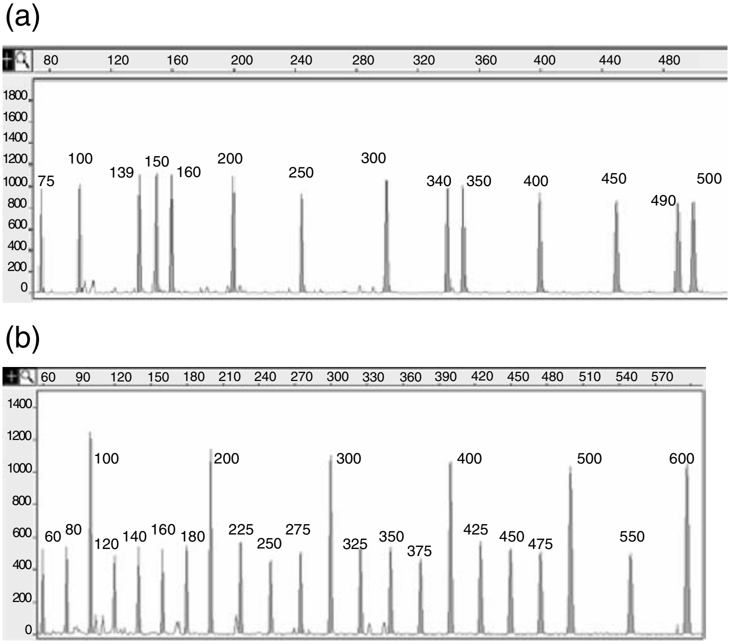


Figure 6.6 Internal-lane size standards are used to size the PCR products precisely. Two commonly used internal-lane size standards are (a) the GeneScan™-500 (Applied Biosystems) and (b) the ILS600 (Promega) (see plate section for full-colour version of this figure)

with a fluorescent dye, and the fragments are detected along with the amplified PCR products during capillary electrophoresis [39]. Commonly used commercial internal-lane size standards are the GeneScan™-500 standards that can be labelled with either ROX™ or LIZ™ dyes (Applied Biosystems) and the ILS600 (Promega Corporation) (Figure 6.6)

Because the internal-lane size standard is analysed along with each PCR any differences between runs that could affect the migration rates during electrophoresis, such as temperature, do not impact significantly on the analysis [40]. The software generates a size calling curve from the internal-lane size standards – the data point of the unknown fragments are compared to the size calling curve. Different algorithms have been developed to measure the size of DNA molecules, the most common one is the local Southern method [41] (Figure 6.7).

After analysing the raw data with the software, the end result is an electropherogram with a series of peaks that represent different alleles: the size, peak height and peak area is also measured by the software (Figure 6.8). The final stage of generating a STR profile is to assign specific alleles to the amplified PCR products. Each peak in the profile is given a number that is a description of the structure of that allele – this is straightforward when naming simple repeats but is more problematic with complex repeat sequences [6].

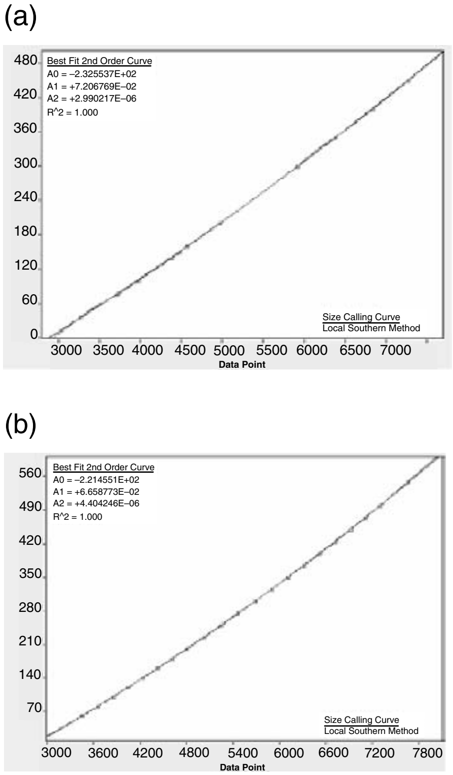
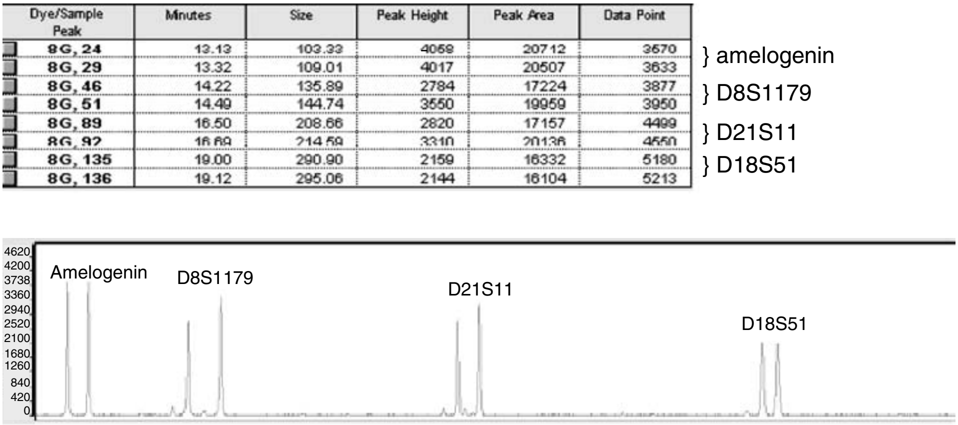


Figure 6.7 During electrophoresis the computer software records the fluorescence levels at regular time points and these are recorded as data points. The DNA fragments that make up the internal-lane size standards are plotted against the data points. An example of the sizing curves that are produced from (a) the GeneScan^{TM}-500 standard (Applied Biosystems) and (b) the ILS600 (Promega) are shown using the local Southern method to generate the size calling curve

The loci used in forensic casework have been well characterized and multiple alleles have been sequenced to determine the allelic structure and verify that the size of the peaks is a good indicator of the alleles they represent. However, because the migration of PCR products and internal-lane size standard varies slightly with factors such as temperature and the electrophoretic conditions, and because some STR alleles differ by only one base pair, the use of allelic ladders that contain all the common alleles (Figure 6.9) at each locus has been adopted by the forensic community to ensure accurate profiling [22, 42]. Unlike the internal-lane size standards the allelic ladders cannot be analysed in the same injection as the samples but are run periodically during the analysis of a batch of samples.

When assigning the alleles, the unknown peaks are compared to the allelic ladder and should fall within a one base-pair window that is +/− 0.5 bp of the allelic ladder size – if the unknown alleles differ by more than this then they are classified as off-ladder (OL)



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point	
8G, 24	13.13	103.33	4058	20712	3570	} amelogenin
8G, 29	13.32	109.01	4017	20507	3633	
8G, 46	14.22	135.89	2784	17224	3877	} D8S1179
8G, 51	14.49	144.74	3550	19959	3950	
8G, 89	16.50	208.66	2820	17157	4499	} D21S11
8G, 92	16.69	214.59	3310	20136	4550	
8G, 135	19.00	290.90	2159	16332	5180	} D18S51
8G, 136	19.12	295.06	2144	16104	5213	

Figure 6.8 The green loci from a profile produced using the AmpF*l*STR® SGM Plus® kit. The size of each peak has been calculated along with the peak heights and areas. The first amelogenin peak was detected after 13.13 minutes (which is when data point 3570 was taken) and is estimated to be 103.33 bp long, the peak area is 20712 rfu and the peak height 4058 rfu

and require further analysis. This comparison of unknown peaks to the allelic ladder can be done manually or by using the Genotyper® (Figure 6.10) or GeneMapper™ *ID* software (Applied Biosystems), which will compare all the unknown alleles in the profile to the allelic ladder.

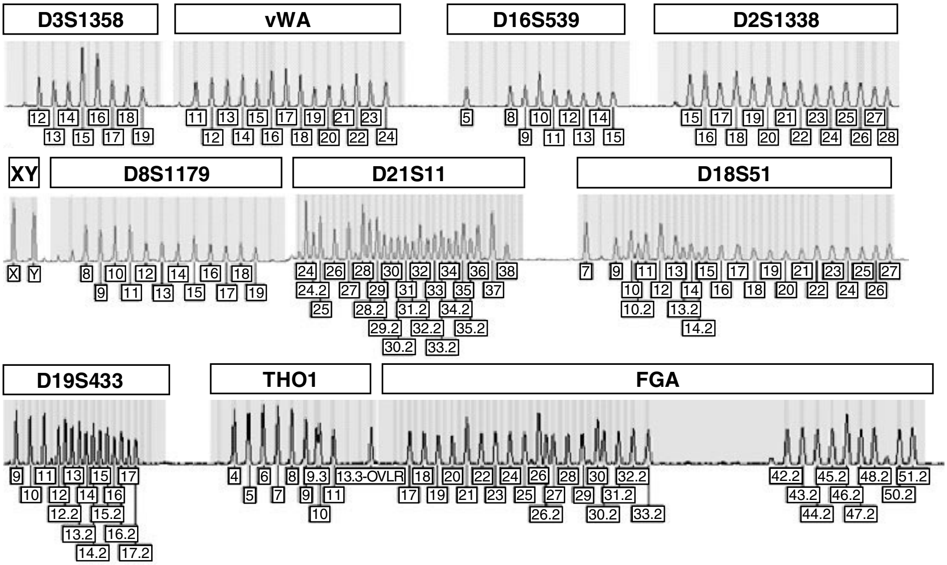


Figure 6.9 The allelic ladder of the AmpF*l*STR® SGM Plus® kit contains all the common alleles (see plate section for full-colour version of this figure)

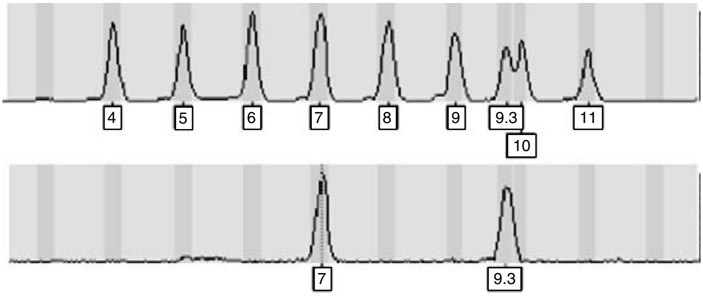


Figure 6.10 The comparison of an unknown allele with the allelic ladder allows the TH01 alleles to be classified as 7 and 9.3. The size of the unknown allele and the allele in the allelic ladder are not identical but fall within 0.5 bp of each other. The 0.5 bp match windows are indicated by the shaded areas.

The final result is a profile where alleles have been assigned to all of the peaks in the profile (Figure 6.11).

The STR profiles should be identical regardless of the laboratory where the analysis took place or the variations in the methodology that may have been used to generate the profile, such as different DNA extraction and quantification techniques and capillary electrophoresis platforms. Loci that are included in different commercial kits should also produce identical results (Table 6.2).

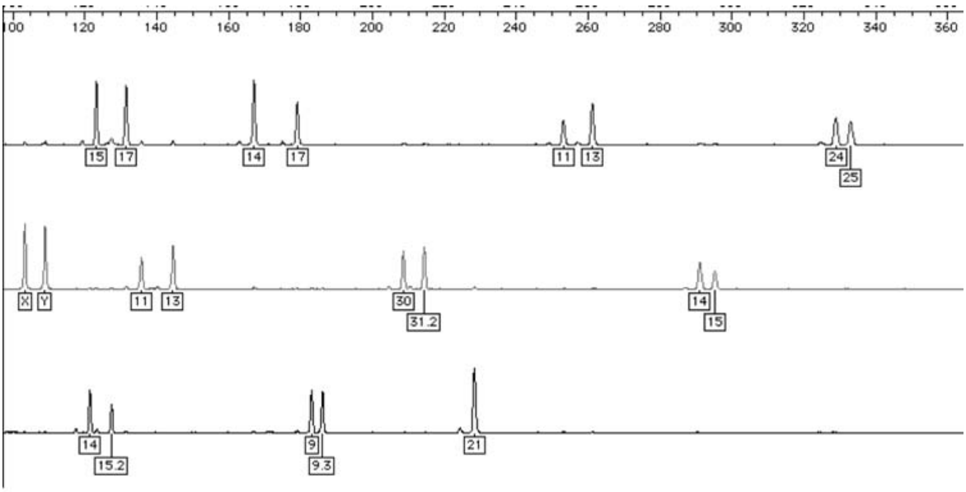


Figure 6.11 The Genotyper® software compares the peaks within a profile to the allelic ladder and assigns alleles. If the peaks in the profile deviate more than ±0.5 bp from the allelic ladder they are designate 'off ladder' (see plate section for full-colour version of this figure)

Table 6.2 Profiles have been generated from the same DNA sample using three commercial kits, the AmpF*l*STR® SGM Plus®, AmpF*l*STR® Identifiler® and the PowerPlex® 16. The alleles that are detected in the loci that are common between the kits are all identical

	Profile					
Locus	SGM Plus®		Identifiler®		PowerPlex® 16	
Amelogenin	X	Y	X	Y	X	Y
D3S1358	15	17	15	17	15	17
vWA	14	17	14	17	14	17
D16S359	11	13	11	13	11	13
D8S1179	11	13	11	13	11	13
D21S11	30	31.2	30	31.2	30	31.2
D18S51	14	15	14	15	14	15
TH01	9	9.3	9	9.3	9	9.3
FGA	21	21	21	21	21	21
D13S317			10	14	10	14
CSF1PO			9	12	9	12
D7S820			8	10	8	10
TPOX			11	11	11	11
D5S818			11	13	11	13
D2S1338	24	25	24	25		
D19S433	14	15.2	14	15.2		
Penta D					12	13
Penta E					12	14

Further reading

Butler J.M. (2005) *Forensic DNA Typing: Biology and Technology and Genetics of STR Markers*, second edition. Academic Press, London.

Butler J.M. (2006) Genetics and genomics of core short tandem repeat loci used in human identity testing. *Journal of Forensic Sciences* **51**, 253–265.

WWW resource

Ruitberg C.M., Reeder D.J., and Butler J.M. (2001) STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Research* **29**, 320–322. (http://www.cstl.nist.gov/div831/strbase/)

References

1. Clayton, T.M., *et al.* (1995) Identification of bodies from the scene of a mass disaster using DNA amplification of short tandem repeat (STR) loci. *Forensic Science International* **76**, 7–15.
2. Hagelberg, E., *et al.* (1991) Identification of the skeletal remains of a murder victim by DNA analysis. *Nature* **352**, 427–429.
3. Jeffreys, A.J., *et al.* (1992) Identification of the skeletal remains of Josef Mengele by DNA analysis. *Forensic Science International* **56**, 65–76.

4. Gill, P., *et al.* (2004) An assessment of whether SNPs will replace STRs in national DNA databases. *Science and Justice* **44**, 51–53.
5. Ruitberg, C.M., *et al.* (2001) STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Research* **29**, 320–322.
6. Gill, P., *et al.* (1997) Considerations from the European DNA profiling group (EDNAP) concerning STR nomenclature. *Forensic Science International* **87**, 185–192.
7. Andersen, J., *et al.* (1996) Report on the third EDNAP collaborative STR exercise. *Forensic Science International* **78**, 83–93.
8. Budowle, B., *et al.* (1998) CODIS and PCR-based short tandem repeat loci: law enforcement tools.. In *Second European Symposium on Human Identification*, Promega Corporation, pp. 73–88.
9. Gill, P., *et al.* (1997) Report of the European DNA profiling group (EDNAP): an investigation of the complex STR loci D21S11 and HUMFIBRA (FGA). *Forensic Science International* **86**, 25–33.
10. Gill, P., *et al.* (1994) Report of the European DNA profiling group (EDNAP) – towards standardization of short tandem repeat (Str) loci. *Forensic Science International* **65**, 51–59.
11. Gill, P., *et al.* (2000) Report of the European Network of Forensic Science Institutes (ENSFI): formulation and testing of principles to evaluate STR multiplexes. *Forensic Science International* **108**, 1–29.
12. Fregeau, C.J., *et al.* (2003) AmpFl STR (R) Profiler PIUS (TM) short tandem repeat DNA analysis of casework samples, mixture samples, and nonhuman DNA samples amplified under reduced PCR volume conditions (25 mu L). *Journal of Forensic Sciences* **48**, 1014–1034.
13. Krenke, B.E., *et al.* (2002) Validation of a 16-locus fluorescent multiplex system. *Journal of Forensic Sciences* **47**, 773–785.
14. LaFountain, M.J., *et al.* (2001) TWGDAM validation of the AmpFlSTR Profiler Plus and AmpFlSTR COfiler STR multiplex systems using capillary electrophoresis. *Journal of Forensic Sciences* **46**, 1191–1198.
15. Moretti, T.R., *et al.* (2001) Validation of short tandem repeats (STRs) for forensic usage: Performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples. *Journal of Forensic Sciences* **46**, 647–660.
16. Moretti, T.R., *et al.* (2001) Validation of STR typing by capillary electrophoresis. *Journal of Forensic Sciences* **46**, 661–676.
17. Cotton, E.A., *et al.* (2000) Validation of the AMPFlSTR (R) SGM Plus (TM) system for use in forensic casework. *Forensic Science International* **112**, 151–161.
18. Thomson, J.A., *et al.* (1999) Validation of short tandem repeat analysis for the investigation of cases of disputed paternity. *Forensic Science International* **100**, 1–16.
19. Wallin, J.M., *et al.* (1998) TWGDAM validation of the AmpFISTR (TM) Blue PCR amplification kit for forensic casework analysis. *Journal of Forensic Sciences* **43**, 854–870.
20. Clayton, T.M., *et al.* (1995) Further validation of a quadruplex STR DNA typing system: A collaborative effort to identify victims of a mass disaster. *Forensic Science International* **76**, 17–25.
21. Kimpton, C., *et al.* (1994) Evaluation of an automated DNA profiling system employing multiplex amplification of 4 tetrameric Str loci. *International Journal of Legal Medicine* **106**, 302–311.
22. Lygo, J.E., *et al.* (1994) The validation of short tandem repeat (STR) loci for use in forensic casework. *International Journal of Legal Medicine* **107**, 77–89.
23. Kimpton, C.P., *et al.* (1996) Validation of highly discriminating multiplex short tandem repeat amplification systems for individual identification. *Electrophoresis* **17**, 1283–1293.
24. Sparkes, R., *et al.* (1996) The validation of a 7-locus multiplex STR test for use in forensic casework.2. Artefacts, casework studies and success rates. *International Journal of Legal Medicine* **109**, 195–204.
25. Sparkes, R., *et al.* (1996) The validation of a 7-locus multiplex STR test for use in forensic casework.1. Mixtures, ageing, degradation and species studies. *International Journal of Legal Medicine* **109**, 186–194.

26. Budowle, B., *et al.* (1999) Population data on the thirteen CODIS core short tandem repeat loci in African Americans, US Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. *Journal of Forensic Sciences* **44**, 1277–1286.
27. Collins, P.J., *et al.* (2004) Developmental validation of a single-tube amplification of the 13 CODIS STR loci, D2S1338, D19S433, and amelogenin: The AmpFlSTR (R) Identifiler (R) PCR amplification kit. *Journal of Forensic Sciences* **49**, 1265–1277.
28. Sullivan, K.M., *et al.* (1993) A rapid and quantitative DNA sex test – fluorescence-based Pcr analysis of X-Y homologous gene amelogenin. *Biotechniques* **15**, 636–638.
29. Ziegle, J.S., *et al.* (1992) Application of automated DNA sizing technology for genotyping microsatellite loci. *Genomics* **14**, 1026–1031.
30. Lins, A.M., *et al.* (1996) Multiplex sets for the amplification of polymorphic short tandem repeat loci – silver stain and fluorescence detection. *Biotechniques* **20**, 882–889.
31. Sprecher, C.J., *et al.* (1996) General approach to analysis of polymorphic short tandem repeat loci. *Biotechniques* **20**, 266–267.
32. Buel, E., *et al.* (1998) Capillary electrophoresis STR analysis: comparison to gel-based systems. *Journal of Forensic Sciences* **43**, 164–170.
33. Butler, J.M., *et al.* (1995) Application of dual internal standards for precise sizing of polymerase chain-reaction products using capillary electrophoresis. *Electrophoresis* **16**, 974–980.
34. Heiger, D.N., *et al.* (1990) Separation of DNA restriction fragments by high-performance capillary electrophoresis with low and zero cross-linked polyacrylamide using continuous and pulsed electric-fields. *Journal of Chromatography* **516**, 33–48.
35. Butler, J.M., *et al.* (2004) Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. *Electrophoresis* **25**, 1397–1412.
36. Madabhushi, R.S. (1998) Separation of 4-color DNA sequencing extension products in noncovalently coated capillaries using low viscosity polymer solutions. *Electrophoresis* **19**, 224–230.
37. Rosenblum, B.B., *et al.* (1997) Improved single-strand DNA sizing accuracy in capillary electrophoresis. *Nucleic Acids Research* **25**, 3925–3929.
38. Butler, J. (2001) *Forensic DNA Typing: Biology and Technology behind STR Markers*. Academic Press.
39. Edwards, A., *et al.* (1991) DNA typing and genetic-mapping with trimeric and tetrameric tandem repeats. *American Journal of Human Genetics* **49**, 746–756.
40. Hartzell, B., *et al.* (2003) Response of short tandem repeat systems to temperature and sizing methods. *Forensic Science International* **133**, 228–234.
41. Elder, J.K., and Southern, E.M. (1983) Measurement of DNA length by gel-electrophoresis.2. Comparison of methods for relating mobility to fragment length. *Analytical Biochemistry* **128**, 227–231.
42. Smith, R.N. (1995) Accurate size comparison of short tandem repeat alleles amplified by PCR. *Biotechniques* **18**, 122–128.

7 Assessment of STR profiles

DNA profiles generated from casework samples require some experience to interpret. Guidelines have evolved to assist with the interpretation of STR profiles, ensuring that the results are robust and consistent; this is especially important when dealing with samples that contain very small amounts of DNA, degraded DNA or mixtures of profiles that come from two or more individuals – all situations that complicate interpretation. This chapter explores a number of artefacts that can occur in DNA profile. Some casework scenarios that can lead to complex profiles are also considered.

Stutter peaks

During the amplification of an STR allele it is normal to generate a stutter peak, that is one repeat unit smaller or larger that the true allele; smaller alleles are formed in the majority of cases [1]. Stutter peaks are formed by strand slippage during the extension of the nascent DNA strand during PCR amplification (Figure 7.1) [2, 3].

Even in good quality profiles there will be some stutter peaks; these are recognizable and do not interfere with the interpretation of the profile. Threshold limits are normally used to aid in the identification and interpretation of stutter peaks, so, for example, while the degree of stutter varies between loci, they are typically less than 15 % of the main peak [4, 5] – understanding stutter peaks is especially important when interpreting mixtures.

Different STR loci have varying tendencies to stutter. This is dependent on the structure of the core repeats: shorter di- and trinucleotide repeats are more prone to stutter than are tetra- and pentanucleotide repeats and this is one of the reasons that all the autosomal STRs that have been adopted by the forensic community have tetra and pentanucleotide core repeats (Figure 7.2). STRs with simple core repeats tend to have higher stutter rates than compound and complex repeats.

Split peaks (+/− A)

The *Taq* polymerase that is used to drive the polymerase chain reaction adds nucleotides to the newly synthesized DNA molecule in a template-dependent manner. However, it

An Introduction to Forensic Genetics W. Goodwin, A. Linacre and S. Hadi
© 2007 John Wiley & Sons, Ltd

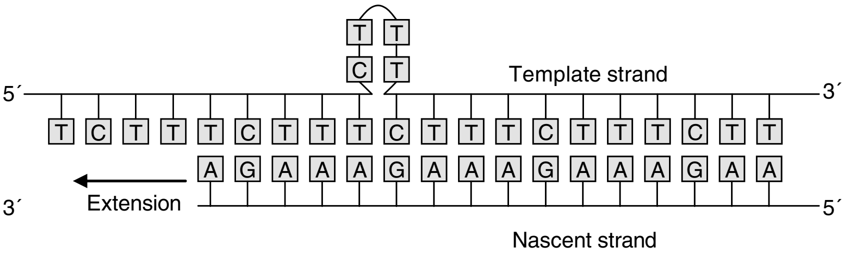

T
T
C
T
Template strand
5´
3´
T C T T T C T T T C T T T C T T T C T T
A G A A A G A A A G A A A G A A
3´
Extension
5´
Nascent strand

Figure 7.1 During PCR, slippage between the template and the nascent DNA strands leads to the copied strand containing one repeat less than the template strand

also has an activity, called terminal transferase, whereby it adds a nucleotide to the end of the amplified molecule which is non-template-dependent [6]. Approximately 85 % of the time an adenine residue is added (Figure 7.3).

It is important that the vast majority of PCR products have the non-template nucleotide added, otherwise a split peak is observed in the DNA profile (Figure 7.4). Split peaks are usually caused either by the sub-optimal activity of the *Taq* polymerase or by too much template DNA in the PCR.

In order to minimize the formation of split peaks in a profile, at the end of the cycling stage of the PCR, the reaction is incubated at 65–72 °C for between 45–60 minutes, allowing the *Taq* polymerase to complete the non-template addition of all the PCR products.

The interpretation of profiles with split peaks is possible because the peak with the nucleotide added is taken as being the correct peak. Problems can occur when alleles are present that differ by only one base pair; the THO1 9.3 allele for example could be confused with the THO1 allele 10. In most cases a profile with a high degree

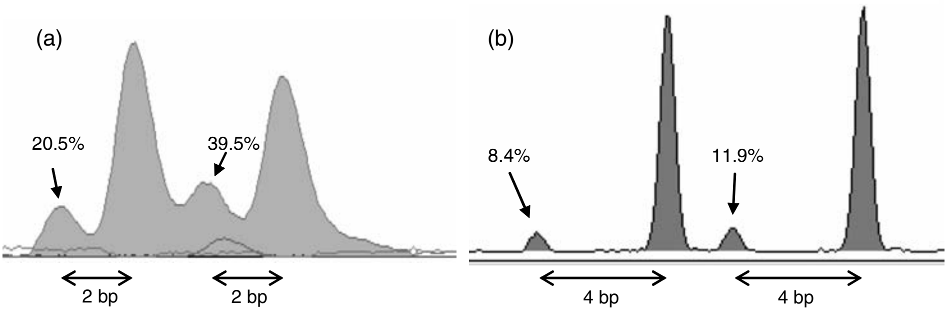


Figure 7.2 Stutter peaks are formed during slippage of the *Taq* polymerase during replication of the template strand. The slippage results in amplification products one repeat unit shorter than the template. The stutter peaks are normally less than 15 % of the true amplification product. Panel (a) shows a dinucleotide repeat, which is prone to high levels of slippage, the stutter peaks are indicated by the arrow and their size relative to the main peak is shown (based on peak area). Panel (b) is a tetranucleotide repeat, which displays lower levels of stutter

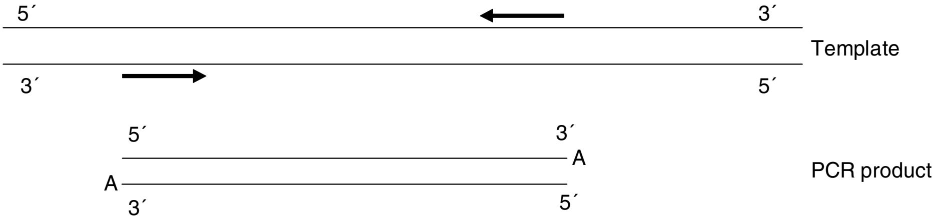


Figure 7.3 The *Taq* polymerase adds a nucleotide to the 3′ end of the newly synthesized strand. The non-template addition is usually an adenine and results in a PCR product that is one base pair longer than the template ($N + 1$). The arrowed lines represent the forward and reverse primers

of split peaks would have to be re-analysed to minimize the possibility of incorrect interpretation.

Pull-up

In Chapter 6 the matrix file was introduced – this file contains information about the levels of spectral overlap that exist with the dyes that have been used to label the PCR products. This information is used by the Genescan® and GeneMapper™ *ID* software to produce peaks that are made up of one colour. If the matrix file is not of good quality then this correction is not perfect and the peaks in the resulting profile are composed of more than one colour; this phenomenon is called pull-up. Pull-ups are easy to recognize as a smaller sized product will appear at exactly the same size as the real STR allele. Pull-up can also occur when there has been over amplification, even if the matrix file is of good quality (Figure 7.5a).

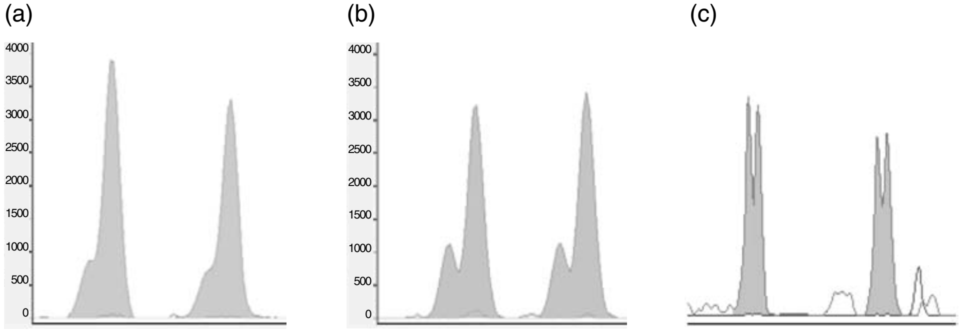


Figure 7.4 Split peaks are seen in profiles when the non-template addition does not occur with all of the PCR products. The three examples show decreasing amounts of non-template addition with panel (a) showing an example where the vast majority of PCR product has the non-template addition through to panel (c), where only 50 % of the PCR product has the non-template addition

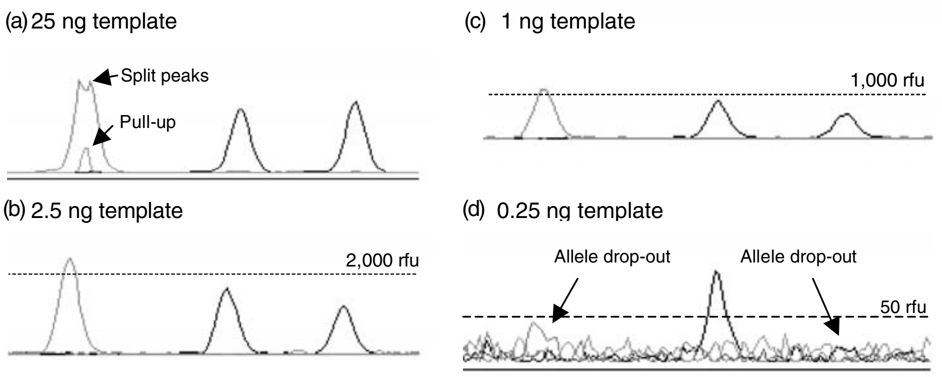


Figure 7.5 If the reaction is overloaded with DNA, (a) the peaks are still present but artefacts such as pull-ups and split peaks are more pronounced. When the template is within the optimal range (b and c) the peaks are well balanced and easy to interpret. When the PCR does not have enough template to amplify (d), then locus and allelic drop-out can occur (see plate section for full-colour version of this figure)

Template DNA

Commercial STR kits have been optimized to amplify small amounts of template DNA, commonly between 0.5 and 2.5 ng, which represents approximately 166 and 833 copies of the haploid human genome. It is not always possible to add the optimum amount of DNA to a PCR when the sample size is limited.

Overloaded profiles

Overloading the PCR can also lead to a profile that is difficult to interpret. If the CCD camera is saturated then the peak height/area is no longer a good indicator of the amount of product and this can lead to problems in assessing peak balance and can make the interpretation of mixtures difficult. Overloaded profiles also tend to have a noisy baseline, increased levels of stuttering, split peaks and pull-ups (Figure 7.6).

Low copy number DNA

At many crime scenes it may be possible to infer surfaces with which the perpetrator has had physical contact, for example the handle of a gun, a knife, a ligature, a door handle or a steering wheel. These areas can be swabbed to collect any epithelial cells that have been shed during the contact [7–10]. The amounts of DNA extracted can be extremely low but in some circumstances it is possible to get a full DNA profile from less than 100 pg of template DNA: the normal range of template DNA is between 500 and 2500 pg (2.5 ng). To analyse such small quantities of DNA the number of amplification cycles is increased to 34. The standard number of cycles in the amplification using commercial kits is between 28 – 32 cycles. Empirical studies have shown that above 34 cycles the

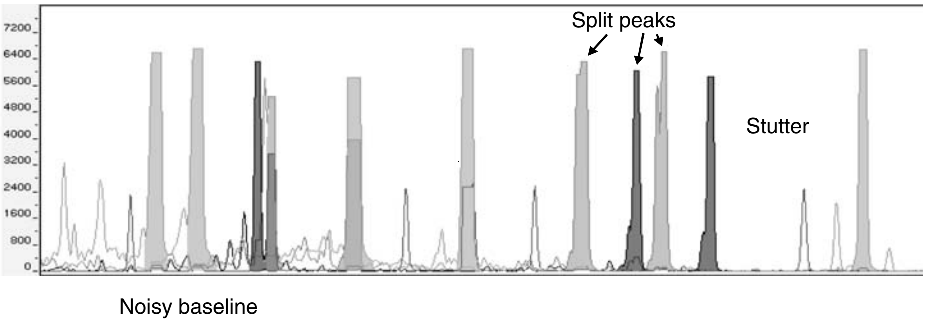


Figure 7.6 A heavily overloaded profile. All the peaks shown have a flat top indicating that they are off-scale, the baseline is very noisy, several split peaks are evident and the peaks are very broad, which can lead to sizing problems. There are also some pronounced stutter peaks (see plate section for full-colour version of this figure)

amount of artefacts that are detected outweigh the benefit of higher levels of artefact [11]. Extreme care has to be taken when interpreting the LCN profiles [11, 12]. A number of features can be seen when amplifying low amounts of template DNA. These are: allele drop-out, and drop-in; severe peak imbalance; locus drop-out (Figure 7.5d); and increased stutter [12–14]. Allele drop-out occurs when through chance events one allele in a heterozygous locus is preferentially amplified; this can give the false impression that the profile at a particular locus is homozygous. To minimize the possibility of this occurring the PCR must be repeated at least two times and only alleles that appear consistently can be called (Figure 7.7).

This phenomenon also leads to a peak imbalance that is much higher than when using higher amounts of template DNA. Allele drop-in is also a common phenomenon when amplifying low amounts of template DNA. The drop-in alleles are spurious amplification products and are not amplified in the duplicate or triplicate reactions but can still confuse the interpretation of the profile. Locus drop-out, particularly of the larger STR loci, can also occur; this reduces the amount of information from the profile but does not confuse the interpretation. At present, there is no clear consensus in the scientific community about the use of LCN PCR [15].

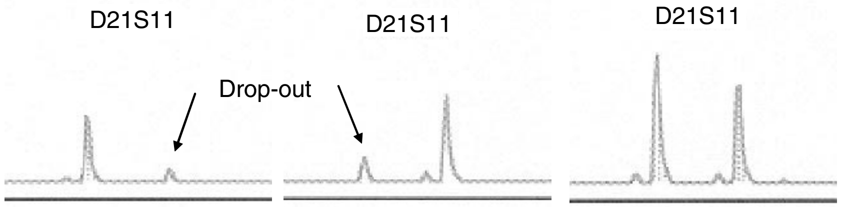


Figure 7.7 Three separate PCR analyses of a DNA extract can lead to different results. When dealing with very low template numbers, allelic dropout is relatively common and the true genotype can only be ascertained through multiple amplifications – and even then the results can be contentious

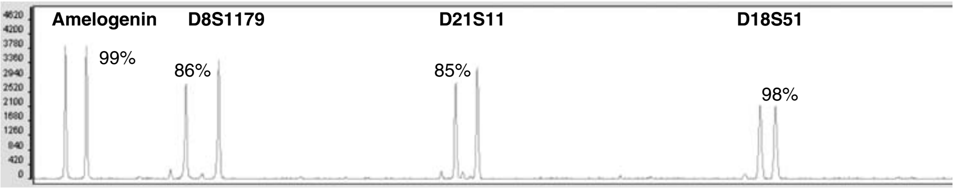


Figure 7.8 The profile shows the green loci from the AmpF*l*STR® SGM Plus® kit. The peak area of the smallest peak at each locus is shown as a percentage of the larger peak. The size of the peaks is proportional to the amount of PCR product – this can be gauged by measuring the peak height or, more usually, the peak area

Peak balance

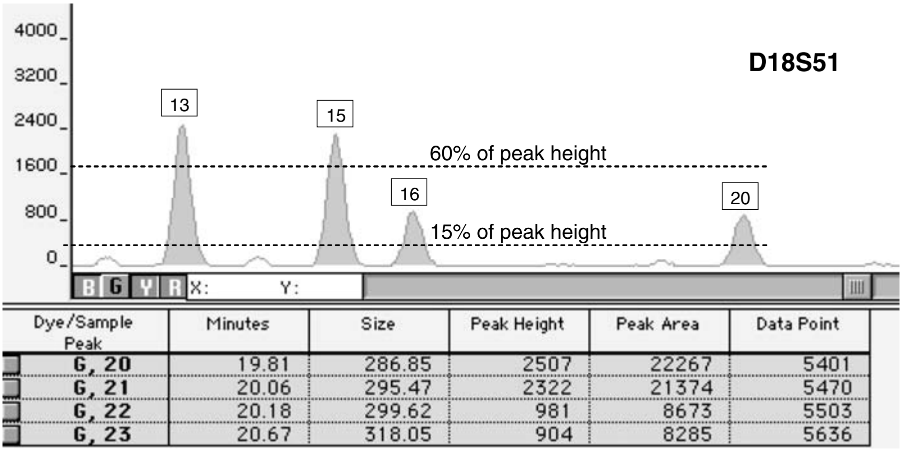
STR loci that are used in forensic analysis are commonly heterozygous, producing two peaks in the profile. In a perfect profile the two peaks that are produced are balanced 1:1 in terms of peak height and area but in reality this is very rare and one peak will be larger than the other (Figure 7.8). The variations in peak height can be due to chance events, where one allele is more efficiently amplified than another. In good quality DNA extracts, the smaller peak is, on average, approximately 90 % the size of the larger peak [4].

Laboratories will use different values that are based on their own validation studies but commonly require the smaller peak of a locus to be within 60 % of the larger peak [4]. Peak imbalance can be more extreme when profiling degraded DNA and when amplifying low amounts of template DNA. On rare occasions the mutation of a primer binding site will reduce the efficiency of the PCR for one allele, which can result in high levels of peak imbalance and even allele drop-out. The frequency of these mutations is low, ranging between frequencies of 0.01 and 0.001 per locus [16].

Mixtures

Many biological samples that are recovered from a scene of crime will contain a mixture of cellular material from more than one person. Clothing will often contain cellular material from the wearer and may also contain material from an assailant after an assault; the handle of a door or a steering wheel may have been handled by several people – there are many circumstances when mixtures of material can be collected. A mixture in a DNA profile can be recognized by the presence of more than two alleles at any locus within the profile, normally there will be several loci that have three or four alleles present and a loss of peak balance.

Having determined that the profile is mixed, the first task is to assess how many contributors are represented in the profile. Two person mixtures are most commonly seen in forensic casework; with a two person mixture a maximum of four alleles will be present at any locus whereas three person mixtures will contain up to six alleles



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
G, 20	19.81	286.85	2507	22267	5401
G, 21	20.06	295.47	2322	21374	5470
G, 22	20.18	299.62	981	8673	5503
G, 23	20.67	318.05	904	8285	5636

Figure 7.9 A mixture of two individuals will lead to up to four peaks at each locus. The area between the dotted lines represents the zone where the minor component of the mixture can be interpreted. The lower dotted line represents 15 % and 60 % of the major peaks – below 15 % is a zone where stutter peaks from the major alleles can occur and peaks, below 60 % cannot be easily explained by peak imbalance. At this locus the major component can be interpreted as 13–15 while the minor component's genotype is 16 – 20

at a locus. When four alleles are present at a given locus and there is a major and minor component, the interpretation is relatively simple (Figure 7.9). The ratio of peak areas within a locus generally corresponds with the ratio of template molecules [17, 18], peak areas that consider the morphology of the peak as well as the height [16], are commonly used as a guide to interpret mixed profiles [19]. Even in a two-person mixture when there are shared alleles between the major and minor profiles, the interpretation becomes more difficult – especially in mixtures where the minor profile is less than one-third of the level of the major profile [16].

In mixtures where the major component is in large excess, it is often possible to deduce the major profile; however, in such cases it is difficult to get much information from the minor component where the interpretation is complicated by artefacts in the profile such as stutter peaks, and also by the major profile masking the minor profile [20]. Software has been developed that helps with the interpretation of complex mixtures [21].

Degraded DNA

Many samples that are collected from a crime scene may have been exposed to the environment for hours, days, or even longer if the crime scene has gone undetected. When DNA analysis is being used to identify human remains, the remains may be several years old before they are analysed or may have been exposed to severe environmental insult such as high temperatures. In all these circumstances the DNA in the

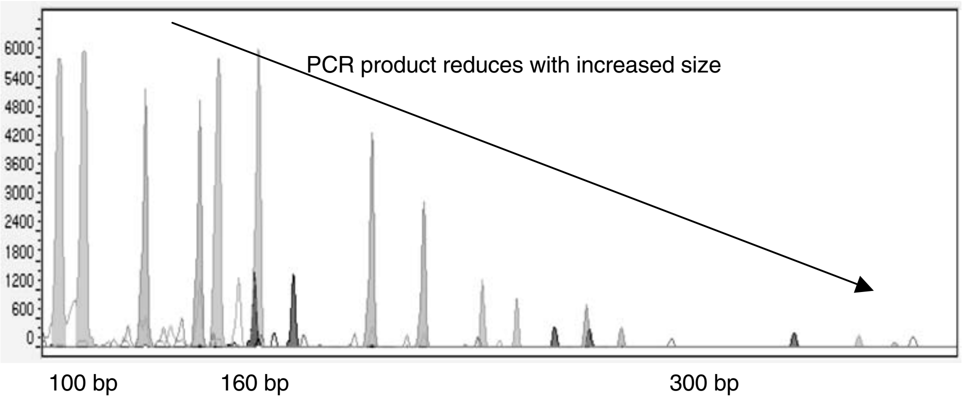


Figure 7.10 The profile was generated using the AmpF*l*STR® Profiler Plus® kit from Applied Biosystems. The DNA was extracted from a bone recovered from a Scottish loch after approximately 30 years. The profile is typical of a degraded profile with a gradual reduction in the amount of product as the amplicons increase in size (see plate section for full-colour version of this figure)

cellular material will not be in pristine condition and will have degraded. This leads to a characteristic DNA profile with over amplification of the smaller loci and the successful amplification declines with the size of the alleles. Figures 7.10 and 7.11 show two examples of degraded DNA sample; the first one is from a bone sample that had been in water for 30 years. The small loci have over amplified whereas the larger loci are barely detectable [22] – the decrease in amplification is gradual as the length of the alleles increases.

In the second example an example of locus drop out can be seen, the first two blue loci, D3S1358 and vWA have amplified successfully but there is no FGA allele. This

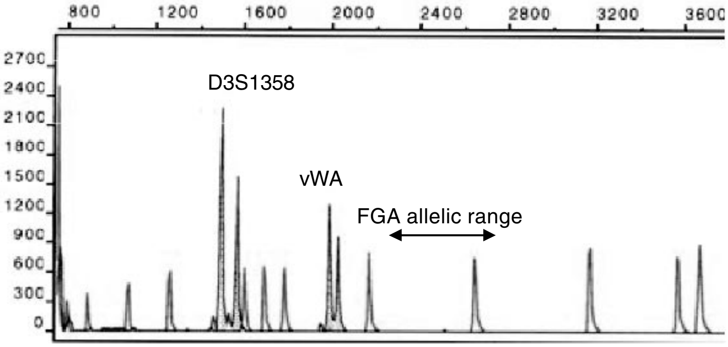


Figure 7.11 The profile was generated using the AmpF*l*STR Blue™ kit from Applied Biosystems. The DNA was extracted from muscle tissue recovered from a plane crash. The muscle had been subjected to high temperatures and the DNA was highly degraded – no amplification products were detected from the FGA locus. The size standard is also shown in by non-shaded peaks (see plate section for full-colour version of this figure)

profile is from human muscle tissue that had been exposed to high temperatures and has degraded to the extent that there is very little or no DNA that is 200 bp or longer [23]. The interpretation of degraded profiles can be difficult and particular attention has to be taken when homozygous loci are detected – are they really homozygous and not heterozygous with one of the alleles having dropped out? When the levels are very low, if there is enough material the PCR is carried out in duplicate, as with LCN PCR, to minimize the possibility of generating an incorrect profile.

To assist with the analysis of degraded DNA, a series of multiplexes have been developed with the primers positioned close to the core repeats of the STRs, thereby minimizing the lengths of the amplicons [24–28].

References

1. Shinde, D., *et al.* (2003) *Taq* DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)(n) and (A/T)(n) microsatellites. *Nucleic Acids Research* **31**, 974–980.
2. Hauge, X.Y., and Litt, M. (1993) A study of the origin of shadow bands seen when typing dinucleotide repeat polymorphisms by the Pcr. *Human Molecular Genetics* **2**, 411–415.
3. Schlotterer, C., and Tautz, D. (1992) Slippage synthesis of simple sequence DNA. *Nucleic Acids Research* **20**, 211–215.
4. Gill, P., *et al.* (1997) Development of guidelines to designate alleles using an STR multiplex system. *Forensic Science International* **89**, 185–197.
5. Corporation, T.P.-E. (1999) *AmpFISTR SGM Plus*™ *PCR Amplification Kit – User's Manual.*
6. Clark, J.M. (1988) Novel non-templated nucleotide addition-reactions catalyzed by procaryotic and eukaryotic DNA-polymerases. *Nucleic Acids Research* **16**, 9677–9686.
7. Bohnert, M., *et al.* (2001) Transfer of biological traces in cases of hanging and ligature strangulation. *Forensic Science International* **116**, 107–115.
8. Esslinger, K.L., *et al.* (2004) Using STR analysis to detect human DNA from exploded pipe bomb devices. *Journal of Forensic Sciences* **49**, 481–484.
9. van Oorschot, R.A.H., and Jones, M.K. (1997) DNA fingerprints from fingerprints. *Nature* **387**, 767–767.
10. Wiegand, P., and Kleiber, M. (1997) DNA typing of epithelial cells after strangulation. *International Journal of Legal Medicine* **110**, 181–183.
11. Gill, P. (2001) Application of low copy number DNA profiling. *Croatian Medical Journal* **42**, 229–232.
12. Gill, P., *et al.* (2000) An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. *Forensic Science International* **112**, 17–40.
13. Kloosterman, A.D., and Kersbergen, P. (2003) Efficacy and limits of genotyping low copy number DNA samples by multiplex PCR of STR loci. *International Congress Series* **1239**, 795–798.
14. Whitaker, J.P., *et al.* (2001) A comparison of the characteristics of profiles produced with the AMPFlSTR® SGM Plus™ multiplex system for both standard and low copy number (LCN) STR DNA analysis. *Forensic Science International* **123**, 215–223.
15. Gill, P., *et al.* (2006) DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures. *Forensic Science International* **160**, 90–101.
16. Clayton, T.M., *et al.* (1998) Analysis and interpretation of mixed forensic stains using DNA STR profiling. *Forensic Science International* **91**, 55–70.
17. Lygo, J.E., *et al.* (1994) The validation of short tandem repeat (Str) loci for use in forensic casework. *International Journal of Legal Medicine* **107**, 77–89.
18. Sparkes, R., *et al.* (1996) The validation of a 7-locus multiplex STR test for use in forensic casework.1. Mixtures, ageing, degradation and species studies. *International Journal of Legal Medicine* **109**, 186–194.

19. Evett, I.W., *et al.* (1998) Taking account of peak areas when interpreting mixed DNA profiles. *Journal of Forensic Sciences* **43**, 62–69.
20. Gill, P., *et al.* (1998) Interpreting simple STR mixtures using allele peak areas. *Forensic Science International* **91**, 41–53.
21. Bill, M., *et al.* (2005) PENDULUM–a guideline-based approach to the interpretation of STR mixtures. *Forensic Science International* **148**, 181–189.
22. Goodwin, W., *et al.* (2003) The identification of a US serviceman recovered from the Holy Loch, Scotland. *Science and Justice* **43**, 45–47.
23. Goodwin, W., *et al.* (1999) The use of mitochondrial DNA and short tandem repeat typing in the identification of air crash victims. *Electrophoresis* **20**, 1707–1711.
24. Dixon, L.A., *et al.* (2006) Analysis of artificially degraded DNA using STRs and SNPs – results of a collaborative European (EDNAP) exercise. *Forensic Science International* **164**, 33–44.
25. Gill, P., *et al.* (2006) The evolution of DNA databases – recommendations for new European STR loci. *Forensic Science International* **156**, 242–244.
26. Coble, M.D., and Butler, J.M. (2005) Characterization of new MiniSTR loci to aid analysis of degraded DNA. *Journal of Forensic Sciences* **50**, 43–53.
27. Drabek, J., *et al.* (2004) Concordance study between Miniplex assays and a commercial STR typing kit. *Journal of Forensic Sciences* **49**, 859–860.
28. Butler, J.M., *et al.* (2003) The development of reduced size STR amplicons as tools for analysis of degraded DNA. *Journal of Forensic Sciences* **48**, 1054–1064.

8 Statistical interpretation of STR profiles

Once it has been established that two DNA profiles are the same, the significance of the match has to be estimated. This requires some knowledge of population genetics and some statistical analysis of the data. This chapter will briefly cover the fundamental concepts involved with estimating the frequency of an STR profile in a given population.

Population genetics

It is necessary from the outset to define what is meant by a population. In the context of forensic genetics a population can be described as a group of people sharing common ancestry. In forensic terms the classification of a population within a country is usually quite broad and many subgroups that can differ in language, culture and religion are placed together and classified as, for example, Caucasian, sub-Saharan African and East Asian.

The Hardy–Weinberg law

Population genetics can be defined as the study of factors affecting the allele and genotype frequencies of different genetic loci in a population. The Hardy–Weinberg Law (HW law), also called the Hardy–Weinberg principle, provides a simple mathematical representation of the relationship of genotype and allele frequencies within an ideal population [1, 2] and is central to forensic genetics. The HW law states that within a randomly mating population the genotype frequencies at any single genetic locus remain constant. When a population is obeying the HW law it is said to be in Hardy–Weinberg equilibrium (HWE). Importantly, when a population is in HWE, the genotype frequencies can be predicted from the allele frequencies. This relationship can be represented in a Punnett square (Figure 8.1).

The polymorphic STR loci used in forensic genetics have multiple alleles; however, the genotype frequency of a homozygote can be calculated using p^2 and that of

An Introduction to Forensic Genetics W. Goodwin, A. Linacre and S. Hadi
© 2007 John Wiley & Sons, Ltd

	A ($p = 0.6$)	B ($q = 0.4$)
A ($p = 0.6$)	AA ($p^2 = 0.6^2 = 0.36$)	AB ($pq = 0.6 \times 0.4 = 0.24$)
B ($q = 0.4$)	AB ($pq = 0.6 \times 0.4 = 0.24$)	BB ($q^2 = 0.4^2 = 0.16$)

Figure 8.1 A Punnet square showing the relationship between alleles A and B, along with all the possible resulting genotypes. If allele A occurs at a frequency (p) of 0.6 and allele B occurs at frequency (q) of 0.4, then it is possible to estimate that the population will contain individuals with genotypes AA, AB and BB at frequencies of 0.36, 0.48 and 0.16 respectively. Each homozygote appears only once, hence p^2, or q^2. The heterozygote will be represented twice, hence $2pq$

heterozygotes can be calculated using $2pq$, removing the need to construct elaborate Punnet squares.

Deviation from the Hardy–Weinberg equilibrium

The HW law states that certain conditions must be met. These are:

- the population is infinitely large;
- random mating occurs within the population;
- the population is free from the effects of migration;
- there is no natural selection;
- no mutations occur.

Clearly no human population will meet these criteria and they will deviate from HWE to a greater or lesser extent.

Infinitely large population

A consequence of finite population size is that the frequency of alleles will change through a process known as random genetic drift, where the frequency of any given allele will increase or decrease through chance events. The effect of genetic drift is more pronounced in smaller populations [3]. However, most populations are sufficiently large for allele frequencies not to be significantly affected by genetic drift. Even in relatively small isolated human populations, it has been shown that alleles that are present at a frequency of more than 1 % are rarely lost in recently diverged populations [4, 5].

Random mating

Humans clearly do not mate completely randomly. However, because STR genotypes do not have any impact on a person's phenotype, such as height, strength or intelligence, selection of an STR through sexual selection is unlikely and has not been demonstrated.

No migration

Human history is full of migrations and this obviously can lead to changes in the gene pools of populations. If two distinct populations are living in the same geographical area and they have different allele frequencies, each population can be in HWE. If the two different populations are not recognized within the larger population and are not treated as separate populations, there can appear to be deviation from the HWE; this is known as the Wahlund effect [6–8]. If random admixture occurs between the two populations, the admixed population would be in HWE after one generation. In reality, where two populations have differences in language, culture or religion, admixture is normally a much longer process.

Natural selection

At some loci in the human genome the effect of selective pressures can be detected, for example lactase persistence that is present in populations where milk has been a sustained part of the diet [9, 10]. Mutations that can confer disease resistance can also exhibit strong selection effects. The mutation CCR5-Δ32 allele, that is thought to offer protection against the haemorrhagic plague that led to vast numbers of Europeans dying between 1347 to 1670 AD, occurs at a frequency of almost zero in Asian, African and American Indian populations, whereas it is present at a frequency of 0.16 (16 %) in European populations [11]. However, the loci that are used for forensic testing are not located within functionally important regions of the genome and there is no evidence that they are under selective pressure.

Mutation

Mutation at STR loci is relatively rapid and it is the instability at these loci that leads to their high levels of polymorphisms – a trait that makes them valuable genetic markers. However, the mutation rates of STRs are still relatively low at less than 0.2 % per generation and do not have a significant effect on the allelic frequencies within a gene pool [12–15].

Statistical tests to determine deviation from the Hardy–Weinberg equilibrium

Given that no human population can meet the requirements of the HW law can we then use it to calculate genotype proportions based on allele frequencies? The answer from most forensic scientists is yes – because we can empirically measure the predicted genotype frequencies under HWE and detect if there is a significant amount of deviation.

Many statistical tests have been developed to calculate the deviation of the allelic frequencies from HWE. These include the goodness-of-fit test (also called the chi square test), homozygosity test, likelihood ratio test and the exact tests [16]. However, when analysing polymorphic STR loci these tests do not have the required sensitivity because there are many undetected genotypes, and numerous genotypes, that are detected at very low frequencies at each locus. The multi-locus exact test was developed and can detect deviation from HWE when a large dataset is tested [17, 18]. Significant deviations from HWE have not been detected in the vast majority of populations. An exact test will not detect variations from HWE in small datasets, unless the deviation is extreme, and therefore conclusions from performing the exact test should not be over interpreted.

Estimating the frequencies of STR profiles

In forensic DNA analysis the HWE is used along with an allele frequency database to calculate genotype frequencies. An allelic frequency database is constructed by measuring the occurrence of alleles within the defined population. It has been recommended that a database of at least 200 alleles per locus (or 100 individuals) be used for a particular population when using the database for generating the statistical estimates of the strength of DNA evidence [19]. The larger the database, the more representative of the population it will be, and current practice dictates that several hundred individuals should be sampled when creating an allelic frequency database. These people should not be direct relations, therefore siblings or mother and child, etc., combinations should not be incorporated into an allele frequency database.

Using the HWE, the expected genotype frequency at each locus is calculated using the observed allele frequencies. Using these frequencies along with the above HWE equations we can calculate the frequency of a STR profile. If we take the profile that was analysed in Chapter 6, the genotype proportions for each locus are calculated using p^2 for the homozygote and $2pq$ for the heterozygote loci (Table 8.1). The overall profile frequency is calculated by multiplying the genotype frequency at each locus. This multiplication is termed the product rule – it is possible because the inheritance of alleles at each locus is independent of the other loci.

There have been some challenges to the approach presented above, namely that the inaccurate estimation of allelic frequencies can lead to inaccurate profile frequency estimates. To overcome this problem several methods have been employed that take into consideration the limitations in allele frequency estimates.

Corrections to allele frequency databases

Allelic frequencies are calculated by measuring a number of alleles in the target population. The more alleles that are measured as a part of the allelic frequency database the more accurate it will be. However, it is impractical to measure all of the alleles in

Table 8.1 The profile frequency is estimated using the principles of the Hardy–Weinberg law and an allele frequency database that was constructed using 400 alleles. Because the loci are all on different chromosomes there is no genetic linkage and the product rule can be used, multiplying each genotype frequency to calculate the overall profile frequency

Locus	Allele	Allele frequency	HWE	Genotype frequency
D3S1358	15	0.2825	$2pq$	0.1257
	17	0.2225		
vWA	14	0.0850	$2pq$	0.0425
	17	0.2500		
D16S539	11	0.2975	$2pq$	0.1041
	13	0.1750		
D2S1338	24	0.1000	$2pq$	0.0240
	25	0.1200		
D8S1179	11	0.0625	$2pq$	0.0434
	13	0.3475		
D21S11	30	0.2625	$2pq$	0.0551
	31.2	0.1050		
D18S51	14	0.1675	$2pq$	0.0477
	15	0.1425		
D19S433	14	0.3275	$2pq$	0.0164
	15.2	0.0250		
TH01	9	0.1375	$2pq$	0.0963
	9.3	0.3500		
FGA	21	0.1775	p^2	0.0315
	21	0.1775		
			Profile frequency	$7.579\ 10^{-14}$

a large population and the frequencies are only estimates, prone to inaccuracies due to the limited size of the database. For common alleles the impact is small but with rare alleles, which can easily be under represented in a frequency database, the impact of limited sampling can have a large effect. It should be noted that the deficiencies in the frequency databases can also lead to over representation of allele frequencies but, as a general principle, when we are estimating the significance of forensic evidence the emphasis is not to over state the strength of the evidence. Different approaches have been taken to overcome the limitations of allele frequency databases. These include the allele frequency ceiling principle, the Balding size bias correction [20], allowing for the effects of subpopulations [21] and using a maximum profile frequency [22].

Allele ceiling principle

Very rare alleles may not appear at all in the frequency database. If a rare allele not previously represented on the frequency database is detected in a crime scene sample then the frequency of the allele would be 0 – which cannot be the case! A mechanism must be put in place to deal with this situation. One approach is to set a minimal allele frequency. The minimum frequency values that are used vary from country to country but are typically around 0.01 (1 %). Any allele occuring with a frequency of less than 0.01 will be adjusted to this figure. An alternative approach is to use a minimal allele count, for example five alleles being the smallest number of alleles that is considered: the allele frequency is simply calculated using the formula $5/2N$, where N is the number of individuals in the database [23].

Simple correction for sampling bias

Allele frequency databases are relatively small when compared with the populations from which they are drawn and therefore there remain sampling uncertainties. A simple method for addressing such uncertainties, which are inherent in allele frequency databases, is suggested by Balding [20]. The allelic information in the evidential material is incorporated into the database to adjust for the potential under-representation of alleles. When there are matching DNA profiles there must be two DNA profiles: one from the crime scene and one from the reference sample. The alleles from these profiles are added to the allelic frequency database. By adding both profiles we are making the assumption that the material found at the crime scene did not come from the suspect. If we look at the profile in Table 8.1, at the vWA locus is a heterozygous locus with alleles 14 and 17; these have frequencies of 0.0850 and 0.2500 respectively. By multiplying the allele frequency with the total number of alleles in the database, we can calculate that the numbers of observed alleles in the database are 34/400 for allele 14 and 100/400 for allele 17. We now have two profiles to add to the database; we have seen a total of four new alleles: 14, 17 in the crime scene sample and also 14, 17 in the suspect's sample. These can be added to the database and the frequency recalculated. The database now has 36 observations of allele 14 out of a total of 404 observed alleles, which leads to an allele frequency of 0.090. Similarly, for allele 17 we now have 102/404, which gives us an allele frequency of 0.2525. This procedure is repeated for each heterozygous locus.

In Table 8.1 the FGA locus is homozygous and in the original database we have 71/400 observations but now need to add four more observations (21, 21 and 21, 21) to both the frequency of allele 21 and the total number of alleles, so the new frequency is $75/404 = 0.1856$. The profile is recalculated using this correction method in Table 8.2.

The Balding correction for size bias has the greatest impact when the database is made from a small number of alleles or when the allele is rare. If the allele is common and the database is large, the effect is negligible.

The above methods both compensate for the limitations of allele frequency databases that are caused by sampling effects. Other more complex methods, such as calculating the confidence 95 % interval, can be employed but are not widely used [23, 24].

Subpopulations

In addition to correcting for sampling effect, it may also be necessary to allow for the presence of subpopulations when calculating profile frequencies. Even within

Table 8.2 The profile frequency has been recalculated from Table 8.1 using the Balding correction for sampling bias. The impact of this correction factor is greatest on the rare alleles

Locus	Alleles	Allele frequency	Allele count	Corrected allele frequency	HWE	Genotype proportion
D3S1358	15	0.2825	113	115/404 = 0.2847	$2pq$	0.1282
	17	0.2225	89	91/404 = 0.2252		
vWA	14	0.0850	34	36/404 = 0.0891	$2pq$	0.0450
	17	0.2500	100	102/404 = 0.2525		
D16S539	11	0.2975	119	121/404 = 0.2995	$2pq$	0.1068
	13	0.1750	70	72/404 = 0.1782		
D2S1338	24	0.1000	40	42/404 = 0.1040	$2pq$	0.0257
	25	0.1200	48	50/404 = 0.1238		
D8S1179	11	0.0625	25	27/404 = 0.0668	$2pq$	0.0466
	13	0.3475	139	141/404 = 0.3490		
D21S11	30	0.2625	105	107/404 = 0.2649	$2pq$	0.0577
	31.2	0.1050	42	44/404 = 0.1089		
D18S51	14	0.1675	67	69/404 = 0.1708	$2pq$	0.0499
	15	0.1425	57	59/404 = 0.1460		
D19S433	14	0.3275	131	133/404 = 0.3292	$2pq$	0.0196
	15.2	0.0250	10	12/404 = 0.0297		
TH01	9	0.1375	55	57/404 = 0.1411	$2pq$	0.0992
	9.3	0.3500	140	142/404 = 0.3515		
FGA	21	0.1775	71	75/404 = 0.1856	p^2	0.0345
	21	0.1775	71	75/404 = 0.1856		
				Profile frequency		$1.4225\ 10^{-13}$

populations of the same broad ethnic group, the population is not homogeneous but comprises related subpopulations. The subpopulations form because people do not mate randomly, but tend, for example, to have children with people from the same geographical area or same social group. Allelic databases are normally composed of samples that have been drawn from the general population, and not from one subpopulation, and therefore provide us with an average estimate of the allele frequencies in the whole population. The effect of subpopulations has been demonstrated as leading to errors in the estimation of profile frequencies [25]. In a subpopulation there is a higher degree of relatedness between individuals than there is to the whole population, i.e. a higher probability that two individuals would have some genetic markers in common through descent from a common ancestor (identical by descent) than by a random match (identical by state) [26]. To incorporate this substructure factor into the profile frequency calculations, a theta value (θ) is used to describe the degree of differentiation between subpopulations (the amount of inbreeding) [27]. The level of population substructure, and therefore the theta values at the STR loci, have been demonstrated to be low [23, 24, 28, 29]. In general a theta value of 0.01 is used for seemingly homogeneous populations, while for more isolated/differentiated populations a theta value of 0.03 has been recommended [23]. To calculate the profile frequencies that allow for subpopulations the following equations are used commonly used [21]:

For homozygotes:

$$\text{Profile frequency} = \frac{[2\theta + (1-\theta)p_i]\,[3\theta + (1-\theta)p_i]}{(1+\theta)(1+2\theta)} \tag{8.1}$$

For heterozygotes:

$$\text{Profile frequency} = \frac{[2\theta + (1-\theta)p_i]\left[\theta + (1-\theta)p_j\right]}{(1+\theta)(1+2\theta)} \tag{8.2}$$

We can use this to recalculate the profile frequency presented in Table 8.1: the calculations for the vWA and FGA locus are shown below with a theta value of 0.01.

vWA

$$\frac{2[0.01 + (1-0.01)0.0850][0.01 + (1-0.01)0.2500]}{(1+0.01)(1+(2\times 0.01))}$$

$$\frac{2[0.0942][0.2575]}{(1.01)(2.02)}$$

$$\frac{0.0485}{1.0302} = 0.0471$$

FGA

$$\frac{[(2\times 0.01) + (1-0.01)0.1775][(3\times 0.01 + (1-0.01)0.1775]}{(1+0.01)(1+(2\times 0.01))}$$

$$\frac{[0.1957][0.2057]}{(1.01)(2.02)}$$

$$\frac{0.0403}{1.0302} = 0.0391$$

The impact of a theta value of 0.01 on this particular profile is a modest three fold increase in the profile frequency, whereas a theta value of 0.03 leads to a frequency that is over 20 times more common – but still exceedingly rare (Table 8.3). It should be noted that the impact of applying theta to a profile frequency calculation differs between profiles.

The current practice in most legal systems is to use a theta value of between 0.01 and 0.03, apart from in exceptional circumstances where very high levels of inbreeding may have occurred.

Profile ceiling principle

In some countries, such as the UK, the approach has been to use a match probability of 1 in a billion (1 000 000 000). This approach is highly conservative [22]. It does have the advantage that individual profile frequencies do not have to be calculated because the value used is much lower than the most common profile frequency, even if conservative corrections are incorporated [24].

Table 8.3 The effect of different correction methods on the profile frequency calculated in Table 8.1. With this profile, applying a minimum allele frequency of 0.0125 would have no impact because the rarest allele frequency is 0.025

Calculation method	Profile frequency	Fold reduction relative to uncorrected frequency
Uncorrected	7.58×10^{-14}	
Size bias	1.42×10^{-13}	2 (1.88)
Subpopulation: $\theta = 0.01$	2.44×10^{-13}	3 (3.29)
Subpopulation: $\theta = 0.03$	1.62×10^{-12}	21
Profile ceiling: 1 in 1 billion	1.00×10^{-9}	13,195

Which population frequency database should be used?

In some cases, the ethnic origins of material recovered from the crime scene are known: for example, if a woman has been sexually assaulted she can normally describe the assailant as white, black, Asian, etc. In such a case, for example, if the assailant was described as white, then it would be logical to use a white Caucasian allele frequency database to calculate the profile frequency. In other contexts, there may be no information about who could have left the material at the crime scene. In countries or regions having substantial populations with different ethnic backgrounds, a common practice is for the profile frequency to be calculated using an allele database for each major population group, and to use the most conservative profile frequency. If we take the example from Table 8.1, the allele frequency data used is from a white Caucasian database (USA); if we recalculate with allele frequency data representing an African American population we get a profile frequency of 3.36×10^{-16}, which is over 200-times less frequent than when we use the Caucasian frequency data. In this case it is clear that the Caucasian data provides a frequency estimate that is more conservative.

Conclusions

The methods that are employed for the correction of profile frequencies vary widely between different judicial systems and even different laboratories within the same judicial system. The allele ceiling principle, Balding correction, 95 % confidence interval to correct for sampling error, accounting for population sub-division using theta, and the profile ceiling principle have all been used in forensic casework to calculate profile frequencies. For the profile that we have been using as an example in this chapter, the effect of the different correction methods can be seen in Table 8.3.

It should be noted that the impact of the different correction methods will vary depending on the individual profile and the size of the allele frequency database. The end result of analysing a profile is to produce a profile frequency, which is an estimate. Incorporating one or more of the correction factors into the profile frequency estimates reduces the chances of overstating the DNA evidence.

Further reading

Balding, D.J. (2005) *Weight-of-evidence for Forensic DNA Profiles*. John Wiley & Sons, Ltd, Chichester, pp. 56–81.

Buckleton, J., Triggs, C.M., and Walsh, S.J. (2005) *Forensic DNA Evidence Interpretation*. CRC Press, pp. 341–347.

References

1. Hardy, G. (1908) Mendelian proportions in a mixed population. *Science* **28**, 49–50.
2. Stern, C. (1943) The Hardy–Weinberg law. *Science* **1997**, 137–138.
3. Helgason, A., *et al.* (2003) A reassessment of genetic diversity in Icelanders: Strong evidence from multiple loci for relative homogeneity caused by genetic drift. *Annals of Human Genetics* **67**, 281–297.
4. Destro-Bisol, G., *et al.* (2000) Microsatellite variation in Central Africa: An analysis of intrapopulational and interpopulational genetic diversity. *American Journal of Physical Anthropology* **112**, 319–337.
5. Pardo, L.M., *et al.* (2005) The effect of genetic drift in a young genetically isolated population. *Annals of Human Genetics* **69**, 288–295.
6. Chakraborty, R. (1987) Biochemical heterozygosity and phenotypic variability of polygenic traits. *Heredity* **59**, 19–28.
7. Smouse, P.E., *et al.* (1983) Multiple-locus departures from panmictic equilibrium within and between village gene pools of Amerindian tribes at different stages of agglomeration. *Genetics* **104**, 133–153.
8. Sinnock, P. (1975) Wahlund effect for 2-locus model. *American Naturalist* **109**, 565–570.
9. Tishkoff, S.A., *et al.* (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics* **39**, 31–40.
10. Wooding, S.P. (2007) Following the herd. *Nature Genetics* **39**, 7–8.
11. Duncan, S.R., *et al.* (2005) Reappraisal of the historical selective pressures for the CCR5-Delta 32 mutation. *Journal of Medical Genetics* **42**, 205–208.
12. Brinkmann, B., *et al.* (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *American Journal of Human Genetics* **62**, 1408–1415.
13. Dauber, E.M., *et al.* (2003) Mutation rates at 23 different short tandem repeat loci. In *Progress in Foresnsic Genetics* **9**, 565–567.
14. Yan, J.W., *et al.* (2006) Mutations at 17 STR loci in Chinese population. *Forensic Science International* **162**, 53–54.
15. Whittle, M.R., *et al.* (2004) Updated Brazilian genetic data, together with mutation rates, on 19 STR loci, including D10S1237. *Forensic Science International* **139**, 207–210.
16. Emigh, T.H. (1980) A comparison of tests for Hardy–Weinberg Equilibrium. *Biometrics* **36**, 627–642.
17. Guo, S.W., and Thompson, E.A. (1992) Performing the exact test of Hardy–Weinberg Proportion for Multiple Alleles. *Biometrics* **48**, 361–372.
18. Wigginton, J.E., *et al.* (2005) A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics* **76**, 887–893.
19. Brinkmann, B. (1998) Overview of PCR-based systems in identity testing. In *Forensic DNA Profiling Protocols*. Humana Press, pp. 105–120.
20. Balding, D.J. (1995) Estimating products in forensic identification using DNA profiles. *Journal of the American Statistical Association* **90**, 839–844.
21. Balding, D.J., and Nichols, R.A. (1994) DNA profile match probability calculation – how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International* **64**, 125–140.
22. Foreman, L.A., and Evett, I.W. (2001) Statistical analysis to support forensic interpretation for a new ten-locus STR profiling system. *International Journal of Legal Medicine*, 147–155

23. National Research Council (1996) *The Evaluation of Forensic DNA Evidence*. National Academy Press.
24. Gill, P.*, et al.* (2003) A comparison of adjustment methods to test the robustness of an STR DNA database comprising of 24 European populations. *Forensic Science International* **131**, 184–196.
25. Curran, J.M.*, et al.* (2003) What is the magnitude of the subpopulation effect? *Forensic Science International* **135**, 1–8.
26. Triggs, C.M., and Buckleton, J.S. (2002) Logical implications of applying the principles of population genetics to the interpretation of DNA profiling evidence. *Forensic Science International* **128**, 108–114.
27. Buckleton, J.S.*, et al.* (2006) How reliable is the sub-population model in DNA testimony? *Forensic Science International* **157**, 144–148.
28. Foreman, L.A., and Lambert, J.A. (2000) Genetic differentiation within and between four UK ethnic groups. *Forensic Science International* **114**, 7–20.
29. Foreman, L.A.*, et al.* (1998) Regional genetic variation in Caucasians. *Forensic Science International* **95**, 27–37.

9 The evaluation and presentation of DNA evidence

The final stage in any criminal case is the presentation of the evidence to the court. The way in which DNA evidence is presented has been, and still can be, a contentious subject. The evaluation process and the wording of the court reports and statements will be affected by both the judicial system and the prevailing approach to DNA evidence; this varies considerably between different countries. This chapter is designed as an introduction to the field and will guide the reader through the basics of the evaluation of DNA evidence.

Hierarchies of propositions

Any statement on the strength of the DNA evidence must be considered in the context of the case. DNA evidence should not be considered in isolation as it is affected by many factors like the type of biological material, method and time of deposition and the substrate on which it was deposited.

There are three hierarchies of propositions in relation to biological material that can be considered in a criminal trial [1,2]:

(1) Source level: from which individual did the biological material originate?

(2) Activity level: what activity led to the deposition of the biological material?

(3) Offence level: did the suspect commit the offence?

The hierarchies can be applied when, for example, considering a bloodstain found on the clothing of a suspect alleged to have committed an assault. The DNA evidence may assist in determining the most likely source of the stain. The size, shape and position of the stain may assist with determining the activity associated with the stain, i.e. how was it transferred and does this support an allegation of kicking or punching. The third hierarchical statement is that of offence: did the suspect commit the offence? The first hierarchical level can be addressed by DNA analysis and in many cases the second

An Introduction to Forensic Genetics W. Goodwin, A. Linacre and S. Hadi

© 2007 John Wiley & Sons, Ltd

hierarchical level can be addressed to some extent by the forensic scientist, for example, by the interpretation of blood spatter patterns, but the third level is the provenance of the court and at no time should an expert witness comment whether the defendant is guilty of the offence. This is clearly the task of the court to consider [3, 4].

To answer the first question, there are currently three approaches to the evaluation of DNA evidence. The three approaches are termed:

the frequentist approach;

the likelihood approach; and

the Bayesian approach.

The starting point for each of these approaches is the frequency of the DNA profile (see Chapter 8).

The frequentist approach

The profile frequency is presented as a random match probability, which can be taken as the reciprocal of the profile frequency.

$$\text{Random match probability} = \frac{1}{\textit{profile frequency}} \tag{9.1}$$

Before the advent of DNA profiling, the results of blood groups and protein polymorphisms were expressed as random match probabilities, so for example, a report might state that ‘approximately 1 in 250 unrelated people will share this blood type’. It was natural to use this same wording with the advent of DNA analysis. In simple terms the frequentist approach describes the chance of a coincidental random match. Random match probability (also called random occurrence ratio) is the probability of a person, selected at random, having the same profile as the defendant [5].

If we take a profile with a frequency of 0.000 001 the random match probability will be:

$$\text{Random match probability} = \frac{1}{0.000\,001} = 1 \text{ in } 1\,000\,000$$

The use of multiplex STR kits can lead to match probabilities of hundreds of billions—far larger numbers than there are people on the planet. In the UK the approach to such large numbers has been to employ a ceiling principle so that a figure of 1 in 1 billion is always quoted when describing a match based on a full AmpF*l*STR® SGM Plus® profile. There is little scientific merit to this approach but it is considered pragmatic. An example of a statement that is used in criminal reports in the UK is presented in the match probability statement.

Match probability statement

DNA analysis of the blood stain from the crime scene gave a full DNA profile that matched that of the suspect. If this blood did not come from the suspect then the STR profile must match by chance. It is estimated that the chance of obtaining these matching profiles if the blood came from a random person unrelated to the suspect is in the order of 1 in a billion (a billion is a thousand million).

Many countries do not use the ceiling value of 1 in a billion and quote the calculated random match probability. In some countries, when the value gets to a certain point the evidence is no longer presented as a random match probability, rather it is presented as coming from the suspect. For example, in the US, random match probabilities of greater than 1 in 260 billion may be considered as identity by the US reporting scientist [6].

Statement of uniqueness

Based on the results of the DNA profiling, Mr J. Doe is the source of the DNA obtained from the evidentiary material, to a reasonable degree of scientific accuracy.

The frequentist approach has the advantage that when dealing with small numbers the phraseology can be understood. It should be noted that part of the duty of the forensic scientist is to make the strength of the evidence understood by the jury or a judge and, therefore, to that extent the frequency approach succeeds. Thus, quoting a random match probability figure of 1 in 1 million, for instance, is relatively simple for the jury to understand and picture. Simplistically, if there are 1 million other people it would be expected that one person in that population would have the same DNA profile. There are however a number of disadvantages with the approach. Assume a full AmpF*l*STR® SGM Plus® DNA profile where the match probability is greater than 1 in 1 billion of the population. This leads to the observation that there are not 1 billion people in the UK so what population is being described? Indeed there are only 6.5 billion people on the planet at present. Such logical problems with the frequentist approach can be addressed by using a likelihood ratio.

Likelihood ratios

A likelihood ratio is the ratio of two competing hypotheses. In terms of a criminal case, it is the ratio of the prosecution hypothesis (H_p) and the defence hypothesis (H_d). The likelihood approach is a more logical way to interpret and present the profile frequency information as it considers an alternative scenario.

Three logical principles for the interpretation of DNA evidence and its quantification that have been suggested [7] are as follows:

(1) When evaluating DNA evidence two assumptions should be considered.

(2) Probability of occurrence of the evidence under each of the two assumptions should be quantified.

(3) The ratio of the probabilities under two assumptions should be quantified and considered.

By considering the two propositions where one is the alternative of the other (the prosecution proposition compared to the defence proposition), then the probability of the evidence if the prosecution case is true can be determined. If the DNA profile of a crime scene sample matches a suspect's DNA profile then there can be two explanations:

(1) Prosecution hypothesis (H_p): the DNA profile originated from the suspect.

(2) Defence hypothesis (H_d): the DNA profile did not originate from the defendant but originated from another person.

The likelihood ratio is described in equation (9.2). It is the probability (Pr) of the DNA evidence (E) given the hypothesis put forward by either the prosecution (H_p) or the defence (H_d):

$$\text{Likelihood ratio} = \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \tag{9.2}$$

The prosecution hypothesis is that the defendant left the biological material at the crime scene. The DNA profiles of the defendant and the evidence from the crime scene match, and therefore it is certain under the prosecution's hypothesis that the defendant left the material, hence $\Pr(E|H_p) = 1$. The probability of occurrence under the defence hypothesis is equal to the probability of observing the profile if its source was somebody other than the suspect in the population to which the defendant belongs. The ratio of the two probabilities is given by:

$$\text{Likelihood ratio} = \frac{1}{\textit{profile frequency}} \tag{9.3}$$

Let us assume that a certain crime scene stain yielded a DNA profile that has matched that of a suspect who is being prosecuted. The frequency of the matching DNA profile was 1 in a million. The likelihood ratio in this case would be:

$$\text{Likelihood ratio} = \frac{1}{0.000\,001} = 1\,000\,000$$

Because with DNA evidence the probability under the prosecution's hypothesis is equal to 1, the value is the same as when calculating a random match probability;

however, the way that the value is expressed differs. The likelihood ratio would normally be expressed in a report with a statement, for example, as in the likelihood ratio statement.

Likelihood ratio statement

DNA analysis of the blood stain from the crime scene gave a full DNA profile that matched that of the suspect. If this blood did not come from the suspect then the STR profile must match by chance. The results of the DNA analysis are approximately 1 million times more likely if the DNA came from the suspect than if the DNA came from a random unrelated male in the population.

The figure being quoted in a likelihood ratio is the odds in favour of the proposition put forward by the prosecution. When the statement reads that 'it is 1 million times more likely that the DNA came from the accused than if it came from any unrelated male', the figure of a million is not a probability but is an odds value, i.e. how many times more likely it is that the DNA matched the crime scene stain if it originates from the suspect, compared to coming from any other unrelated male. The frequency approach has a problem when the chance of a match, or the match probability, exceeds the total sample size. This is not the case when quoting odds in favour as odds can reach near infinity. In a horse race with only five horses, one horse may have odds against of 10 to 1 and here the odds outweigh the number of possibilities.

The disadvantage with this statement is that it can seem to be cumbersome when presented to a jury. It is easy to make an error and state the probability that the evidential material came from the suspect instead of the probability that the DNA profile obtained from the evidential material matches that of the suspect.

Using the current multiplex STR kits that analyse between 10 and 15 STR loci, the match probabilities and therefore likelihood ratios are extremely high. In order to avoid some of these complications in presenting huge numbers, and also because it is easier for the scientist to express verbally the weight of the evidence, it has been suggested [3, 7] that verbal scales might be used for likelihood ratios as presented in the table below:

Table 9.1 Verbal scales for likelihood ratios

Likelihood ratios	Verbal equivalent
1–10	Limited support for prosecution hypothesis
10–100	Moderate support for prosecution hypothesis
100–1000	Moderately strong support for prosecution hypothesis
1000–10 000	Strong support for the prosecution hypothesis
>10 000	Very stong support for prosecution hypothesis

The use of verbal equivalents is itself a contentious issue because of its subjective nature and also because some believe that it is encroaching on the role of the jury. Further, there is no rationale for the boundaries, for example, to discriminate between a likelihood ratio of 99 and one of 101.

The Bayesian approach

The Bayesian approach is favoured by many forensic scientists but has not gained widespread usage in the presentation of DNA evidence. The approach builds upon the likelihood but allows non-scientific data to be introduced in the form of prior odds. The non-scientific data will update the likelihood ratio to produce the final odds either in favour of, or against, the proposition put forward by the prosecution or defence (equation 9.4).

$$\frac{\Pr H_\mathrm{p}}{\Pr H_\mathrm{d}} \times \frac{\Pr(E|H_\mathrm{p})}{\Pr(E|H_\mathrm{d})} = \frac{\Pr(H_\mathrm{p}|E)}{\Pr(H_\mathrm{d}|E)} \tag{9.4}$$

(or prior odds × likelihood ratio = posterior odds)

Consider the case against Dennis Adams (*R* v *Adams* [1996] Cr. App. R., Part 3) in the UK where he was accused of a rape that occurred in 1991. The trial was in 1994 when the DNA profiling methodology was based upon VNTR analysis that predated STR typing. The DNA evidence put forward by the prosecution was that the DNA profile occurred in 1 in 200 million of the population. Adams pleaded not guilty; he had an alibi and was not identified at an identity parade. The defence expert produced numerical values for the crime being committed by a local man, for the possibility of not being identified at the parade, and for the alibi. All these prior probabilities were multiplied to determine the prior probability. The defence argued that the evidence (genetic and non-genetic) indicated the innocence of the accused. At the trial the judge allowed this to happen and directed the jury that they could use the Bayesian figure if they wished. Adams was found guilty.

The use of Bayesian evaluation of DNA evidence in the UK legal system has not been accepted and has led to successful appeals against convictions, including the above example, when the Appeal Court took the view that the use of Bayesian statistics trespassed on areas exclusively and peculiarly those of the jury. The relationship between different pieces of evidence was for the jury to decide and the mathematical formula might be applied differently by a different set of jurors. Jurors should evaluate the evidence by the joint application of their common sense and knowledge of the world to the material before them. A significant ruling was laid down in the English Courts following another appeal; (*R* v *Doheny and Adams* [1996] EWCA Crim 728). A number of the relevant points can be summarized:

(1) The scientist should give the frequency of the occurrence with which the DNA profile is likely to be found in the population.

(2) It might be appropriate, if the scientist has the necessary data and statistical expertise, to say how many people might be found to have matching profiles in the United Kingdom or in a limited sub-group of individuals (the idea is to give the jury an estimate of how many people in the relevant section of the population are expected to have a matching profile and, therefore, could be the source of the stain).

(3) The jury would then decide, on all the information available, whether the stain originated from the suspect or some other individual with a matching profile.

(4) To help the jury, the judge might direct them along the following lines:

> ...if you accept the evidence that indicates there are only four or five (or what ever figure) men in the UK population from whom the stain could have originated and the suspect is one of them, are you sure the suspect left the stain or is it possible it was one of the other individuals in the small group who has a matching profile.

Two fallacies

When presenting evidence two errors can be committed if the wording used is not precise — care should be taken to avoid committing the prosecutor's or defendant's fallacy.

Prosecutor's fallacy

This fallacy in describing the strength of the evidence is also called 'transposed conditional' [8]. If a horse was described as a four legged animal it does not transpose that every four legged animal is a horse. Similarly the statement 'the probability of gaining this DNA profile IF it came from someone other than the suspect is 1 in 1 million', does not mean that 'the probability that the evidence came from someone other than the suspect is 1 in 1 million". The first statement considers the probability of the evidence given the hypothesis and is correct, but the second statement considers the probability of the hypothesis given the evidence, and is a clear case of the prosecutor's fallacy.

In the case of Andrew Deen (*R* v *Deen* [1994] *The Times*, January 10th, 1994) in the UK, the DNA analyst incorrectly defined match probability as the 'probability of the semen having originated from someone other than Andrew Deen'. In the examination in chief, the DNA analyst said 'the likelihood of (the source of the semen) being any other man but Andrew Deen is one in three million'. It is clear that the analyst transposed the condition of the hypothesis and it was one of the reasons that the conviction was quashed by the Court of Appeal.

As another example, consider the following dialogue from *R* v *Doheny and Adams* (*R* v *Doheny and Adams* [1996] EWCA Crim 728):

Q. Is it possible that the semen could have come from a different person from the person who provided the blood samples?

A. It is possible but it is so unlikely as to really not be credible. I can calculate; I can estimate the chances of this semen having come from a man other than the provider of the blood sample. I can work out the chances as being less than 1 in 27 million.

Instead of estimating the probability of semen (crime scene sample) matching blood (of the suspect) if the suspect was innocent, the DNA analyst was estimating the probability of the semen matching suspect's blood if the suspect left the semen stain. For the population of the UK a match probability of 1 in 27 million means that other persons could have a matching profile and without other supporting evidence may not in itself provide sufficient evidence against the defendant. This however is a matter for the court to address and not the scientist. The prosecution expert witness thus inadvertently enhanced and misrepresented the probative value of the evidence [8]. It is therefore imperative that in order to avoid the prosecutor's fallacy, the scientists write their report carefully and while answering any questions in the court keep their match probability or likelihood statements conditioned on 'if the defendant was innocent'.

Defendant's fallacy

If the match probability for a DNA profile was 1 in 27 million as it was in the case against Doheny, the defence could argue, for example, that three people in the UK might match the crime scene profile, the probability of the defendant being the donor of the crime scene sample is therefore only 1/3, which is insufficient for proof beyond reasonable doubt. The issue with this statement is that nothing is known of these three people; whether they exist, where they live, what age they are, what sex they are, and what opportunity they might have to leave their DNA at the scene.

Comparison of three approaches

The high statistical values that are attached to DNA profiles might seem intimidating and can unduly enhance the probative value of DNA evidence. This had led to heated debate over the way in which the evidence should be presented to a court.

The frequentist approach is straightforward and understandable by both a jury and a judge. For the reporting officer it is straightforward to state in court and the opportunity for transposing the conditional and stating the prosecutor's fallacy is less than with the other two approaches. A disadvantage of the approach is that it does not consider two propositions where one is the alternative of the other. The likelihood ratio is a logical approach, it considers an alternative hypothesis. The Bayesian approach is the most logical way to incorporate all evidence in a case; it considers alternate hypotheses but it is difficult to calculate and conceptualize.

Further reading

Balding, D.J. (2005) *Weight-of-evidence for Forensic DNA Profiles*. John Wiley & Sons, Ltd, Chichester, pp. 145–156.

Buckelton, J., Triggs, C.M., and Walsh S.J. (2005) *Forensic DNA Evidence Interpretation*. CRC Press, pp. 27–63.

Evett, I. W., and Weir, B. S. (1998) *Interpreting DNA Evidence – Statistical Genetics for Forensic Scientists*. Sinauer Associates, pp. 217–246.

References

1. Cook, R.*, et al.* (1998) A model for case assessment and interpretation. *Science and Justice* **38**, 151–156.
2. Evett, I.W.*, et al.* (2000) More on the hierarchy of propositions: exploring the distinction between explanations and propositions. *Science and Justice* **40**, 3–10.
3. Evett, I.W.*, et al.* (2000) The impact of the principles of evidence interpretation on the structure and content of statements. *Science and Justice* **40**, 233–239.
4. Taroni, F., and Aitken, C.G.G. (2000) DNA evidence, probabilistic evaluation and collaborative tests. *Forensic Science International* **108**, 121–143.
5. Meester, R., and Sjerps, M. (2004) Why the effect of prior odds should accompany the likelihood ratio when reporting DNA evidence. *Law Probablity and Risk* **3**, 51–62.
6. DNA Advisory Board (2000) Statistical and population genetic issues affecting the evaluation of the frequency of occurrence of DNA profiles calculated from pertinent population databases. *Forensic Science Communications* **2**. http://www.fbi.gov/hq/tab/fsc/backissu/july2000/dnastat.htm
7. Evett, I.W., and Weir, B.S. (1998) *Interpreting DNA Evidence*. Sinauer Associates, Inc.
8. Balding, D.J., and Donnelly, P. (1994) The prosecutor's fallacy and DNA evidence. *Criminal Law Review*, 711–721.

10 Databases of DNA profiles

Several countries have developed national DNA databases that contain large numbers of DNA profiles – the UK and the USA national DNA databases now both contain the DNA profiles of over 3 million individuals. DNA databases that store STR profiles have emerged as a powerful tool in the investigation of crime. The effective use of the DNA database, in particular in the UK, has acted as a catalyst for the establishment and expansion of DNA databases in other countries, including the USA and many European countries that now have databases with hundreds of thousands of profiles stored in them. This chapter will examine the development and application of the UK national DNA database, which is the first and most extensive database of its kind. It will briefly examine the development of databases worldwide.

The UK National DNA database (NDNAD)

The UK NDNAD was established in 1995 [1], shortly after STR profiling using six STR loci (the SGM) was introduced into criminal casework.

Rationale for criminal databases in the UK

There are several justifications for the time, effort and money that a criminal DNA database consumes:

- Criminals tend to re-offend – 90% of rapists have had a previous conviction; 50% of armed robbers have a previous conviction.
- The severity of crimes often increases — in many instances criminal activity starts at a young age with many criminals committing their first offence between 16–19 years of age.
- A small number of criminals can be responsible for a large number of crimes – linking these crimes together can aid police investigations. This is particularly the case for burglaries, auto crimes, and serious cases such as sexual assaults.

An Introduction to Forensic Genetics W. Goodwin, A. Linacre and S. Hadi
© 2007 John Wiley & Sons, Ltd

Legislation

The UK DNA database did not require specific statutes for its establishment although the police service launched the national DNA database at the same time that the provisions of the Criminal Justice and Public Order Act 1994 came into force on 10th April 1995. Subsequent legislation has increased the scope of samples that may be collected and retained on the NDNAD (see next section).

Legislation in England and Wales

1994 The Criminal Justice and Public Order Act

Within the UK the Police and Criminal Evidence Act 1984 (PACE), which governs the taking of samples from persons suspected of criminal activity, was amended to reclassify saliva and mouth swabs as non-intimate, thus allowing the samples to be collected without consent and without the need for a medical practitioner.

1997 The Criminal Evidence (Amendment) Act

This allowed non-intimate samples to be collected from inmates currently in prison but convicted of an offence prior to the establishment of the NDNAD.

2001 Extension to the Police and Criminal Evidence Act 1984 (PACE)

This allowed samples to be retained indefinitely irrespective of whether the person was acquitted at trial and from samples obtained from volunteers taking part in mass screens, provided that these volunteers gave their consent.

2003 Extension to the Criminal Justice Act

Section 63 of the Police and Criminal Evidence Act (1984) was amended to allow the police to take a non-intimate sample from a person in police detention who has been arrested for, charged with, informed they will be reported for, or convicted of, a recordable offence. These powers came into force in 2004.

Criteria for entry onto the UK NDNAD

The original criterion for addition of a sample from an individual to the National DNA Database was that the person had been arrested for an offence punishable by imprisonment. If the person was found not guilty at a subsequent trial, or the case was discontinued, then their profile would be removed. In 2001 the Criminal Evidence Act allowed samples to be retained on the NDNAD, even if the individual was not found guilty. The regulations were further relaxed in 2003 with the Extension to the

Criminal Justice Act. Calls have been made in the UK by police chiefs and politicians for everybody to be entered onto the NDNAD – this prospect is still some way in the future.

Technology Underlying the NDNAD

The development of STR profiling was essential for the successful implementation of a large-scale DNA database. Attempts had been made to construct databases of VNTR profiles, and these did produce some successes. However, the difficulty of comparing VNTR profiles was a major limitation. STR profiles can be digitized very easily and this has allowed for the effective computerization of DNA profiles.

The UK NDNAD was established using the SGM multiplex, which analysed six STR loci and the amelogenin locus. The match probability of SGM was 1 in 10^8 of the population, which for a population of 58 million within the UK was deemed acceptable. However, when six loci were used there were a number of coincidental matches (see below).

Adventitious Hits

In 1995 Raymond Easton was asked to donate a DNA profile as part of an investigation into a domestic dispute. Four years later a burglary at a home approximately 200 kilometres from where Raymond Easton lived generated a DNA profile that was compared to the NDNAD. This profile matched that of Raymond Easton and he was accused of the crime. A match probability of 1 in 37 million was reported. At the time of the burglary, Raymond Easton was suffering from Parkinson's disease and was unable to walk more than 10 metres unaided. This was an example of an adventitious hit. The test was conducted using the SGM loci, but the chance that a similar adventitious cold hit will occur has been reduced greatly by extending the test to ten loci.

In 1999, the six-locus SGM test was changed to the ten locus AmpF*l*STR® SGM Plus® test. The chance that two DNA profiles from unrelated people will match at all ten loci is less than 1 in 1 billion. To date no two people have been found to match at all ten loci; matches to two or more people can occur if a partial DNA profile is searched against the NDNAD.

Operation of the NDNAD

The NDNAD has two main sets of data: profiles generated from evidence that has been collected from crime scenes (263 000 at the end of 2005 [2]) and profiles generated from individuals (3.45 million at the end of 2005 [2]).

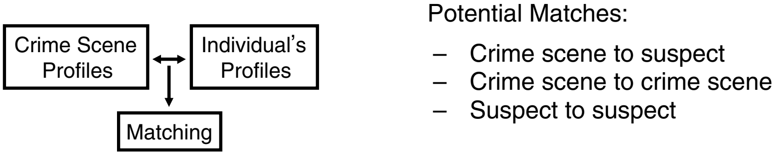


Figure 10.1 Following entry onto the database the new samples are searched against all other samples on the NDNAD. Suspect-to-suspect matches will only occur when individuals have given incorrect details to the police about their identity unless a coincidental match occurs – to date no coincidental matches have been reported with a full SGM Plus DNA profile

A biological sample from a scene will be collected by the scene-of-crime officers and submitted for DNA analysis. The resulting DNA profile will be compared to those currently held on the NDNAD and if there is a match then this will be reported back to the police force that collected the sample (Figure 10.1). A fresh sample from the individual to which there was a match will be collected and the DNA analysis will be repeated.

While the intention had been to use the NDNAD to match samples from serious crimes such as sexual assaults and murders, the addition of samples from high volume crime such as burglary resulted in an increase of DNA profiles on the NDNAD. In an average year the NDNAD produces around 40 000 crime scene to individual matches: the majority of these are high volume crime but there are invariably matches to more serious crimes such as murder, rape and assaults [2]. With such a large number of DNA profiles held on the NDNAD there is currently a 45% chance that a DNA profile obtained from an incident will match a DNA profile on the NDNAD [2].

In the UK, approximately 1 in 20 people are on the NDNAD; this includes 8% of the male population [2,3]. Ethical concerns have been raised that the NDNAD discriminated against vulnerable sections of society – 75% of young black males between the ages of 15 to 34 are on the database whereas only 22% of white males in the same age bracket are found on the NDNAD.

Familial searching

Familial searching was devised by the Forensic Science Service of the UK and is used when there is not a full DNA profile match between the crime scene and the NDNAD samples but a match is achieved at 15 or more alleles and the perpetrator most likely lived in the vicinity of the incident. While the person on the NDNAD can not be the donor of the sample obtained from the incident, it is highly likely that the originator of the sample is a relative of this person.

Familial searching

Craig Harman was the first person to be convicted of an offence following a link between a sample taken from a scene and a relative of the perpetrator. In March 2003, Craig Harman, then 19, was walking over a footbridge spanning the M3 motorway to the west of London when he dropped a brick onto passing traffic. The brick struck and broke the windscreen of a lorry, causing a fatal injury to the driver. The brick was examined for the presence of biological material, and fingerprints, and a DNA profile obtained. The DNA profile did not match in full with any person on the NDNAD but 16 of the 20 alleles matched a genetic relative of Craig Harman. A separate match between the sample obtained from the brick and a sample taken from Craig Harman after he was linked to the crime through the database resulted in a match and Craig Harman pleaded guilty to manslaughter.

Cold cases

Since the advent of PCR-based techniques it is now possible to obtain DNA profiles from old case samples. The application of low copy number (LCN) PCR has further increased the chance of obtaining DNA profiles from highly degraded material. Cases, such as those of murder, that have remained open from dates prior to the introduction of DNA typing can now be re-examined using either standard DNA testing or LCN in combination with the NDNAD. The current technology has allowed numerous cases to result in a conviction and therefore closure.

Cold cases

In 1969 Roy Tutill was a 14-year-old boy whose body was found in woodland near Leatherhead. He had been sexually assaulted and strangled. Samples collected from the body and the clothing of Roy Tutill were examined but blood group testing failed to give any satisfactory results.

In 2001 the UK FSS retested the medical swab extracts using SGM Plus and produced a partial DNA profile that was compared to the NDNAD. The DNA profile matched that taken from a Mr Brian Field who, 2 years earlier, had been stopped by police on a drink-driving offence and had donated a DNA profile. Further work was performed by the UK FSS on samples from the trousers of Roy Tutill that had been kept in the freezer and this gave a full DNA profile that matched Field. Field denied the charges at his first court appearance but pleaded guilty to murder when he appeared at the Old Bailey in November 2001.

Caution must be excercised when examining samples collected by crime scene operators prior to the advent of PCR-based techniques as it is unlikely that those handling the items will have taken the standard precautions to minimize contamination that are now standard practice.

International situation

Following the success of the operation in the UK, other countries developed their own DNA databases. For many countries there was a need to enact special legislation leading to delays in the implementation of DNA databases [3].

New Zealand implemented a DNA database in 1996 along similar lines to that of the UK. The population is significantly smaller but as a percentage of the population New Zealand is second only to the UK in terms of the number of DNA profiles held on its database. Australia and South Africa were also rapid in developing DNA databases.

In mainland Europe, almost all countries have established DNA databases although all are limited in comparison to the UK version. The Netherlands and Austria established their version of a DNA database in 1997, with Germany following one year later and Finland and Norway in 1999 [5].

Two countries in the Middle East, Kuwait and the United Arab Emirates, are both currently developing plans that would see the entire population analysed and placed on a DNA database.

US DNA database

The US Army established a database of their own in 1992 to identify missing persons in operation Desert Storm and this experience helped to pave the way for a national database within the US. In 1994 the US congress passed the DNA Identification Act (Public Law 103 322) which enabled the establishment of the Combined DNA Index System (CODIS). The CODIS, which is the federally held DNA database, has expanded very quickly and comprises the National DNA Index System (NDIS), the State DNA Index System (SDIS) and the Local DNA Index System (LDIS). The information about each sample that is loaded onto the CODIS database includes a laboratory identifier, a specimen identifier, information to classify and review the integrity of the DNA record, and the DNA profile itself. CODIS links local, state and federal crime laboratories. The FBI selected 13 STR loci (CODIS loci) for developing the database. Like the UK NDNAD there are two main segments called ‘indices’ of CODIS:

- The Forensic Index contains DNA profiles from crime scene samples.
- The Offender Index contains DNA profiles of individuals convicted of certain categories of violent crime, though now many states are expanding their databases and are profiling persons arrested for all felonies.

Other CODIS indices are:

- unidentified human remains;
- relatives of missing persons.

All 50 US states now have databases of which only 13 obtain DNA samples for databasing for all felonies. At the moment there are about 180 DNA laboratories around the USA that are designated and accredited as CODIS laboratories. These laboratories are validated according to the standards of FBI and are authorized to submit the DNA profile information into CODIS.

The situation in the US as of late 2006 is:

total number of profiles: 3 676 971

total forensic profiles: 148 068

total convicted offender profiles: 3 528 903

When compared with the UK, the USA is a much larger jurisdiction but due to lack of funding, coherent structure and variable legal approaches, there are lengthy delays in DNA profiling of casework samples that has led to massive backlogs. The President of the USA announced the 'President's DNA Initiative' in 2003 in order to enhance and streamline the use of DNA as a forensic tool and also signed an act to enhance the facilities for DNA databasing [4]. The main aims of this initiative are to clear the backlogs quickly and also to improve the capacity of the forensic laboratories for databasing the samples besides promoting research and development in the field.

Cross-border databases

Criminals tend to operate in their own country but there are circumstances when crimes will be committed in more that one country. In order for criminal databases to be effective in these circumstances there is a need to share data. Interpol has been instrumental in facilitating cross-border comparisons of DNA profiles. The STR loci commonly used in the forensic community were combined to make the Interpol Standard Set of Loci (ISSOL); these have since been expanded from seven loci to ten loci [6]. Other organizations, such as the European DNA Profiling Group (EDNAP), are working towards the standardization of DNA profiling such that an organization in one country will be able to access DNA data in the database of another country. The biggest obstacle to cross-border data sharing is now political rather than technical.

WWW resources

Interpol (DNA front page): www.interpol.int/Public/Forensic/DNA/
Federal Bureau of Investigation (CODIS Information): www.fbi.gov/hq/lab/codis/

Association of Chief Police Officers of England, Wales and Northern Ireland: (National DNA Database report): http://www.acpo.police.uk/policies.asp
GeneWatch UK: www.genewatch.org

References

1. Werrett, D.J. (1997) The National DNA Database. *Forensic Science International* **88**, 33–42.
2. Parliamentary Office of Science and Technology. (2006) *The National DNA Database*.
3. Harbison, S.A. *et al*. (2001) The New Zealand DNA databank: its development and significance as a crime solving tool. *Science and Justice* **41**, 33–37.
4. President's DNA Initiative. Advancing justice Through DNA Technology. (Vol. 2007)
5. Schneider, P.M. and Martin, P.D. (2001) Criminal DNA databases: the European situation. *Forensic Science International* **119**, 232–238.
6. Gill, P. *et al*. (2006) The evolution of DNA databases – Recommendations for new European STR loci. *Forensic Science International* **156**, 242–244.

11 Kinship testing

The application of DNA profiling to kinship analysis is widespread and offers an easy means of establishing biological relationships. Not surprisingly, paternity testing is the most common form of kinship testing, with hundreds of thousands of tests being performed worldwide each year [1]. Since the first DNA based kinship test in 1985 [2], DNA analysis has been applied to larger numbers of kinship tests, to the testing of more complex relationships and to the identification of highly compromised human remains.

Paternity testing

PCR-based STR profiling has now become the standard tool and the PowerPlex® 16 (Promega) and AmpF*l*STR® Identifiler® (Applied Biosystems) STR kits that can analyse 15 loci simultaneously are routinely used (see Chapter 6). Laboratories that undertake kinship testing often have over 20 genetic markers at their disposal, including STR markers on the X and Y chromosomes [3], that allow for the testing of complex relationships [4]. The sensitivity of STR analysis, while not essential for most forms of paternity testing, allows samples to be routinely collected using buccal swabs [1] and has expanded the possible scenarios where it can be used, for example, the analysis of low amounts of DNA recovered from foetal cells [5, 6] (Figure 11.1).

The methodology used to produce DNA profiles for paternity testing is identical to the analysis of material recovered from crime scenes (see Chapters 4 to 7). The interpretation of results is more complex than when comparing profiles from crime scenes and suspects. If the tested man does not possess the alleles that have been inherited from the biological father we can conclude that he cannot be the biological father. However, because mutations between the father and child could lead to a false exclusion at any given loci [1, 7–10], it is standard practice is to require an exclusion at three or more loci before a test is declared negative.

If we cannot exclude the tested man as being the biological father then we have to assign a value to indicate the significance of non-exclusion. Likelihood ratios (see Chapter 9), which consider two competing, and mutually exclusive hypotheses are

An Introduction to Forensic Genetics W. Goodwin, A. Linacre and S. Hadi
© 2007 John Wiley & Sons, Ltd

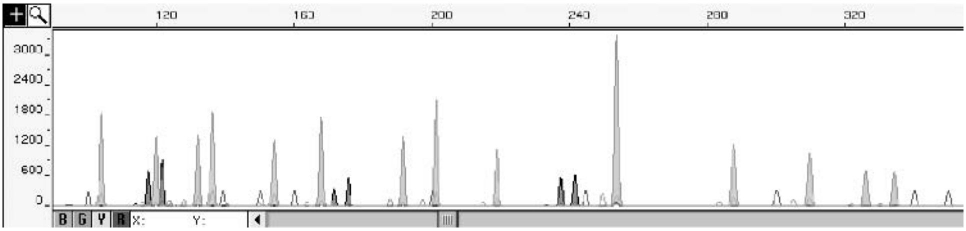


Figure 11.1 A STR profile generated from foetal cells recovered from amniotic fluid in early pregnancy. In this case it was possible to determine that pregnancy had resulted from a rape, and allowed an informed decision to be made on whether or not to have the foetus aborted. The profile was generated using the AmpF*l*STR® Profiler Plus® STR kit (Applied Biosystems) (see plate section for full-colour version of this figure)

used. The hypotheses are:

$$\frac{\text{The tested man is the biological father}}{\text{The tested man is not the biological father}} \qquad \begin{matrix} = H_{\mathrm{p}} \\ = H_{\mathrm{d}} \end{matrix}$$

The symbols H_{p} and H_{d} were introduced in Chapter 9 when comparing the proposition or hypothesis put forward by the prosecution (H_{p}) compared with the hypothesis put forward by the defence (H_{d}), although in many civil cases the terms prosecution and defence are not appropriate. This likelihood is called a paternity index (PI) and can be assessed using equation (11.1).

$$\textit{Paternity index} = \frac{\Pr(G_{\mathrm{c}}|G_{\mathrm{m}}, G_{\mathrm{tm}}, H_{\mathrm{p}})}{\Pr(G_{\mathrm{c}}|G_{\mathrm{m}}, G_{\mathrm{tm}}, H_{\mathrm{d}})} \tag{11.1}$$

where Pr is probability.

To calculate this likelihood ratio, we compare the probability of the child's genotype (G_{c}) given the mother's (G_{m}) and tested man's genotype (G_{tm}), if the tested man is the biological father (H_{p}) and the probability of the child's genotype given the mother's and tested man's genotype, if the tested man is not the biological father (H_{d}).

The numerator and denominator are conditional on the genotypes of the mother, child and tested man. They can be derived using a 'Punnet square'.

Punnett square

The equations are not difficult to understand, particularly if derived from a Punnett square and converting to text form. Consider the case where the mother is genotype a, b and child is genotype b, c and the alleged father is c, d. If he is the father then the mother must pass on allele b, and the father must pass on allele c. If he is not the father then the mother must still pass on allele b but some other man must pass on allele c. This is given in the Punnett square below.

		Alleles from alleged father	
		c	d
Alleles from mother	a	a, c	a, d
	b	b, c	b, c

If the alleged father is the biological father then this can happen one in four ways, with a probability of 0.25.

If the alleged father is not the biological father, then the mother must pass on allele b with probability of 0.5 and the chance that a male other than the alleged father is the father is dependent upon the frequency of allele c (p_c) in the population. This gives a likelihood ratio of:

$$PI = \frac{0.25}{0.5p_c} = \frac{1}{2p_c}$$

The same process can be used for any of the possible combinations. Consider the version where the alleged father is homozygous (b, b) and the mother heterozygous (a, b) and the child is heterozygous (a, b)

		Alleles from alleged father	
		b	b
Alleles from mother	a	a, b	a, b
	b	b, b	b, b

If the alleged father is the biological father then this can happen in two ways, with a probability of 0.5.

If the alleged father is not the biological father then the mother must pass on allele b with probability of 0.5 and the chance that a male other than the alleged father is the father is dependent upon the frequency of allele b (p_b) in the population. This gives a likelihood ratio of:

$$PI = \frac{0.5}{0.5p_b} = \frac{1}{p_b}$$

Consider a case when the mother is a, b, the child is a, b and the alleged father is a, c.

		Alleles from alleged father	
		a	c
Alleles from mother	a	a, a	a, c
	b	a, b	b, c

Allele a or b could be passed from the mother. Note that if she passed on allele a then this would be an exclusion and therefore it would need to be allele b that is passed from mother to child if the man is the biological father. Considering the numerator (H_p) genotype a, b occurs in only one of four ways (0.25). Considering the denominator (H_d) the mother passed on either allele a (0.5) or allele b (0.5) and the chance that either event took place, allele a or allele b, is the sum of the probabilities. This results in the equation below:

$$PI = \frac{0.25}{0.5p_a + 0.5p_b} = \frac{1}{2(p_a + p_b)}$$

In Table 11.1 all the potential combinations of alleles from a mother, child and tested man are shown along with the resulting numerator, denominator and PI equation.

Table 11.1 The numerator and denominator that should be used when calculating a paternity index are determined by the genotypes of the child (G_C), mother (G_M), and tested man (G_{TM}). The alleles are represented by i, j, k and l where $i \neq j \neq k \neq l$. Reproduced from Lucy, 2006 [11] p. 174, with permission from John Wiley & Sons (originally based on Evett and Weir, 1998 [12] p. 168)

G_C	G_M	G_{TM}	Numerator	Denominator	PI
A_iA_i	A_iA_i	A_iA_i	1	p_i	$^1/_{p_i}$
		A_iA_j	$^1/_2$	p_i	$^1/_{2p_i}$
		A_jA_k	0	p_i	0
	A_iA_j	A_iA_i	$^1/_2$	$^{p_i}/_2$	$^1/_{p_i}$
		A_iA_j	$^1/_4$	$^{p_i}/_2$	$^1/_{2p_i}$
		A_iA_k	$^1/_4$	$^{p_i}/_2$	$^1/_{2p_i}$
		$A_j\ A_k$	0	$^{p_i}/_2$	0
A_iA_j	A_iA_i	A_jA_j	1	P_j	$^1/_{p_j}$
		A_iA_j	$^1/_2$	P_j	$^1/_{2p_j}$
		A_jA_k	$^1/_2$	P_j	$^1/_{2p_j}$
		A_kA_l	0	P_j	0
	A_iA_j	A_iA_i	$^1/_2$	$^{(P_i+P_j)}/_2$	$^1/_{(p_i+p_j)}$
		A_iA_j	$^1/_2$	$^{(P_i+P_j)}/_2$	$^1/_{p_i+p_j}$
		A_iA_k	$^1/_4$	$^{(P_i+P_j)}/_2$	$^1/_{2(p_i+p_j)}$
		A_jA_k	$^1/_4$	$^{(P_i+P_j)}/_2$	$^1/_{2(p_i+p_j)}$
		A_kA_l	0	$^{(P_i+P_j)}/_2$	0
	A_iA_k	A_jA_j	$^1/_2$	$^{P_j}/_2$	$^1/_{p_j}$
		A_iA_j	$^1/_4$	$^{P_j}/_2$	$^1/_{2p_j}$
		A_jA_k	$^1/_4$	$^{P_j}/_2$	$^1/_{2p_j}$
		A_jA_l	$^1/_4$	$^{P_j}/_2$	$^1/_{2p_j}$
		A_kA_l	0	$^{P_j}/_2$	0

We can apply the formula in Table 11.1 to the paternity case presented in Table 11.2.

Table 11.2 The result of a paternity test using the Powerplex® 16 STR Kit (Promega). The alleles that the child could have inherited from the mother are underlined and the alleles that are from the biological father are shown in bold. The ^{i,j,k}, ^{l} symbols correspond to symbols in Table 11.1. The allele frequencies were taken from Marino *et al.* (2006) [13]

	Child (G_C)	Mother (G_M)	Tested Man (G_{TM})	Num	Denom	PI	$P_{i/j}$	PI
D3S1358	**15**^{i} – **15**^{i}	14^j– 15^i	**15**^{i}– 19^k	$^1/_4$	$P_i/2$	$1/2p_i$	0.3239	1.54
VWA	**17**^{j}– 18^i	16^k– 18^i	**17**^{j}– 18^i	$^1/_4$	$P_j/2$	$1/2p_j$	0.2715	1.84
D16S359	11^i– **12**^{j}	11^i– 13^k	**12**^{j}– 13^k	$^1/_4$	$P_j/2$	$1/2p_j$	0.2773	1.80
D8S1179	10^i– 13^j	10^i– 13^j	**10**^{i}– **10**^{i}	$^1/_2$	P_i + P$_j/2$	$1/p_i + p_j$	0.0630 0.3033	2.73
D21S11	30^i– **32.2**^{j}	30^i– 31^k	27^l– **32.2**^{j}	$^1/_4$	$P_j/2$	$1/2p_j$	0.1245	4.02
D18S51	13^i– **14**^{j}	13^i– 14^j	12^k– **14**^{j}	$^1/_4$	$P_i + P_j/2$	$1/2(p_i + p_j)$	0.1326 0.2063	1.48
THO1	9^i– **9.3**^{j}	9^i– 9.3^j	6^k– **9.3**^{j}	$^1/_4$	$P_i + P_j/2$	$1/2(p_i + p_j)$	0.1407 0.2624	1.24
FGA	18^i– **23**^{j}	18^i– 25^k	**23**^{j}– **23**^{j}	$^1/_2$	$P_j/2$	$1/p_j$	0.1440	6.94
D13S317	8^i– **13**^{j}	8^i– 11^k	11^k– **13**^{j}	$^1/_4$	$P_j/2$	$1/2p_j$	0.1444	3.46
CSF1PO	11^i– **11**^{i}	11^i– 13^j	**11**^{i}– 13^j	$^1/_4$	$P_i/2$	$1/2p_i$	0.2916	1.71
D7S820	9^i– **9**^{i}	9^i– 10^j	**9**^{i}– 11^k	$^1/_4$	$P_i/2$	$1/2p_i$	0.0998	5.01
TPOX	**8**^{j}– 10^i	10^i– 11^k	**8**^{j}– **8**^{j}	$^1/_2$	$P_j/2$	$1/p_j$	0.5243	1.91
D5S818	**11**^{i}– **12**^{j}	11^i– 12^j	**11**^{i}– **12**^{j}	$^1/_2$	$P_i + P_j/2$	$1/p_i + p_j$	0.3618 0.2992	1.51
Penta D	**13**^{j}– 15^i	12^k– 15^i	12^l– **13**^{j}	$^1/_4$	$P_j/2$	$1/2p_j$	0.1726	2.90
Penta E	10^i– **18**^{j}	10^i– 10^i	16^k– **18**^{j}	$^1/_2$	P_j	$1/2p_j$	0.0304	16.4
						Combined PI	**2,920,823**	

The combined PI is calculated by applying the product rule and multiplying the PI from each locus in this case the PI is 2 920 823. This can be represented by this statement:

Statement of positive paternity

The results of the DNA testing are 2 920 823 times more likely if the tested man is the biological father of the child than if the biological father is another man, unrelated to the tested man.

The significance of likelihood ratios can be difficult for lay people to evaluate and the results are often presented as a probability of paternity, making the results more accessible. To calculate a probability of paternity requires Bayesian analysis and takes into consideration non-genetic evidence: the likelihood ratio (LR) is multiplied by

Table 11.3 The impact of prior probabilities on the probability of paternity is shown with two paternity indexes: one with a value of 1 000 and the other taken from the above example, with a value of 2 920 823

	Paternity index	
Prior Odds	1 000	2 920 823
0.0001	0.090 917 356	0.996 588 329
0.0010	0.500 250 125	0.999 658 09
0.0100	0.909 918 107	0.999 966 107
0.1000	0.991 080 278	0.999 996 919
0.5000	0.999 000 999	0.999 999 658
0.7500	0.999 666 778	0.999 999 886
0.9000	0.999 888 901	0.999 999 962

the prior odds of paternity that are determined by non-genetic evidence, such as the testimony of the woman. It can be calculated using equation (11.2).

$$\text{Probability of paternity} = \frac{\text{LR} \times \Pr(H_\text{p}|\text{non-genetic evidence})}{\text{LR} \times \Pr(H_\text{p}|\text{non-genetic evidence}) + [1 - \Pr(H_\text{p}|\text{non-genetic evidence})]} \tag{11.2}$$

Taking the above paternity test it is possible to turn the likelihood ratio into a probability of paternity for any prior odds of paternity; for example:

Prior probability = 0.1

$$\text{Probability of paternity} = \frac{2\,920\,823 \times 0.1}{(2\,920\,823 \times 0.1) + (1 - 0.1)} = 0.999996919$$

When this figure is used to report the results of a test it is often quoted as a percentage, which is more accessible to non-scientists. In this case the probability of paternity would be quoted as 99.9997%.

The value that is attributed to the prior odds of paternity is, of course, subjective. In civil cases, the value of 0.5 is commonly used, although there is little scientific merit to this value. In criminal cases, probabilities of paternity are often not presented because it is the duty of the jury/judge to assess the prior odds of paternity. If results are presented as a probability of paternity, a range of values calculated using different prior odds is often quoted (Table 11.3).

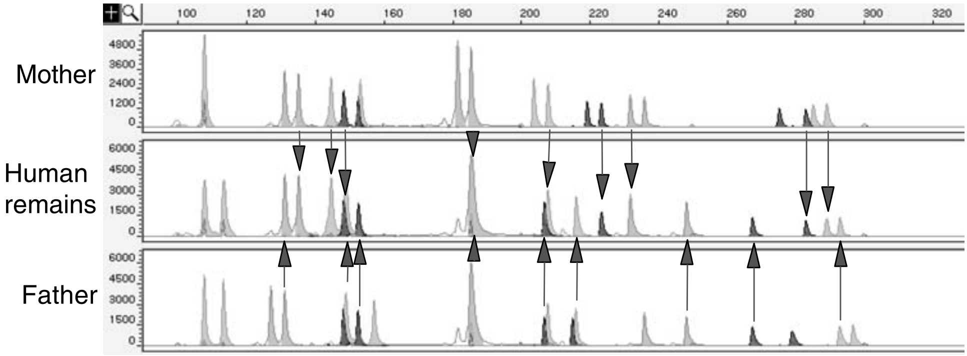


Figure 11.2 The identification of human remains recovered from an air crash [35]. Blood samples were provided by the mother and father who were missing a son. Alleles in the profile of human remains could have come from the mother and father (indicated by the arrows). The profiles were generated using the AmpF*l*STR® Profiler Plus® STR kit (Applied Biosystems) (see plate section for full-colour version of this figure)

With low paternity indexes the impact of prior odds can be significant. However, with the possibility of analysing a large number of STR loci, the PIs are typically in the millions and the posterior probability of paternity is therefore extremely high, even when the prior odds are very low. In the paternity test presented above, even with the prior odds as low as 0.001, the probability of paternity is still 99.9966 %.

In addition to the standard paternity testing where the mother, child and alleged father are available, testing can also be carried out when the mother is not available [14–17]. More complex relationships can be examined, such as determination of sibship [18] and paternity tests to discriminate between close relatives [12, 19–21]. Calculations can also incorporate correction factors to allow for deficiencies in allele frequency databases, in particular, the effects of subpopulations [22, 23] (see Chapter 8). Fortunately, computer programs have been developed to deal with both routine and complex scenarios [24–30].

Identification of human remains

The first application of DNA analysis to the identification of human remains was in 1987, when skeletal remains were profiled using single nucleotide polymorphisms in the DQα locus [31, 32]. Unfortunately, this system did not have high powers of discrimination and it was not until the early 1990s that DNA profiling was successfully applied to the identification of human remains [33, 34]. As DNA profiling technology and methodology have evolved to be more robust and powerful, it has been applied to increasingly complex situations including the identification of people killed in air crashes [29, 35–38]; fire [39–42]; terrorist attacks [43–46]; natural disasters [47] and war [48–51]. STRs are the most commonly used tool but mitochondrial DNA (see Chapter 13) and SNPs (see Chapter 12) have also been employed on occasion.

The matching of human remains can be through comparison to personal objects that belonged to the missing person, such as combs and toothbrushes [52], or by comparison to close family members (Figure 11.2).

In cases that involve hundreds of victims, the statistical analysis becomes very complex. Because of the high number of pair-wise comparisons that are made between the victims and relatives, the potential for coincidental matches that result in false positives and ultimately misidentifications is significant [47, 52, 53]. The existence of relatives within the population of victims also complicates the analysis [29, 47] and there are limitations as to what can be achieved.

Further reading

Balding, D.J. (2005) *Weight-of-evidence for Forensic DNA Profiles*. John Wiley & Sons, Ltd, Chichester, pp. 82–134.

Buckleton, J., Triggs, C.M., and Walsh, S.J. (2005) *Forensic DNA Evidence Interpretation*. CRC Press, pp. 341–437.

References

1. American Association of Blood Banks (2004) Annual report summary for testing in 2004. http://www.aabb.org/content/Accreditation/Parentage_Testing_Accreditation_Program/ptprog.htm
2. Jeffreys, A.J., *et al.* (1985) Positive Identification of an immigration test-case using human DNA fingerprints. *Nature* **317**, 818–819.
3. Szibor, R., *et al.* (2003) Use of X-linked markers for forensic purposes. *International Journal of Legal Medicine* **117**, 67–74.
4. Junge, A., *et al.* (2006) Mutations or exclusion: an unusual case in paternity testing. *International Journal of Legal Medicine* **120**, 360–363.
5. Makrydimas, G., *et al.* (2004) Early prenatal diagnosis by celocentesis. *Ultrasound in Obstetrics and Gynecology* **23**, 482–485.
6. Makrydimas, G., *et al.* (2004) Prenatal paternity testing using DNA extracted from coelomic cells. *Fetal Diagnosis and Therapy* **19**, 75–77.
7. Brinkmann, B., *et al.* (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *American Journal of Human Genetics* **62**, 1408–1415.
8. Gunn, P.R., *et al.* (1997) DNA analysis in disputed parentage: the occurrence of two apparently false exclusions of paternity, both at short tandem repeat (STR) loci, in the one child. *Electrophoresis* **18**, 1650–1652.
9. Leopoldino, A.M., and Pena, S.D.J. (2003) The mutational spectrum of human autosomal tetranucleotide microsatellites. *Human Mutation* **21**, 71–79.
10. Chakraborty, R., and Stivers, D.N. (1996) Paternity exclusion by DNA markers: effects of paternal mutations. *Journal of Forensic Sciences* **41**, 671–677.
11. Lucy, D. (2005) *Introduction to Statistics for Forensic Scientists*. John Wiley and Sons, Ltd, Chichester.
12. Evett, I.W., and Weir, B.S. (1998) *Interpreting DNA Evidence*. Sinauer Associates, Inc.
13. Marino, M., *et al.* (2006) Population genetic analysis of 15 autosomal STRs loci in the central region of Argentina. *Forensic Science International* **161**, 72–77.
14. Chakraborty, R., *et al.* (1994) Paternity evaluation in cases lacking a mother and nondetectable alleles. *International Journal of Legal Medicine* **107**, 127–131.

15. Poetsch, M., *et al.* (2006) The problem of single parent/child paternity analysis – Practical results involving 336 children and 348 unrelated men. *Forensic Science International* **159**, 98–103.
16. Thomson, J.A., *et al.* (2001) Analysis of disputed single-parent/child and sibling relationships using 16 STR loci. *International Journal of Legal Medicine* **115**, 128–134.
17. Brenner, C.H. (1993) A note on paternity computation in cases lacking a mother. *Transfusion* **33**, 51–54.
18. Wenk, R.E., *et al.* (1996) Determination of sibship in any two persons. *Transfusion* **36**, 259–262.
19. Ayres, K.L., and Balding, D.J. (2005) Paternity index calculations when some individuals share common ancestry. *Forensic Science International* **151**, 101–103.
20. Fung, W.K., *et al.* (2002) Power of exclusion revisited: probability of excluding relatives of the true father from paternity. *International Journal of Legal Medicine* **116**, 64–67.
21. Wenk, R.E., and Chiafari, F.A. (2000) Distinguishing full siblings from half-siblings in limited pedigrees. *Transfusion* **40**, 44–47.
22. Ayres, K.L. (2000) Relatedness testing in subdivided populations. *Forensic Science International* **114**, 107–115.
23. Fung, W.K., *et al.* (2003) Testing for kinship in a subdivided population. *Forensic Science International* **135**, 105–109.
24. Fung, W.K. (2003) User-friendly programs for easy calculations in paternity testing and kinship determinations. *Forensic Science International* **136**, 22–34.
25. Riancho, J.A., and Zarrabeitia, M.T. (2003) A Windows-based software for common paternity and sibling analyses. *Forensic Science International* **135**, 232–234.
26. Brenner, C.H. (1997) Symbolic kinship program. *Genetics* **145**, 535–542.
27. Fung, W.K., *et al.* (2006) On statistical analysis of forensic DNA: theory, methods and computer programs. *Forensic Science International* **162**, 17–23.
28. Brenner, C.H. DNA.VIEW. (http://dna-view.com/)
29. Leclair, B., *et al.* (2004) Enhanced kinship analysis and STR-based DNA typing for human identification in mass fatality incidents: The Swissair Flight 111 disaster. *Journal of Forensic Sciences* **49**, 939–953.
30. Buckleton, J., *et al.* (2005) *Forensic DNA Evidence Interpretation*. CRC Press.
31. Stoneking, M., *et al.* (1991) Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. *American Journal of Human Genetics* **48**, 370–382.
32. Blake, E., *et al.* (1992) Polymerase chain-reaction (Pcr) amplification and human-leukocyte antigen (Hla)-Dq-Alpha oligonucleotide typing on biological evidence samples – casework experience. *Journal of Forensic Sciences* **37**, 700–726.
33. Hagelberg, E., *et al.* (1991) Identification of the skeletal remains of a murder victim by DNA analysis. *Nature* **352**, 427–429.
34. Sajantila, A., *et al.* (1991) The polymerase chain-reaction and postmortem forensic identity testing – application of amplified D1S80 and HLA-DQ-Alpha loci to the identification of fire victims. *Forensic Science International* **51**, 23–34.
35. Goodwin, W., *et al.* (1999) The use of mitochondrial DNA and short tandem repeat typing in the identification of air crash victims. *Electrophoresis* **20**, 1707–1711.
36. Olaisen, B., *et al.* (1997) Identification by DNA analysis of the victims of the August 1996Spitsbergen civil aircraft disaster. *Nature Genetics*. **15**, 402–405.
37. Hsu, C.M., *et al.* (1999) Identification of victims of the 1998 Taoyuan Airbus crash accident using DNA analysis. *International Journal of Legal Medicine* **113**, 43–46.
38. Ballantyne, J. (1997) Mass disaster genetics. *Nature Genetics*. **15**, 329–331.
39. Clayton, T.M., *et al.* (1995) Further validation of a quadruplex STR DNA typing system: A collaborative effort to identify victims of a mass disaster. *Forensic Science International* **76**, 17–25.
40. Clayton, T.M., *et al.* (1995) Identification of bodies from the scene of a mass disaster using DNA amplification of short tandem repeat (STR) loci. *Forensic Science International* **76**, 7–15.
41. Meyer, H.J. (2003) The Kaprun cable car fire disaster – aspects of forensic organisation following a mass fatality with 155 victims. *Forensic Science International* **138**, 1–7.

42. Calacal, G.C., *et al.* (2005) Identification of exhumed remains of fire tragedy victims using conventional methods and autosomal/Y-chromosomal short tandem repeat DNA profiling. *American Journal of Forensic Medicine and Pathology* **26**, 285–291.
43. Kahana, T., *et al.* (1997) Suicidal terrorist bombings in Israel – identification of human remains. *Journal of Forensic Sciences* **42**, 260–264.
44. Budimlija, Z.M., *et al.* (2003) World trade center human identification project: experiences with individual body identification cases. *Croatian Medical Journal* **44**, 259–263.
45. Holland, M.M., *et al.* (2003) Development of a quality, high throughput DNA analysis procedure for skeletal samples to assist with the identification of victims from the world trade center attacks. *Croatian Medical Journal* **44**, 264–272.
46. Budowle, B., *et al.* (2005) Forensic aspects of mass disasters: strategic considerations for DNA-based human identification. *Legal Medicine* **7**, 230–243.
47. Brenner, C.H. (2006) Some mathematical problems in the DNA identification of victims in the 2004 tsunami and similar mass fatalities. *Forensic Science International* **157**, 172–180.
48. Huffine, E., *et al.* (2001) Mass identification of persons missing from the break-up of the former Yugoslavia: structure, function, and role of the International Commission on Missing Persons. *Croatian Medical Journal* **42**, 271–275.
49. Holland, M.M., *et al.* (1993) Mitochondrial-DNA sequence-analysis of human skeletal remains – identification of remains from the Vietnam War. *Journal of Forensic Sciences* **38**, 542–553.
50. Boles, T.C., *et al.* (1995) Forensic DNA testing on skeletal remains from mass graves – a pilot project in Guatemala. *Journal of Forensic Sciences* **40**, 349–355.
51. Lleonart, R., *et al.* (2000) Forensic identification of skeletal remains from members of Ernesto Che Guevara's guerrillas in Bolivia based an DNA typing. *International Journal of Legal Medicine* **113**, 98–101.
52. Brenner, C.H., and Weir, B.S. (2003) Issues and strategies in the DNA identification of World Trade Center victims. *Theoretical Population Biology* **63**, 173–178.
53. Gornik, I., *et al.* (2002) The identification of war victims by reverse paternity is associated with significant risks of false inclusion. *International Journal of Legal Medicine* **116**, 255–257.

12 Single nucleotide polymorphisms

One of the most significant outcomes of the Human Genome Project has been the identification of large numbers of single nucleotide polymorphisms (SNPs) [1–3]. The application of SNPs to forensic analysis is currently limited to some specialist cases. However, with advances both in our knowledge of SNPs and in the technology used to detect the polymorphisms, SNP analysis may play an increasingly important role in the future.

SNPs – occurrence and structure

'SNPs are single base pair positions in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals in some population(s), wherein the least frequent allele has an abundance of 1% or greater' [4]. The structure of a SNP is very simple, an example is shown in Figure 12.1.

SNPs are found in the human genome about once in every 1000 bp [1–3, 5]. Given that the human genome is 3.2 billion bp long, we can estimate that there will be approximately 1 million differences between two genomes that are due to SNPs: this represents approximately 85% of human genetic variation.

The biallelic state of the vast majority of polymorphisms intrinsically limits the information that can be gained from the analysis of any given SNP, and this has been the major factor limiting their application to forensic analysis: between 50 and 80 SNPs are required to achieve the same levels of discrimination as the current STR based methods [6, 7]. However, the vast numbers of SNPs within the genome (currently over 10 million SNPs have been placed in public databases [1, 2]) can compensate for the limited information carried by any individual SNP, and make them a tempting polymorphism to exploit. The technology that is being used for SNP detection is evolving and SNP analysis is becoming possible in many forensic laboratories.

Detection of SNPs

There are many techniques available for the resolution of SNPs. In the 1970s it was established that particular enzymes produced by bacteria can be used to cut the DNA

© 2007 John Wiley & Sons, Ltd

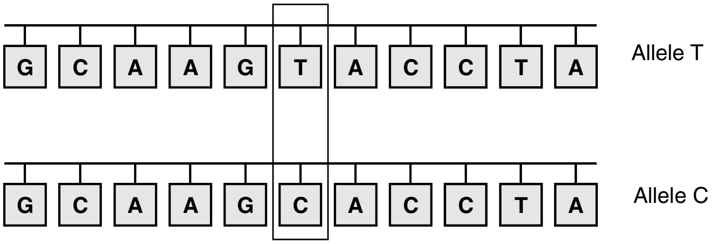


Figure 12.1 SNPs are created when the DNA replication enzymes make a mistake as they copy the cell's DNA during meiosis. The enzyme incorporates the wrong nucleotide approximately once every 9 to 10 million bases. In the vast majority of cases, SNP are biallelic and only have two different alleles. In the alleles shown above, the thymine nucleotide has been replaced by a cytosine

molecule by recognizing specific sequences [8]. Restriction digestion can be used to genotype SNPs when the SNP either creates or destroys a particular restriction enzyme recognition sequence [9, 10] but the method is limited for forensic casework because it needs a large amount of DNA and is a long and laborious process.

Sanger sequencing

Sanger sequencing, also known as chain-termination sequencing, was developed in the late 1970s and is a milestone in the development of molecular biology [11]. The sequencing takes advantage of the biochemistry of DNA replication. The first stage of the analysis is to amplify the target region using PCR; amplified products are then used as the template in a sequencing reaction. The DNA sequencing reaction is similar to PCR amplification and the reaction mixture is very similar, containing the thermophilic *Taq* DNA polymerase and deoxynucleotide triphosphates (dNTPs). It differs from PCR in that only one primer is used and, in addition to the dNTPs, there are four fluorescently labelled dideoxyribonucleotides (ddNTPs); each ddNTP is labelled with a different coloured dye [12]. The ddNTPs do not contain the hydroxyl group on the 3′ carbon, which prevents any extension of the DNA molecule [13] (Figure 12.2).

The concentration of dNTPs is higher than ddNTPs and therefore in most cases a dNTP is added. The ddNTPs are incorporated at random intervals along the molecule. This produces a range of different sized molecules. The products of the sequencing reaction are analysed using capillary gel electrophoresis systems, such as the ABI PRISM® 310 Genetic Analyzer, that separates DNA to single base pair resolution and can simultaneously detect four different fluorescent labels (Figure 12.3).

Sequencing is not a practical option for the analysis of SNPs in a forensic context. Most SNPs are widely dispersed around the genome and a separate reaction is required for each SNP. An exception is the mitochondrial genome, where a number of SNPs are concentrated into a small area and can be analysed in a small number of reactions (see Chapter 13). Sequencing has also been a powerful method to type SNPs within rapidly evolving regions of DNA in the HIV virus [14, 15].

5´ 3´
AGCTGTAAGTCTATACGTATCGTTAGTGCCTTGACTATGTCCGTA – Template
CGGAACTGATACAGGCAT – Primer
ACGGAACTGATACAGGCAT – 1
CACGGAACTGATACAGGCAT – 2
TCACGGAACTGATACAGGCAT – 3
ATCACGGAACTGATACAGGCAT – 4
AATCACGGAACTGATACAGGCAT – 5
CAATCACGGAACTGATACAGGCAT – 6
GCAATCACGGAACTGATACAGGCAT – 7
AGCAATCACGGAACTGATACAGGCAT – 8

Figure 12.2 A primer anneals to the template strand. This is extended by *Taq* polymerase until a ddNTP is incorporated. The ddNTPs are incorporated at random, which leads to a collection of extension molecules that differ from each other by one nucleotide (shown above labelled 1 to 8). The four ddNTPs are labelled with different fluorescent dyes that are detected during capillary electrophoresis (see plate section for full-colour version of this figure)

SNP detection for forensic applications

Restriction digestion and sequence analysis are not viable methods to use for most forensic cases that might require the analysis of 50 to 80 SNPs dispersed around the genome. A number of methods have evolved that can be applied to the detection of

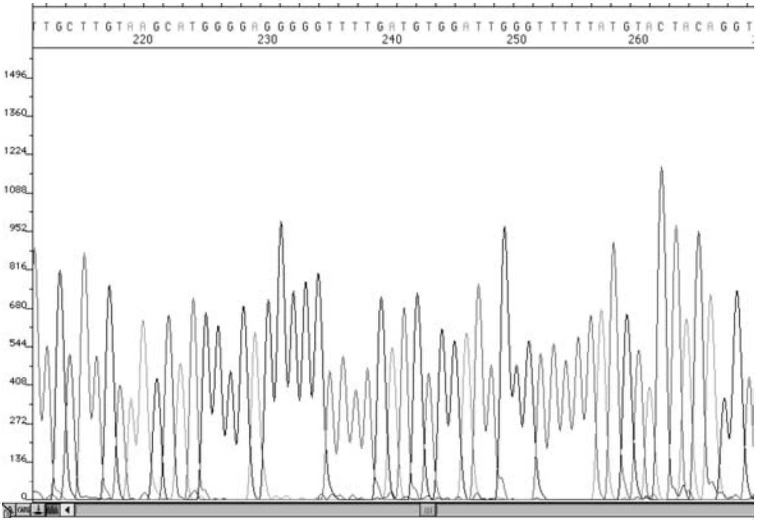


Figure 12.3 The sequence of a region of the mitochondrial genome. The sequencing software interprets the sequence data and 'calls' the bases. This information is provided above the sequencing peaks (see plate section for full-colour version of this figure)

multiple SNPs. Methods that are based around the concepts of either primer extension or primer hybridization are the most widely used.

Primer extension

Primer extension is a robust method for discriminating between different alleles and several methodologies have been developed [16]. One of the commonly used methods is the mini-sequencing reaction [17]. The basis of the reaction is very similar to Sanger sequencing. The first part of the procedure is to amplify the target region using PCR. An internal primer then anneals to the denatured PCR product; the 3′ end of the primer is adjacent to the polymorphic site. The primer is then extended by *Taq* polymerase but only ddNTPs that are labelled with fluorescent dyes are provided; the primer is only extended by one nucleotide. The extended primer can be analysed using capillary gel electrophoresis and the colour of the detected peak allows the SNP to be characterized (Figure 12.4). A widely used commercial kit called SNaPshot^{TM} (Applied Biosystems) is based on this methodology.

By using different sized primers and different fluorescent tags for each of the four bases, a large number of SNPs can be simultaneously detected [18].

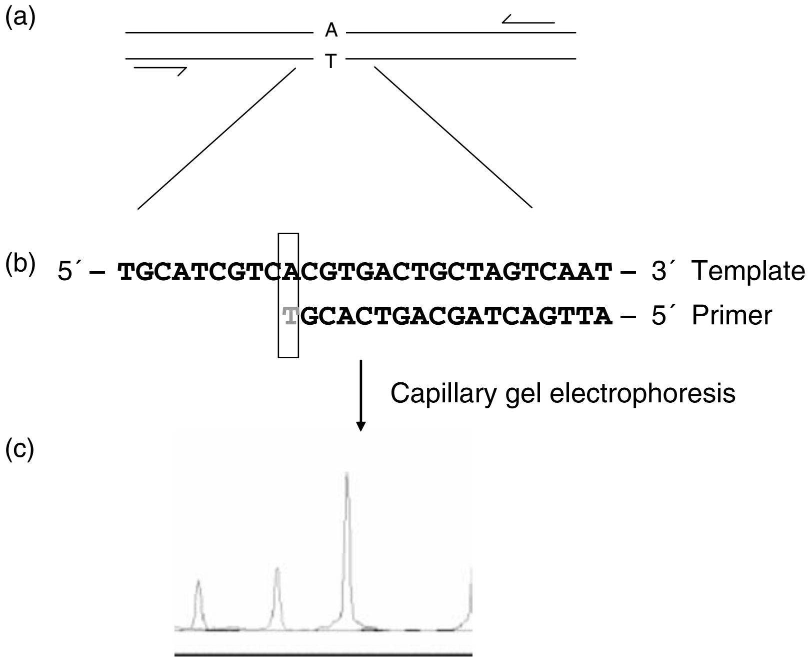

(a)
A
T
(b)
5´ – TGCATCGTCACGTGACTGCTAGTCAAT – 3´ Template
TGCACTGACGATCAGTTA – 5´ Primer
Capillary gel electrophoresis
(c)

Figure 12.4 The primer extension assay. (a) The target sequence is amplified using PCR and the products are used as the template in the extension assay; (b) an internal primer hybridizes to the target adjacent to the SNP and a single fluorescently labelled ddNTP is added by *Taq* polymerase; (c) the reaction is analysed by capillary electrophoresis (see plate section for full-colour version of this figure)

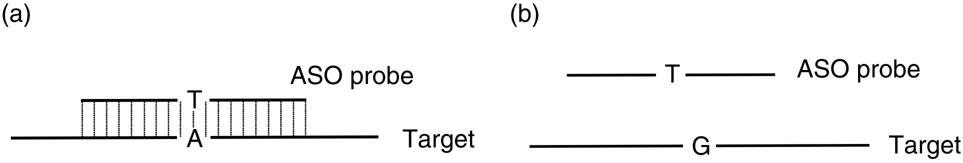


Figure 12.5 Allele specific hybridization. Allele specific oligonucleotide (ASO) probes that include the SNP are hybridized with the target DNA. (a) Under highly stringent conditions only perfectly matched sequences will form stable interactions; (b) with one mismatch in sequence the ASO will not hybridize

Variations on the primer extension technique include pyrosequencing [19, 20]; microarrays, where the extension primers are attached to a silicon chip [21, 22]; and allele specific extension, when the primer is only extended if it is 100% complementary to the target sequence [23].

Allele specific hybridization

Under stringent conditions, even one nucleotide mismatching between a template and primer can differentiate between two alleles (Figure 12.5).

There is a large number of methods that exploit the hybridization of probes, including reverse dot blots [24]; Taqman® assays [25]; LightCycler® assays [26]; molecular beacons [27, 28]; and GeneChips® [29] (see Further Reading for details).

Forensic applications of SNPs

A vast amount of data is available on the different SNPs in the human genome and one of the biggest tasks when applying SNPs to forensic applications is to select the most appropriate SNPs from the overwhelming numbers that are available. The choice of SNPs is very much dependent on the application.

Forensic identification

The vast majority of forensic DNA analysis involves the characterization of biological material recovered from the scene of a crime. Several panels of SNPs have been developed that are designed to provide maximum discrimination powers for forensic identification [18, 30, 31]. These contain SNPs that are polymorphic in all major population groups. A panel, containing 52 SNPs was developed by the SNPforID Consortium [18]. Using this panel of SNPs produced match probabilities that ranged from 5.0×10^{-19} in an Asian population to 5.0×10^{-21} in a European population. When applied to paternity testing, average paternity indices of between 336 000 in Asian populations and 550 000 in European populations, were achieved.

However, even with the high discrimination power, the effort involved in analysing 50 SNPs is greater than when undertaking standard STR analysis. The major attraction of using SNPs with the current technology is that SNP analysis can provide results from highly degraded DNA when conventional STR profiling has failed [30, 32].

Prediction of the geographical ancestry

In many cases, the identification of the population group from which a crime scene sample has come from can be valuable intelligence for the investigating agencies: was the person who left the material at the crime scene of Caucasian, Asian, African, mixed ancestry? Panels consisting of mtDNA SNPs and Y SNPs have already been found useful for this purpose [33, 34] but are intrinsically limited by the fact that they can only provide information on either the maternal or paternal ancestry. Autosomal SNPs that have different frequencies in different major population groups can provide valuable information on geographic ancestry [35]. Many of the SNPs selected for this purpose are associated with coding regions that have been subjected to selection pressures. These include pigmentation genes and genes involved with the metabolism of xenobiotics. The pigmentation genes, in addition to providing information on geographic ancestry, can also give information on phenotype of the person who deposited the biological material at a crime scene, including skin, hair and eye colour [36–38].

SNPs compared to STR loci

Current STR-based multiplex kits like AMPF*l*STR Identifiler® and PowerPlex® 16 can amplify 15 STR loci and the amelogenin locus. Using the current technology it is difficult to co-amplify and detect any more STR loci. Also the size of the amplicon for each STR is quite large. The great advantage STRs have over SNPs is their power of discrimination due to the large number of alleles they have in comparison with biallelic SNPs. In contrast to STRs, around four-times more SNPs are required to reach the discrimination power equivalent to STR loci. Another major disadvantage to using SNPs is that mixtures of two or more people might be either problematic or impossible to interpret since SNPs are biallelic markers. Also current DNA databases consist of profiles comprising STR loci and therefore SNPs cannot be used in that context. At the same time it is possible to analyse hundreds of SNP loci and, due to their structure, the amplicon size can be much smaller, typically less than 100 bp. This allows the detection of DNA templates that are highly degraded and may generate data when standard STR typing fails to generate a result. A comparison between STR and SNP markers is shown in Table 12.1.

In the foreseeable future, STRs will be the most commonly used genetic polymorphism analysed. They are tried and tested in most judicial systems and also form the basis of most forensic DNA databases. Even so, the use of SNPs in forensic genetics is likely to increase in the coming years and may at some point in the future replace the

Table 12.1 A comparison of the properties of SNPs and STRs

	STR	SNP
Frequency of occurrence	Once every 15 Kb	Once every 1 Kb
Typical rate of mutation	10^{-3}	10^{-8}
Typical number of alleles	Between 5 and 20	2
Potential to multiplex	Currently a maximum of 15 STR loci examined at one time	Difficult to amplify more than 50 SNPs in one reaction
Number of loci required to have a P_M of 1 in 1 billion	10	∼60
Method of detection	Capillary gel electrophoresis (CGE)	CGE, microarrays, mass spectroscopy
Automation potential	Medium	High
Artefacts	Amplification of STRs can produce artefacts such as stutter and split peaks.	No stutter artefacts associated with the amplification of the SNPs
Amount of DNA required	∼0.5 to 1 ng	∼100 pg
Size of amplicon	Amplicon sizes typically between 100 and 400 bp	Amplicon sizes can be less than 100 bp
Mixtures	Interpretation of mixtures of STR loci is possible	Mixtures of SNP loci can be highly problematic to interpret
Predicting geographical origin	Limited ethnic identification from STR loci	Some SNPs can be associated with particular ethnic groups
Phenotypic information	No possibility of inferving phenotype	Possible to predict some hair colours, eye colour, skin colour.

analysis of STR polymorphisms. The application of SNPs to specialized applications, for example, SNP based blood grouping [31, 39] and molecular autopsy (looking for mutations that can explain sudden death [40, 41]), is likely to become more widespread.

Further reading

Kwok, P.Y. (2001) Methods for genotyping single nucleotide polymorphisms. *Annual Review of Genomics and Human Genetics* **2**, 235–258.

References

1. Miller, R.D., *et al*. (2005) High-density single-nucleotide polymorphism maps of the human genome. *Genomics* **86**, 117–126.
2. Sachidanandam, R., *et al*. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933.

3. Thorisson, G.A., and Stein, L.D. (2003) The SNP Consortium website: past, present and future. *Nucleic Acids Research* **31**, 124–127.
4. Brookes, A.J. (1999) The essence of SNPs. *Gene* **234**, 177–186.
5. Collins, F.S., *et al.* (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945.
6. Gill, P. (2001) An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *International Journal of Legal Medicine* **114**, 204–210.
7. Krawczak, M. (2001) Forensic evaluation of Y-STR haplotype matches: a comment. *Forensic Science International* **118**, 114–115.
8. Danna, K., and Nathans, D. (1971) Studies of Sv40 DNA.1. Specific cleavage of Simian Virus 40 DNA by restriction endonuclease of *hemophilus influenzae*. *Proceedings of the National Academy of Sciences of the United States of America* **68**, 2913–2917.
9. Denaro, M., *et al.* (1981) Ethnic variation in Hpa-I endonuclease cleavage patterns of human mitochondrial-DNA. *Proceedings of the National Academy of Sciences of the United States of America–Biological Sciences* **78**, 5768–5772.
10. Riehn, R., *et al.* (2005) Restriction mapping in nanofluidic devices. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 10012–10016.
11. Sanger, F., *et al.* (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463–5467.
12. Rosenblum, B.B., *et al.* (1997) New dye-labeled terminators for improved DNA sequencing patterns. *Nucleic Acids Research* **25**, 4500–4504.
13. Prober, J.M., *et al.* (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**, 336–341.
14. Albert, J., *et al.* (1993) Forensic evidence by DNA sequencing. *Nature* **361**, 595–596.
15. Metzker, M.L., *et al.* (2002) Molecular evidence of HIV-1 transmission in a criminal case. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 14292–14297.
16. Kwok, P.Y. (2001) Methods for genotyping single nucleotide polymorphisms. *Annual Review of Genomics and Human Genetics* **2**, 235–258.
17. Syvanen, A.C., *et al.* (1990) A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein-E. *Genomics* **8**, 684–692.
18. Sanchez, J.J., *et al.* (2006) A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis* **27**, 1713–1724.
19. Ronaghi, M., *et al.* (1998) A sequencing method based on real-time pyrophosphate. *Science* **281**, 363–365.
20. Ronaghi, M., *et al.* (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry* **242**, 84–89.
21. Shumaker, J.M., *et al.* (1996) Mutation detection by solid phase primer extension. *Human Mutation* **7**, 346–354.
22. Fan, J.B., *et al.* (2000) Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Research* **10**, 853–860.
23. Pastinen, T., *et al.* (2000) A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays. *Genome Research* **10**, 1031–1042.
24. Saiki, R.K., *et al.* (1986) Analysis of enzymatically amplified beta-globin and Hla-Dq-Alpha DNA with allele-specific oligonucleotide probes. *Nature* **324**, 163–166.
25. Livak, K.J., *et al.* (1995) Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting Pcr product and nucleic-acid hybridization. *Pcr-Methods and Applications* **4**, 357–362.
26. Lareu, M., *et al.* (2001) The use of the LightCycler for the detection of Y chromosome SNPs. *Forensic Science International* **118**, 163–168.
27. Tyagi, S., *et al.* (1998) Multicolor molecular beacons for allele discrimination. *Nature Biotechnology* **16**, 49–53.
28. Tyagi, S., and Kramer, F.R. (1996) Molecular beacons: probes that fluoresce upon hybridization. *Nature Biotechnology* **14**, 303–308.

29. Wang, D.G., *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082.
30. Dixon, L.A., *et al.* (2005) Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes. *Forensic Science International* **154**, 62–77.
31. Inagaki, S., *et al.* (2004) A new 39-plex analysis method for SNPs including 15 blood group loci. *Forensic Science International* **144**, 45–57.
32. Dixon, L.A., *et al.* (2006) Analysis of artificially degraded DNA using STRs and SNPs – results of a collaborative European (EDNAP) exercise. *Forensic Science International* **164**, 33–44.
33. Brion, M., *et al.* (2005) Introduction of an single nucleodite polymorphism-based 'Major Y-chromosome haplogroup typing kit' suitable for predicting the geographical origin of male lineages. *Electrophoresis* **26**, 4411–4420.
34. Wetton, J.H., *et al.* (2005) Inferring the population of origin of DNA evidence within the UK by allele-specific hybridization of Y-SNPs. *Forensic Science International* **152**, 45–53.
35. Frudakis, T., *et al.* (2003) A classifier for the SNP-based inference of ancestry. *Journal of Forensic Sciences* **48**, 771–782.
36. Bastiaens, M., *et al.* (2001) The melanocortin-1-receptor gene is the major freckle gene. *Human Molecular Genetics* **10**, 1701–1708.
37. Duffy, D.L., *et al.* (2007) A three–single-nucleotide polymorphism haplotype in Intron 1 of OCA2 explains most human eye-color variation. *American Journal of Human genetics* **80**, 241–252.
38. Grimes, E.A., *et al.* (2001) Sequence polymorphism in the human melanocortin 1 receptor gene as an indicator of the red hair phenotype. *Forensic Science International* **122**, 124–129.
39. Doi, Y., *et al.* (2004) A new method for ABO genotyping using a multiplex single-base primer extension reaction and its application to forensic casework samples. *Legal Medicine* **6**, 213–223.
40. Ackerman, M.J., *et al.* (2001) Molecular autopsy of sudden unexplained death in the young. *American Journal of Forensic Medicine and Pathology* **22**, 105–111.
41. Levo, A., *et al.* (2003) Post-mortem SNP analysis of CYP2D6 gene reveals correlation between genotype and opioid drug (tramadol) metabolite ratios in blood. *Forensic Science International* **135**, 9–15.

13 Lineage markers

Genetic lineage markers comprise polymorphisms that are present on the maternally inherited mitochondrial genome and the paternally inherited Y chromosome. The analysis of lineage markers is limited in most forensic casework because they do not possess the power of discrimination of autosomal markers. Even so, there are some features of both mtDNA and the Y chromosome that make them valuable forensic markers.

Mitochondria

The mitochondria are organelles that exist in the cytoplasm of eukaryotic cells. They carry out the vital job of producing approximately 90% of the energy required by the cell through the process of oxidative phosphorylation.

Inheritance of the mitochondrial genome

Mitochondria contain their own genome (mtDNA) which is maternally inherited [1, 2]. This was discovered in the 1950s after unusual patterns of inheritance of certain phenotypes were explained by the existence of extra nuclear genomes that did not obey Mendel's laws of inheritance.

During fertilization of an ovum, the sperm penetrates the egg and the sperm midpiece, which contains between 50–75 mitochondria, enters the egg along with the head [3]. The egg has around 1000-times more mitochondria than the sperm [3]. Although some paternal mtDNA enters the ovum it is actively removed [4]. The process is not always completely effective and very rare cases of paternal mtDNA inheritance have been documented [5].

Copy number

The mtDNA genome is present in multiple copies – individual cells can contain hundreds of mitochondria and a single human mitochondrion can contain several copies of the genome [6–8]. Somatic cells, therefore, have thousands of copies of the mitochondrial genome and approximately 1% of total cellular DNA comprises mtDNA [9, 10]. This compares with only two copies per cell of the nuclear genome.

An Introduction to Forensic Genetics W. Goodwin, A. Linacre and S. Hadi
© 2007 John Wiley & Sons, Ltd

The mtDNA genome

The human mitochondrial genome is a 16 569 bp circular molecule. It encodes for 22 transfer RNAs (tRNAs), 13 proteins and two ribosomal RNAs (the 12S and 16S rRNA) [11, 12]. The majority of mitochondrial proteins is encoded by the nuclear genome as, over hundreds of millions of years, following the formation of the symbiotic relationship between eubacteria and eukaryote cells, most of the genes have been transferred from the mitochondrial to the nuclear genome [13]. Analysis of the human mtDNA genome revealed a very economic use of the DNA and there are very few non-coding bases within the genome except in a region called the D-loop. The D-loop is the region of the genome where the initial separation, or displacement, of the two strands of DNA during replication occurs. The regulatory role of the D-loop has led to the other name by which it is known – the control region. It is approximately 1100 bp long.

Polymorphisms in mtDNA

The mtDNA genome accumulates mutations relatively rapidly as compared with the nuclear genome [14]. The high mutation rate[†] is due in part to the exposure of the mtDNA to reactive oxygen species that are produced as by-products in oxidative phosphorylation [15]. Direct analysis of mother-to-children transmissions has estimated that a mutation in the hypervariable regions is passed from mother to child approximately once in every 30 to 40 events. In the vast majority of cases where a mutation is detected, there is only one base change between the mother and child [16, 17].

Hypervariable regions

In most forensic investigations the aim of DNA profiling is to differentiate between individuals, therefore the most polymorphic regions are analysed. Following the sequencing of the human mtDNA genome it was apparent that the D-loop was not under the same functional constraints as the rest of the genome. Some blocks within the control region are highly conserved but large parts are not. Two main regions are the focus of most forensic studies, these are known as hypervariable sequence regions I and II (HVS-I and HVS-II) and they contain the highest levels of variation within the mtDNA genome. Both the hypervariable blocks are approximately 350 bp long. A third hypervariable region, HVS-III has also been used in some cases. Within the hypervariable regions the rate of mutation is not constant and some sites are hotspots for mutation, while others show much lower rates of change [18–20].

Because the polymorphic sites are concentrated within relatively small regions of the mtDNA genome, they can be analysed using PCR amplification followed by Sanger sequencing [21]. Many of the methods used for SNP detection can also be used (see Chapter 12).

[†] Note: mutations in the hypervariable regions are normally referred to as 'a base substitution' as they do not have an effect on any of the products encoded by the mtDNA – for simplicity the term 'mutation' will be used throughout this chapter.

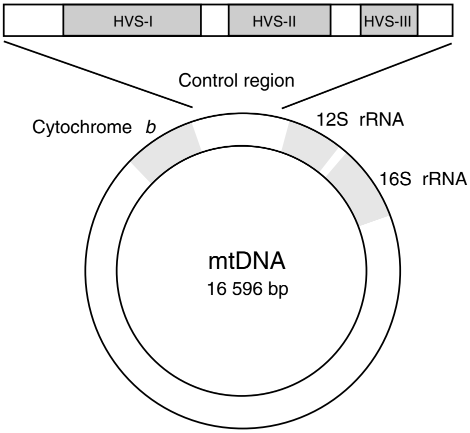


Figure 13.1 The mitochondrial genome is circular and 16 569 bp long. It encodes for 13 proteins, 22 transfer RNAs and two ribosomal RNAs. The polymorphic hypervariable sequence regions I, II and III (HVS-I, HVS-II and HVS-III) are located within the control region. Other regions of the genome that are utilized in forensic casework for species identification are the coding regions for the 12S and 16S ribosomal RNAs and cytochrome *b* gene

Applications of mtDNA profiling

There are several scenarios where mtDNA is a valuable genetic marker. These are related to two properties of mtDNA – the high copy number and the maternal inheritance. The high copy number is valuable when the amount of cellular material available for analysis is very small. Crime scene material that is commonly profiled using mtDNA includes hair shafts [22, 23] and faecal samples [24]. mtDNA is also useful for the analysis of human remains that are highly degraded and not amenable to standard STR typing [25, 26]. The maternal inheritance is a useful trait for human identification when there are no direct relatives to use as a reference sample – the identification of some of the Romanov family using Prince Philip as a reference sample provides a powerful illustration (Figure 13. 2) of the use of maternal inheritance [27].

A series of historical cases has followed that demonstrates the application of mtDNA when linking relatives to human remains [28–30].

Homoplasmy and heteroplasmy

Normally an individual contains only one type of mtDNA – this is termed homoplasmy. Mutations will inevitably occur within some of the thousands of copies of mtDNA within a cell and if these mutated copies of the genome were passed on to future generations, a mixture of different mtDNA genomes would occur. The process that maintains homoplasmy as the norm is not precisely understood but at some point a

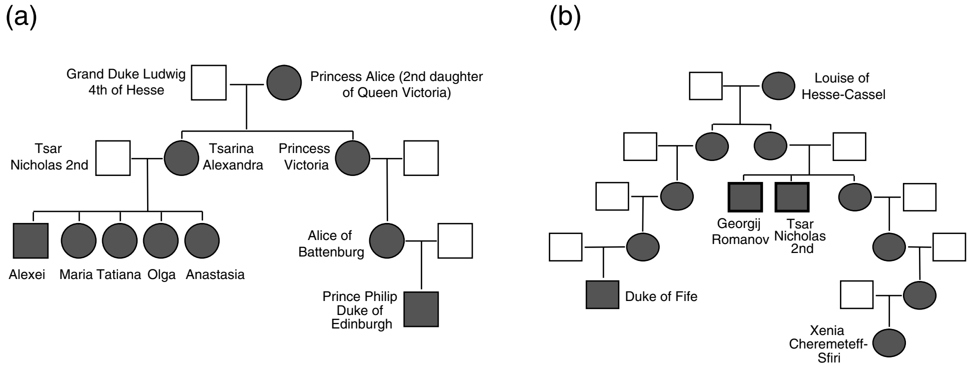

(a)
Grand Duke Ludwig 4th of Hesse
Princess Alice (2nd daughter of Queen Victoria)
Tsar Nicholas 2nd
Tsarina Alexandra
Princess Victoria
Alice of Battenburg
Prince Philip Duke of Edinburgh
Alexei
Maria
Tatiana
Olga
Anastasia
(b)
Louise of Hesse-Cassel
Georgij Romanov
Tsar Nicholas 2nd
Duke of Fife
Xenia Cheremeteff-Sfiri

Figure 13.2 The family tree of the Romanov royal family. (a) The maternal lineage of the Tsarina and her children, which provides a direct link to Prince Philip. (b) The maternal lineage of Tsar Nicholas, linking him to two living maternal relatives, the Duke of Fife and Xenia Cheremeteff-Sfiri. The squares represent males and the circles represent females, the transmission of the relevant mtDNA type is indicated by shading

genetic bottleneck occurs before the formation of a mature oocyte [31]. The bottleneck allows only a few mtDNA molecules to pass into the oocyte during its formation [32, 33], thereby reducing the possibility of passing on a mixture of wild type and mutant genomes.

It is, however, possible to find more than one type of mtDNA within a cell – this is known as heteroplasmy and it arises when a mother passes on a normal version of her mtDNA genome (wild type) and also a version of the genome that contains a mutation. An individual will therefore posses two versions on their mtDNA, maybe only differing by one base. Two factors, the severity of the bottleneck and subsequent genetic drift, determine the relative levels of wild to mutated mtDNA [16, 17]. Heteroplasmy can be

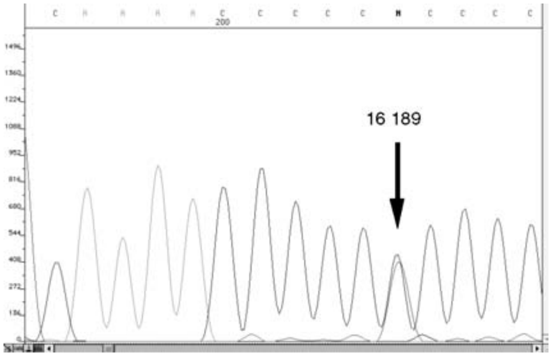


Figure 13.3 The above sequence shows the presence of heteroplasmy at position 16 189. Two bases, a C and an A, are present in approximately equal amounts (see plate section for full-colour version of this figure)

Table 13.1 An example where an mtDNA profile has been generated from the HVS-I of five bones that were found in close proximity. The mtDNA profiles of three women who were maternal relatives of three missing individuals are also shown. Maternal reference 1 clearly matches the bone profiles, while maternal references 2 and 3 can be excluded as potential maternal relatives as they have different mtDNA types. In this particular case the mtDNA profiling helped to establish the identification of the human remains, and that the bones all came form the same person [36]

Sample	HVS-I sequence			
Right femur	16 189C	16 223T	16 271C	16 278T
Left femur	16 189C	16 223T	16 271C	16 278T
Right pelvis	16 189C	16 223T	16 271C	16 278T
Left ulna	16 189C	16 223T	16 271C	16 278T
Left tibia	16 189C	16 223T	16 271C	16 278T
Maternal Reference 1	16 189C	16 223T	16 271C	16 278T
Maternal Reference 2	Same as Cambridge Reference Sequence			
Maternal Reference 3	16,278T	16,293G	16,311C	

stable through several generations before one of the mtDNA versions becomes fixed [16, 17, 29, 33, 34]. When heteroplasmy is detected, the two types of mtDNA genomes normally only differ at one position [22, 23].

Interpretation and evaluation of mtDNA profiles

mtDNA is used for both associating crime scene samples with individuals and also in the identification of human remains. In both cases the profile that has been generated from the unknown sample has to be compared to a reference profile. In the case of a crime scene investigation, the reference sample will be from a suspect. In the case of human identification, a sample taken from a maternal relative or a personal artefact such as a toothbrush can be used [35].

The first step is to turn the information into a more manageable form. The data after sequencing consists of upwards of 350 DNA bases from both HVS-I and HVS-II and around 150 bases from HVS-III. Once the sequencing data have been checked to ensure that there is confidence in the sequence data and no errors, it is compared to the Cambridge Reference Sequence (CRS) [11, 12]. The CRS was the first complete sequence of the mtDNA genome to be published in 1981. Differences between the questioned sequence and the CRS are noted and only these differences are recorded. Table 13.1 shows an example of where a HVS-I profile generated from a set of human remains is compared to profiles generated from three reference samples. The mtDNA profile is called a haplotype.

Declaring a match is straight forward but exclusions can be more problematic. When a questioned sample and a reference sample differ at only one position the likelihood of that one base difference occurring though a mutation has to be assessed. In such an instance the results are usually classified as inconclusive – when there are two or more differences between a questioned and known sample it is normally classified as an exclusion [21].

If a match is declared, the statistical significance of the match has to be assessed. The mtDNA genome is inherited as a single locus and this limits the evidential value of the marker in forensic cases. Haplotype frequencies have to be measured directly by counting the occurrence of a particular haplotype in a database and reporting the size of the database. When databases are relatively small, for example 100, many of the less common haplotypes that are within a population will not be represented. There are mechanisms that compensate for the limitations of reference databases, such as minimum haplotype frequencies, and employing standard error calculations and correction factors to allow for subpopulations [37, 38].

When reporting the results of mtDNA analysis, the caveats associated with mtDNA have to be clearly explained so that there is no confusion with 'standard' (autosomal STR typing) analysis. In particular that '*it is inherited only from one's mother, and therefore all individuals who are related by a maternal link will have the same mtDNA profile*', and that '*it varies less between individuals, and therefore more individuals chosen at random from the population will have the same mtDNA profile*' should be made very clear.

The Y chromosome

In humans the Y chromosome is approximately 60 Mb long (million base pairs) long and contains just 78 genes [39]. The SRY gene (sex-determining region Y) located on the Y chromosome encodes a protein that triggers the development of the testes and through an extended hormonal pathway causes a developing foetus to become male [40].

With the exception of two regions, PAR 1 and 2 (PAR = pseudoautosomal region), located at the tips of the chromosome, no recombination occurs during meiosis. The remaining 95% of the Y chromosome is non-recombining, male specific, and is passed from father to son unchanged, except when mutations occur. The lack of recombination may be the reason why there are relatively few genes on the Y chromosome. If there is no chromosome crossing over, mutations within genes have little chance to be repaired or rectified and hence will be passed onto the next generation.

Y chromosome polymorphisms

The Y chromosome contains a large number of polymorphisms including variable number [41] and short tandem repeats (VNTRs and STRs), insertions, deletions and single nucleotide polymorphisms (SNPs).

The first STR locus to be identified on the Y chromosome was DYS19 [42]. Since then hundreds of Y chromosome STR have been described. The development of Y STR typing has mirrored the development of the autosomal STRs, and multiplexes have been developed with increasing numbers of robust and highly discriminating Y STR multiplexes [43–46]. The growth in interest in Y STR loci has led to numerous

Table 13.2 The Y chromosome STR loci that are commonly used in forensic analysis

Minimal haplotype	Extended haplotype	PowerPlex® Y	AmpF*l*STR® Yfiler®
DYS19	DYS19	DYS19	DYS19
DYS385 a/b	DYS385 a/b	DYS385 a/b	DYS385
DYS389 I	DYS389 I	DYS389 I	DYS389 I
DYS389 II	DYS389 II	DYS389 II	DYS389 II
DYS390	DYS390	DYS390	DYS390
DYS391	DYS391	DYS391	DYS391
DYS392	DYS392	DYS392	DYS392
DYS393	DYS393	DYS393	DYS393
	DYS438	DYS437	DYS437
	DYS439	DYS438	DYS438
		DYS439	DYS439
			DYS448
			DYS456
			DYS458
			DYS635
			GATA H4

population studies to establish allele frequency databases. The ‘Y Chromosome Haplotype Reference Database’ was established to collate STR haplotypes (www.yhrd.org). To ensure comparability between datasets, minimal and extended haplotypes were defined. Two commercial kits, the PowerPlex® Y (Promega Corporation) and the AmpF*l*STR® Yfiler® (Applied Biosystems) incorporate all of the extended haplotype loci (Table 13.2).

Forensic applications of Y chromosome polymorphisms

That the Y chromosome is only found in males makes it a valuable tool, in particular for the analysis of male and female mixtures after sexual assaults when differential DNA extraction is not possible; Y STR analysis has been successful with female:male ratios of up to 2000:1 [47]. The presence of male DNA has also been detected when vaginal swabs are analysed, even when no spermatozoa have been detected – either through the assailant being azoospermic (1–2% of rape cases) or through the deterioration of the spermatozoa [44,48]. The Y STRs can also be used to detect the presence of two male profiles – the interpretation of the mixtures depends on the presence of major and minor contributors [47].

In addition to using Y chromosome testing for the identification of evidential samples, it has also been used for paternity testing and is particularly valuable in deficient cases, where the alleged father is not available for testing. In these cases, any male relative who is paternally related to the alleged father can be used as a reference. An extreme example of where this has been used is the paternity analysis that linked the third US president, Thomas Jefferson to the child of one of his slaves, Sally Hemings [49]. Cases involving human identification have also used the Y chromosome as a tool to link remains to paternal family members, and as with deficient paternity cases the use of the Y chromosome is particularly advantageous when there are no parents or children to use as reference material; it also simplifies the sorting of the material following mass disasters [50]. The mutation rate in Y STR loci is similar to autosomal STRs, at approximately 2.8×10^{-3} [51, 52]. The Y chromosome will accumulate mutations as it is passed through the patrilineal line and direct comparison between males on the same lineage may result in a false exclusion if mutations are not considered.

The non-random distribution of the Y chromosome among global populations, due largely to the widespread practice of patrilocality [53, 54] (where the female moves to the male's birth place/residence after marriage), makes it a useful tool for inferring the geographical origins of biological material recovered from a crime scene and human remains [55]. In some cultures, where the male name is passed onto male children, there is also the potential of attributing surnames to Y profiles [56, 57].

Interpretation and evaluation of Y STR profiles

When the Y chromosome profiles from a reference and an unknown sample match, the significance of the match has to be assessed. The first step is to assess the frequencies of the Y STR haplotypes in the population of interest. The simplest method is to report the frequency of the Y STR haplotype in the population, known as the counting method. The figure quoted is entirely dependent upon the size of the database and is normally based on frequency databases that are constructed for the major ethnic groups represented within individual countries, although comparisons can also be made to the combined data in the yhrd databases with over 40 000 haplotypes (representing at least the minimal haplotype). So, for example, a match can be reported as 'the haplotype has been seen twice in 400 UK Caucasian individuals'.

Difficulties arise in the interpretation of the Y chromosome. This is primarily caused by the patrilineal inheritance and clustering of male family members in relatively small geographic areas. This geographical clustering of male relatives coupled with the limited size of the haplotype frequency databases (many haplotypes are seen only once) makes the estimation of profile frequencies hazardous. An alternative method for assessing the significance of a match is to use a likelihood ratio and to incorporate population subdivisions with the increased potential for common co-ancestry [38]. Regardless of the method used to calculate the matching frequency, when presenting the results of Y chromosome analysis, as with mtDNA, there is a need clearly to state how the use of Y STR typing varies from that of autosomal markers and that there will be other males in the population with the same Y STR haplotype.

Further reading

Jobling, M., Hurles, M., and Tyler-Smith, C. (2004) *Human Evolutionary Genetics: Origins, Peoples and Disease.* Garland Science.

Cavalli-Sforza, L.L., Menozzi, P., and Piazza, A. (1996) *The History and Geography of Human Genes.* Princeton University Press.

WWW resources

The Y Chromosome Haplotype Reference Database: http://www.ystr.org/index.html

EMPOP - Mitochondrial DNA Control Region Database: http://www.empop.org/

MITOMAP: http://www.mitomap.org/

References

1. Giles, R.E., *et al.* (1980) Maternal inheritance of human mitochondrial-DNA. *Proceedings of the National Academy of Sciences of the United States of America–Biological Sciences* **77**, 6715–6719.
2. Shitara, H., *et al.* (1998) Maternal inheritance of mouse mtDNA in interspecific hybrids: segregation of the leaked paternal mtDNA followed by the prevention of subsequent paternal leakage. *Genetics* **148**, 851–857.
3. AnkelSimons, F., and Cummins, J.M. (1996) Misconceptions about mitochondria and mammalian fertilization: implications for theories on human evolution. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 13859–13863.
4. Thompson, W.E., *et al.* (2003) Ubiquitination of prohibitin in mammalian sperm mitochondria: possible roles in the regulation of mitochondrial inheritance and sperm quality control. *Biology of Reproduction* **69**, 254–260.
5. Schwartz, M., and Vissing, J. (2002) Paternal inheritance of mitochondrial DNA. *New England Journal of Medicine* **347** (8), 576–580.
6. Garrido, N., *et al.* (2003) Composition and dynamics of human mitochondrial nucleoids. *Molecular Biology of the Cell* **14**, 1583–1596.
7. Cavelier, L., *et al.* (2000) Analysis of mtDNA copy number and composition of single mitochondrial particles using flow cytometry and PCR. *Experimental Cell Research* **259**, 79–85.
8. Margineantu, D.H., *et al.* (2002) Cell cycle dependent morphology changes and associated mitochondrial DNA redistribution in mitochondria of human cell lines. *Mitochondrion* **1**, 425–435.
9. Robin, E.D., and Wong, R. (1988) Mitochondrial-DNA molecules and virtual number of mitochondria per cell in mammalian-cells. *Journal of Cellular Physiology* **136**, 507–513.
10. Bogenhag, D., and Clayton, D.A. (1974) Number of mitochondrial deoxyribonucleic-acid genomes in mouse L and human Hela cells – quantitative isolation of mitochondrial deoxyribonucleic acid. *Journal of Biological Chemistry* **249**, 7991–7995.
11. Anderson, S., *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465.
12. Andrews, R.M., *et al.* (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics* **23**, 147–147.
13. Dyall, S.D., *et al.* (2004) Ancient invasions: from endosymbionts to organelles. *Science* **304**, 253–257.
14. Brown, W.M., *et al.* (1979) Rapid evolution of animal mitochondrial-DNA. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 1967–1971.
15. Hashiguchi, K., *et al.* (2004) Oxidative stress and mitochondrial DNA repair: implications for NRTIs induced DNA damage. *Mitochondrion* **4**, 215–222.
16. Parsons, T.J., *et al.* (1997) A high observed substitution rate in the human mitochondrial DNA control region. *Nature Genetics* **15**, 363–368.
17. Forster, L., *et al.* (2002) Natural radioactivity and human mitochondrial DNA mutations. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 13950–13954.
18. Tamura, K., and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10**, 512–526.
19. Meyer, S., *et al.* (1999) Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics* **152**, 1103–1110.

20. Excoffier, L., and Yang, Z.H. (1999) Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. *Molecular Biology and Evolution* **16**, 1357–1368.
21. Carracedo, A., *et al.* (2000) DNA commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. *Forensic Science International* **110**, 79–85.
22. Melton, T., and Nelson, K. (2001) Forensic mitochondrial DNA analysis: two years of commercial casework experience in the United States. *Croatian Medical Journal* **42**, 298–202.
23. Melton, T., *et al.* (2005) Forensic mitochondrial DNA analysis of 691 casework hairs. *Journal of Forensic Sciences* **50**, 73–80.
24. Hopwood, A.J., *et al.* (1996) DNA typing from human faeces. *International Journal of Legal Medicine* **108**, 237–243.
25. Goodwin, W., *et al.* (1999) The use of mitochondrial DNA and short tandem repeat typing in the identification of air crash victims. *Electrophoresis* **20**, 1707–1711.
26. Holland, M.M., *et al.* (1993) Mitochondrial-DNA sequence-analysis of human skeletal remains – identification of remains from the VietnamWar. *Journal of Forensic Sciences* **38**, 542–553.
27. Gill, P., *et al.* (1994) Identification of the remains of the Romanov family by DNA Analysis. *Nature Genetics* **6**, 130–135.
28. Stone, A.C., *et al.* (2001) Mitochondrial DNA analysis of the presumptive remains of Jesse James. *Journal of Forensic Sciences* **46**, 173–176.
29. Ivanov, P.L., *et al.* (1996) Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II. *Nature Genetics* **12**, 417–420.
30. Stoneking, M., *et al.* (1995) Establishing the identity of Anderson Anna Manahan. *Nature Genetics* **9**, 9–10.
31. Marchington, D.R., *et al.* (1998) Evidence from human oocytes for a genetic bottleneck in an mtDNA disease. *American Journal of Human Genetics* **63**, 769–775.
32. Sekiguchi, K., *et al.* (2003) Inter- and intragenerational transmission of a human mitochondrial DNA heteroplasmy among 13 maternally related individuals and differences between and within tissues in two family members. *Mitochondrion* **2**, 401–414.
33. Chinnery, P.F., *et al.* (2000) The inheritance of mitochondrial DNA heteroplasmy: random drift, selection or both? *Trends in Genetics* **16**, 500–505.
34. Lagerstrom-Fermer, M., *et al.* (2001) Heteroplasmy of the human mtDNA control region remains constant during life. *American Journal of Human Genetics* **68**, 1299–1301.
35. Brenner, C.H., and Weir, B.S. (2003) Issues and strategies in the DNA identification of World Trade Center victims. *Theoretical Population Biology* **63**, 173–178.
36. Goodwin, W., *et al.* (2003) The identification of a US serviceman recovered from the Holy Loch, Scotland. *Science and Justice* **43**, 45–47.
37. Balding, D.J. (2005) *Weight-of-evidence for Forensic DNA Profiles*, John Wiley & Sons.
38. Buckleton, J., *et al.* (2005) *Forensic DNA Evidence Interpretation*, CRC Press.
39. Skaletsky, H., *et al.* (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–U822.
40. Sinclair, A.H., *et al.* (1990) A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature* **346**, 240–244.
41. Jobling, M.A., *et al.* (1998) Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). *Human Molecular Genetics* **7**, 643–653.
42. Roewer, L., *et al.* (1992) Simple repeat sequences on the human Y-chromosome are equally polymorphic as their autosomal counterparts. *Human Genetics* **89**, 389–394.
43. Butler, J.M., *et al.* (2002) A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers. *Forensic Science International* **129**, 10–24.
44. Gusmao, L., *et al.* (1999) Y chromosome specific polymorphisms in forensic analysis. *Legal Medicine* **1**, 55–60.
45. Mulero, J.J., *et al.* (2006) Development and validation of the AmpFlSTR (R) Yfiler (TM) PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. *Journal of Forensic Sciences* **51**, 64–75.

46. Krenke, B.E., *et al.* (2005) Validation of a male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex. *Forensic Science International* **148**, 1–14.
47. Prinz, M., *et al.* (1997) Multiplexing of Y chromosome specific STRs and performance for mixed samples. *Forensic Science International* **85**, 209–218.
48. Sibille, I., *et al.* (2002) Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa. *Forensic Science International* **125**, 212–216.
49. Foster, E.A., *et al.* (1998) Jefferson fathered slave's last child. *Nature* **396**, 27–28.
50. Corach, D., *et al.* (2001) Routine Y-STR typing in forensic casework. *Forensic Science International* **118**, 131–135.
51. Kayser, M., *et al.* (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *American Journal of Human Genetics* **66**, 1580–1588.
52. Kayser, M., and Sajantila, A. (2001) Mutations at Y-STR loci: implications for paternity testing and forensic analysis. *Forensic Science International* **118**, 116–121.
53. Oota, H., *et al.* (2001) Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nature Genetics* **29**, 20–21.
54. Burton, M.L., *et al.* (1996) Regions based on social structure. *Current Anthropology* **37**, 87–123.
55. Jobling, M.A. (2001) Y-chromosomal SNP haplotype diversity in forensic analysis. *Forensic Science International* **118**, 158–162.
56. Jobling, M.A. (2001) In the name of the father: surnames and genetics. *Trends in Genetics* **17**, 353–357.
57. Sykes, B. and Irven, C. (2000) Surnames and the Y chromosome. *American Journal of Human Genetics* **66**, 1417–1419.

Appendix 1 Forensic parameters

Match probability

The probability that the two randomly selected individuals will have identical genotype.

Formula:

$$p_{\mathrm{M}} = \sum_{k=1}^{m} p_k^2$$

where, p_{M} is the match probability, p_k represents the frequency of each distinct genotype, m is the number of distinctive genotypes. The combined probability of match over several loci is the product of the value for all the loci.

Power of discrimination

The probability that two randomly selected individuals will have different genotypes.

Formula:

For a single locus the formula is :$p_{\mathrm{D}} = 1 - p_{\mathrm{M}}$

For several loci the formula is :$P_{\mathrm{Dcomb}} = 1 - \prod_{i=1}^{n} (1 - P_{\mathrm{Di}})$

where, p_{D} is the power of discrimination of a single locus, p_{M} is the match probability of a single locus, P_{Dcomb} is the power of discrimination of several loci, P_{Di} is the individual locus power of discrimination. The "Π" sign stands for multiplication.

Power of exclusion

The fraction of the individuals that is different from that of a randomly selected individual. It can also be defined as the power of a locus to exclude a person being the biological father. Thus the value differs in each case. The average for a locus is the power for a single locus.

An Introduction to Forensic Genetics W. Goodwin, A. Linacre and S. Hadi
© 2007 John Wiley & Sons, Ltd

Formula:

$$\mathrm{PE} = h^2(1 - 2hH^2)$$

where, h is the heterozygosity and H is the homozygosity at the locus.

For several loci the formula is:

$$PE_{\mathrm{comb}} = 1 - \prod_{l=1}^{L}(1 - \mathrm{PE}_l)$$

where L is the number of the loci, PE_l is the exclusion probability for the lth locus and the Π sign stands for multiplication.

Polymorphic information content

Indicates the polymorphic level of a locus.

Formula:

$$\mathrm{PIC} = 1 - \sum_{i=1}^{n} p_i^2 - \left(\sum_{i=1}^{n} p_i^2\right)^2 + \sum_{i=1}^{n} p_i^4$$

where, p_i is the frequency of each distinct allele, and n is the number of distinct alleles.

Appendix 2 Useful web links

Professional bodies and agencies

ANZFSS – The Australian and New Zealand Forensic Science Society, Inc. http://www.anzfss.org.au/index.htm

ENSFI – European Network of Forensic Science Institutes. http://www.enfsi.org/

FBI – Federal Bureau of Investigation: FBI Laboratory. http://www.fbi.gov/hq/lab/labhome.htm

FSS – The Forensic Science Service (UK). http://www.forensic.gov.uk/

ISFG – International Society for Forensic Genetics: http://www.isfg.org/

INTERPOL – The International Criminal Police Organization. http://www.interpol.int/

NIFS – National Institute of Forensic Science Australia. http://www.nifs.com.au/home.html

The Forensic Science Society (UK). http://www.forensic-science-society.org.uk/

Statistical analysis

Forensic Mathematics: Contains information of the kinship software with DNA·View^{TM} and articles/discussions focused on the statistical/mathematical interpretation of DNA profiles. http://dna-view.com

GDA: a statistical software package that computes linkage and Hardy–Weinberg disequilibrium, some genetic distances, and provides method-of-moments estimators for hierarchical F-statistics. http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php

Powerstats: A Microsoft Office Excel-based tool for calculating descriptive statistics and forensic parameters for STR loci. http://www.promega.com/geneticidtools/powerstats/

© 2007 John Wiley & Sons, Ltd

ENFSI DNA WG STR Population Database: calculates the profile frequency of a SGM Plus profile using 24 European allele frequency databases. http://www.str-base.org/index.php

Genetic markers and population databases

STRBase – Short Tandem Repeat DNA Internet DataBase: contains a large amount of information on STR polymorphisms. http://www.cstl.nist.gov/biotech/strbase/

YHRD – Y Chromosome Haplotype Reference Database: a searchable database of Y chromosome STR haplotypes. http://www.ystr.org/index.html

EMPOP – Mitochondrial DNA Control Region Database: a collection of searchable mtDNA control region haplotypes from all over the world. http://www.empop.org/

MITOMAP –A compendium of polymorphisms and mutations of the human mitochondrial DNA. http://www.mitomap.org/

ALFRED – The ALlele FREquency Database: contains allele frequency data for a wide range of genetic polymorphisms in different populations. http://alfred.med.yale.edu/alfred/index.asp

Commercial providers

All these companies provide a wide range of products in addition to the ones noted.

Applied Biosystems: suppliers of kits and equipment for the analysis of STR and SNP polymorphisms. http://www.appliedbiosystems.com/

Promega Corporation: suppliers of kits for STR analysis and a wide variety of molecular biology products. http://www.promega.com/

Qiagen: suppliers of widely used DNA isolation kits. http://www1.qiagen.com/

Whatman®: suppliers of FTA® card. http://www.whatman.com/

Glossary

Allele: alternative forms of a gene or section of DNA at a given genetic locus.

Allelic Drop Out: non detection of an allele at a given locus. This results in only one of the two alleles being detected at a heterozygous locus.

Allelic Ladder: a mixture of all the common alleles at a given locus. The allelic ladder allows comparison with the unknown alleles and assists in allelic designation.

Amplifiable fragment length polymorphisms (AMP-FLPs): polymorphic loci where alleles differ in the number of tandem core repeats. Alleles are typically between 500 and 1,000 bp in length. An example of an AMP-FLP is the locus D1S80.

Autosome: a non-sex chromosome. In humans there are the 22 pairs of autosomal chromosomes; these do not include the X and Y sex chromosomes.

bp: base pair – two complementary nucleotides in double stranded DNA. Adenine pairs with thymine and guanine pairs with cytosine.

Chromosome: a single molecule of double stranded DNA associated with proteins to form highly ordered structure. Chromosomes are located in the cell nucleus of eukaryotes and are visible with light microscopy only during cell division when they become highly condensed.

CODIS loci: the FBI defines a set of 13 STR markers for use in forensic analysis.

Diploid: presence of two chromosomal sets in a cell and therefore containing two copies of the genome.

DNA polymerase: an enzyme that catalyses the formation of a complimentary DNA strand in the $5'$ to $3'$ direction acting on a template DNA strand.

Electrophoresis: separation of charged molecules through a matrix. DNA is negatively charged and will migrate from the cathode (-ve) to the anode (+ve) when an electric current is applied across the matrix.

Euchromatin: part of the chromosome that is loosely packed in the interphase of the cell cycle. Most of the transcribed regions of the genome are located within the euchromatin.

An Introduction to Forensic Genetics W. Goodwin, A. Linacre and S. Hadi
© 2007 John Wiley & Sons, Ltd

Gene: a functional part of DNA that encodes a protein or RNA molecule.

Gene Frequency: the relative frequency (proportion) of a gene in a population.

Genome: the entire haploid complement of DNA in a cell or organism. The human genome comprises approximately 3.2 billion bases.

Genotype: the particular set of alleles present in each cell. At any one locus the two alleles define the genotype such that if there are only 2 alleles then there are 3 possible genotypes (AA, AB, BB).

Haploid: presence of one set of chromosomes in a cell and therefore containing one copy of the genome.

Hardy-Weinberg Law (HW law): a law stating that in an ideal population, the frequencies of an allele will remain constant from one generation to the next.

Heterochromatin: a highly condensed part of the chromosome which remains tightly packed throughout the cell cycle and is predominantly non-coding.

Heterozygous: having different alleles at any particular locus.

Homozygous: presence of two identical alleles at a given locus.

Locus: the physical position of a gene or section of DNA on a chromosome. The plural of locus is loci.

Locus Drop Out: the non detection of both alleles at a given locus.

Low Copy Number PCR: a more sensitive method of DNA profiling where the number of cycles is increased in order to amplify amounts of DNA typically less than 100 picograms.

Mitochondria: small semi-autonomous organelles of the cell which are primarily responsible for energy production. They have their own circular genome that is 16 569 bp in length.

Mutation: alteration in the DNA sequence. The most common form of a mutation is a single base transition (see below).

Oligonucleotide: short sequence of single stranded DNA.

Phenotype: physical form of an organism resulting from genetic traits and environmental factors.

Polymerase Chain Reaction (PCR): a process of enzymatic amplification of DNA *in vitro*.

Posterior Odds: in terms of a court case these are the odds of guilt of a defendant after the presentation of all evidence. It is the posterior odds that the judge or jury consider when coming to a verdict.

Primer: oligonuclceotide that binds to complimentary single stranded DNA and acts as a priming site for the initiation for the synthesis of the complementary strand by DNA polymerase.

Prior Odds: in terms of a court case these are the odds of guilt of a defendant before the presentation of forensic evidence and may be considered as based on the non-scientific evidence.

Purine: nitrogenous bases found in nucleic acids. The purine bases are adenine and guanine.

Pyrimidine: nitrogenous bases found in nucleic acids. The pyrimidine bases are cytosine and thymine.

Restriction enzyme: enzyme with endonucleic activity that cuts DNA at specific sequences. An example is *Eco* RI that cuts within the sequence GAATTC.

Short tandem repeat (STR): polymorphic region of DNA where alleles differ in the number of tandemly arranged core repeats. STR alleles typically range in size between 100 bp and 400 bp. Also known as microsatellites.

Single Nucleotide Polymorphism (SNP): the occurrence of two (or more) alleles at a single base position within the genome. SNPs are typically biallelic: transitions (see below) are more common than transversions (see below). To be considered as a population level SNP polymorphism the rare allele should be at a frequency greater than 1%.

Southern blot: a technique that is used to transfer DNA from a gel onto a nylon membrane.

Telomere: the terminal regions at the tips of the chromosomes. Satellite DNA forms much of the telomeric regions.

Theta (θ) Statistic: a measurement of inbreeding within a population. Theta is often used interchangeably with F_{ST} in forensic genetics.

Transition: the change in a base from a purine (see above) to a purine (A/G) or a pyrimidine (see above) to a pyrimidine (C/T).

Transversion: the change in a base from a purine to a pyrimidine (A or G to C or T) or a pyrimidine to a purine (C or T to A or G).

Variable number tandem repeat (VNTR): polymorphic region of DNA where alleles differ in the number of tandemly arranged core repeats. VNTR alleles can range in size from around 500 bp to over 20 kb. Also known as minisatellites.

Abbreviations

A: Adenine – A purine base. One of the four bases of DNA

AMP-FLPs: Amplified fragment length polymorphisms

bp: Base pair

C: Cytosine – a pyrimidine base. One of the four bases of DNA

CODIS: Combined DNA Index System

DNA: Deoxyribonucleic acid.

EDTA: Ethylenediaminetetraacetic acid

FBI: Federal Bureau of Investigation of the USA

FSS: Forensic Science Service of the UK

G: Guanine – a purine base. One of the four bases of DNA

kb: Kilobase, a string of a thousand DNA bases.

LINE: Long interspersed elements

LTR: Long terminal repeat

MLP: Multi-locus probe

PCR: Polymerase chain reaction

RNA: Ribonucleic acid.

SINE: Short interspersed elements

SLP: Single locus probe

SNP: Single nucleotide polymorphism

STR: Short terminal repeat

T: Thymine – a pyrimidine. One of the four bases found in DNA.

VNTR: Variable number tandem repeat

An Introduction to Forensic Genetics W. Goodwin, A. Linacre and S. Hadi
© 2007 John Wiley & Sons, Ltd

Units of measurement

Prefix	Symbol	10^n	Numerical representation
giga	G	10^9	1 000 000 000
mega	M	10^6	1 000 000
kilo	k	10^3	1 000
milli	m	10^{-3}	0.001
micro	μ	10^{-6}	0.000 001
nano	n	10^{-9}	0.000 000 001
pico	p	10^{-12}	0.000 000 000 001

Index

α-amylase 22
ABI PRISM® genetic analyzers 55, 116
ABO blood grouping system 2
absorption–inhibition ABO typing 2
adventitious hits 99
agarose gels 33–4, 35
allele
  ceiling principle 79–80, 83
  drop-out 68, 69, 73
  specific hybridization 119
allele frequency databases 78–83
  allele ceiling principle 79–80, 83
  Balding size bias correction 80, 81, 83
  profile ceiling principle 82, 83
  selection 83
  subpopulations 80–2, 83
allelic ladders 58–61
amelogenin locus 53
3-aminophthalhydrazide 19–20
AMP-FLPs *see* amplified fragment length polymorphisms
AmpF/STR® *see* Identifiler®; PowerPlex® 16 system; SGM Plus®
amplified fragment length polymorphisms (AMP-FLPs) 39
AmpliTaq Gold® 40–1, 44

Balding size bias correction 80, 81, 83
Bayesian approach
  evaluation process 92–3, 95
  kinship testing 109–11
biological samples 17–25
  collection 3, 19, 20–1
  contamination 19, 20
  identification and characterization 19–20
  presumptive testing 21–3
  reference samples 23
  sexual and physical assault 21
  sources 17–19
  storage 23–4
blood
  biological samples 3, 17–20, 21, 23–4
  DNA quantification 32
  grouping 2, 121
  polymerase chain reaction 44, 46
bones 31–2, 72–3, 129
buccal swabs 23, 46

Cambridge Reference Sequence (CRS) 129
capillary electrophoresis (CE) 54–5, 57–8, 118
CE *see* capillary electrophoresis
Chelex® 100 resin 27–9, 31, 34, 45
chi square tests 78
chromosomal organization 7–9
CODIS *see* Combined DNA Index System
cold cases 101–2
Combined DNA Index System (CODIS) 52, 102–3
contamination
  biological samples 19, 20
  cold cases 102
  polymerase chain reaction 45–8
Crick, F. 7
CRS *see* Cambridge Reference Sequence

D-loop 126
DAPI 33, 35
ddNTPs *see* dideoxynucleotide triphosphates
defendant's fallacy 94
degraded DNA 71–3, 101–2
deionized formamide 55
deoxynucleotide triphosphates (dNTPs) 8, 42, 116
deoxyribonucleic acid (DNA)
  extraction 27–32
    challenging samples 30–2
    Chelex® 100 resin 27–9, 31, 34, 45
    FTA® paper 29–30
    general principles 27–30
    hair shafts 31
    hard tissues 31–2
    phenol–chloroform-based 29, 31
    seminal fluid 30
    silica based methods 28–9, 31, 36
  quantification 32–6
    agarose gels 33–4, 35

deoxyribonucleic acid (DNA) (*Contd.*)
    fluorescence spectrophotometry 35
    hybridization 35
    real-time polymerase chain reaction 35–6
    ultraviolet spectrophotometry 34
  replication 40
  structure 7–15
    chromosomal organization 7–9
    coding and regulatory sequence 9–10
    extragenic DNA 10
    forensic genetics 11
    genetic diversity 11
    human genome 9–10, 11
    short tandem repeats 13, 14
    single nucleotide polymorphisms 13–14
    tandem repeats 10, 12–13, 14
  transposons 10
deoxyribose 8
4′,6-diamidino-2-phenylindole 33, 35
dideoxynucleotide triphosphates (ddNTPs) 116–17
dinucleotide repeats 66
dithiothreitol (DTT) 30
DNA *see* deoxyribonucleic acid
DNA IQ™ Isolation System 36
dNTPs *see* deoxynucleotide triphosphates
DTT *see* dithiothreitol

EDNAP *see* European DNA Profiling Group
epithelial cells 17–19, 22, 33
ethidium bromide 33, 35
European DNA Profiling Group (EDNAP) 103
evaluation process 87–95
  Bayesian approach 92–3, 95
  defendant's fallacy 94
  frequentist approach 88–9, 94–5
  hierarchies of propositions 87–93
  likelihood ratios 89–92, 95
  prosecutor's fallacy 91, 93–5
exact tests 78
extragenic DNA 10

familial searching 100–1
flanking DNA 12
fluorescence
  labelling 54
  spectrophotometry 35
formamide 55
frequentist approach 88–9, 94–5
FTA® paper 23–4, 29–30

GeneMapper™ *ID* 56, 59, 67
GeneScan® 56, 58, 67
genetic diversity 11
Genotyper® 59–60
geographical
  ancestry 120
  clustering 132

hair
  follicles 17–19, 23
  shafts 31
haplotype frequencies 130–1, 132
Hardy–Weinberg equilibrium (HWE) 75–83
  allele frequency databases 78–83
  deviations 76–8
  infinitely large populations 76–7
  migration effects 77
  mutations 77
  natural selection 77
  random mating 77
  statistical tests 77–8
heteroplasmy 127–9
heterozygosity 138
hierarchies of propositions 87–93
  Bayesian approach 92–3, 95
  frequentist approach 88–9, 94–5
  likelihood ratios 89–92, 95
homoplasmy 127–9
homozygosity 78, 138
human
  genome 9–10, 11
  karyotype 9
  remains 1, 111–12
    lineage markers 129
    short tandem repeats 71–3
HWE *see* Hardy–Weinberg equilibrium
hybridization-based methods 35
hypervariable regions 126–7, 129

Identifiler® 105, 120, 52, 61
inbreeding 81–2
infinitely large populations 76–7
inhibition of PCR 44–5
internal-lane size standards 56–8
Interpol Standard Set of Loci (ISSOL) 103
interspersed repeats 10
ISSOL *see* Interpol Standard Set of Loci

Jefferson, Thomas 131
Jeffreys, Alec 2–3

Kastle–Meyer (KM) 21
kinship testing 1, 105–14
  absorption–inhibition ABO typing 2
  Bayesian approach 109–11
  complex relationships 111

  human remains 111–12
  likelihood ratios 105, 109–10
  lineage markers 131
  paternity testing 105–11
  Punnett squares 106–7
  subpopulations 111
KM *see* Kastle–Meyer

Landsteiner, Karl 2
Lattes, Leone 2
LCN *see* low copy number
leuco-malachite green (LMG) 21
likelihood ratios
  evaluation process 89–92, 95
  kinship testing 105, 109–10
  short tandem repeats 78
  statements 91
lineage markers 125–35
  applications 127–30, 131–2
  copy number 125
  hypervariable regions 126–7, 129
  inheritance 125
  interpretation 129–30, 132
  mitochondrial genome 125–30
  polymorphisms 126, 130–2
  Y chromosome 125, 130–2
LINEs *see* long interspersed elements
LMG *see* leuco-malachite green
local Southern method 57–8
locus drop-out 68, 69
long interspersed elements (LINEs) 10
long terminal repeats (LTRs) 10
low copy number (LCN) PCR
  contamination 44
  degraded DNA 101–2
  short tandem repeats 68–70, 73
LTRs *see* long terminal repeats
luminol 19–20

match probability 137
  statements 88–9
migration effects 77
mitochondrial genome 111, 117, 125–30
mixed profiles 70–1
molecular autopsy 121
Mullis, Kary 3–4
multi-locus exact tests 78
multiplex development 51–3
muscle tissue 72–3
mutations 77

national DNA databases 97–104
  adventitious hits 99
  cross-border 103
  United Kingdom 97–102
  United States 102–3
natural selection 77
non-template-dependence 66–7

off-ladder peaks 58–9, 60
overloaded profiles 68

PACE *see* Police and Criminal Evidence Act
paternity index 106, 110–11
paternity testing 1, 105–11
  absorption–inhibition ABO typing 2
  lineage markers 131
PCR *see* polymerase chain reaction
peak balance 69, 70
personal objects 112, 129
phenol–chloroform-based DNA extraction 29, 31
phosphodiester bonds 8
PicoGreen® 35
Pitchfork, Colin 1
Police and Criminal Evidence Act (PACE) 98
polymerase chain reaction (PCR) 39–50
  components 40–2
  contamination 45–8
  DNA extraction 30
  DNA quantification 32, 35–6
  DNA replication 40
  historical material 3–4, 39–40
  hybridization-based methods 35
  inhibition 44–5
  laboratory set-up 46–8
  lineage markers 126
  low copy number 44, 68–70, 73, 101–2
  post-PCR 47–8
  pre-PCR 46–7
  primers 41–2
  procedure 42–4
  reagents and buffers 42
  real-time 35–6
  sensitivity 45–6
  Taq polymerase 40–1, 42–5
  template DNA 40
  see also short tandem repeats
polymorphic information content 138
population genetics *see* Hardy–Weinberg equilibrium
positive paternity statements 109
post-PCR 47–8
power of discrimination 137
power of exclusion 137–8
PowerPlex® 16 system
  kinship testing 105, 109, 120
  short tandem repeats 52, 61

PowerPlex® Y system 131
pre-PCR 46–7
presentation of evidence *see* evaluation process
presumptive testing 21–3
primers 41–2, 53
  extension 36, 118–19
prior odds 92, 110–11
probe hybridization 3
profile
  ceiling principle 82, 83
  frequencies 79, 88, 90
Profiler Plus® 72, 106
prosecutor's fallacy 91, 93–5
prostate specific antigen (PSA) 22
protein markers 2
proteinase K 30, 31
PSA *see* prostate specific antigen
pull-up 67–8
Punnett squares 75–6, 106–7

Quadraplex (QUAD) 52
Quantiblot® 35

random match probabilities 88–9
random mating 77
re-offenders 97
real-time polymerase chain reaction (RT-PCR) 35–6
reference samples 1–2, 23
regulatory sequence 9–10
repetitive DNA 10
restriction
  digestion 116
  enzymes 2–3
Romanov royal family 128
RT-PCR *see* real-time polymerase chain reaction

saliva 17–19, 22
salting-out procedure 28–9, 31, 36
samples *see* biological samples
sampling bias 80, 81, 83
Sanger sequencing 2, 116–17, 118, 126
seminal fluid
  biological samples 3, 17–19, 21–2
  DNA extraction 30
  DNA quantification 32–3
sex chromosomes *see* X chromosome; Y chromosome
SGM 52, 99
SGM Plus®
  evaluation process 88–9
  national DNA databases 99, 101
  short tandem repeats 52, 59, 61, 70
short interspersed elements (SINEs) 10
short tandem repeats (STRs) 4, 51–85
  allele frequency databases 78–83
  allelic ladders 58–61
  corrections 78–83
  degraded DNA 71–3
  detection of polymorphisms 54–5
  DNA structure 13, 14
  drop-out 68, 69, 73
  evaluation process 92
  Hardy–Weinberg equilibrium 75–83
  internal-lane size standards 56–8
  interpretation 56–61, 65–74
  kinship testing 111
  lineage markers 130–2
  low copy number DNA 68–70, 73
  mixed profiles 70–1
  multiplex development 51–3
  national DNA databases 99
  overloaded profiles 68
  peak balance 69, 70
  population genetics 75–83
  primers 53
  pull-up 67–8
  single nucleotide polymorphisms 120–1
  split peaks 65–7
  standardization 52
  structure 51
  stutter peaks 65, 69
  template DNA 68
silica based DNA extraction 28–9, 31, 36
silver staining 54
SINEs *see* short interspersed elements
single nucleotide polymorphisms (SNPs) 115–23
  allele specific hybridization 119
  detection 115–19
  DNA structure 13–14
  forensic applications 117–20
  forensic identification 119–20
  geographical ancestry 120
  historical material 4
  kinship testing 111
  lineage markers 126, 130
  occurrence and structure 115
  primer extension 118–19
  Sanger sequencing 116–17, 118, 126
  short tandem repeats 120–1
SNaPshot™ 118
SNPs *see* single nucleotide polymorphisms
Southern blotting 2–3, 57–8
split peaks 65–7
statements
  likelihood ratios 91
  match probability 88–9

- positive paternity 109
- uniqueness 89

storage of biological samples 23–4
STRs *see* short tandem repeats
stutter peaks 65, 69
subpopulations
- allele frequency databases 80–2, 83
- kinship testing 111

SYBR® Green 36

tandem repeats 10, 12–13, 14
- *see also* short tandem repeats; variable number tandem repeats

Taq polymerase
- polymerase chain reaction 36, 40–1, 42–5
- short tandem repeats 65–7
- single nucleotide polymorphisms 117, 118–19

TaqMan® system 36
template DNA 40, 68
terminal transferase 66–7
tetranucleotide repeats 66
theta values 81–2, 83
transposed conditional 91, 93–5

ultraviolet spectrophotometry 34
uniqueness statements 89
United Kingdom NDNAD 97–102
- cold cases 101–2
- entry criteria 98–9
- familial searching 100–1
- legislation 98
- operation 99–100
- rationale 97
- technology 99

variable number tandem repeats (VTNRs) 2–3, 4, 12–13
- evaluation process 92
- lineage markers 130
- national DNA databases 99
- polymerase chain reaction 39

verbal scales 91–2
VTNRs *see* variable number tandem repeats
Yfiler® 131

Watson, J. 7

X chromosome
- amelogenin locus 53
- structure 9

Y chromosome
- amelogenin locus 53
- lineage markers 125, 130–2
- structure 9

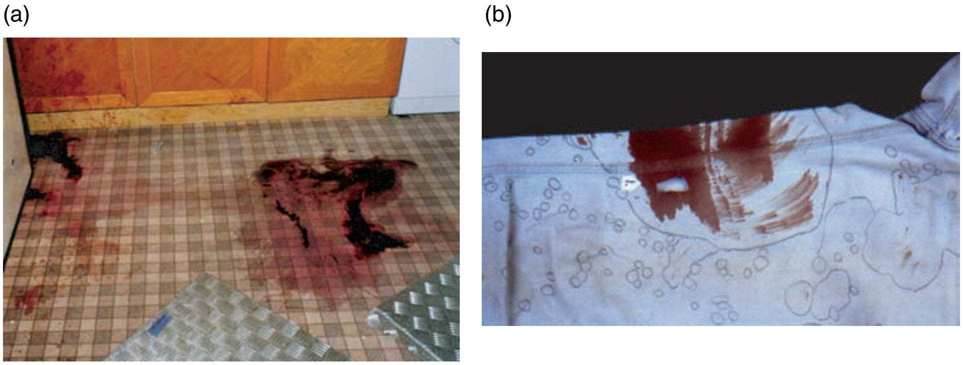


Plate 3.1 Blood is the most common form of biological material that is recovered from crime scenes. (a) Large volumes of blood can be collected using a swab, if the blood is liquid then a syringe or pipette can be used (picture provided by Allan Scott, University of Central Lancashire) (b) Blood on clothing is normally collected by swabbing, or cutting out the stain (picture provided by Elizabeth Wilson)

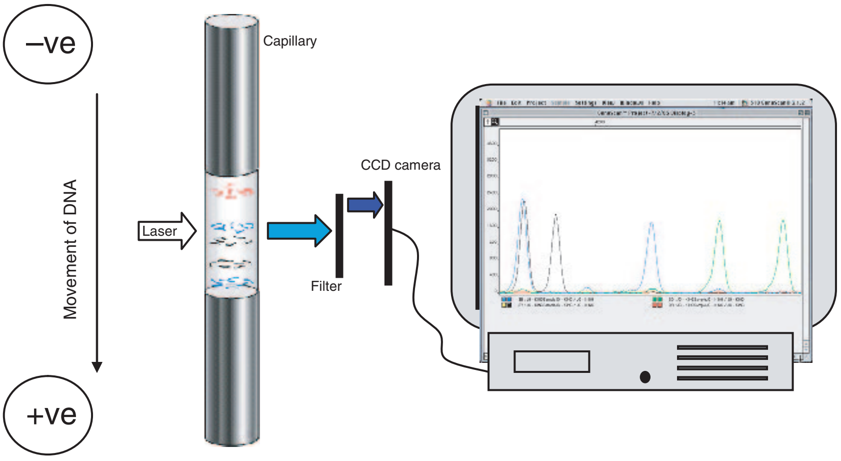


Plate 6.4 During electrophoresis an argon laser is shone through the window in the capillary. As the labelled PCR products migrate through the gel towards the anode they are separated based on their size. When the laser hits the fluorescent label on the PCR products, the lable is excited and emits fluorescent light that passes though a filter to remove any background noise, and then on to a charged coupled device camera that detects the wavelength of the light and sends the information to a computer where software records the profile

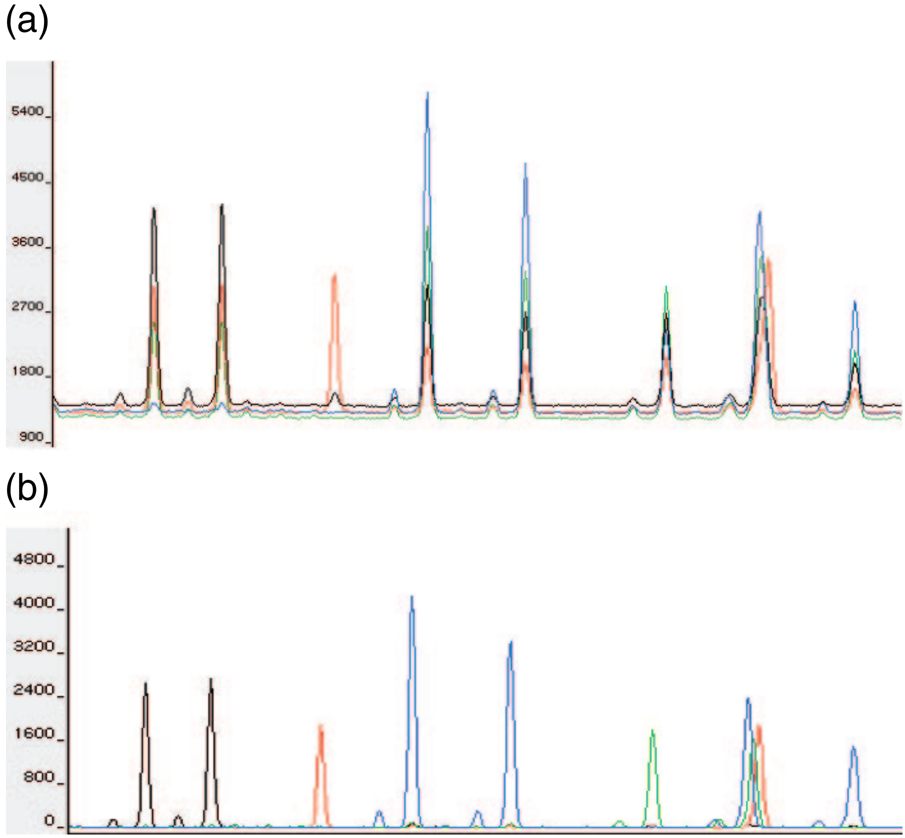


Plate 6.5 The application of a matrix file, using the GeneScan® or GeneMapper™ *ID* software removes the spectral overlap from the raw data (a) to produce peaks within the profile that are composed of only one colour (b)

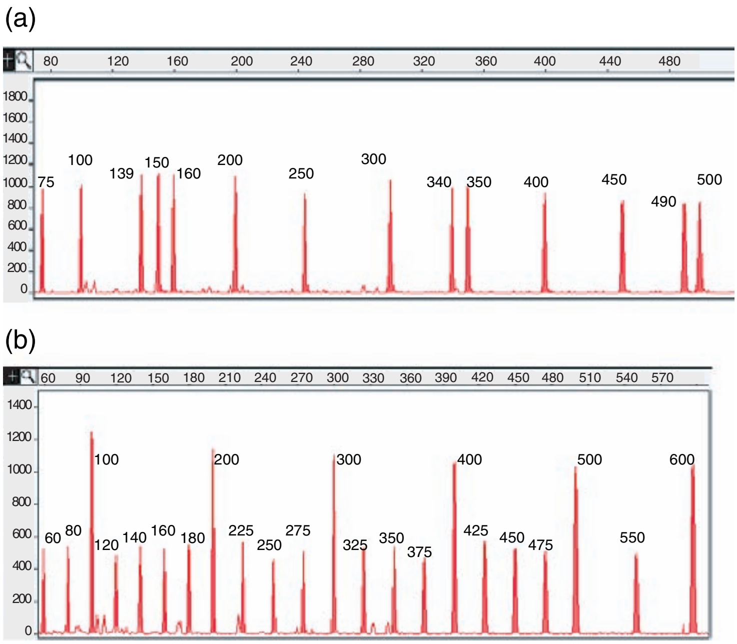


Plate 6.6 Internal-lane size standards are used to size the PCR products precisely. Two commonly used internal-lane size standards are (a) the GeneScan®-500 (Applied Biosystems) and (b) the ILS600 (Promega)

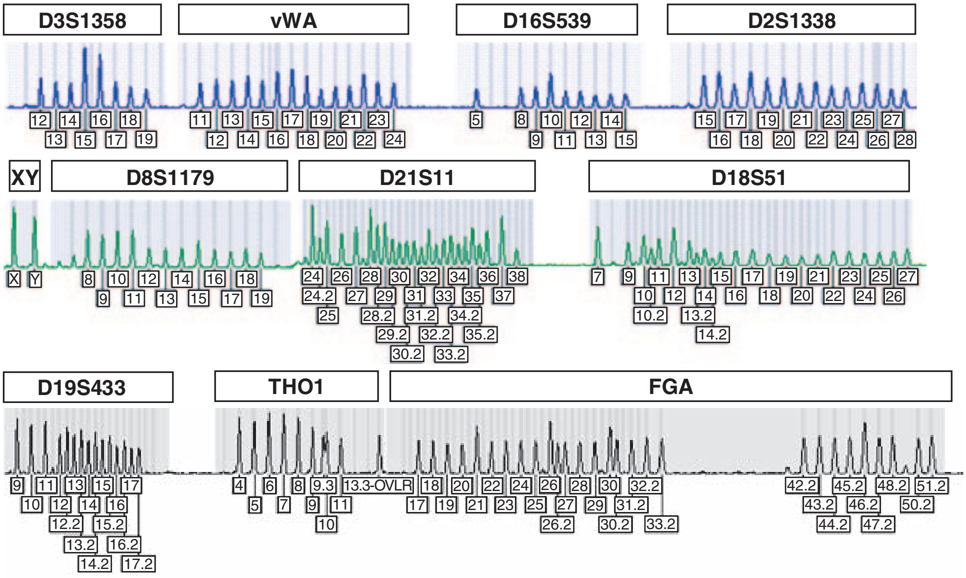


Plate 6.9 The allelic ladder of the AmpF*l*STR® SGM Plus® kit contains all the common alleles

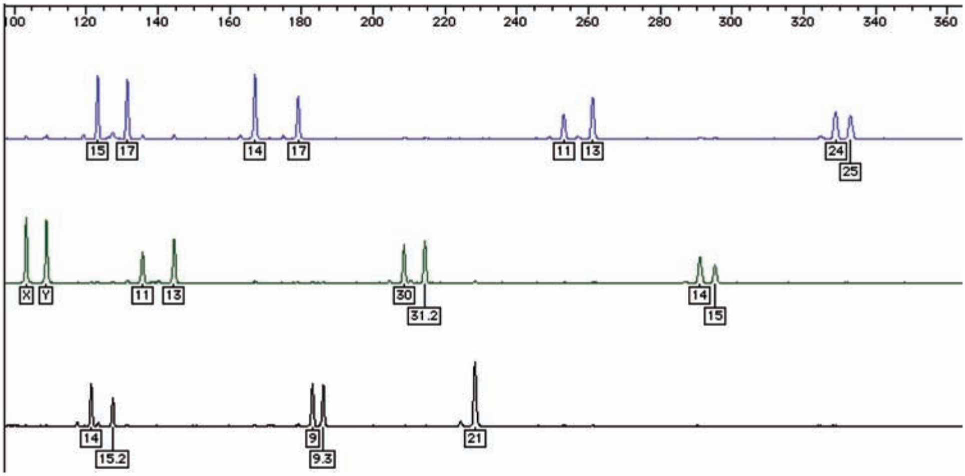
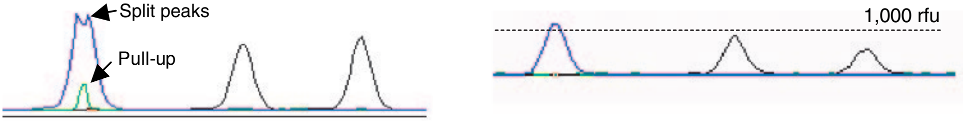
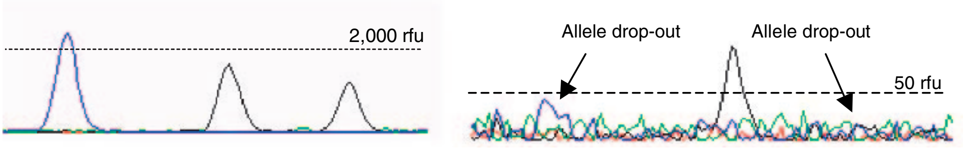


Plate 6.11 The Genotyper® software compares the peaks within a profile to the allelic ladder and assigns alleles. If the peaks in the profile deviate more than ±0.5 bp from the allelic ladder they are designate 'off ladder'


(a) 25 ng template

(c) 1 ng template


(b) 2.5 ng template

(d) 0.25 ng template

Plate 7.5 If the reaction is overloaded with DNA, (a) the peaks are still present but artefacts such as pull-ups and split peaks are more pronounced. When the template is within the optimal range (b and c) the peaks are well balanced and easy to interpret. When the PCR does not have enough template to amplify (d), then locus and allelic drop-out can occur

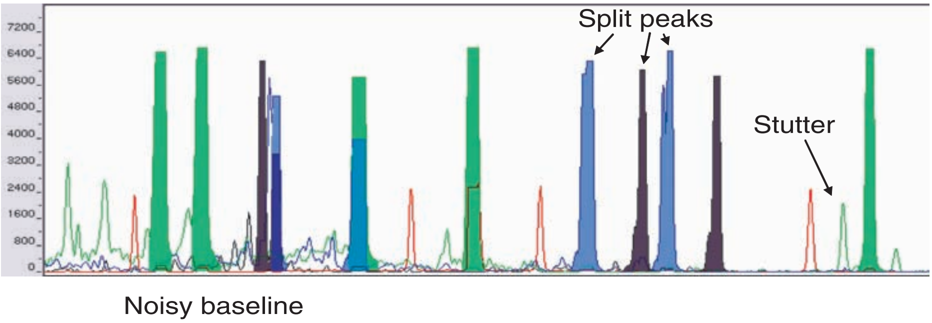


Plate 7.6 A heavily overloaded profile. All the peaks shown have a flat top indicating that they are off-scale, the baseline is very noisy, several split peaks are evident and the peaks are very broad, which can lead to sizing problems. There are also some pronounced stutter peaks

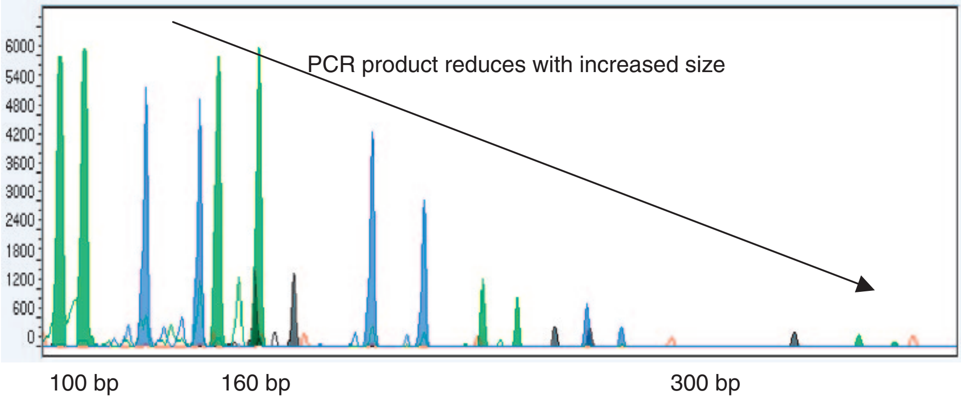


Plate 7.10 The profile was generated using the AmpF*l*STR® Profiler Plus® kit from Applied Biosystems. The DNA was extracted from a bone recovered from a Scottish loch after approximately 30 years. The profile is typical of a degraded profile with a gradual reduction in the amount of product as the amplicons increase in size

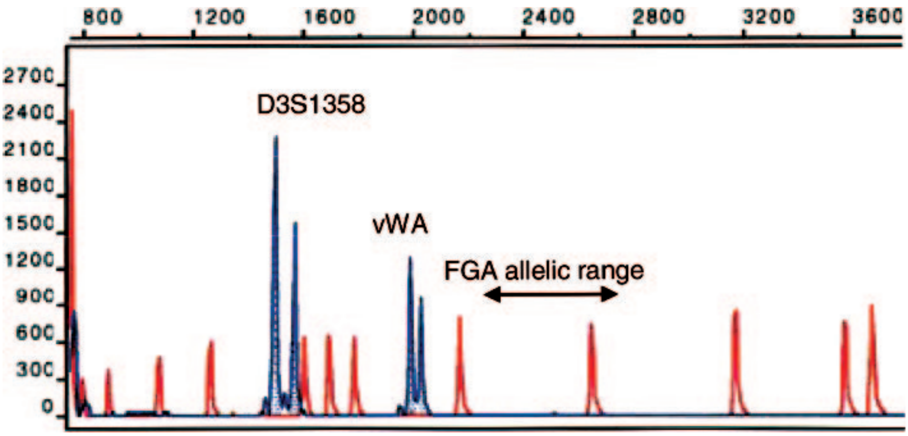


Plate 7.11 The profile was generated using the AmpF*l*STR Blue™ kit from Applied Biosystems. The DNA was extracted from muscle tissue recovered from a plane crash. The muscle had been subjected to high temperatures and the DNA was highly degraded – no amplification products were detected from the FGA locus. The size standard is also shown in by non-shaded peaks

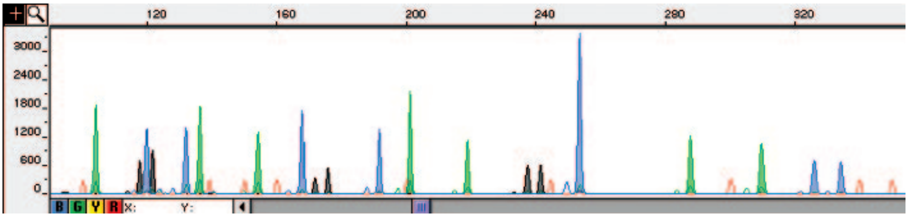


Plate 11.1 A STR profile generated from foetal cells recovered from amniotic fluid in early pregnancy. In this case it was possible to determine that pregnancy had resulted from a rape, and allowed an informed decision to be made on whether or not to have the foetus aborted. The profile was generated using the AmpF*l*STR® Profiler Plus® STR kit (Applied Biosystems)

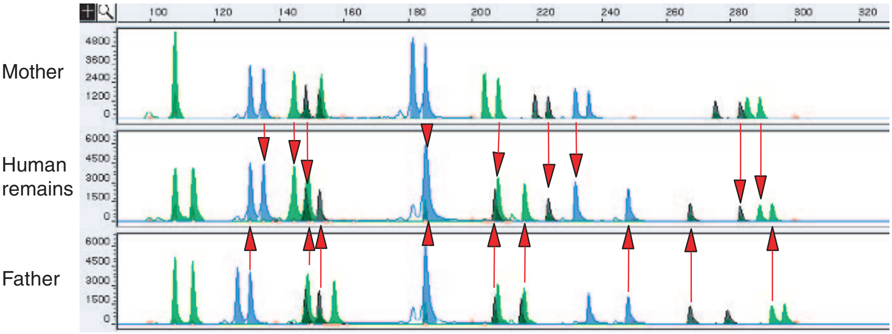


Plate 11.2 The identification of human remains recovered from an air crash [35]. Blood samples were provided by the mother and father who were missing a son. Alleles in the profile of human remains could have come from the mother and father (indicated by the arrows). The profiles were generated using the AmpF*l*STR® Profiler Plus® STR kit (Applied Biosystems)

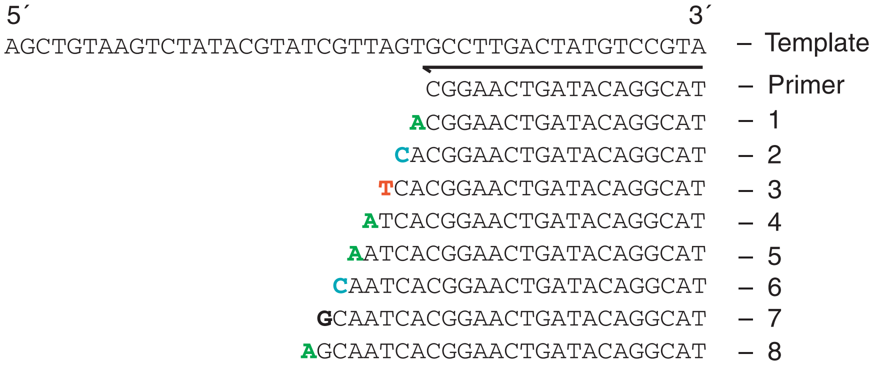

5´
3´
AGCTGTAAGTCTATACGTATCGTTAGTGCCTTGACTATGTCCGTA
– Template
CGGAACTGATACAGGCAT
– Primer
ACGGAACTGATACAGGCAT – 1
CACGGAACTGATACAGGCAT – 2
TCACGGAACTGATACAGGCAT – 3
ATCACGGAACTGATACAGGCAT – 4
AATCACGGAACTGATACAGGCAT – 5
CAATCACGGAACTGATACAGGCAT – 6
GCAATCACGGAACTGATACAGGCAT – 7
AGCAATCACGGAACTGATACAGGCAT – 8

Plate 12.2 A primer anneals to the template strand. This is extended by *Taq* polymerase until a ddNTP is incorporated. The ddNTPs are incorporated at random, which leads to a collection of extension molecules that differ from each other by one nucleotide (shown above labelled 1 to 8). The four ddNTPs are labelled with different fluorescent dyes that are detected during capillary electrophoresis

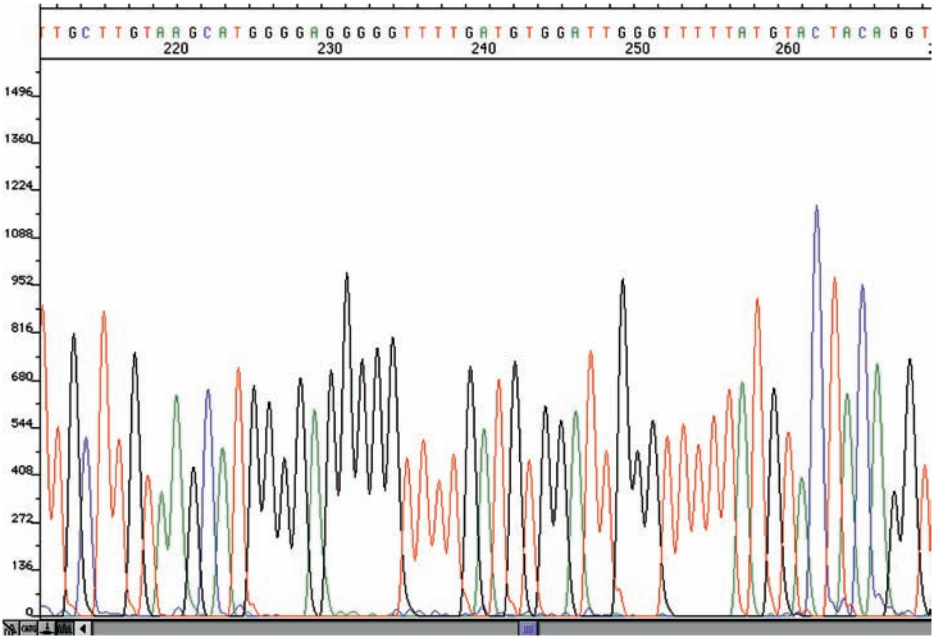


Plate 12.3 The sequence of a region of the mitochondrial genome. The sequencing software interprets the sequence data and 'calls' the bases. This information is provided above the sequencing peaks

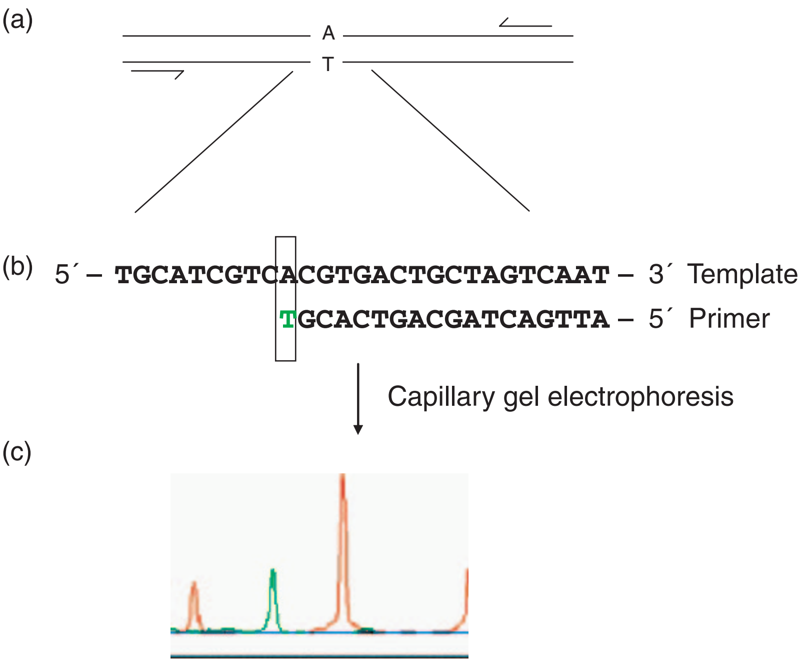

(a)
A
T
(b)
5´ – TGCATCGTCACGTGACTGCTAGTCAAT – 3´ Template
TGCACTGACGATCAGTTA – 5´ Primer
Capillary gel electrophoresis
(c)

Plate 12.4 The primer extension assay. (a) The target sequence is amplified using PCR and the products are used as the template in the extension assay; (b) an internal primer hybridizes to the target adjacent to the SNP and a single fluorescently labelled ddNTP is added by *Taq* polymerase; (c) the reaction is analysed by capillary electrophoresis

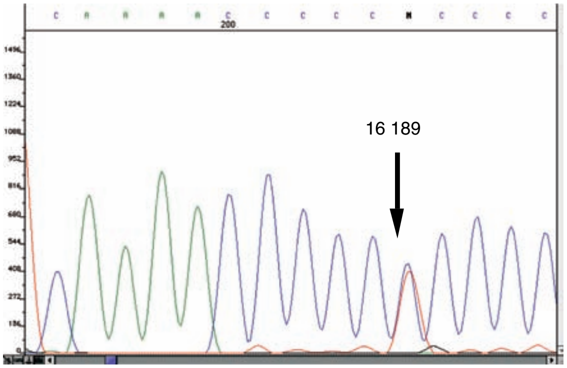


Plate 13.3 The above sequence shows the presence of heteroplasmy at position 16 189. Two bases, a C and an A, are present in approximately equal amounts