Resource

Genome-based analysis of the nonhuman primate *Macaca fascicularis* as a model for drug safety assessment

Martin Ebeling,[1] Erich Küng,[2] Angela See,[3] Clemens Broger,[4] Guido Steiner,[1] Marco Berrera,[1] Tobias Heckel,[2] Leonardo Iniguez,[3] Thomas Albert,[3] Roland Schmucki,[1] Hermann Biller,[4] Thomas Singer,[2] and Ulrich Certa[2,5]

[1]*Translational Research Sciences, F. Hoffmann-La Roche AG, Pharmaceutical Research and Early Development (pRED), 4070 Basel, Switzerland;* [2]*Global Non-clinical Safety, F. Hoffmann-La Roche AG, Pharmaceutical Research and Early Development (pRED), 4070 Basel, Switzerland;* [3]*Roche NimbleGen, Inc., Madison, Wisconsin 53719, USA;* [4]*Research Informatics, F. Hoffmann-La Roche AG, Pharmaceutical Research and Early Development (pRED), 4070 Basel, Switzerland*

The long-tailed macaque, also referred to as cynomolgus monkey (*Macaca fascicularis*), is one of the most important nonhuman primate animal models in basic and applied biomedical research. To improve the predictive power of primate experiments for humans, we determined the genome sequence of a *Macaca fascicularis* female of Mauritian origin using a whole-genome shotgun sequencing approach. We applied a template switch strategy that uses either the rhesus or the human genome to assemble sequence reads. The sixfold sequence coverage of the draft genome sequence enabled discovery of about 2.1 million potential single-nucleotide polymorphisms based on occurrence of a dimorphic nucleotide at a given position in the genome sequence. Homology-based annotation allowed us to identify 17,387 orthologs of human protein-coding genes in the *M. fascicularis* draft genome, and the predicted transcripts enabled the design of a *M. fascicularis*–specific gene expression microarray. Using liver samples from 36 individuals of different geographic origin we identified 718 genes with highly variable expression in liver, whereas the majority of the transcriptome shows relatively stable and comparable expression. Knowledge of the *M. fascicularis* draft genome is an important contribution to both the use of this animal in disease models and the safety assessment of drugs and their metabolites. In particular, this information allows high-resolution genotyping and microarray-based gene-expression profiling for animal stratification, thereby allowing the use of well-characterized animals for safety testing. Finally, the genome sequence presented here is a significant contribution to the global "3R" animal welfare initiative, which has the goal to reduce, refine, and replace animal experiments.

[Supplemental material is available for this article.]

Drug discovery and development are labor- and cost-intensive processes that can last up to 20 yr from concept to market. The most critical concerns before entry into clinical development of a novel compound are the balance between risk and benefit for the patient. International drug safety agencies like the US Food and Drug Administration (FDA) or the European Medicines Agency (EMA) apply standardized testing procedures and requirements for submission of new medicines. Animal experiments designed to predict parameters such as toxicity or pharmacokinetics are a key prerequisite of the drug approval process. Rodents, dogs, mini-pigs, and in particular, nonhuman primates are the main species used in translational drug safety research and risk assessment (Boelsterli 2003).

Compared with rodents or dogs, nonhuman primates have a closer evolutionary relationship to humans, exhibit greater physiological similarity, and have the added benefit of being capable of completing memory tests originally designed for humans (Capitanio and Emborg 2008). Only a small number of species, such as macaques of the Cercopithecidae family of Old World monkeys, are well suited and established as translational models for drug testing. These primates share a common ancestor with humans that is estimated to have lived about 32 million years ago (Perelman et al. 2011). From this family, the species *Macaca mulatta*, also known as rhesus monkey, and *Macaca fascicularis*, the long-tailed macaque, are the most common and best-studied nonhuman primate animal models today (Ferguson et al. 2007; Gibbs et al. 2007). Although closely related, these two species show distinct phenotypic differences, morphology, behavior, and physiology. In 1978, India banned all rhesus monkey exports to breeding centers across the world, and since then usage of this species in drug safety testing has declined. As an alternative to rhesus, several commercial breeding centers in Indonesia, China, the Philippines, and Mauritius are able to provide sufficient numbers of captive bred long-tailed macaques originating from wild-trapped founders. The natural range of *M. fascicularis* monkeys spans the mainland of southern Asia, Indonesia, the Philippines, and more recently Mauritius, where a small number of founder animals was imported on a trading ship during the 15th century (Ferguson et al. 2007). Rhesus monkeys inhabit predominantly the mainland of China, Vietnam, Laos, Nepal, Thailand, northern India, and Pakistan. The wide geographic distribution of both species and considerable interspecies hybridization in shared habitats point to populations that are genetically and phenotypically quite diverse (Tosi et al. 2002). Thus, genotyping

[5]Corresponding author.
E-mail ulrich.certa@roche.com.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.123117.111
Freely available online through the *Genome Research* Open Access option.

21:1746–1756 © 2011 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/11; www.genome.org

and phenotypic characterization of animals is desirable prior to studies aimed at predicting the safety of novel medicines in humans.

A cornerstone of primate research with high impact for biomedical research was the publication of the first genome draft of the rhesus monkey, *Macaca mulatta,* in 2007 (Gibbs et al. 2007). This multicenter effort was mainly focused on highlighting differences between Old World monkeys and great apes like the chimpanzee, which differs from humans by only ~1.23% based on genome-wide sequence identity comparison (Waterston et al. 2005). This genome draft enabled the search for macaque-specific genes related to physiology and phenotypes, as well as the assembly of signaling cascades and pathways involved in the immune response to pathogens such as simian immunodeficiency virus (SIV). In addition, the mapping of random sequence reads obtained from additional animals with different geographic origin to the genome draft allowed assessment of population diversity at single nucleotide resolution. The genome assembly also enabled the design of novel rhesus genome based mRNA expression microarrays, which were applied to the analysis of human influenza virus infection (Gibbs et al. 2007). Compared with earlier expressed sequence tag (ESTs)-based chip designs (Magness et al. 2005), availability of the rhesus draft genome allowed selection of unique probes, resulting in reduced cross-hybridization and improved performance.

Drug safety studies carried out under good laboratory practice (GLP) require relatively large numbers of animals that match the experimental requirements with respect to parameters like age, weight, or gender. *M. fascicularis* today is the most widely used primate species for drug safety testing in pharmaceutical companies or contract research organizations (CROs). In addition, validated assays for measuring blood parameters and safety biomarkers such as liver enzymes or cytokines are available. Depending on the molecule and mode of action, repro-toxicity testing in nonhuman primates is frequently part of the required safety package. Long-tailed macaques are particularly advantageous in this area, since this species has no seasonal fertility, unlike rhesus, whose females are receptive only once a year (Weinbauer et al. 2008).

We have generated the first draft genome of the long-tailed macaque *M. fascicularis* using a whole-genome shotgun (WGS) sequencing approach employing two independent deep sequencing technologies. We identified about 2.1 million candidate single-nucleotide polymorphisms (SNPs) and used homology-based annotation for transcript identification to design a *M. fascicularis*-specific gene expression microarray. With focus on drug safety, we performed a phylogenetic analysis of the SLCO drug transporter family and transcript profiling of cytokines and cytochrome p450s in liver samples from 36 naïve monkeys within the context of a global gene expression analysis.

Results

Genome sequencing and assembly

As a DNA source for genome sequencing we selected a 3-yr-old female *M. fascicularis* monkey born in Mauritius, because of limited genetic variability in this geographically isolated population. The organization of the long-tailed macaque genome into 2 × 20 autosomes and two sex chromosomes (Chu and Bender 1961) and the availability of the related rhesus draft genome (RheMac2), together with the well-annotated human genome (HG18), permitted a direct whole-genome shotgun (WGS) sequencing strategy, omitting construction of bacterial artificial chromosome (BAC) or fosmid libraries. We generated long read libraries using 454 Life Sciences (Roche) FLX pyrosequencing technology and short read libraries for Applied Biosystems (Life Technologies) Sequencing by Oligonucleotide Ligation and Detection (SOLiD) using the same DNA source. A total of 59 standard 454-FLX sequencing runs yielded ~550,000 reads per run with 250 nucleotides average length. Eight additional runs with the improved *Titanium* chemistry added ~1,100,000 reads per run with 350 nucleotides average length. All reads with a unique best-mapping location were included in template-based assembly using either the rhesus genome draft or the human reference genome as template. Out of the 38,189,378 reads generated, 28,327,570 (74.2%) mapped uniquely to the rhesus genome draft; 4,188,914 (11.0%) reads had multiple matches without a clear best match; and 5,672,894 reads (14.9%) failed to map. Out of these "left-over" reads, 610,803 (10.8%) could be uniquely mapped to the human genome. We expected that these reads would map to well-conserved sequences such as protein-coding genes. The remaining 5,062,091 reads are expected to match to intergenic or intronic sequences in the human genome that are missing in rhesus, as BLAST searches of several hundred randomly selected reads have shown (data not shown).

It is well appreciated that different sequencing technologies suffer from different intrinsic error biases (Harismendy et al. 2009). Therefore, we applied the DNA ligation-based SOLiD technology for sequence quality refinement. Compared with 454-FLX, this technique generates about 400 million reads per sequencing run with a read length of 50 nucleotides. We added about 182 million SOLiD reads to the 454-based genome draft, which raised the sequence coverage from three- to sixfold. Figure 1 shows the coverage distribution from the combined 454 and SOLiD reads mapped to the rhesus genome draft. It follows a Poisson-like distribution with a mean coverage of sixfold per residue in agreement with a total of 17.6 billion bases sequenced for a genome size of 2,879,309,005 bases.

It is important to note that our template-based approach circumvents the problem of de novo assembly. With the exception of gap closure, we did not attempt to correct any assembly errors in the rhesus template. Therefore, the value of the genome draft presented here is primarily the direct comparison between the long-tailed macaque and rhesus. In addition, we consider the numerous sequence stretches that were not available before for any macaque species as a useful resource for the identification or completion of protein-coding genes. In contrast to conventional genome assembly

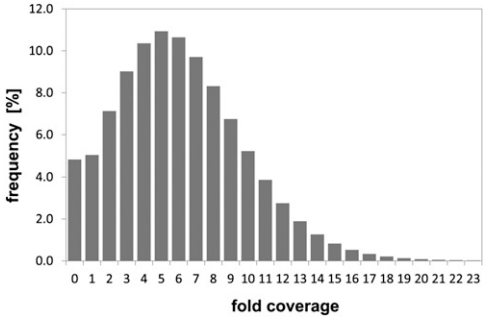

Figure 1. Coverage histogram of the *M. fascicularis* genome draft. The histogram shows the coverage distribution from the combined 454 and SOLiD reads of the *M. fascicularis* genome mapped to the reference genome (RheMac2). The frequency of nucleotide positions was plotted against the sequence coverage. Coverage exhibited a Poisson-like distribution with a mean of sixfold sequence coverage. The *lowest* bin in the histogram represents reference positions with zero aligned reads. About 80% of these noncovered positions are annotated as being repetitive in the RheMac2 draft.

strategies, our template-based sequencing strategy did not generate "scaffolds" or "contigs," and as a further consequence, the discovery of *M. fascicularis* specific duplications, genomic rearrangements, or *Alu* and LINE elements that are not represented in the template genomes is inherently excluded. The mapping efficiency and coverage of our WGS sequencing strategy depends largely on sequence properties and quality of the template used for assembly (Fig. 2). In nonrepetitive regions, mapping of sequencing reads is relatively dependable and almost all regions including annotated exons are covered (Fig. 2A). Mapping efficiency drops significantly in regions that are repetitive in the rhesus genome and, as a result, gaps appear in the *M. fascicularis* draft sequence (Fig. 2B). In numerous cases, the length of 454-FLX reads reduced the gap size in repetitive regions that are flanked by unique DNA. For nonrepetitive sequence stretches that failed to map to the rhesus genome sequence, switching to the human genome as template allowed us to reduce in size or even close numerous gaps. This template switch approach allowed, for instance, the unambiguous identification of four exons of the *M. fascicularis* gene for transmembrane serine protease 9 (TMPRSS9), which reside inside a large gap of the rhesus genome draft (Fig. 2C). On a genomic scale, the template switch strategy allowed mapping of about 50 million bases to 60,583 regions that are absent in the rhesus template.

Comparative analysis of the *M. fascicularis* draft genome sequence

To approach the divergence between *M. fascicularis* and *M. mulatta* at the genomic level we performed pairwise comparisons of homologous chromosomes resulting in an average, overall DNA sequence identity of 99.21%. Single base mismatches constitute 0.47% diversity (A vs. G and C vs. T transitions 0.31%; purine vs. pyrimidine transversions 0.16%). The remaining 0.32% consists of small deletions or insertions. The sequence identity between *M. fascicularis* and the human genome is 92.83% based on alignments of homologous genomic fragments. This value is marginally lower than the reported identity of 93.54% for *M. mulatta* and *H.sapiens* (Gibbs et al. 2007). This first genome-based comparison between *M. fascicularis* and *M. mulatta* confirms the close evolutionary relationship of the two species deduced from shared SNPs, mitochondrial, and EST sequences (Magness et al. 2005; Street et al. 2007). This finding is consistent with a relatively recent split of these primates about 2.8 million years ago, as determined by phylogenetic analysis of genomic DNA sequences (Perelman et al. 2011).

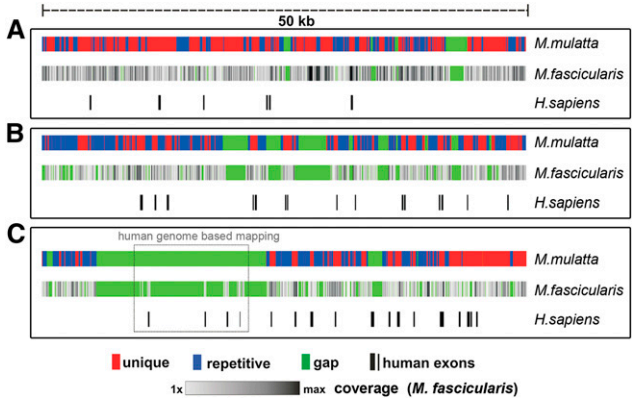

Figure 2. Representative examples of mapping efficiencies of *M. fascicularis* WGS reads to rhesus or human template genomes with different repeat and gap content. (*A*) Graphical comparison of a 50-kb fragment on chromosome 1 with low repeat and gap content. Unique sequences in rhesus in general have better coverage in *M. fascicularis* than repetitive segments, and gaps cannot be closed as an inherent feature of the WGS approach. For rhesus (*top*), unique stretches are shown in red, repetitive DNA in blue, and gaps in green. For the corresponding *M. fascicularis* section (*middle*) the local coverage is indicated from onefold (light gray) to ≥ sixfold (black), and gaps are shown in green. For human (*bottom*), only exons are shown as reference. (*B*) Display of a chromosome 1 region with increased gap and repeat content, where average coverage in *M. fascicularis* is significantly reduced due to ambiguous mapping. In some cases, the long 454 reads reduce gap size relative to rhesus. (*C*) Human genome-based identification of exons in a 20-kb rhesus genome gap based on homology and conservation of intron/exon boundaries in primates.

Prediction of transcripts and analysis of orthologous relationships

The *M. fascicularis* draft genome allows prediction of protein-coding genes and transcripts orthologous to human sequences, which is critical to predict cross-reactivity and pharmacology of drugs developed for human therapy. The transcriptome of the human genome is by far the best characterized among all species. A total of 19,022 genes encode 30,864 unique mRNAs according to the human RefSeq database release 44 (Pruitt et al. 2007). We used these human sequences for automated identification of orthologs in the *M. fascicularis* genome, assuming conservation of splice donor and acceptor sites in both species. This approach yielded a total of 17,387 protein-coding genes and 27,870 putative transcripts.

Next, we assessed the conservation of protein-coding transcripts by three-way alignments of *M. fascicularis*, *M. mulatta*, and *H. sapiens* mRNAs. We developed an automated pipeline to identify human mRNA sequences that have a corresponding predicted mRNA in both *M. fascicularis* and *M. mulatta*. We applied rigorous filters to remove incomplete sequences to select only one transcript per gene, as well as mRNAs that are similar in length with respect to the human reference. We identified 10,919 mRNAs that fulfilled these criteria in all three species, comparable to the 10.376 genes used in the human/chimpanzee/rhesus analysis of Gibbs et al. (2007). When comparing *M. fascicularis* and rhesus mRNA sequences, we found that 42.9% of the 5′ UTRs, 10.2% of the CDSs, and 14.1% of the 3′ UTRs were 100% identical. The elevated conservation of the 5′ UTRs is a result of the shorter average length. Evaluation of the nonidentical sequences demonstrated that the vast majority of these mRNAs show very high sequence identities with modes around 99.5%, 99.8%, and 99.6% for 5′ UTRs, CDSs, and 3′ UTRs, respectively (Fig. 3). As expected, the degree of sequence identity decreased when compared with the human set of orthologs. Here, we found modes around sequence identities of 94.6% for 5′ UTRs, 97.9% for CDSs, and 93.4% for 3′ UTRs for rhesus and corresponding identities of 94.6%, 97.7%, and 92.7% for the long-tailed macaque (Fig. 3). These results can be related to the transcriptome divergence patterns between human and chimpanzee with average sequence identities of ~98.7% for 5′ UTRs, 99.3% for CDSs, and 97.9% for 3′ UTRs (Sakate et al. 2007).

Phylogenetic analysis of the SLCO (solute carriers for organic anions) gene family

We next dissected the evolutionary relationship between *M. fascicularis*, *M. mulatta*, and *H.sapiens* in more detail using the solute carrier protein (SLC) superfamily as an example with pharmacological relevance. The SLC superfamily consists of 55 gene families with at least 362 protein-coding genes that mediate transport of a wide spectrum of amphipathic organic solutes across

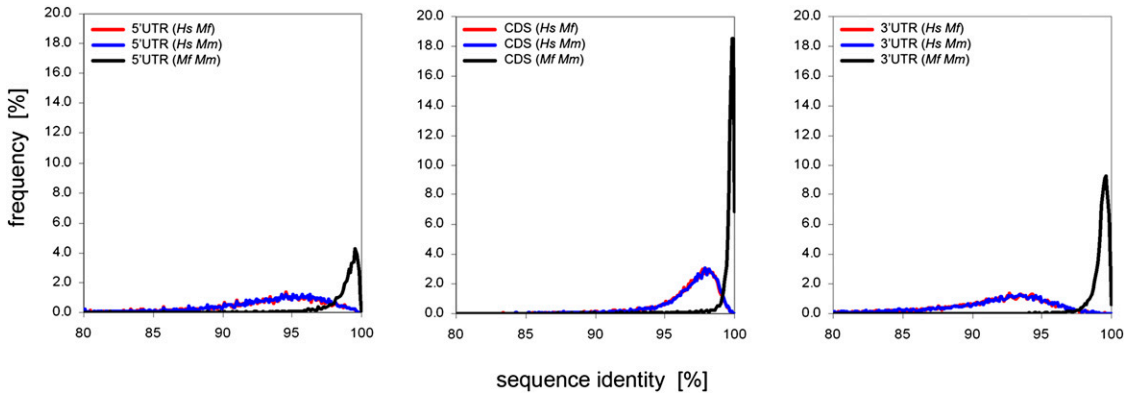

Figure 3. Sequence identities between orthologous transcripts of *M. fascicularis*, *M. mulatta*, and *H. sapiens*. The 5′ UTR (*left*), CDS (*middle*), and 3′ UTR (*right*) of 10,919 orthologous mRNAs were considered separately for the calculation of pairwise sequence identities. The relative number of 1:1 orthologous sequences was plotted against the sequence identities. Frequency plots of sequence identities <100% between *M. fascicularis* and *M. mulatta* (*Mf Mm*), *H. sapiens* and *M. fascicularis* (*Hs Mf*), and *H. sapiens* and *M. mulatta* (*Hs Mm*) transcripts are displayed. Note that the peak sequence identities for the UTRs are significantly lower between humans and macaques than for the coding regions.

cells and organs (Hagenbuch and Meier 2003). In particular, the SLCO subfamily is of high importance for drug safety research, because its members are involved in the transport of organic molecules into the liver for detoxification by cytochrome p450 and other metabolizing enzyme systems. Therefore, a comparison between human and *M. fascicularis* SLCOs is critical to predict the translational value of monkey drug safety data for humans. The human SLCO gene family forms a separate evolutionary branch with 11 annotated members. One member, SLCO1B3, has undergone gene duplication, giving rise to an additional copy named LST-3B. We identified *M. mulatta* and *M. fascicularis* orthologs for all human SLCO family members for phylogenetic analysis (Fig. 4A). Beyond the split of human and macaque evolution (bold lines in Fig. 4A) no major diversification within the macaques appears to have taken place. Interestingly, the dendrogram highlights clear differences in the evolutionary rate of individual SLCO genes; SLCO1C1, for example, is the most conserved gene, whereas SLCO6A1 has undergone pronounced macaque-specific evolution and shows the largest divergence from the shared ancestral gene. Steroid hormones are the main substrates of the SLCO1B3 transporter, and this gene underwent duplication in primates, followed by independent evolution (Fig. 4A). Interestingly, a minor allele of human SLCO1B3 carries two nonsynonymous SNPs (T334G and G699A) causing S112A and M233I amino acid changes (Hamada et al. 2008). Using recombinant expression of the human variants it has been shown that the individual mutations have no major impact on testosterone binding, whereas double mutants have lost substrate binding ability (Hamada et al. 2008). The *M. fascicularis* and *M. mulatta* orthologs of SLCO1B3 both carry these inactivating mutations. Based on sequence alignments of all macaque SLCO transporters with human orthologs, we found that all family members bear amino acid changes ranging from 9% (SLCO6A1) to 0.75% (SLCO1C1). Results of a detailed relationship analysis of the *M. fascicularis* SLCO family with rhesus and human are shown in Figure 4B. Based on sequence data, it is not possible to predict whether amino acid changes in the *M. fascicularis* SLCOs modulate substrate recognition and transport. However, the available *M. fascicularis* SLCO sequences allow transfection into eukaryotic cell lines for the analysis of drug transport and substrate specificity, an approach successfully applied to dissect human SLCO1B3 polymorphisms at the functional level (Hamada et al. 2008).

Single-nucleotide polymorphisms (SNPs)

SNPs are valuable genetic markers for assessment of interindividual variability in large populations. The human HapMap project, for instance, generated large collections of SNPs for genome-wide association studies (GWAS) and haplotyping as an alternative to whole-genome sequencing (Altshuler et al. 2005). In addition, GWAS can reveal haplotypes correlating with drug response, metabolism, or other parameters of pharmacological relevance. *M. fascicularis* animals used for biomedical studies are usually F_2 offspring of wild-caught founders, and genetic variability is a serious concern in drug safety studies, in particular for animals of Asian origin, where intensive interspecies hybridization is ongoing (Tosi et al. 2002; Kanthaswamy et al. 2008).

Any comprehensive analysis of SNPs in *M. fascicularis* will require the sequencing of multiple animals. However, during the assembly process of the draft genome we realized the possibility to predict SNPs at positions where the individual is heterozygous. For genome-wide discovery of candidates, we required the presence of each allele in more than two sequencing reads and in at least a quarter of all reads covering the dimorphic nucleotide position. In addition, we only considered regions with coverage of up to 12 reads. We also ignored regions with more than one SNP within 50 bases. These filters aim to eliminate artifacts from any unrecognized duplication events. Following this procedure, we identified 2,068,743 candidate SNPs with about two-thirds corresponding to A-to-G or T-to-C transitions (Supplemental Fig. S1; Supplemental Table S1), which is a typical frequency in mammalian genomes (Harismendy et al. 2009). For validation of the SNP detection method we randomly picked 28 SNPs and amplified their flanking regions using DNA of the sequenced animal as template, followed by standard Sanger sequencing. All candidate SNPs were confirmed by this independent sequencing technology (Supplemental Table S2). More efficient high-throughput tools such as microarrays can be developed to determine genome-wide allele frequencies in large monkey-breeding cohorts or wild populations. The discovery of distinct genotypes may further reveal associations with pharmacology, drug response, or metabolism.

We note here that the population-wide analysis of genetic diversity of long-tailed macaques on Mauritius offers the unique opportunity to study the rapid expansion of an isolated primate

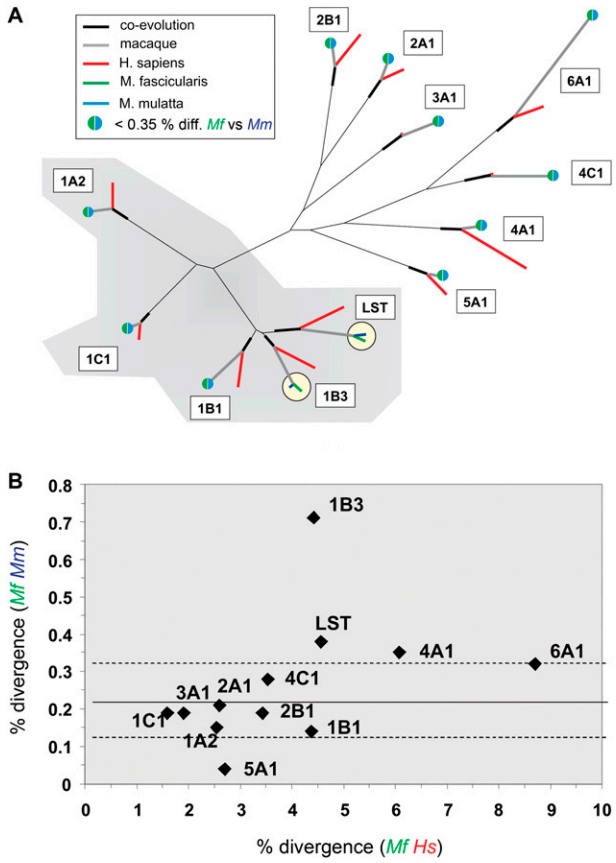

Figure 4. SLCO solute carrier gene family evolution in primates (*H. sapiens, M. mulatta*, and *M. fascicularis*). (*A*) Phylogenetic tree of SLCO divergence based on DNAML maximum likelihood analysis from the PHYLIP software package. Human orthologs of SLCO-encoding genes were identified in the draft genomes of *M. fascicularis* and *M. mulatta*, and the calculated sequence relationship is shown in "substitution events per residue" units. Thin black lines denote long-distance relationships, and bold lines are used to display closer relationships, shown at 10-fold magnification for better resolution. Gray lines and red lines mark separate evolution in macaques and humans, respectively. The drug transporters *SLCO1B3* and *LST* (RefSeq NM_001009562) show the highest degree of sequence diversity within macaques (marked by yellow circles). Differences below 0.35% within the macaques are marked by blue and green symbols. For simplicity, "*SLCO*" was omitted for labeling of individual family members. (*B*) Independent diversification of SLCO genes in primates. SLCO sequence divergence between *M. fascicularis* and human (*Mf* and *Hs*; *x*-axis), and between *M. fascicularis* and *M. mulatta* (*Mf* and *Mm*; *y*-axis) orthologs are displayed. *SLCO6A1* shows the highest divergence between *M. fascicularis* and humans, and *SLCO1B3* is the most diverse gene of the SLCO family within macaques. The black line indicates the average value and the hatched lines indicate ±SD. Note the different scales of the *x*- and *y*-axis.

population, following a bottleneck of genetic diversity that is relevant for the understanding of human evolution and diversity (Edgar et al. 2002; Zenger et al. 2003).

Expression profiling in *M. fascicularis* liver samples

To profile gene expression, we used the protein-coding transcripts predicted from the genome draft to produce a *M. fascicularis*-specific microarray by maskless photolithography (Singh-Gasson et al. 1999). We created a multiplex format with 12 arrays each containing 135,000 oligonucleotide features with three to six 60-mer probes per transcript.

For the biological validation of the microarrays, we analyzed *M. fascicularis* monkey-tissue expression profiles from liver, spleen, kidney, testes, and bladder samples. A set of genes with liver and testis-specific and robust expression was selected from the human NCBI Unigene EST database (build #224) and compared with *M. fascicularis* orthologs. These genes exhibited expression profiles in good concordance with the EST data, including housekeeping genes that showed comparable expression in all tissues (Supplemental Fig. S2).

Following biological validation, we applied the microarrays for global gene expression profiling of liver samples from 36 naïve long-tailed macaques from the Philippines, a Chinese colony, and Mauritius. We anticipated that sampling from several geographic locations would allow an overall estimate of the variability of gene expression in this most important organ for drug metabolism and clearance of toxins. We included all samples and transcripts for the statistical analysis of expression variability and we identified 93.5% of all expressed genes as low variance genes (LVGs), while 718 genes (6.5%) were scored as high-variance genes (HVGs) (Fig. 5A). Notably, variable genes were found across the entire detection range of the microarray technology and not constrained to low expressed genes, which would indicate poor sample integrity or inconsistent detection of low signal intensities. In fact, our algorithm classified highly abundant transcripts such as liver-specific metalloproteinase MT1B or uridine phosphorylase 2 (UPP2) as HVGs (Fig. 5A). Gene Ontology term-based classification of the HVG panel revealed a statistically significant enrichment of genes involved in amino acid, lipid, and sugar metabolism (Ashburner et al. 2000). A principal component analysis (PCA) with geographical origin and HVG expression as variables revealed a clear separation between the animals from Philippine breeders and primates coming from Mauritius and a Chinese colony (Fig. 5B). These features of the HVG panel in *M. fascicularis* liver point to an adaptation of gene expression to specific environmental peculiarities or diet in a specific geographical location. At this point, we cannot conclude whether genetic differences or dynamic adaptation drives expression of these genes in a particular geographic range.

Expression mode of immune response genes

Apart from genes with metabolic functions, we detected in the HVG panel a significant number of immune response genes such as cytokines, HLA transcripts, and chemokines. In preclinical and clinical drug safety studies, the plasma levels of the cytokines IL1B, IL2, IL4, IL6, IL8, IL10, IL12A, IFNG, and TNF together with the chemokine CCL2 are measured, and an increase above a critical threshold indicates activation of the innate or adaptive immune system by toxic insults. Despite the high conservation of the mammalian immune system, responses in nonhuman primates are, in general, poor predictors of human defense reactions in clinical trials, which can lead to fatal outcomes (Suntharalingam et al. 2006). Possible reasons include molecular differences in downstream signaling cascades or simply poor cross-reactivity of detection kits and assays that are usually designed and validated for human cytokines. To address potential cross-reactivity of these markers, we compared the protein sequences of the human biomarker panel mentioned above with the *M. fascicularis* and *M. mulatta* orthologs. The overall protein sequence identity with the human orthologs was at least 90%. Interestingly, some ELISA kits for human TNF and IL10 fail to cross-react with *M. fascicularis* samples despite remarkably high-sequence conservation of 97.3% or 95.9% identity, respectively. Low expression levels are an al-

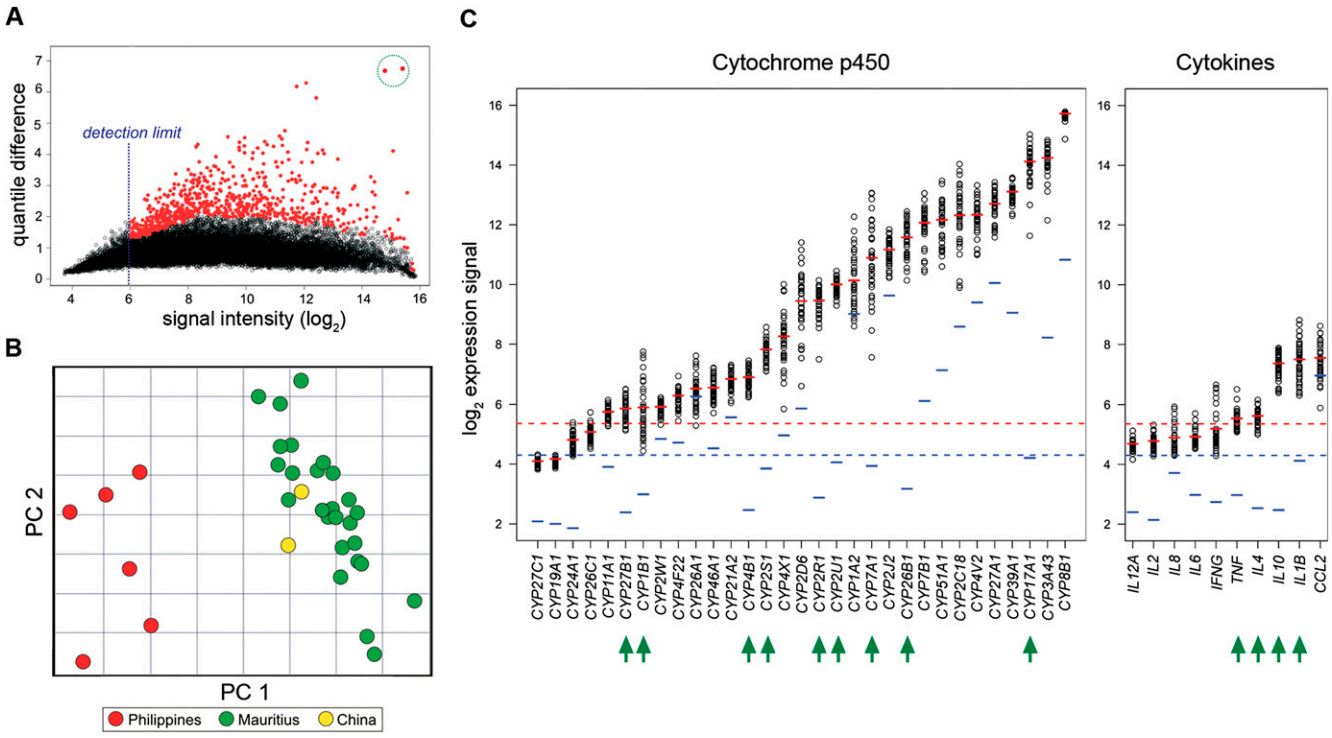

Figure 5. Microarray-based gene expression profiling in *M. fascicularis*. (*A*) Global variability of liver gene expression in 36 naive animals. Low-variance genes (LVGs; black dots) and high-variance genes (HVGs; red dots) were identified based on calculation of quantile differences of probe intensities. The 90% quantile of the $\log_2$ probe signal intensities is plotted on the *x*-axis, and the difference of the 90% and the 10% quantiles are plotted on the *y*-axis. The detection limit is marked by a dotted line. The metalloproteinase *MT1B* and the uridine phosphorylase 2 (*UPP2*) data points are denoted by a hatched green circle. (*B*) HVG-based clustering of animals according to geographical origin. Principal component analysis based on all HVG gene expression signals discriminates between animals from the Philippines (red dots) and animals from Mauritius (green dots) or a Chinese breeder (yellow dots). (*C*) Variability in baseline gene expression of cytochrome p450 isoforms and a panel of cytokines and response-related genes routinely used for drug safety assessment in humans. Scatter plots of expression levels ($\log_2$ values) of unambiguously annotated *M. fascicularis* cytochrome p450 genes (*left*) as well as key cytokines and the chemokine *CCL2* (*right*). Data are sorted according to expression levels in ascending order from *left* to *right*. Black circles indicate the expression signals of individual animals. The mean expression signal per gene is depicted by red bars for *M. fascicularis* and blue bars for *H. sapiens*. The detection limit of the microarray platforms is indicated by dotted lines (red: *M. fascicularis* NimbleGen array; blue: Affymetrix human array). Green arrows denote differences in baseline expression levels in the two species.

ternative explanation for poor detection, and the occurrence of several cytokine genes in the HVG panel prompted us to compare the liver expression levels of the diagnostic marker set at single animal resolution (Fig. 5C, right). The levels of *IL2, IL6, IL8*, and *IL12A* were close to the detection limit of the microarray platform. *IFNG, IL1B*, and *CCL2* showed considerable interanimal variation spanning at least three $\log_2$ units. For comparison with humans, we used publically available microarray data from a human liver transplant study as reference for expression and only *CCL2* was significantly expressed in this data set (de Jonge et al. 2009). The observed variability in baseline expression of primate cytokines suggests the need for quantification of predose levels of cytokines to eliminate animals with abnormal levels, especially in GLP toxicity studies with focus on immunological endpoints.

Cytochrome p450 expression

The human cytochrome p450 protein superfamily is mainly involved in the oxidation of endogenous, as well as exogenous, chemical substrates. Often such a biotransformation process requires multiple steps involving several p450s, each with a defined substrate preference. Of the 57 human CYP450s identified to date (Nelson et al. 2004), only nine isoforms (CYP1A1, CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP2E1, and CYP3A4/5/7) are capable of metabolizing at least 75% of all drugs or toxins in humans (Evans and Relling 1999). Single-nucleotide polymorphisms in several human p450s alter the properties of the protein, resulting in slow or fast metabolizers (Ingelman-Sundberg 2004). Similar biochemical consequences can be expected from significant differences in basal gene expression, causing therapeutic failure and adverse liver and renal reactions. The genome of *M. fascicularis* encodes more than 50 cytochrome p450 genes, including the major key enzymes mentioned above. Due to high-sequence similarity within this family, we could design specific microarray probes for 31 cytochrome p450 genes to assess variability of gene expression (Fig. 5C). In general, the absolute expression levels of the majority of this cytochrome p450 panel show relatively low interanimal variation. Interestingly, low interanimal variation of p450 expression was also shown at the protein level by immunoblotting in liver microsome samples from five *M. fascicularis* monkeys (Shimada et al. 1997). Moreover, a comparison of this panel with published human microarray data (de Jonge et al. 2009) reveals a well-conserved expression pattern between the two species (correlation coefficient $r = 0.73$). On the other hand, the expression levels of six *M. fascicularis* cytochromes (*CYP1B1, CYP4X1, CYP2D6, CYP1A2, CYP7A1*, and *CYP2C18*) showed significant interanimal variation spanning at

least two $\log_2$ units. Nine cytochromes show considerably different basal expression levels in humans and *M. fascicularis* monkeys (Fig. 5C, green arrows). *CYP17A1*, for instance, has high and stable expression in *M. fascicularis*, while this transcript is hardly detectable in humans. Thus, the gene expression levels of certain cytochromes p450 can complicate the interpretation of primate drug metabolism experiments with respect to their translational relevance for humans.

Table 1. Key features of novel *M. fascicularis* cytochrome p450 alleles

Gene name	Sequence identity (%)		Tissue expression preference (normalized counts)			
	M. fascicularis - rhesus	*M. fascicularis* - human	Heart	Liver	Lung	Spleen
CYP11B1	99.8	98.2	0.003	0	0.001	0
CYP11B2	99.7	94.3	0	0	0.001	0
CYP1B1	99.6	94.4	0.165	0.004	0.076	0.234
CYP20A1	99.8	97.2	0.005	0.032	0.089	0.010
CYP21A2	97.8	94.5	0.006	0.074	0.002	0.008
CYP24A1	99.4	97.1	0	0.002	0.001	0
CYP26A1	100	99.6	0.006	0.143	0.029	0.007
CYP26B1	100	99.2	0.044	0.028	0.414	0.001
CYP26C1	100	97.2	0	0	0	0
CYP27B1	100	96.2	0.001	0.002	0.006	0.003
CYP27C1	99.5	98.6	8.457	6.770	9.765	1.944
CYP2R1	99.4	97.6	0.056	0.111	0.085	0.014
CYP2U1	100	98.9	0.366	0.232	0.337	0.055
CYP2W1	99.4	94.6	0	0	0.001	0
CYP46A1	100	99.6	0.009	0	0	0
CYP4V2	100	97.1	0.918	1.683	3.054	0.339
CYP7A1	99.4	96.2	0.033	0.713	0.014	0.004
CYP7B1	100	96.8	0.026	0.074	0.010	0

Novel *M. fascicularis* CYP450 genes absent in public databases (SwissProt, TrEMBL, QFbase) were predicted as orthologs of known human p450 genes. The overall identity between the predicted protein sequences for *M. fascicularis* and rhesus as well as for *M. fascicularis* and human is shown. The digital expression levels (clone frequencies) of *M. fascicularis* CYP450 genes for heart, liver, lung, and spleen represent the average values of data from independent SAGE expression libraries constructed for each of the eight monkeys of Mauritian origin. Digital gene expression is shown as clone frequency relative to an abundant housekeeping gene (*RPL18*) in each tissue.

Discovery of novel cytochrome p450 homologs in macaques

The critical importance of cytochrome p450s in drug metabolism has prompted us to identify all human cytochrome p450 orthologs in *M. fascicularis*. This gene family is complex and characterized by many gene duplication events producing closely related paralogs, pseudogenes and truncated gene fragments (Nelson et al. 2004). Currently, the SwissProt database contains 10 annotated *M. fascicularis* p450s, whereas other public databases, like TrEMBL (Bairoch et al. 2004) or QFbase (Osada et al. 2008), contain p450 sequences or sequence fragments that point to the existence of more CYP members not yet characterized in the long-tailed macaque. Using BLAST searches with the human CYPs as query sequence, we unambiguously identified 18 novel p450 enzymes in the draft genome sequence of *M. fascicularis* (Table 1; Supplemental Appendix 1). The protein identity of these novel macaque variants is between 94% and 99% compared with human and almost 100% between *M. mulatta* and *M. fascicularis*. An exception is CYP21A1 with low identity to *M. mulatta* (97.8%). A gene-duplication event has generated a *CYP21A1*-related pseudogene in humans that is missing in the rhesus draft genome. As a consequence, genomic *M. fascicularis* reads derived either from the genuine *CYP21A1* gene or from the highly similar pseudogene will map to a single locus in the rhesus template. This generates a chimeric sequence composed of reads from both, *CYP21A1* and its pseudogene, which explains the low identity of 97.8%. *M. fascicularis* reads derived from *CYP20A1* mapped to two chromosomes in the rhesus draft genome template: Fragments from the 5′-end mapped to chromosome 12, while reads from the 3′-end mapped to chromosome 6. An assembly error in the rhesus genome probably led to a split of the *CYP20A1* gene and assignment to two different chromosomes. To test this possibility, we designed PCR primers derived from the 5′- and 3′-flanking regions of the *CYP20A1* gene. Amplification of a single 1.6-kb fragment proved localization of the *CYP20A1* gene on a contiguous chromosomal locus in *M. fascicularis*, allowing correct assembly of the coding sequence (Supplemental Appendix 1).

To verify expression of the newly annotated *M. fascicularis* cytochrome p450 variants, we used them as templates to map the reads from digital gene-expression SAGE libraries derived from four *M. fascicularis* tissues (heart, liver, lung, spleen) (Table 1). The observed expression levels of individual p450 variants vary over a range of four orders of magnitude. According to this approach, all genes with the exception of *CYP26C1* are transcribed in at least one of the tissues analyzed, and most isoforms are expressed in more than one tissue. The low expression levels detected for *CYP11B1* (11β-hydroxylase) and *CYP11B2* (aldosterone synthase) are explained by the fact that they are predominantly expressed in adrenal glands being involved in cortisol biosynthesis (Mornet et al. 1989). Recombinant expression and in silico modeling can now be applied to test or predict substrate specificity of the novel *M. fascicularis* p450 variants described here.

Discussion

In this study, we report the genome sequence of the long-tailed macaque, *M. fascicularis*, as one of the most important primate animal models in drug safety assessment. We utilized the gained knowledge to perform and analyze a number of experiments, demonstrating the need to fully characterize the relevance of this organism as a drug-safety model. We obtained a genome draft with sixfold coverage using a WGS approach with rhesus and human genomes as templates for assembly. This enabled genome-wide discovery of about 2.1 million candidate SNPs, which are important tools for genotyping and animal stratification. Furthermore, we were able to design the first genome-based *M. fascicularis* microarray, interrogating about 20,000 transcripts. We assessed global variation of gene expression in a collection of 36 individual drug-naïve liver samples. Knowledge of the *M. fascicularis* genome content and insight into gene expression will certainly improve the design and outcome of primate studies and thereby the safety of human drugs.

SNPs in the *M. fascicularis* genome

The *M. fascicularis* draft genome from a single individual enabled the identification of about 2.1 million candidate SNPs. This is an important resource for genome-wide variation association studies

and thereby significantly extends the power and resolution of genotyping approaches in population genetics or genotype–phenotype association studies. Recently, it has been shown that a relatively small panel of SNPs can be used to distinguish macaque populations according to their geographic origin (Ferguson et al. 2007). A drastically increased number of SNPs, however, will allow the identification of a plethora of markers that are necessary to study regional populations, animal ancestry and origins, and even hybrids between *M. fascicularis* and *M. mulatta*. In addition, a larger set of SNPs will facilitate the generation of *M. fascicularis*-specific microarrays for genome-wide genotype–phenotype association studies. In human populations, GWAS with more than 900,000 SNPs lead, for instance, to the identification of a polymorphism in the *SCN1A* gene that is associated with short-term memory (Papassotiropoulos et al. 2011). The identification and characterization of more SNPs in *M. fascicularis* protein-encoding genes is expected to reveal additional genotypes associated with drug response or toxicity. This could potentially enable genotyping of primates before they enter GLP-toxicity studies that are critical for transition of novel drugs into clinical development.

Comparative *M. fascicularis* transcriptome analysis

Information on diversity and specific expression of genes in *M. fascicularis* monkeys will become increasingly important for gene function studies, in general, and for drug metabolism studies, in particular. A comparison of 10,919 mRNAs between *M. fascicularis* and orthologs in rhesus revealed sequence identities above 97% in protein-coding sequences and their flanking UTRs. Human–macaque comparisons indicated high conservations in the CDSs and considerably higher divergence in the 5′- and 3′-flanking regions. Interestingly, a similar situation was described for human–chimpanzee comparisons, although the conservation in the flanking regions was considerably higher than for human–macaque (Watanabe et al. 2004; Sakate et al. 2007). Hence, our data represent an important contribution to the hypothesis that specific regulatory elements, microRNA target sites, epigenetic imprints, and other genetic factors in the flanking UTRs, rather than the primary protein sequences, cause phenotypic and physiological differences in the primate lineage. This evolutionary model was originally proposed in 1975 by King and Wilson, based on limited protein and nucleic acid comparisons in human and chimpanzee (King and Wilson 1975). More recently, this model was further supported by interspecies gene-expression analysis in the livers of humans, chimpanzees, orangutans, and rhesus macaques, demonstrating a human-specific increase in transcription-factor expression causing significant differences in gene expression (Gilad et al. 2006).

Analysis of cytochrome p450 and cytokine expression in *M. fascicularis*

Virtually all *M. fascicularis* cytochrome p450 isoforms identified in the draft genome show a relatively low degree of sequence diversity compared with their human homologs (Supplemental Appendix 1). These changes may alter the catalytic properties of the *M. fascicularis* cytochromes. In this context, it has been shown that some *M. fascicularis* p450 enzymes purified from microsomes are in general more active than their human equivalents (Shimada et al. 1997). Alternatively, different p450 expression levels in *M. fascicularis* and human liver can modulate drug metabolite profiles. About one-third of the p450s analyzed in this study are differently expressed, including the key enzyme CYP1B1. An integrative approach taking the expression level and sequence polymorphisms into account will likely lead to an improved understanding of qualitative and quantitative differences in drug metabolism between humans and *M. fascicularis* monkeys.

Elevated cytokine release in response to drug administration constitutes a serious concern for drug safety. In 2006, the T-cell stimulatory, superagonist anti-CD28 antibody TGN1412 caused a rapid onset of a massive cytokine storm with multi-organ failure in all volunteers participating in a phase I clinical trial (Suntharalingam et al. 2006). In preclinical *M. fascicularis* studies, TGN1412 was pharmacologically active and caused only moderate, dose-dependent increases in cytokine serum levels according to the Investigational Medicinal Product Dossier (TeGenero 2005). However, administration of 50 mg/kg of antibody resulted in elevated and quite variable blood IL-6 levels ranging from 24 to 390 pg/mL; this range was, on average, 17-fold higher than in the control group. These results point toward the elucidation of factors causing interanimal variation. We have shown by transcript profiling that cytokine expression shows significant interanimal variation in naïve monkeys. Thus, stratification of animals according to their basal cytokine expression profile before their assignment to experimental groups would lead to more standardized results.

RNAi therapeutics

Therapeutic small interfering RNAs (siRNAs) represent an emerging class of novel medicines (Castanotto and Rossi 2009). Once delivered to a target organ by appropriate delivery vehicles, siRNAs have the potential to silence virtually any gene, leading to efficient degradation of target mRNAs. However, due to the short length of siRNA (21 nucleotides) there is considerable concern about unspecific cross-reactivity caused by annealing to off-target transcripts. Selection of optimal therapeutic siRNA candidates is usually performed based on knock-down efficiency in cell lines, followed by in vivo efficacy and safety assessment in animal models. Knowledge of the target sequence and the possibility of genome-wide off-target site prediction in *M. fascicularis* will contribute to the design of target-specific siRNA therapeutics and more specific research tools.

M. fascicularis as a model for human drug safety

M. fascicularis is the most frequently used nonhuman primate species for drug safety evaluation, backed up by a huge body of scientific knowledge covering physiology, anatomy, drug metabolism, and other categories of biomedical interest. The lack of alternative primate species for toxicity assessment underlines the need for rational refinement of experiments, with a focus on human relevance. The genome sequence of *M. fascicularis* and its predicted proteome is therefore an essential source of mechanistic information about drug targets, biomarker expression, safety of RNAi therapeutics, signaling cascades, and drug metabolism. Translational aspects of animal experiments for humans are currently under debate among researchers and animal welfare organizations. More rational species selection and data interpretation are the focus of the international 3R initiative, which has the objective to reduce, refine, and replace animal experiments (Hartung 2010). Institutions like the European Center of the Validation of Alternative Methods (ECVAM) and all major pharmaceutical companies adhere to this program. The availability of full-genome sequences of all model organisms would certainly aid selection of the appropriate test species, and the genome of *Macaca fascicularis* is an important contribution toward completion of this goal.

Methods

Animal samples

All tissue samples used in this study were taken from untreated animals of GLP drug-safety studies in accordance with current animal welfare standards. Liver, spleen, kidney, bladder, and testis tissue samples used for microarray validation were derived from three male long-tailed macaques of Mauritian origin. The 36 liver tissue samples for microarray analysis originated from *M. fascicularis* breeding centers located in the Philippines (three females and three males), in China (one female and one male), or in Mauritius (14 females and 14 males). The organs used for confirmation of cytochrome p450 expression by transcriptome sequencing are from four male and four female animals of Mauritian origin. Details (gender, weight, age, origin) of all animals and their suppliers are on record and were part of the data submitted to public databases.

DNA and RNA sequencing

Genomic DNA (gDNA) was isolated using a commercial kit (QIAamp DNA Mini Kit; QIAGEN, Inc.) from the liver tissue of a 3-yr-old female *M. fascicularis* monkey from Mauritius. A total of 3–5 μg of gDNA was fractionated by a mechanical sheering process ("nebulization") to an average length of 400–800 bp (454-FLX Standard), 300–1500 bp (454-FLX Titanium), or 150 bp (SOLiD), and purified by size-exclusion chromatography. Briefly, following adapter ligation, single-fragment bead libraries were constructed and amplified by emulsion PCR prior to sequencing using 454-FLX (Roche/454) or SOLiD technology (Life Technologies/ABI). All reactions were carried out using commercial kits as recommended by the suppliers. The sequencing instruments (454-FLX and SOLiD) capture high-resolution images with single-bead resolution at each cycle of the sequencing process. Following several quality-assurance steps, these images were converted into machine readable (SFF, CSFASTA) or analysis program-compatible (FASTA) sequence data formats using standard software supplied by the manufacturers. Image data comparison and analysis was performed in parallel using *Beowulf* computer clusters, resulting in a significant reduction of computing time from 1 wk to about 10 h per instrument run.

RNA-sequencing was performed according to the manufacturer's protocols using the SOLiD SAGE Kit (Life Technologies/ABI). A total of 5–10 μg of total RNA was extracted from eight independent tissue samples, resulting in 32 libraries from lung, heart, liver, and spleen. Poly(A)-selected SAGE cDNA libraries of 3′ expression tags averaging 27 bp in length were constructed and subjected to SOLiD sequencing.

Mapping of sequence reads and genome assembly

The 454 reads were mapped to the rhesus reference genome (RheMac2) (Gibbs et al. 2007) using a fast word-matching algorithm. The method was optimized for rapid comparison of closely related sequences and relied on perfect matches of nonoverlapping words between read and template. Briefly, using a word size of 16 bases, for each read we required at least three words to match perfectly between read and template, otherwise the read was labeled as not mapped. Reads that failed to map to the rhesus template were mapped to the human genome (HG18). In this case, we used a shorter word size of 12 bases to take into account the expected lower sequence identity. Reads whose two best mappings differed by a word-count of less than three were considered as mapping multiple times and were excluded from further analysis. They typically corresponded to repetitive elements. Many reads covering repetitive elements could still be mapped to unique locations. SOLiD reads were mapped onto the rhesus template or to the human genome, using the ABI SOLiD SAGE Corona Lite analysis pipeline tool (version 0.40r2.0) permitting two mismatches per read.

The 454 and SOLiD reads that mapped uniquely to the rhesus genome were assembled using the rhesus sequence as template. Reads were assembled using the Roche-454 Newbler program (version 2.3), with no trimming of input sequences, no grouping of duplicate reads, and minimum overlap length of 20 bases (-notrim -ud -mL 20). Consensus contigs were merged based on sequence overlaps using PHRAP (de la Bastide and McCombie 2007).

Reads mapped onto the human genome were then inserted into the *M. fascicularis* draft in a procedure described as follows: For each incomplete region of the rhesus genome, the two human fragments that are homologous to the two rhesus genomic fragments flanking the missing sequence were located. All of the *M. fascicularis* reads that were mapped onto the intervening human genomic fragment were assembled using the human sequence as template. The resulting continuous sequence was transferred into the *M. fascicularis* genome draft in the region corresponding to rhesus' incomplete sequence.

Reads derived from RNA sequencing were mapped using ABI's SOLiD SAGE software (version 1.10), where the tag length was set to 22 bases and a maximum of two mismatches was allowed. Typically, this procedure yielded an average mapping efficiency of ~45% for all available sequencing reads.

Transcriptome prediction and comparative alignments

Transcriptome predictions were obtained using a comparative genomics approach. For each human gene, protein-coding transcripts from the NCBI RefSeq database, release 44 (Pruitt et al. 2007), were mapped to the corresponding human genomic locus using SIM4 (Florea et al. 1998) to identify the exon coordinates. Subsequently, exon coordinates have been fixed to apply RefSeq exon annotation and to accommodate noncanonical splice sites. Based on the alignment of human, *M. fascicularis*, and rhesus genomic DNA, orthologous exons were then predicted in the two macaque species and assembled into full-length transcripts.

Human mRNA sequences that have a corresponding predicted mRNA in both *M. fascicularis* and rhesus were used for 5′ UTR, CDS, and 3′ UTR identity calculations. Here, we have considered only one mRNA for each gene and, aiming at selecting complete mRNA predictions, the length difference between the aligned sequences and reference human mRNA was restricted to 10% of the reference sequence length. From 10,919 mRNA alignments, pairwise sequence identities were calculated.

SNP detection

The Newbler assembly software (Roche/454) produced detailed alignment files with stacked reads for all of the assembled regions. In the sequenced animal, each SNP allele is ideally present in 50% of the corresponding reads. Here, for SNP candidate detection, we required that each allele is covered by at least two reads and is present in at least 25% of the reads covering the position. Gaps, ambiguous bases, or trimorphic positions were not considered. We only considered regions with coverage of up to 12 reads. We also ignored regions with more than one SNP within 50 bases. These criteria lead to identification of about 2.1 million SNPs, with two-thirds showing the preferred A/G and C/T alleles. For validation of the approach, the SNP region of interest was amplified by PCR and sequenced by Sanger-based capillary sequencing using standard procedures.

Microarray design and selection

Two versions of Gene Expression Microarrays were designed for whole-transcriptome profiling. Version I was based on the predicted transcriptome using the threefold coverage draft genome, and version II was designed on the final genome version with sixfold coverage. Version II microarrays contained essentially the same number of annotated transcripts as version I, but fewer errors due to improved coverage. In addition, version II microarrays contained six 60-mer probes per probe set, whereas version I arrays contained only three to five probes per probe set. Probes for the microarray were selected based on uniqueness scores and 24-bp composition rules (Selzer et al. 2005). About 88% of all microarray probes were SNP free. The version II microarray contained 20,047 probe sets that measured the expression levels of 16,896 genes, including splice variants.

Microarray-based gene expression profiling

Total RNA from deep-frozen tissue pieces ($3 \times 3 \times 3$ mm^3) was extracted using the RNeasy Mini kit combined with DNase treatment on a solid support (QIAGEN, Inc.). RNA quality assessment and quantification was performed using microfluidic chip analysis on an Agilent 2100 Bioanalyzer (Agilent Technologies, Inc.). For the biological validation of version I microarrays, 10 μg of total RNA were reverse transcribed using the SuperScript double-stranded cDNA synthesis kit (Life Technologies/Invitrogen, Inc.). After precipitation, double-stranded cDNA was labeled with Cy3 by using the Roche NimbleGen One color DNA Labeling Kit (Roche NimbleGen, Inc.). For expression profiling of the 36 liver samples on version II microarrays, 50 ng of liver total RNA was reverse transcribed and amplified by using the NuGEN WT-Ovation Pico RNA Amplification System (NuGEN Technologies, Inc.), followed by labeling as described above. Microarrays were hybridized with 4 μg of Cy3-labeled cDNA for 16–20 h at 42°C and were washed and dried according to the manufacturer's instruction. For validation of array version I, three technical replicates per tissue sample were used to gain statistical power (data not shown). Microarray data was collected by confocal scanning using the Roche NimbleGen MS200 Microarray Scanner at 2-μm pixel resolution (Roche NimbleGen, Inc.). NimbleGen probe intensities were subjected to background correction, quantile normalization (Bolstad et al. 2003), and Robust Multi-Array Analysis (RMA) (Irizarry et al. 2003a,b) as implemented in the NimbleScan Software, version 2.6 (Roche NimbleGen, Inc.). Averaged gene-level signal intensities were summarized into gene calls and $\log_2$ transformed. For version I microarrays, gene calls were averaged across technical replicates. Data analysis was performed using Partek Genomics Suites (Partek, Inc.) and the R software for statistical computing and graphics (R-Development-Core-Team 2008). The detection limit of NimbleGen microarrays was determined at the 95th percentile with respect to the signals of the random probes present on the array.

Human reference microarray data analysis

For comparison of *M. fascicularis* with human liver expression values we used a published data set from NCBI Gene Expression Omnibus (accession no. GSE12720) (de Jonge et al. 2009). Data was generated using Affymetrix HG-U133 Plus 2.0 GeneChips and we applied in-house quality control algorithms to select high-quality expression data. In addition, we excluded any data of living donors whose liver has been manipulated or HCV infected, as well as data from diseased donors with hepatocellular damage. Only four out of 63 available samples passed these filter criteria, and we used RMA normalized mean signals to generate the $\log_2$-converted expression data. The detection limit of Affymetrix microarrays was determined as the 95th percentile relative to the signals from probe sets that failed to map to unique locations in the human genome and served as negative controls.

Global assessment of gene expression variability

For the detection of genes with high interanimal variation, we used an algorithm based on the 10% and 90% quantiles of all individual probe intensity distributions. Briefly, gene expression values ($\log_2$) for all transcripts and all 36 monkeys were distributed into 40 bins, and an upper threshold for the expected variability was calculated for each bin. This procedure yielded 718 high-variance genes (HGVs) using the selected input parameters of the algorithm. The remaining genes were classified as low-variance genes (LVGs).

Principal component analysis (PCA)

PCA of the expression data set was performed with SIMCA-P+ version 12.0.1 (Umetrics AB) using expression values of HVGs. GeneOntology enrichment analysis of HVGs was estimated using Fisher's exact test.

Data access

The *Macaca fascicularis* genome draft sequence has been submitted to the European Nucleotide Archive of EMBL-EBI (http://www.ebi.ac.uk/ena/). The 21 chromosomes have the accession numbers FR874244–FR874264. The microarray data from this study have been deposited at the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.gov/geo/) (Edgar et al. 2002) under accession number GSE30184. The *Macaca fascicularis* candidate SNPs reported in this study have been deposited in the NCBI dbSNP database, build 135 (http://www.ncbi.nlm.nih.gov/projects/SNP/). The Submitter SNP (ss) accession numbers that were assigned by dbSNP are listed in Supplemental Table S1.

Acknowledgments

We thank Dr. Michelle Browner for scientific discussions and significant contributions to the final version of the paper. We thank Judith Knehr for efficient resequencing of the SNP amplicons and Dr. Jean-Christophe Hoflack for supplying well-documented *M. fascicularis* liver samples. We are grateful to Professor Gerhard Weinbauer for discussions on macaque physiology and use of this species in biomedical research. We thank Drs. Michael Otteneder and Lutz Müller for discussions on drug safety and Dr. Laura Burleigh for thorough editing of the manuscript and scientific input. Finally, we appreciate the continued interest of Dr. Jean-Jacques Garaud in the publication of this work and managerial support. This work was completely funded by F. Hoffmann-La Roche AG Basel, Switzerland.

Authors' contributions: M.E., E.K., and A.S. performed the research. L.I. and T.A. designed the microarray. C.B., G.S., R.S., T.H., and M.B. analyzed the data. T.S. provided scientific input for drug safety related topics of this work and support for this genome sequencing project. U.C. initiated and managed the project. U.C., M.B., and T.H. wrote the paper.

References

Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly JM, Donnelly P, Gibbs RA, Yang H, Zeng C, Gabriel SB, et al. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the

unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25–29.
Bairoch A, Boeckmann B, Ferro S, Gasteiger E. 2004. Swiss-Prot: juggling between evolution and stability. *Brief Bioinform* **5:** 39–55.
Boelsterli UA. 2003. Animal models of human disease in drug safety assessment. *J Toxicol Sci* **28:** 109–121.
Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19:** 185–193.
Capitanio JP, Emborg ME. 2008. Contributions of non-human primates to neuroscience research. *Lancet* **371:** 1126–1135.
Castanotto D, Rossi JJ. 2009. The promises and pitfalls of RNA-interference-based therapeutics. *Nature* **457:** 426–433.
Chu EH, Bender MA. 1961. Chromosome cytology and evolution in primates. *Science* **133:** 1399–1405.
de Jonge J, Kurian S, Shaked A, Reddy KR, Hancock W, Salomon DR, Olthoff KM. 2009. Unique early gene expression patterns in human adult-to-adult living donor liver grafts compared to deceased donor grafts. *Am J Transplant* **9:** 758–772.
de la Bastide M, McCombie WR. 2007. Assembling genomic DNA sequences with PHRAP. *Curr Protoc Bioinformatics* **17:** 11.4.1–11.4.15.
Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30:** 207–210.
Evans WE, Relling MV. 1999. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **286:** 487–491.
Ferguson B, Street SL, Wright H, Pearson C, Jia Y, Thompson SL, Allibone P, Dubay CJ, Spindel E, Norgren RB Jr. 2007. Single nucleotide polymorphisms (SNPs) distinguish Indian-origin and Chinese-origin rhesus macaques (*Macaca mulatta*). *BMC Genomics* **8:** 43. doi: 10.1186/1471-2164-8-43.
Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* **8:** 967–974.
Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316:** 222–234.
Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP. 2006. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* **440:** 242–245.
Hagenbuch B, Meier PJ. 2003. The superfamily of organic anion transporting polypeptides. *Biochim Biophys Acta* **1609:** 1–18.
Hamada A, Sissung T, Price DK, Danesi R, Chau CH, Sharifi N, Venzon D, Maeda K, Nagao K, Sparreboom A, et al. 2008. Effect of SLCO1B3 haplotype on testosterone transport and clinical outcome in caucasian patients with androgen-independent prostatic cancer. *Clin Cancer Res* **14:** 3312–3318.
Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, et al. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* **10:** R32. doi: 10.1186/gb-2009-10-2-r32.
Hartung T. 2010. Lessons learned from alternative methods and their validation for a new toxicology in the 21st century. *J Toxicol Environ Health B Crit Rev* **13:** 277–290.
Ingelman-Sundberg M. 2004. Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past, present and future. *Trends Pharmacol Sci* **25:** 193–200.
Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. 2003a. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31:** e15. doi: 10.1092/nar/gng015.
Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. 2003b. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4:** 249–264.
Kanthaswamy S, Satkoski J, George D, Kou A, Erickson BJ, Smith DG. 2008. Interspecies hybridization and the stratification of nuclear genetic variation of Rhesus (*Macaca Mulatta*) and long-tailed Macaques (*Macaca Fascicularis*). *Int J Primatol* **29:** 1295–1311.
King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188:** 107–116.
Magness CL, Fellin PC, Thomas MJ, Korth MJ, Agy MB, Proll SC, Fitzgibbon M, Scherer CA, Miner DG, Katze MG, et al. 2005. Analysis of the *Macaca mulatta* transcriptome and the sequence divergence between *Macaca* and human. *Genome Biol* **6:** R60. doi: 10.1186/gb-2005-6-7-r60.
Mornet E, Dupont J, Vitek A, White PC. 1989. Characterization of two genes encoding human steroid 11 beta-hydroxylase (P-450(11) beta). *J Biol Chem* **264:** 20961–20967.
Nelson DR, Zeldin DC, Hoffman SM, Maltais LJ, Wain HM, Nebert DW. 2004. Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. *Pharmacogenetics* **14:** 1–18.
Osada N, Hashimoto K, Kameoka Y, Hirata M, Tanuma R, Uno Y, Inoue I, Hida M, Suzuki Y, Sugano S, et al. 2008. Large-scale analysis of *Macaca fascicularis* transcripts and inference of genetic divergence between *M. fascicularis* and *M. mulatta*. *BMC Genomics* **9:** 90. doi: 10.1186/1471-2164-9-90.
Papassotiropoulos A, Henke K, Stefanova E, Aerni A, Muller A, Demougin P, Vogler C, Sigmund JC, Gschwind L, Huynh KD, et al. 2011. A genome-wide survey of human short-term memory. *Mol Psychiatry* **16:** 184–192.
Perelman P, Johnson WE, Roos C, Seuanez HN, Horvath JE, Moreira MA, Kessing B, Pontius J, Roelke M, Rumpler Y, et al. 2011. A molecular phylogeny of living primates. *PLoS Genet* **7:** e1001342. doi: 10.1371/journal.pgen.1001342.
Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35:** D61–D65.
R-Development-Core-Team. 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
Sakate R, Suto Y, Imanishi T, Tanoue T, Hida M, Hayasaka I, Kusuda J, Gojobori T, Hashimoto K, Hirai M. 2007. Mapping of chimpanzee full-length cDNAs onto the human genome unveils large potential divergence of the transcriptome. *Gene* **399:** 1–10.
Selzer RR, Richmond TA, Pofahl NJ, Green RD, Eis PS, Nair P, Brothman AR, Stallings RL. 2005. Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosomes Cancer* **44:** 305–319.
Shimada T, Mimura M, Inoue K, Nakamura S, Oda H, Ohmori S, Yamazaki H. 1997. Cytochrome P450-dependent drug oxidation activities in liver microsomes of various animal species including rats, guinea pigs, dogs, monkeys, and humans. *Arch Toxicol* **71:** 401–408.
Singh-Gasson S, Green RD, Yue Y, Nelson C, Blattner F, Sussman MR, Cerrina F. 1999. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat Biotechnol* **17:** 974–978.
Street SL, Kyes RC, Grant R, Ferguson B. 2007. Single nucleotide polymorphisms (SNPs) are highly conserved in rhesus (*Macaca mulatta*) and cynomolgus (*Macaca fascicularis*) macaques. *BMC Genomics* **8:** 480. doi: 10.1186/1471-2164-8-480.
Suntharalingam G, Perry MR, Ward S, Brett SJ, Castello-Cortes A, Brunner MD, Panoskaltsis N. 2006. Cytokine storm in a phase 1 trial of the anti-CD28 monoclonal antibody TGN1412. *N Engl J Med* **355:** 1018–1028.
TeGenero AG. 2005. *TGN1412 Investigational Medicinal Product Dossier*. Medicines and Healthcare products Regulatory Agency (UK).
Tosi AJ, Morales JC, Melnick DJ. 2002. Y-chromosome and mitochondrial markers in *Macaca fascicularis* indicate introgression with Indochinese *M. mulatta* and a biogeographic barrier in the Isthmus of Kra. *Int J Primatol* **23:** 161–178.
Watanabe H, Fujiyama A, Hattori M, Taylor TD, Toyoda A, Kuroki Y, Noguchi H, BenKahla A, Lehrach H, Sudbrak R, et al. 2004. DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* **429:** 382–388.
Waterston RH, Lander ES, Wilson RK, Mikkelsen TS, Hillier LD, Eichler EE, Zody MC, Jaffe DB, Yang S-P, Enards W, et al. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87.
Weinbauer GF, Niehoff M, Niehaus M, Srivastav S, Fuchs A, Van Esch E, Cline JM. 2008. Physiology and endocrinology of the ovarian cycle in Macaques. *Toxicol Pathol* **36:** 7S–23S.
Zenger KR, Richardson BJ, Vachot-Griffin AM. 2003. A rapid population expansion retains genetic diversity within European rabbits in Australia. *Mol Ecol* **12:** 789–794.

Received March 9, 2011; accepted in revised form July 11, 2011.

Genome-based analysis of the nonhuman primate *Macaca fascicularis* as a model for drug safety assessment

Martin Ebeling, Erich Küng, Angela See, et al.

Genome Res. 2011 21: 1746-1756 originally published online August 23, 2011
Access the most recent version at doi:10.1101/gr.123117.111