UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 263

Analysis of Nucleotide Variations in Non-human Primates

ANN-CHARLOTTE RÖNN

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2007

ISSN 1651-6206
ISBN 978-91-554-6904-7
urn:nbn:se:uu:diva-7904

Dissertation presented at Uppsala University to be publicly examined in Rudbecksalen, Rudbecklaboratoriet, Dag Hammarskjölds väg 20, Uppsala, Friday, June 1, 2007 at 13:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English.

Abstract

Rönn, A-C. 2007. Analysis of Nucleotide Variations in Non-human Primates. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine* 263. 41 pp. Uppsala. ISBN 978-91-554-6904-7.

Many of our closest relatives, the primates, are endangered and could be extinct in a near future. To increase the knowledge of non-human primate genomes, and at the same time acquire information on our own genomic evolution, studies using high-throughput technologies are applied, which raises the demand for large amounts of high quality DNA.

In study I and II, we evaluated the multiple displacement amplification (MDA) technique, a whole genome amplification method, on a wide range of DNA sources, such as blood, hair and semen, by comparing MDA products to genomic DNA as templates for several commonly used genotyping methods. In general, the genotyping success rate from the MDA products was in concordance with the genomic DNA. The quality of sequences of the mitochondrial control region obtained from MDA products from blood and non-invasively collected semen samples was maintained. However, the readable sequence length was shorter for MDA products.

Few studies have focused on the genetic variation in the nuclear genes of non-human primates. In study III, we discovered 23 new single nucleotide polymorphisms (SNPs) in the Y-chromosome of the chimpanzee. We designed a tag-microarray minisequencing assay for genotyping the SNPs together with 19 SNPs from the literature and 45 SNPs in the mitochondrial DNA. Using the microarray, we were able to analyze the population structure of wild-living chimpanzees.

In study IV, we established 111 diagnostic nucleotide positions for primate genera determination. We used sequence alignments of the nuclear epsilon globin gene and apolipoprotein B gene to identify positions for determination on the infraorder and Catarrhini subfamily level, respectively, and sequence alignments of the mitochondrial 12S rRNA (MT-RNR1) to identify positions to distinguish between genera. We designed a microarray assay for immobilized minisequencing primers for genotyping these positions to aid in the forensic determination of an unknown sample.

Keywords: SNP, genotyping, primate, chimpanzee, whole genome amplification, multiple displacement amplification, minisequencing, microarray, Y-chromosome, mitochondria

Ann-Charlotte Rönn, Department of Medical Sciences, Molecular Medicine, Akademiska sjukhuset, Uppsala University, SE-75185 Uppsala, Sweden

ISSN 1651-6206
ISBN 978-91-554-6904-7
urn:nbn:se:uu:diva-7904 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-7904)

List of publications

This thesis is based on the following paper, which will be referred to in the text by roman numerals:

I Rönn A-C, Andrés O, Bruford MW, Crouau-Roy B, Doxiadis G, Domingo-Roura X, Roeder AD, Verschoor E, Zischler H, Syvänen A-C. Multiple displacement amplification for generating an unlimited source of DNA for genotyping nonhuman primate species. International Journal of Primatology 27:1145-1169 (2006)

II Andrés O, Rönn A-C, Ferrando, Bosch M, Domingo-Roura X. Sequence quality is maintained after multiple displacement amplification of non-invasively obtained macaque semen DNA. Biotechnology Journal 1:466-469 (2006)

III Andrés O, Rönn A-C, Bonhomme M, Kellermann T, Crouau-Roy B, Doxiadis G, Verschoor EJ, Goossens B, Domingo-Roura X, Bruford MW, Bosch M, Syvänen A-C. A microarray system for Y-chromosomal and mitochondrial SNP analysis in chimpanzee populations. Submitted.

IV Rönn A-C, Andrés O, López-Giráldez F, Johnsson-Glans C, Verschoor EJ, Domingo-Roura X, Bruford MW, Bosch M, Syvänen A-C A microarray-system for forensic identification of primate species subject to bushmeat trade. Manuscript.

Contents

Abbreviations

BAC	Bacterial artificial chromosome
bp	Base pairs
ddNTP	dideoxyribonucleotide triphosphate
DGGE	Denaturing gradient gel electrophoresis
DHPLC	Denaturing high performance liquid chroma-toghraphy
DNA	Deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
DOP-PCR	Degenerate oligonucleotide primer-PCR
Gb	Giga bases
Indel	Insertion/deletion
LINE	Long interspersed element
Mb	Mega bases
MDA	Multiple displacement amplification
MSY	Male specific part
mtDNA	Mitochondrial DNA
PAR	Pseudoautosomal region
PCR	Polymerase chain reaction
PEP	primer-extension preamplification
RNA	Ribonucleic acid
SINE	Shiort interspersed element
SNP	Single nucleotide polymorphism
STR	Short tandem repeat

Introduction

The idea of evolution was not new when Charles Darwin and Alfred Russel Wallace presented their papers on the subject in the 1850's. Evidence that life on earth was old, changing and that many species had become extinct was known by geologists and palaeontologists. However, in his work "On the origin of species" Darwin proposed a mechanism behind the theory of evolution, natural selection, provided support in the form of collected data from his voyages, and became famous for his work. Darwin recognised that all living things share an ancestry and that evolutionary processes over millions of years have led to the vast number of species on this earth.

In the late 19th century, Gregor Mendel studied distinct traits in pea plants and the laws by which they were inherited in very predictable manners. The importance of his research was not recognised at first by the scientific community, but was rediscovered in the 20th century and provided a mechanism for variation on the level of genes, and genetics was born. The theories of Darwin and Mendel have become fundamental for the modern biology research.

The DNA molecule

The genetic information of an organism is contained in the chemical compound deoxyribonucleic acid (DNA), a molecule that consists of four different bases; adenine (A), cytosine (C), guanine (G) and thymine (T), linked by a phosphate-deoxyribose backbone. Long polymers, or strands, of DNA are base-paired to a complementary DNA strand by hydrogen bonds of the pairing bases, which are always A to T and C to G. Together the two strands form a DNA double helix structure. Because of an asymmetry in the bonding of the backbone, the DNA strand is said to have direction, which is denoted 5' or 3', and the two strands in the double helix are in opposite directions to each other, and thus are complementary as well as reversed in direction to each other.

The DNA double helix structure was discovered by James D. Watson and Francis Crick in 1953 (Watson, Crick, 1953), famously announced over a pint in the local pub on what must have been one of the most creative afternoons in history since Archimedes decided to take a bath. They were influenced by the x-ray diffraction images taken by Rosalind Franklin and they

knew how the bases were paired in the molecule. Watson and Crick, together with Maurice Wilkins, received the Nobel Prize for medicine in 1962 for the discovery of the molecular structure of nucleic acids and its significance for information transfer in living material.

Primates

About 350 different primate species are currently recognised. This number has oscillated because of new discoveries, extinctions and different species categorisations made by different scientists. A common taxonomic categorisation is to divide the primates into four infraorders. In decreasing phylogenetic relatedness to humans are the Catharrhini, including Hominoids and Old world monkeys, the Platyrrhini, including New World monkeys, the Tarsiiformes, represented by one genus living on Southeast Asian islands, and the Strepsirrhini, including Lemuriformes, Loriformes and Chiromyiformes. The primate species from these groups are further characterised into families and genera, often also divided into super- and sub-categories (Groves, 2001).

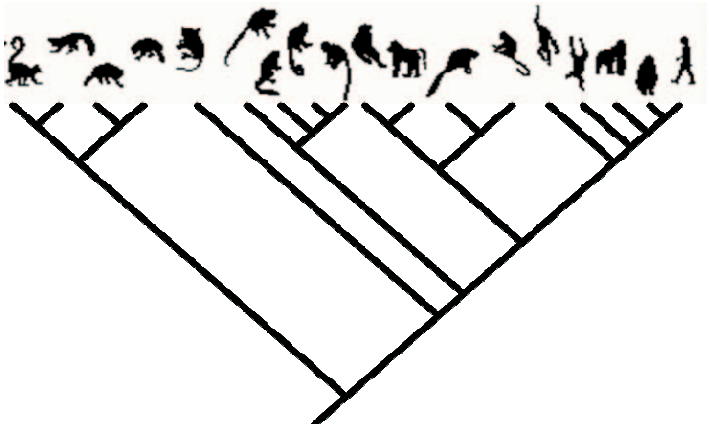

Figure 1. The primate phylogeny, counting the branches from left to right; Strepsirrhini, including lemurs, loriformes and galagos; the Tarsier; New World monkeys, including tamarins, marmosets, squirrel monkeys, capuchins, howler monkeys and spider monkeys; Old World monkeys, including vervet monkeys, macaques, baboons, colobus and langurs; gibbons; orang-utans, gorillas, chimpanzees and bonobos (on the same branch), and humans.

Currently, according to the 2006 Red List of endangered species (www.iucnredlist.com) from the World Conservation Union (IUCN), 68 of the primate species are considered endangered or critically endangered, and

this includes all of the great ape species. Deforestation and habitat loss through commercial logging has been the main cause of the African ape population decline historically. However, the illegal bushmeat trade has surpassed the habitat loss as the primary threat to great apes, facilitated by the infrastructure of new roads and transports linking forests and hunters to consumers. Effects of the bushmeat trade are evident in areas of forest that are suitable for habitat but contain very low densities of animals (Walsh *et al.*, 2003).

The bushmeat trade also increases the human exposure to cross-species infections (zoonosis). As an example, the original reservoirs for the human immunodeficiency virus type I (HIV-I) have been traced to the chimpanzee, commonly hunted for food in areas of west equatorial Africa (Gao *et al.*, 1999). Apart from being a serious threat to the great ape populations, the Ebola virus is also transmitted by ill or dead great apes to humans causing fatal epidemics in hemorrhaging fevers (Walsh *et al.*, 2003) (www.who.int). The risk of zoonosis is present in all contact with animals; however, there are an increased number of epidemics emerging from pathogens not previously posing a threat to humans. The recent epidemic of the severe acute respiratory syndrome (SARS) exemplifies how serious such an outbreak of a fast evolving infectious disease might become. While there is a need for more research on the mechanisms behind the increased number of emerging infectious diseases, efforts to suppress these pathogens from spreading in human populations must be made (Maillard, Gonzalez, 2006).

Primate genomes

Genes and genomes are in constant change by evolution, giving rise to variation within as well as between species. A central aim of evolutionary and biomedical genetics is to correlate this molecular variation with phenotypic changes. To trace the history of a genome and pinpoint important molecular changes in a certain evolutionary lineage, it is necessary to obtain sequence information of orthologous genes and long, contiguous DNA segments stemming from organisms exhibiting various degrees of phylogenetic relatedness.

Encouraged by, and benefiting from, the success of the Human Genome Project in 2001 (Lander *et al.*, 2001; Venter *et al.*, 2001), the first draft of the complete sequence of the chimpanzee genome soon followed, and in 2005 an extensive analysis of the draft sequenced was published (Mikkelsen *et al.*, 2005). In 2007, the rhesus macaque was the second of the non-human primate genomes to be sequenced, analysed and compared to the previously finished primate genomes (Gibbs *et al.*, 2007). A low coverage assembly of the bushbaby has been made publicly available as part of the Mammalian Genome Project initiative, which is also sequencing a mouse lemur and a

tarsier. Recently, the National Human Genome Research Institute announced that the orang-utan and the marmoset genomes have been assembled and comparisons with the other primate genomes are being performed.

The human genome consists of 22 autosomal chromosome pairs and the sex chromosomes, X and Y. However, through chromosomal rearrangement over time, the number of autosomal chromosomes varies between primate species. Ten of the autosomal chromosomes in humans are involved in major rearrangements compared to their chimpanzee homologs. This has been a factor in the human speciation, as these rearrangements tend to reduce recombination, which is required for gene flow in a population. Thus, a reproductive barrier was created, and over time all gene flow stopped and gave rise to the *Homo* and *Pan* lineages (Navarro, Barton, 2003).

The Y-chromosome

The Y-chromosome is male-specific in mammals. Switching-on of the SRY gene on this chromosome during embryonic development ultimately "makes a male" (Sinclair *et al.*, 1990). Over 95 % of the human Y-chromosome is male-specific and does not recombine with the, partly, homologous X-chromosome, and thus, it is suitable for studies of male lineages. The euchromatic part consists of a male-specific part (MSY) and pseudoautosomal regions (PARs) at the telomeric ends that are able to recombine with the X-chromosome during meiosis. The genes on the human Y-chromosome code for only about 60 proteins and a high amount of repetitive sequence and palindromic structures have made it a difficult task to sequence and assemble the chromosome (Kirsch *et al.*, 2005; Ross *et al.*, 2005; Skaletsky *et al.*, 2003). The Y-chromosome also consists of large amounts of highly repetitive non-coding heterochromatic sequence, mostly dismissed as non-functional, and largely un-sequenced to date.

Euchromatic sequence consists of palindromic self-recombining ampliconic sequences or non-recombining X-degenerate sequence. Comparisons of the human and chimpanzee Y-chromosomes have revealed a rapid evolution of this chromosome towards gene decay for the chimpanzee chromosome. Large inversions exist between the two species, and the heterochromatic and euchromatic sequence parts are arranged very differently along the chromosome arms in (Hughes *et al.*, 2005).

The mitochondrial genome

For studies on maternal lineages, the maternally inherited genome of the cell organelle mitochondria is often used. The lack of genetic recombination of its haploid mitochondrial DNA (mtDNA) makes it ideal for studies on evolution of populations. Mutation rate is high in mtDNA and has been estimated to 6-17 times higher than nuclear DNA (Sigurgardottir *et al.*, 2000).

Some heteroplasmy has been observed. Heteroplasmy is the coexistence of more than one mitochondrial genome in a cell, and occurs due to mutations in mtDNA that is transferred in mitosis. Heteroplasmy from a difference in nucleotide length is believed to be common in animal cells (e.g. (Lunt *et al.*, 1998), whereas a difference in nucleotide composition is rare (e.g (Taylor *et al.*, 2003). As there are 100-10,000 copies of mtDNA present in each cell, one can argue that the effect of hetetoplasmy can be neglected in many genetic studies for stochastic reasons.

The mtDNA in humans (Anderson *et al.*, 1981), and most animals, is circular, intronless and contains 13 protein-coding genes, 22 transfer RNAs , 2 ribosomal RNAs and the non-coding D-loop region, consisting of hypervariable reion I and II. The gene-dense mitochondrial genome is often used in fields such as forensics (Allen *et al.*, 1998) evolutionary studies (Ingman *et al.*, 2000). To date, the mtDNA of 23 non-human primates mtDNA have been sequenced.

DNA sequence variation

Repetitive elements

Only about 1.5% of the 3 billion bp long human genome is protein-coding. At least half of the genome is made up of repetitive sequences, often viewed as “junk” (Lander *et al.*, 2001). Repetitive elements can, however, be used as markers for studies on evolutionary processes, such as selection or mutation. Most of the repeats are transposable elements, such as long interspersed elements (LINEs) and short interspersed elements (SINEs), whose transcribed RNA can be inserted into a new site of the genome after utilising the reverse transcriptase encoded by the LINEs (Dewannieux *et al.*, 2003). One class of SINEs are the *Alu*-elements, which are non-functioning copies of a gene coding for a cytoplasmic RNA involved in protein secretion (Ullu, Tschudi, 1984). The *Alu*-element is exclusively present in primates and can be utilised to trace the evolutionary history of the primates as families of this repeated element has expanded in the primate genomes. About 1 million copies of this 300 bp fragment are present in the human genome, or 10 % of the total genome (Lander *et al.*, 2001).

Another important repeat in genetic studies are short tandem repeats (STRs, or ”microsatellites”), traditionally used as markers in population studies. STRs are tandem repeats of 2-6 bp units, with a total fragment size of 100-500 bp. They are multi-allelic in populations, a feature that makes them very useful as a lot of information can be retrieved from only a few STRs (Edwards *et al.*, 1991). Assays of STR analyses can be implemented in closely related species, provided the flanking sequences of the STRs are not too divergent. A technical difficulty in using STRs is the occurrence of alle-

lic dropouts, or “null-alleles”, causing heterozygotes to be falsely determined as homozygotes in the analysis.

Single nucleotide variation

Single nucleotide polymorphisms (SNPs) are point mutations that have reached a minor allele frequency above 1% in a population, per definition. SNPs are the most common polymorphism in the human genome, found on average every 1200 bp when comparing any two chromosomes. Thus, any two human individuals will be about 99.92% identical at the nucleotide level (Sachidanandam *et al.*, 2001). SNPs are widely used as markers in genetic fields, for example in phamacogenetics, population genetics, forensics and for mapping disease genes. In most monogenetic diseases, rare nucleotide point mutations are the main cause of disease. However, complex diseases involving several genes are more common and in those cases, even SNPs with high minor allele frequencies can serve as markers for susceptibility for disease in association studies (Risch, Merikangas, 1996). As SNPs are bi-allelic they are stable in an evolutionary perspective and do not mutate as often as multi-allelic markers.

In the comparison of the chimpanzee genome draft sequence to the human genome, the single nucleotide divergence was 1.23%, of which 1.06% was fixed in either genome, and differences due to insertions/deletions (indels) was about 3%, of which about 45% involved a single nucleotide (Mikkelsen *et al.*, 2005). In a study based on comparison of the human chromosome 21 and the high-quality BAC sequence of the homologue chimpanzee 22 chromosome the single nucleotide divergence was 1.52% and the indel divergence was 5.07%, of which 39% involved a single nucleotide. When considering only non-repetitive DNA the total sequence divergence was 2.37%. Indels occurred frequently in coding sequences and altered the gene product in 5% of the cases (Wetterbom *et al.*, 2006). Thus, indels must be considered an important part of primate evolution.

The extensive analysis of the chimpanzee genome draft sequence also produced SNP information within and between the two sub-species *Pan troglodytes verus* and *Pan troglodytes troglodytes*. A total of 1.66 million SNPs were identified, of which 1.01 million were heterozygote positions in the *P. t. verus*. The diversity in *P. t. verus* was similar to human populations, and the diversity of *P. t. troglodytes* was about twice as high (Mikkelsen *et al.*, 2005).

Discovery and genotyping of single nucleotide polymorphisms

PCR

Few techniques have revolutionised genetic research as much as the polymerase chain reaction (PCR; Mullis, Faloona, 1987; Saiki *et al.*, 1985). PCR has been used together with essentially all genotyping methods to date, due to its exponential amplification of a region of interest to the researcher. This reduces the complexity of the genome being studied and provides a high amount of DNA of the fragment that is amplified.

In this method, a primer pair of DNA oligonucleotides, each complementary to the DNA strands of the fragment, is designed to anneal to the fragment, when the fragment is denatured to single strands, with their 3' ends directed towards each other. A DNA polymerase extends the primers with deoxynucleotides (dNTPs) and makes a complementary copy of the strand. In the second round of denaturing, annealing and extending, the primers that anneal will reach the end of the fragment of interest, as it was defined by the primer on the opposite strand. The reaction then proceeds exponentially; amplifying the fragment until the primers, dNTPs or the DNA polymerase is exhausted. An important improvement to the technique was the use of a thermostable *Taq* polymerase, which facilitated the cyclic reaction (Saiki *et al.*, 1988).

Apart from being an essential method used together with downstream techniques, methods such as assays for STR marker analysis are based on this principle.

Discovery of single nucleotide polymorphisms

Sequencing

The "Sanger sequencing" method (Murray, 1989; Sanger *et al.*, 1977), has been considered the golden standard for DNA sequence analysis and retrieving new knowledge of genomes and their variation. In this method, a well-balanced addition of fluorescently labelled dideoxynucleotides

(ddNTPs) terminates the elongation of a primer annealed to a DNA template (PCR product) randomly in a cyclic reaction, similar in components to the PCR. The reaction creates sequences of different length in a linear growth rate. The different lengths represent the whole template at the end of the reaction, with a fluorescently labelled ddNTP as the 3'-end that can be detected electrophoretically on a polyacrylamid gel using laser detection. An important step towards a high throughput technique was the use of capillaries for the polyacrylamid matrix. This simplified both hands-on and data handling, as well as saving time being more cost-effective (Drossman *et al.*, 1990; Luckey *et al.*, 1990).

To discover new SNPs, or other polymorphisms, the re-sequenced fragments are aligned with fragments from other individuals. There are several software options for this, for example Sequencher™ (www.gencodes.com), where the SNP is ultimately discovered after visual inspection. Whereas the software automatically detects the presence of sequences of both DNA polarities and assembles them together, it cannot facilitate the detection of indels in autosomal chromosomes. If a site contains an indel polymorphism that is heterozygous in an individual, the sequences from the two chromosomes will overlap and the chromatogram of the sequence peaks will appear "double". Other software, such as Mutation Surveyor™ (www.softgenetics.com), will analyse the SNPs automatically, and will also detect and analyse indels by comparing the "doubled" sequence from a heterozygote individual to a reference sequence from a homozygote. As a rule, sequence data should be inspected visually when discovering new SNPs; however, the use of software to conduct a first "look" can prove timesaving.

New and even more rapid and cost-effective techniques for sequencing have been developed. One of them is the 454 Sequencing™ technology (Margulies *et al.*, 2005), developed from the Pyrosequencing technique (Fakhrai-Rad *et al.*, 2002; Ronaghi *et al.*, 1996) by 454 Life Sciences™. Pyrosequencing is an enzyme mediated sequencing-by-synthesis method that utilises the release of pyrophosphate upon dNTP incorporation, and a subsequent conversion of pyrophosphate to detectable luminescence. In 454 Sequencing, adaptor-ligated DNA fragments are captured on beads in an emulsion where the PCR is performed. The adaptors are used for both PCR and sequencing reactions, performed in a PicoTiterPlate format where the beads are placed. Reagents are flowed over the plate by a fluidic system and once a dNTP is incorporated, the luminescence is detected optically by a CCD camera. In one run of 4.5 hours 20 Mb of about 100 bp long sequences are produced, which then is assembled. The Max Planck Institute for Evolutionary Anthropology in Leipzig announced a collaboration with 454 Life Science in July 2006 to sequence the Neanderthal genome. The Neanderthal, the closest relatives to humans, was extinct 30000 years ago and therefore its sequencing proposes a challenge that would not have been thought possible a few years ago.

Another sequence-by-synthesis system is the Solexa Sequencing Technology (www.illumina.com). The 1G Genome Analysis System produces 1 Gb sequence in one flow cell. In this method, randomly fragmented genomic DNA is ligated to adaptors and hybridised to a flow cell. An abundance of adaptors are bound in the flow cell, causing the fragments of DNA to form a bridge when the adaptors anneal to each other. The fragments are amplified and denatured, until dense clusters of sequence are present on the surface of the flow cell. The fragments are sequenced one base at a time, using fluorescently labelled reversible terminators, from which the block and fluorescence is removed each round after the fluorescence have been detected.

DHPLC

Denaturing high performance liquid chromatography (DHPLC) is a method for detecting sequence variation through the mismatch of a "wild type" sequence to a "mutant" allele-containing sequence. By pooling several PCR amplicons, denaturing and re-annealing the fragments, some newly formed double stranded duplexes will not anneal to its complementary strand if there is at least one chromosome in the pool that contains a "mutant" allele. These formations of heteroduplexes will have less retention through the stationary phase of the column in liquid chromatography due to its conformation. The eluation times can then be monitored by UV-absorbance detection (Oefner, Underhill, 1998).

DHPLC is a fast and cost-effective screening for SNPs over relatively large physical distances. In pools of 20 individuals, one mutant allele could be detected, corresponding to a minor allele frequency of 5% (Wolford *et al.*, 2000). However, after the initial detection by DHPLC, the SNP has to be characterised by sequencing and the position determined.

Multiplexed SNP genotyping

One of the great advantages of genotyping SNPs over other genetic variable elements, such as STRs, is the ability to perform highly multiplexed reactions, and thus acquire a large amount of data per experiment. Thereby the total time, cost and consumption of valuable DNA are lowered. It has been estimated that 30-50 independently inherited autosomal SNPs could achieve the same power to discriminate between to human individuals as 13 STR loci, depending on allele frequencies (Chakraborty *et al.*, 1999). Assays for genotyping SNPs are also more robust and are less likely to suffer from allelic dropouts and false alleles that are often seen when genotyping multiallelic markers, such as STRs.

Allele specific hybridisation

In hybridisation using allele specific oligonucleotides (ASO), allelic discrimination of SNPs is made as oligonucleotides (or probes) form duplexes with the target sequence that contains the SNP (Wallace *et al.*, 1979). A perfect match will hybridise stronger than a mismatch sequence, under stringent conditions. The principle is simple, since it involves no enzymatic reaction, but cannot readily be multiplexed as the hybridisation conditions are sequence dependant (Southern *et al.*, 1999). Real-time PCR and TaqMan (Livak *et al.*, 1995) are based on this principle and have been multiplexed at a low degree with differently labelled probes. Affymetrix GeneChip® (www.affymetrix.com) has oligonuclotides synthesised directly on the surface in systems with a dense amount of probes. Approximately 40 probes for each SNP to be genotyped, all different in sequence, are present on the array. Affymetrix currently holds many patents involving microarrays and oligonucleotides. Their latest array on the market is the 500K Array Set, which consists of two arrays, each genotyping 250,000 SNPs in one assay, replacing the 10K and 100K assays (Matsuzaki *et al.*, 2004a; Matsuzaki *et al.*, 2004b).

Ligation based assays

Genotyping assays based on ligation (OLA) achieve a high specificity due to the DNA ligase and its ability to distinguish between perfect matches and mismatches (Landegren *et al.*, 1988). Only when a probe of the matching sequence is present will the ligase fuse together the junction that is present between two adjacent probes in the assay. This is utilised in the padlock probe approach, where the closing of the junction between probes creates a circle (Nilsson *et al.*, 1994). Since the ligation is highly specific, only one of the alleles will then be amplified with a rolling circle, or a universal primer pair that enables high multiplexing, in the next step (Baner *et al.*, 2003; Baner *et al.*, 1998; Hardenbol *et al.*, 2003).

Extension based assays

The highly multiplexed Illumina GoldenGate® assay (Fan *et al.*, 2003; Oliphant *et al.*, 2002) is based on a combination of extension and ligation. Allele specific primers are annealed and extended on the template DNA, followed by ligation of the extended primers to loci specific primers. The ligated products are then amplified with universal primers.

There are several high throughput applications to single base extension ("minisequencing"). The Illumina Infinium® II has the advantage that it does not require PCR prior to genotyping. Instead, whole genome amplification is performed and the DNA is fragmented and annealed to locus specific

50-mers bound to BeadArrays. Following base extension, the products are fluorescently stained and the intensities detected (Gunderson *et al.*, 2005).

Minisequening on microarrays

The principle of minisequencing (Syvanen *et al.*, 1990) has been combined with several detection methods, including radioactive, fluorescent and mass spectrometric detection. The tag-array minisequencing method (Lindroos *et al.*, 2002; Lovmar *et al.*, 2003) comprises the following steps: Genomic DNA is amplified by multiplexed PCR. The product serves as template in a multiplexed minisequencing reaction in solution, in which minisequencing primers designed to anneal immediately adjacent to the nucleotide position to be detected are extended by a DNA polymerase with one out of four fluorescently labelled ddNTP's. The 5'-end of each minisequencing primer contains a unique 20bp tag-sequence from the Affymetrix GeneChip® Tag collection (Affymetrix, Santa Clara, CA, USA) (Fan *et al.*, 2000), and the extended minisequencing primers are captured by hybridization to complementary oligonucleotides that are immobilized on a microarray (glass-slide) in an "array-of-arrays" conformation (Pastinen *et al.*, 2000). This conformation allows detection of up to 196 nucleotide positions in 80 samples simultaneously on each slide (Figure 2).

Instead of performing the minisequencing reaction in solution and capture the extended primers with tags, the primers can be immobilized on the microarray. In this format, the PCR products are hybridized to the primers and the minisequencing reaction is performed on the microarray. The advantage of having immobilized primers is the decrease in primer-dimer reactions, as they are well separated on the microarray. If the primer-set in an assay contains many primers for SNPs very close to each other on the genome, the primer-to-primer interactions could give unreliable genotyping results. However, it is more laborious than the tag-array format as it involves more washing steps, which also leads to a decrease in signal intensity of the fluorophores.

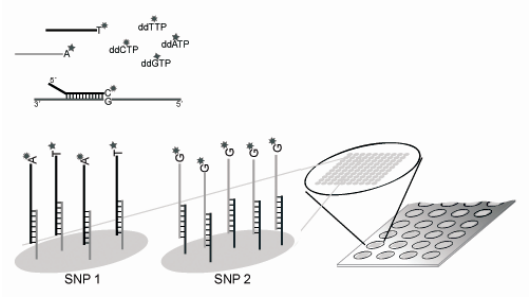

Figure 2 The principles of tag-array minisequencing. Top: The minisequencing reaction is performed in solution, where each minisequencing primer that is annealed to the PCR product is extended with one fluorescently labelled ddNTP by a polymerase. Lower half: Each extended minisequencing primer contains a unique tag-sequence that is complementary to a capture tag-oligonucleotide that has been covalently bound to a glass-microarray in an "array-of-arrays" format. The microarray contains 80 sub-arrays, one for each sample to be analysed, which contains up to 196 "spots", one for each SNP to be analysed (Figure by L. Lovmar).

Whole genome amplification

High throughput technologies increase the demand for high quality DNA. Traditionally, immortalization of the cell through lymphocyte transformation has been a valuable source of primate genetic material. However, this strategy has its limitations, as it requires fresh blood and does not work well for some species. In other types of cell transformations, genomic rearrangements occur frequently. Another approach is to apply a whole genome amplifications method, which should ideally be applicable to all types of genetic material, produce no bias in amplification of loci or alleles, and be cost effective.

PCR based whole genome amplification methods

The primer-extension preamplification (PEP) method of amplifying whole genomes is based on PCR and the *Taq* polymerase. Random 15-base oligonucleotides (femtomers) are used as primers in a cyclic reaction that is estimated to result in at least 30 copies of no less than 78% of the DNA present in a single cell of a haploid human genome (Zhang *et al.*, 1992). The method was improved (I-PEP; Dietmaier *et al.*, 1999) by optimising the protocol, including the addition of the proofreading *Pwo* polymerase.

Degenerate oligonucleotide primer-PCR (DOP-PCR) was originally developed for genome mapping studies. However, it was found to deliver an even amplification of whole genomes (Telenius *et al.*, 1992). The primers used in this method are degenerate in a central hexamer sequence flanked by defined 5' and 3' ends. The PCR is performed at low temperatures in the first cycles, to ensure amplification from multiple and evenly dispersed sites in the genome, and is then performed at higher temperature for the remainder of the cycles. Both DOP-PCR and PEP, and its improved versions, have shown uneven amplification of loci and alleles (Dean *et al.*, 2002; Dietmaier *et al.*, 1999).

The OmniPlex technology (GenomePlex™, Sigma-Aldrich) by Rubicon (www.rubicongenomics.som) is also a PCR based whole genome amplification method (Langmore, 2002). However, it differs from the traditional PCR methods as the genomic DNA is first chemically cleaved into 200-2000 bp fragments that are ligated to adaptor-sequences to create a library of frag-

ments of the genome. The fragments are then PCR amplified producing less than 0.043% locus dropout (Barker *et al.*, 2004).

Multiple displacement amplification

Multiple displacement amplification (MDA) is an isothermal method based on the annealing of random hexamers to genomic DNA, followed by strand-displacement synthesis by the phi29 DNA polymerase (Blanco *et al.*, 1989). It was originally proposed for circular DNA (Lizardi *et al.*, 1998), and was improved by modifying the hexamers (Dean *et al.*, 2001) to resist the 3'-5'proofreading activity of the phi29 DNA polymerase, which possesses a resulting error rate of $<10^{-6}$ (Blanco, Salas, 1985; Esteban *et al.*, 1993). As DNA is synthesized by strand displacement, a gradually increasing number of priming events occur, forming a network of hyper-branched DNA structures (Figure 3).

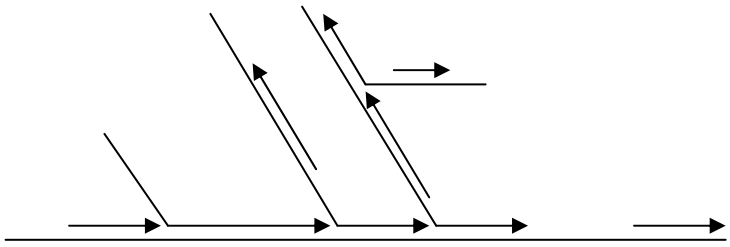

Figure 3 The hexamers anneal to the genomic DNA and are elongated by the phi29 DNA polymerase. A newly synthesised strand encounters another hexamer 5'-end and the polymerase displaces it. A new hexamer will anneal to the newly synthesised, displaced and single stranded DNA.

As the reaction is isothermal and produces high molecular weight molecules, the genomic coverage of MDA is higher than that of DOP-PCR and PEP (Dean *et al.*, 2002). The hyper-branching that produces up to 10 kb long fragments is the strength of this method. However, if the genomic DNA is highly degraded and fragmented, hyper-branching cannot continue and is disrupted by the short templates. Provided good quality DNA, small amounts of DNA input is enough for a well-balanced amplification of loci and alleles (Dean *et al.*, 2002; Hosono *et al.*, 2003; Lasken, Egholm, 2003; Lovmar *et al.*, 2003).

A wide range of applications for MDA has been described. One of the useful applications is direct amplification of cell lysates, as for example buccal swabs and whole blood (Hosono *et al.*, 2003. Successful MDA has also been performed directly on mosquito larva and adult legs {Gorrochotegui-Escalante, 2003 #64) and on residual cells left by incidental contact as fingerprints (Sorensen *et al.*, 2004).

The present study

Aims

- To investigate the feasibility of whole genome amplification using the multiple displacement amplification (MDA) technique for various non-human primate DNA samples, such as hair and semen.
- To develop a microarray system for genotyping SNPs for simultaneous analysis of both paternal and maternal lineages of chimpanzee populations.
- To establish diagnostic nucleotide positions for identifying to which primate genus an unknown sample belongs, and to implement a microarray minisequencing assay for genotyping these positions.

Material and methods

DNA samples

Non-human DNA samples mainly from the collections of the Inprimat consortium (www.inprimat.org) were used in this study. Sample origin is described in detail in the respective papers. In study I, II and IV, DNA was extracted from blood cell lines, and tissues from muscle and liver biopsies obtained during diagnostic necropsies, using standard phenol-chloroform or salting-out methods. DNA from plucked hair was extracted as in (Goossens *et al.*, 1998) and hair follicles were extracted as in (Allen *et al.*, 1998). DNA from non-invasively collected semen was extracted as in Domingo-Roura *et al.* (2004). In total, 148 DNA samples from 22 primate genera were genotyped in study I, as detailed in Table I, paper I. Sixteen non-invasively collected semen samples and 12 blood samples of Japanese macaques were sequenced in study II. In study IV 70 primate samples from 34 genera, covering all four infraorders, and 5 non-primate samples, were genotyped, as detailed in Table 2 in paper IV. In study III, DNA was extracted from 61 male unrelated chimpanzee blood, tissue or rooted hairs using the DNeasy Tissue Kit (Qiagen, Hilden, Germany), as detailed in Supplementary Table 1 in paper III. Several chimpanzee populations were also used in study III; from a zoo population of 18 individuals with a known pedigree DNA was

extracted from peripheral blood lymphocytes or immortalized B cells, and from 59 wild living individuals DNA was extracted from plucked hairs, as described in Goossens *et al.* (2002). DNA was whole genome amplified by the MDA technique in study I, II and III as described in paper I and in Lovmar *et al.* (2003).

Genotyping methods

Single and multiplexed PCRs were performed in all studies, as described in detail in the respective papers I-IV. Genotyping by tag-array minisequencing (Lindroos *et al.*, 2002; Lovmar *et al.*, 2003) was performed in study I and III, and immobilized array minisequencing (Liljedahl *et al.*, 2003) in study IV. In addition to our in-house tag-array minisequencing, several commonly used genotyping assays were performed by collaborating partners in study I. These include denaturing gradient gel electrophoresis (DGGE) analysis and sequencing of MHC class II loci (Doxiadis *et al.*, 2003; Doxiadis *et al.*, 2000; Knapp *et al.*, 1997; Otting *et al.*, 2000), analysis of 21 STR markers (Andrade *et al.*, 2004; Bradley *et al.*, 2000; Clifford *et al.*, 1999; Domingo-Roura *et al.*, 1997; Lathuilliere *et al.*, 2001; Nurnberg *et al.*, 1998; Roeder *et al.*, 2006; Smith *et al.*, 2000), sequencing of the mitochondrial control region (Marmi *et al.*, 2004; Vigilant *et al.*, 1989) and 12S ribosomal RNA gene (Kocher *et al.*, 1989), analysis of an *Alu* repeat (Salem *et al.*, 2003). Sequence analysis of the mitochondrial control region was performed in study II (Marmi *et al.*, 2004; Vigilant *et al.*, 1989). Denaturing high performance liquid chromatography (DHPLC) and sequencing was performed in study III (Oefner, Underhill, 1998).

Primer design

The PCR primers used for the minisequencing assay in study I were adapted from Jiang *et al.*(1998). Universal primer pairs for exonic regions of the genes encoding the brain-derived neurotrophic factor (BDNF), estrogen receptor 1 (ESR1), nerve growth factor B (NGFB), tumor necrosis factor-alfa (TNF-alfa), and prodynorphin (PDYN) were chosen. The primer sequences were redesigned using the NetPrimer software (http://www.premierbiosoft.com/netprimer) to better fit primate genomes and multiple alignments of available primate sequences from GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html) were made with MultAlin (http://prodes.toulouse.inra.fr/multalin.html). The five primer pairs were suitable for multiplexed PCR.

The 25-plexed and 17-plexed PCRs performed in study III were designed using the Autoprimer software (Beckman Coulter Inc., Fullerton, CA, USA) and based on the chimpanzee sequence assembly available at Ensembl in March 2006 (http://www.ensembl.org).

All other primers used in the minisequencing assays, including the minisequencing reaction primers, were designed using the NetPrimer software.

For the minisequencing primers used in study IV, multiple alignments made with CLUSTAL W (Thompson *et al.*, 1994) were displayed in BioEdit version 6.0.7 software (Hall, 1999). As the sequence alignments revealed non-conserved regions close to the nucleotide positions to be genotyped, 65 % of the primers were designed to contain degenerate bases at 1-3 nucleotide positions. This strategy can enhance genotyping success when one primer must fit and anneal to several non-identical sequences.

Study I and II

Background

There is an increasing demand for high-quality primate DNA from the scientific community. As many species are on the brink of extinction, samples are hard to come by and therefore precious. The MDA technique is a most promising method for whole genome amplification, generating micrograms of DNA from nanograms, displaying very low amplification bias and could be of invaluable use for preserving DNA.

In study I, the feasibility of the MDA technique on various sample materials, including blood, cell lines, tissues, hair and semen, was investigated by a comparative analysis of the performance of MDA products and genomic DNA in several genotyping assays. In each of the analyses, the genotyping performance of MDA products was compared to that of the original genomic DNA. As the different genotyping-assays were designed for different primate taxa, a microarray-assay for genotyping nucleotide positions conserved across taxa was designed for analyzing and comparing all samples used in the study .

In study II, sequencing products from mitochondrial DNA extracted from semen and blood were compared with MDA products obtained from the same samples. The number of unresolved bases and read length of sequences of the mitochondrial control region (390bp) were established for both original DNA templates and MDA products.

Results and discussion

DGGE is as an appropriate pre-screening method for multicopy genes, as usually one allele is represented by one band. For all samples analysed, DGGE from original DNA and MDA products provided consistent patterns. In cases where differences arose, only the intensity of the bands varied between MDA products and those generated from the original DNA. Further,

direct sequencing of other variable class II loci, *DPB1, DQA1*, and *DQB1* produced concordant results between original DNA and MDA products.

Three different STR marker sets with DNA samples from different sources: blood, cultured cell samples or hair were tested. Allelic dropouts, i.e. amplification failure of one allele, were more frequent in the MDA products than the original DNA, and samples extracted from hair showed more allelic dropout than samples taken from blood. As a conclusion, MDA performed on good quality DNA yielded high quality genotyping results.

Fragment size analysis of the presence/absence polymorphism of an *Alu*-repeat was equally successful, using the original DNA or the MDA products. This analysis demonstrates the stability and accuracy of MD amplification of high quality DNA.

A high sequence identity of the mitochondrial DNA, ranging from 98.3% for hair samples to 99.8% for semen samples, was achieved when comparing the common sequence read length of MDA products and original DNA.

When performing a paired t-test, no significant difference in the number of unresolved bases in the overlapping regions of obtained sequences could be observed between original DNA template and MDA products ($p=0.0872$, $n=28$). The semen samples were analysed separately and neither this t-test showed a significant difference ($p=0.228$, $n=16$).

The readable sequence length was analysed with a Wilcoxon's signed ranks test, as the data did not follow normal distribution. The test showed a significant difference between the original DNA template and the MDA products ($p<0.001$, $n=28$), indicating that it is possible to obtain longer sequence reads from original DNA. The same trend was observed for the semen samples when analysing them separately ($p=0.023$, $n=16$).

It was demonstrated that it is possible to obtain good quality sequence from non-invasively collected semen samples subjected to MDA, although the sequence read is shorter than that from original DNA.

In general, the success rate of nucleotide detection by the minisequencing assay in MDA products was similar to the original DNA. We observed a tendency towards a lower success rate for hair samples of <10ng DNA input amount, in agreement with the STR data. This was probably caused by degraded starting material. Encouragingly, samples originating from blood or tissues of <1ng DNA input amount were considerably more successful for MDA products than the original DNA. It therefore seems possible to conserve DNA samples of low concentration to a certain degree, bearing in mind that any imbalanced amplification of alleles would not be detected, because the analyzed nucleotides positions are not polymorphic. It has previously been shown that at least 3ng of DNA in the MDA reaction should be used to be certain of balanced amplification (Lovmar *et al.*, 2003) (Table 1).

Table 1. *Success rate of nucleotide detection using tag-array minisequencing. DNA input is the amount of DNA subjected to the MDA reaction.*

	DNA sample source		
DNA input	Blood or cell-line (n=111)	Hair (n=13)	Semen (n=19)
>10ng (n=44)	No. of nucleotides successfully detected		
MDA product	642 (94%)	34 (94%)	72 (100%)
original DNA	671 (98%)	35 (97%)	72 (100%)
1-10ng (n=66)			
MDA product	825 (96%)	29 (54%)	270 (100%)
original DNA	827 (96%)	35 (65%)	270 (100%)
<1ng (n=33)			
MDA product	409 (91%)	116 (81%)	
original DNA	375 (83%)	139 (97%)	

Study III

Background

Due to human activities, chimpanzee populations are diminishing and the species is now endangered. Conservation programs can benefit from analysis of genetic diversity and structure of threatened populations. Multiplexed SNP genotyping by a microarray system would offer a robust and low-cost alternative to STR marker analysis for obtaining this genetic information. By including both Y-chromosomal and mitochondrial DNA SNPs in this microarray system, both paternal and maternal lineages could be monitored simultaneously.

In study III, new SNPs in the chimpanzee Y-chromosome were discovered by DHPLC and mitochondrial DNA SNPs were identified bioinformatically from alignments of sequences found in GenBank. A tag-microarray-assay for the simultaneous genotyping of mitochondrial DNA SNPs and Y-chromosomal SNPs, including 19 SNPs from (Stone *et al.*, 2002), was designed. The microarray was tested in a zoo family material of 18 chimpanzees with known pedigrees. DNA from 59 individuals belonging to wild-living populations from a release project of illegally poached animals seized by the authorities of the Republic of Congo (Goossens *et al.*, 2002), were genotyped using the microarray. The mitochondrial D-loop was also sequenced in these wild populations and together with the genetic data from

the microarray, the population structure and demographic history was analysed.

Results and discussion

The DHPLC analysis identified 23 new SNPs in the introns of the JARID1D and PRKY genes. A tag-microarray-assay was designed for the simultaneous genotyping of 45 mitochondrial SNPs, which were discovered bioinformatically, and 42 Y-chromosomal SNPs that included 19 SNPs from the literature. Thirty-seven mitochondrial SNPs and 35 Y-chromosomal SNPs were successfully genotyped in a zoo family of 18 chimpanzees, and all genotypes followed inheritance according to the pedigrees. In the wild populations, 37 mitochondrial SNPs and 28 Y-chromosomal SNPs were successfully genotyped. An example of four scatterplotts for genotyping four of the SNPs is seen in Figure 4. The lower success rate for the Y-chromosomal SNPs than the mitochondrial SNPs in the wild samples, deriving from hair, is likely due to degradation of the DNA. This limitation could be solved by re-designing some of the longer PCR-fragments (>200 bp).

Results of the population structure and demographic history analyses were similar to the results obtained from STRs marker analyses in these chimpanzee populations (Goossens *et al.*, 2002), which were indicating a high migration rate. Both mitochondrial DNA D-loop sequences and mitochondrial gene SNPs corroborated these observations since lower differentiation for mitochondrial (female) markers than for microsatellites or Y-chromosomal DNA was detected, and females are known to be the migrating sex. However, a surprising lack of genetic structure for lineages belonging to males were found, although this is possibly due to sampling bias as only 12 males were included in the study.

The microarray had some limitations and can be improved for further use. While the lack of male data derived from it is mostly due to a lack of male samples, Y-chromosomal SNPs showed a lower success rate than mitochondrial, as discussed. The array could also be improved further by adding more SNPs to the assay, as the microarray has double the capacity than was fully used in this study.

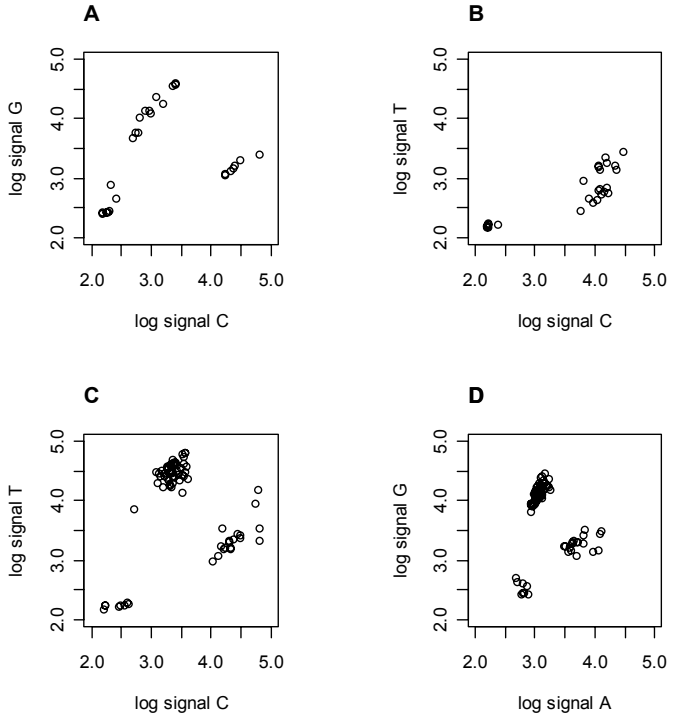

Figure 4. . Examples of scatter plots of fluorescence signals from tag-array minisequencing used for genotyping Y-chromosomal (A, B) and mitochondrial (C, D) SNPs. A) SNP sY19 (defined in Stone et al. (2002) in 19 male samples, average signal/noise ratio (s/n) 150 for the C-allele and 51 for the G-allele; B) SNP PY19.3, that is monomorphic in 19 samples with s/n 82 for the G-allele, detected in reverse polarity as the C-signal. C) SNP 956 in 69 samples, s/n 104 for the C-allele and 184 for the T-allele. D) SNP 7188 in 69 samples, s/n 10 for the A-allele and 41 for the G-allele

Study IV

Background

The hunting pressure proposes a large threat to non-human primates. Body parts encountered by law enforcement often lack morphological features and there is, therefore, a need for tools for species or genera determination on a genetic level.

In study IV, sequences retrieved from GenBank and from this study, were aligned and positions to differentiate between infraorders, Catarrhini subfamilies and most primate genera were identified. For the infraorder and Catarrhini subfamily differentiation the nuclear genes encoding epsilon globin and apolipoprotein B, respectively, was used. For differentiation on the genus level the mitochondrial 12S rRNA was used. A minisequencing assay

to genotype the positions were designed and tested on DNA from species from all infraorders, including many endangered species.

Results and discussion

Nucleotide positions suitable for distinguishing between genera should be highly polymorphic between genera, but conserved within genera. To detect them, they should also be within a region of high conservation to facilitate minisequencing primer design. To screen for the positions in the alignments, DIAPOS software was used (unpublished). Primates are a highly diverse group, and determination on the genera level could only be achieved after first determining the infraorder and Catarrhini subfamily. Although the nucleotides on the different levels were genotyped in the same assay, in both the design and analysis this had to be considered.

In total, 111 diagnostic positions were selected with at lest one flanking region suitable for primer design. These positions were able to differentiate 45 genera. After initial optimizing of the microarray-based minisequencing-assay, 87 minisequencing primers were selected for genotyping. The microarray was tested on known primate samples and 66 primers for 57 positions were successfully genotyped in repeated experiments. The failing primers in the assay were most likely due to nucleotide variation in the primer-binding site. Seventy primate samples and 5 non-primate samples were tested. In summary, 65 of the samples were assigned to the correct infraorder and two were assigned to the wrong one. Of the Catarrhini, 36 of 42 samples were assigned to the correct one. Of the 59 samples that were correctly assigned to infraorder, and where applicable to Catarrhini subfamily, 47 could be assigned at the genus level. All but one of the incorrectly assigned samples belonged to the Cercopithecinae subfamily.

The microarray assay could be improved by the addition of more positions in either mitochondrial or nuclear genes. For the Cercopithecinae it seems necessary to add more positions to determine to which sub-groups a sample belongs to before being able to determine genera.

Of the 47 samples genotyped at the genus level, 32 were correct, and of these, 16 were unambiguously assigned to one genus. Interestingly most of these samples belonged to genera with genus-specific diagnostic positions suggesting that these positions are more useful and easier to interpret than those positions that are highly variable between genera.

Ethical considerations

This study involves genetic analyses of non-human primates, but does not involve any experimentation or euthanasia in non-human primates, whether captive or not, for the sake of the study. The samples in this study have been

collected by members of the Inprimat consortium following local regulations, as the UK Animal Scientific Procedures Act (1986), the Decree 214/1997 and Law 5/1995 from the Generalitat de Catalunya (Catalonian Regional Government) and other similar regulation in each country.

Concluding remarks

In this post-sequencing era in which many scientists study the function and diversity of the human genome, the scientific community is faced with a considerable underrepresentation of molecular information from animals that are intermediate between humans and mammalian animal models, such as rodents. Since non-human primates share a common ancestor with humans exclusive to all other mammalian orders, to increase knowledge and information about non-human primate DNA is crucial in comparative genomics. To understand the underlying genetic basis of diseases found in humans, comparative sequence analysis in other primates is often prerequisite.

Acknowledgements

The work presented in this thesis was performed in the group of Molecular Medicine at the Department of Medical Sciences, Uppsala University.

I would like to thank the whole group for making the work place an enjoyable and friendly one, and for always being so helpful. I think it is rare to really like everyone you work with so much, with no exceptions. I hope the good atmosphere never changes.

Some people deserve my special gratitude, and I am almost afraid to mention anyone for fear of forgetting someone.

To my supervisor, Professor Ann-Christine Syvänen. Thank you for giving the Inprimat project to me. I could not believe my luck! And for the trust you have shown me, from the start, and for sending me all over Europe.

Thank you, all the people of Inprimat. You have all been such a pleasure to work with, and you have taught me a lot. Hope to see you again in the future.

Lillebil, the fixer. Thanks for everything.

Raul and Anki, for help in the lab.

The “old” members. Lovisa, and Mona, you are missed! Ulrika, you did’t go very far, but I missed sharing office with you. Doktor S, I haven’t had time to sit on you sofa for a long while now. It’s the end of an era. All of you have taught me so much!

Lili, thanks for reading my thesis and giving valuable comments. Johanna, thanks for the “nation-hunt”. Good luck in the US, Annika. And thanks for all good discussions and for always being so helpful. Per, for all the help, and comments on the thesis. Andreas, especially for all good talks. And Gulla, thanks for the good times. Cissi, thanks for the good work and many hours you have put into my project. Lars, thanks for all help in general and on my strange computer problems in particular. And thank you Jessica for

reading my thesis. And thank you all the “platform-people”, always so helpful: Marie, Caisa, Tomas, Kristina, Mats, Björn.

And to Aron, thanks for being so amazing when I have needed it the most.

This thesis was funded by the European Commision (INPRIMAT Consortium, contract QLRI-CT-2002-01325), Knut and Alice Wallenberg’s Foundation and the Swedish Research Council for Science and Technology

References

Allen M, Engstrom AS, Meyers S*, et al.* (1998) Mitochondrial DNA sequencing of shed hairs and saliva on robbery caps: sensitivity and matching probabilities. *J Forensic Sci* **43**, 453-464.

Anderson S, Bankier AT, Barrell BG*, et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457-465.

Andrade MC, Penedo MC, Ward T*, et al.* (2004) Determination of genetic status in a closed colony of rhesus monkeys (Macaca mulatta). *Primates* **45**, 183-186.

Baner J, Isaksson A, Waldenstrom E*, et al.* (2003) Parallel gene analysis with allele-specific padlock probes and tag microarrays. *Nucleic Acids Res* **31**, e103.

Baner J, Nilsson M, Mendel-Hartvig M, Landegren U (1998) Signal amplification of padlock probes by rolling circle replication. *Nucleic Acids Res* **26**, 5073-5078.

Barker DL, Hansen MS, Faruqi AF*, et al.* (2004) Two methods of whole-genome amplification enable accurate genotyping across a 2320-SNP linkage panel. *Genome Res* **14**, 901-907.

Blanco L, Bernad A, Lazaro JM*, et al.* (1989) Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J Biol Chem* **264**, 8935-8940.

Blanco L, Salas M (1985) Characterization of a 3'----5' exonuclease activity in the phage phi 29-encoded DNA polymerase. *Nucleic Acids Res* **13**, 1239-1249.

Bradley BJ, Boesch C, Vigilant L (2000) Identification and redesign of human microsatellite markers for genotyping wild chimpanzee (Pan troglodytes verus) and gorilla (Gorilla gorilla gorilla) DNA from faeces. *Conservation Genetics* **1**, 289-292.

Chakraborty R, Stivers DN, Su B, Zhong Y, Budowle B (1999) The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. *Electrophoresis* **20**, 1682-1696.

Clifford SL, Jeffrey K, Bruford MW, Wickings EJ (1999) Identification of polymorphic microsatellite loci in the gorilla (Gorilla gorilla gorilla) using human primers: application to noninvasively collected hair samples. *Mol Ecol* **8**, 1556-1558.

Dean FB, Hosono S, Fang L*, et al.* (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* **99**, 5261-5266.

Dean FB, Nelson JR, Giesler TL, Lasken RS (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* **11**, 1095-1099.

Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* **35**, 41-48.

Dietmaier W, Hartmann A, Wallinger S*, et al.* (1999) Multiple mutation analyses in single tumor cells with improved whole genome amplification. *Am J Pathol* **154**, 83-95.

Domingo-Roura X, Lopez-Giraldez T, Shinohara M, Takenaka O (1997) Hypervariable microsatellite loci in the Japanese macaque (Macaca fuscata) conserved in related species. *Am J Primatol* **43**, 357-360.

Domingo-Roura X, Marmi J, Andres O, Yamagiwa J, Terradas J (2004) Genotyping from semen of wild Japanese macaques (Macaca fuscata). *Am J Primatol* **62**, 31-42.

Doxiadis GG, Otting N, de Groot NG*, et al.* (2003) Evolutionary stability of MHC class II haplotypes in diverse rhesus macaque populations. *Immunogenetics* **55**, 540-551.

Doxiadis GG, Otting N, de Groot NG, Noort R, Bontrop RE (2000) Unprecedented polymorphism of Mhc-DRB region configurations in rhesus macaques. *J Immunol* **164**, 3193-3199.

Drossman H, Luckey JA, Kostichka AJ, D'Cunha J, Smith LM (1990) High-speed separations of DNA sequencing reactions by capillary electrophoresis. *Anal Chem* **62**, 900-903.

Edwards A, Civitello A, Hammond HA, Caskey CT (1991) DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *Am J Hum Genet* **49**, 746-756.

Esteban JA, Salas M, Blanco L (1993) Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *J Biol Chem* **268**, 2719-2726.

Fakhrai-Rad H, Pourmand N, Ronaghi M (2002) Pyrosequencing: an accurate detection platform for single nucleotide polymorphisms. *Hum Mutat* **19**, 479-485.

Fan JB, Chen X, Halushka MK*, et al.* (2000) Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Research* **10**, 853-860.

Fan JB, Oliphant A, Shen R*, et al.* (2003) Highly parallel SNP genotyping. *Cold Spring Harb Symp Quant Biol* **68**, 69-78.

Gao F, Bailes E, Robertson DL*, et al.* (1999) Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* **397**, 436-441.

Gibbs RA, Rogers J, Katze MG*, et al.* (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222-234.

Goossens B, Funk SM, Vidal C*, et al.* (2002) Measuring genetic diversity in translocation programmes: principles and application to a chimpanzee release project. *Animal Conservation* **5**, 225-236.

Goossens B, Waits LP, Taberlet P (1998) Plucked hair samples as a source of DNA: reliability of dinucleotide microsatellite genotyping. *Mol Ecol* **7**, 1237-1241.

Groves CP (2001) *Primate Taxonomy* Smithsonian Institution Press, Washington DC.
Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS (2005) A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet* **37**, 549-554.
Hall TA (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/97/NT. *Nucleic Acids Symposium Series* **41**, 95-98.
Hardenbol P, Baner J, Jain M*, et al.* (2003) Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol* **21**, 673-678.
Hosono S, Faruqi AF, Dean FB*, et al.* (2003) Unbiased whole-genome amplification directly from clinical samples. *Genome Res* **13**, 954-964.
Hughes JF, Skaletsky H, Pyntikova T*, et al.* (2005) Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature* **437**, 100-103.
Ingman M, Kaessmann H, Paabo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708-713.
Jiang Z, Priat C, Galibert F (1998) Traced orthologous amplified sequence tags (TOASTs) and mammalian comparative maps. *Mamm Genome* **9**, 577-587.
Kirsch S, Weiss B, Miner TL*, et al.* (2005) Interchromosomal segmental duplications of the pericentromeric region on the human Y chromosome. *Genome Res* **15**, 195-204.
Knapp LA, Cadavid LF, Eberle ME*, et al.* (1997) Identification of new mamu-DRB alleles using DGGE and direct sequencing. *Immunogenetics* **45**, 171-179.
Kocher TD, Thomas WK, Meyer A*, et al.* (1989) Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proc Natl Acad Sci U S A* **86**, 6196-6200.
Landegren U, Kaiser R, Sanders J, Hood L (1988) A ligase-mediated gene detection technique. *Science* **241**, 1077-1080.
Lander ES, Linton LM, Birren B*, et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.
Langmore JP (2002) Rubicon Genomics, Inc. *Pharmacogenomics* **3**, 557-560.
Lasken RS, Egholm M (2003) Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens. *Trends Biotechnol* **21**, 531-535.
Lathuilliere M, Menard N, Crouau-Roy B (2001) Sequence conservation of nine Barbary macaque (Macaca sylvanus) microsatellite loci: implication of specific primers for genotyping. *Folia Primatol (Basel)* **72**, 85-88.
Liljedahl U, Karlsson J, Melhus H*, et al.* (2003) A microarray minisequencing system for pharmacogenetic profiling of antihypertensive drug response. *Pharmacogenetics* **13**, 7-17.

Lindroos K, Sigurdsson S, Johansson K, Ronnblom L, Syvanen AC (2002) Multiplex SNP genotyping in pooled DNA samples by a four-colour microarray system. *Nucleic Acids Research* **30**, e70.

Livak KJ, Flood SJ, Marmaro J, Giusti W, Deetz K (1995) Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *PCR Methods Appl* **4**, 357-362.

Lizardi PM, Huang X, Zhu Z*, et al.* (1998) Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat Genet* **19**, 225-232.

Lovmar L, Fredriksson M, Liljedahl U, Sigurdsson S, Syvanen AC (2003) Quantitative evaluation by minisequencing and microarrays reveals accurate multiplexed SNP genotyping of whole genome amplified DNA. *Nucleic Acids Res* **31**, e129.

Luckey JA, Drossman H, Kostichka AJ*, et al.* (1990) High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Res* **18**, 4417-4421.

Lunt DH, Whipple LE, Hyman BC (1998) Mitochondrial DNA variable number tandem repeats (VNTRs): utility and problems in molecular ecology. *Mol Ecol* **7**, 1441-1455.

Maillard JC, Gonzalez JP (2006) Biodiversity and emerging diseases. *Ann N Y Acad Sci* **1081**, 1-16.

Margulies M, Egholm M, Altman WE*, et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.

Marmi J, Bertranpetit J, Terradas J, Takenaka O, Domingo-Roura X (2004) Radiation and phylogeography in the Japanese macaque, Macaca fuscata. *Mol Phylogenet Evol* **30**, 676-685.

Matsuzaki H, Dong S, Loi H*, et al.* (2004a) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* **1**, 109-111.

Matsuzaki H, Loi H, Dong S*, et al.* (2004b) Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res* **14**, 414-425.

Mikkelsen TS, Hillier LW, Eichler EE, Consortium tCSaA (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87.

Mullis KB, Faloona FA (1987) Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol* **155**, 335-350.

Murray V (1989) Improved double-stranded DNA sequencing using the linear polymerase chain reaction. *Nucleic Acids Res* **17**, 8889.

Navarro A, Barton NH (2003) Chromosomal speciation and molecular divergence--accelerated evolution in rearranged chromosomes. *Science* **300**, 321-324.

Nilsson M, Malmgren H, Samiotaki M*, et al.* (1994) Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* **265**, 2085-2088.

Nurnberg P, Sauermann U, Kayser M*, et al.* (1998) Paternity assessment in rhesus macaques (Macaca mulatta): multilocus DNA fingerprinting and PCR marker typing. *Am J Primatol* **44**, 1-18.

Oefner PJ, Underhill PA (1998) DNA mutation detection using denaturing high-performance liquid chromatography (DHPLC). In: *Current Protocols in Human Genetics*, pp. Supplement 19, 17.10.11-17.10.12. Wiley & Sons, New York.

Oliphant A, Barker DL, Stuelpnagel JR, Chee MS (2002) BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques* **Suppl**, 56-58, 60-51.

Otting N, de Groot NG, Noort MC, Doxiadis GG, Bontrop RE (2000) Allelic diversity of Mhc-DRB alleles in rhesus macaques. *Tissue Antigens* **56**, 58-68.

Pastinen T, Raitio M, Lindroos K*, et al.* (2000) A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays. *Genome Research* **10**, 1031-1042.

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* **273**, 1516-1517.

Roeder AD, Jeffery K, Bruford MW (2006) A universal microsatellite multiplex kit for genetic analysis of great apes. *Folia Primatol (Basel)* **77**, 240-245.

Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* **242**, 84-89.

Ross MT, Grafham DV, Coffey AJ*, et al.* (2005) The DNA sequence of the human X chromosome. *Nature* **434**, 325-337.

Sachidanandam R, Weissman D, Schmidt SC*, et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-933.

Saiki RK, Gelfand DH, Stoffel S*, et al.* (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487-491.

Saiki RK, Scharf S, Faloona F*, et al.* (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350-1354.

Salem AH, Ray DA, Xing J*, et al.* (2003) Alu elements and hominid phylogenetics. *Proc Natl Acad Sci U S A* **100**, 12787-12791.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467.

Sigurgardottir S, Helgason A, Gulcher JR, Stefansson K, Donnelly P (2000) The mutation rate in the human mtDNA control region. *Am J Hum Genet* **66**, 1599-1609.

Sinclair AH, Berta P, Palmer MS*, 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.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ*, et al.* (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825-837.

Smith DG, Kanthaswamy S, Viray J, Cody L (2000) Additional highly polymorphic microsatellite (STR) loci for estimating kinship in rhesus macaques (Macaca mulatta). *Am J Primatol* **50**, 1-7.

Sorensen KJ, Turteltaub K, Vrankovich G, Williams J, Christian AT (2004) Whole-genome amplification of DNA from residual cells left by incidental contact. *Anal Biochem* **324**, 312-314.

Southern E, Mir K, Shchepinov M (1999) Molecular interactions on microarrays. *Nat Genet* **21**, 5-9.

Stone AC, Griffiths RC, Zegura SL, Hammer MF (2002) High levels of Y-chromosome nucleotide diversity in the genus Pan. *Proc Natl Acad Sci U S A* **99**, 43-48.

Syvanen AC, Aalto-Setala K, Harju L, Kontula K, Soderlund H (1990) A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. *Genomics* **8**, 684-692.

Taylor RW, McDonnell MT, Blakely EL*, et al.* (2003) Genotypes from patients indicate no paternal mitochondrial DNA contribution. *Ann Neurol* **54**, 521-524.

Telenius H, Carter NP, Bebb CE*, et al.* (1992) Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718-725.

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673-4680.

Ullu E, Tschudi C (1984) Alu sequences are processed 7SL RNA genes. *Nature* **312**, 171-172.

Wallace RB, Shaffer J, Murphy RF*, et al.* (1979) Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res* **6**, 3543-3557.

Walsh PD, Abernethy KA, Bermejo M*, et al.* (2003) Catastrophic ape decline in western equatorial Africa. *Nature* **422**, 611-614.

Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738.

Venter JC, Adams MD, Myers EW*, et al.* (2001) The sequence of the human genome. *Science* **291**, 1304-1351.

Wetterbom A, Sevov M, Cavelier L, Bergstrom TF (2006) Comparative genomic analysis of human and chimpanzee indicates a key role for indels in primate evolution. *J Mol Evol* **63**, 682-690.

Vigilant L, Pennington R, Harpending H, Kocher TD, Wilson AC (1989) Mitochondrial DNA sequences in single hairs from a southern African population. *Proc Natl Acad Sci U S A* **86**, 9350-9354.

Wolford JK, Blunt D, Ballecer C, Prochazka M (2000) High-throughput SNP detection by using DNA pooling and denaturing high performance liquid chromatography (DHPLC). *Hum Genet* **107**, 483-487.

Zhang L, Cui X, Schmitt K*, et al.* (1992) Whole genome amplification from a single cell: implications for genetic analysis. *Proc Natl Acad Sci U S A* **89**, 5847-5851.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 263

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-7904

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2007