MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 5328–5331

NEWS AND VIEWS

PERSPECTIVE

Noninvasive genome sampling in chimpanzees

MICHAEL H. KOHN

Department of Ecology and Evolutionary Biology, Rice University, 6100 Main Street, MS 170, Houston, TX 77005, USA

Abstract

The inevitable has happened: genomic technologies have been added to our noninvasive genetic sampling repertoire. In this issue of *Molecular Ecology*, Perry *et al.* (2010) demonstrate how DNA extraction from chimpanzee faeces, followed by a series of steps to enrich for target loci, can be coupled with next-generation sequencing. These authors collected sequence and single-nucleotide polymorphism (SNP) data at more than 600 genomic loci (chromosome 21 and the X) and the complete mitochondrial DNA. By design, each locus was 'deep sequenced' to enable SNP identification. To demonstrate the reliability of their data, the work included samples from six captive chimps, which allowed for a comparison between presumably genuine SNPs obtained from blood and potentially flawed SNPs deduced from faeces. Thus, with this method, anyone with the resources, skills and ambition to do genome sequencing of wild, elusive, or protected mammals can enjoy all of the benefits of noninvasive sampling.

Keywords: chimpanzee, conservation, faeces, genomics, next-generation sequencing, *Pan troglodytes*

Received 15 September 2010; revision received 24 September 2010; accepted 26 September 2010

The noninvasive genetic sampling of faeces (Kohn & Wayne 1997; Waits & Paetkau 2005) has been a wonderful complement to observational studies on chimp and bonobo troops (Gerloff *et al.* 1999; Constable *et al.* 2001), including Jane Goodall's famous chimps of Gombe (Fig. 1). Genetic data have enabled analyses of kinship and reproductive success, which in conjunction with the observations of social behaviour and rank, have taught us much about chimp societies (Vigilant & Guschanski 2009). Noninvasive genetic sampling has allowed researchers to avoid handling the chimps, thus preserving the trust level between researchers and habituated troops. Noninvasive sampling has enabled the gathering of at least some data on nonhabituated troops also (McGrew *et al.* 2004).

Fig. 1 Chimps (*Pan troglodytes*) such as this one from Gombe Streams National Park, Tanzania, should be given a lot of credit in terms of their imagination beyond mere instinct. However, despite the seemingly reflective facial expression, this chimp can hardly imagine what we already had found out about his kind based on noninvasive genetic analyses of their faeces collected from the wild, and what we might learn from their noninvasive genomic sampling next. Photographs credit: Leanne T. Nash, Arizona State University.

Another area that has benefited from noninvasive genetic sampling is the study of the origins of AIDS. Chimps have been accused, and acquitted based on genetic data obtained from faeces, of serving as one of the reservoirs of simian immunodeficiency virus (SIV) (Keele *et al.* 2006; Sharp & Hahn 2010). Contact with chimp bush meat—which, sadly, remains a frequent event—seems to be a plausible route that enabled the virus to jump hosts and to evolve into what in now known as various forms of HIV1. To pinpoint areas and populations of chimps that harbour

Correspondence: Michael H. Kohn, Fax: +1 713 348 5232; E-mail: hmkohn@rice.edu

the likely source strain of SIV required the analysis of huge numbers of samples dispersed across the chimp's geographic range. Such sampling would have been difficult to accomplish with blood or tissue samples.

Noninvasive conservation genetic approaches have relied on the PCR amplification of a few mitochondrial DNA loci, X and/or Y chromosome markers, select nuclear genes, and typically a dozen or so microsatellites and/or single-nucleotide polymorphisms (SNPs). Now, Perry *et al.* (2010) have tapped into the genomic toolbox and demonstrated that the sequencing of hundreds of loci can be performed from the faecal samples. Their approach could be modified to include the entire genome, all protein coding genes or any genomic regions of interest to conservation biologists (e.g., the MHC). It should be noted that at least 17 faecal metagenomic projects are currently underway [Gold, accessed September 2010; (Liolios *et al.* 2010)]. However, metagenomic projects sequence everything in the sample and sort through the sequences later; Perry *et al.* are the first to use next-generation sequencing to sample the endogenous DNA of the animal that dropped the poop (Fig. 2).

Noninvasive genome sequencing requires expertise both at the bench, as is evident in Fig. 2, and with the computer. The sequencing itself is the easiest molecular step involved, and it can be outsourced to commercial or collaborating laboratories. The molecular methods can probably be established in many laboratories, but are clearly more involved than noninvasive genetic sampling protocols employing PCR. Because of the novelty of this work, it is necessary to clearly describe protocols until genomic literacy is more commonplace. Furthermore, any modifications that might cut costs would make this approach more palatable to conservation biologists.

The effort required to conduct projects like these is not fully reflected by the overview of molecular steps shown in Fig. 2. What is not depicted are dozens of bioinformatics steps. Here, Perry *et al.* have set a good example by providing access to some of their scripts, the generation of which can draw considerable time and resources. For instance, the steps preceding one include the choice of genomic regions for which to design baits. This appears to be the main hurdle to replicating this work in other species, as it requires a fully sequenced genome. Perry *et al.* express their confidence that the genomes of 10 000 vertebrate species will be sequenced soon (Haussler *et al.* 2009). However, most of these genome sequences will have less coverage and less-sophisticated annotation than the chimpanzee genome. Genome-enabled species, i.e. those closely related to a spe-

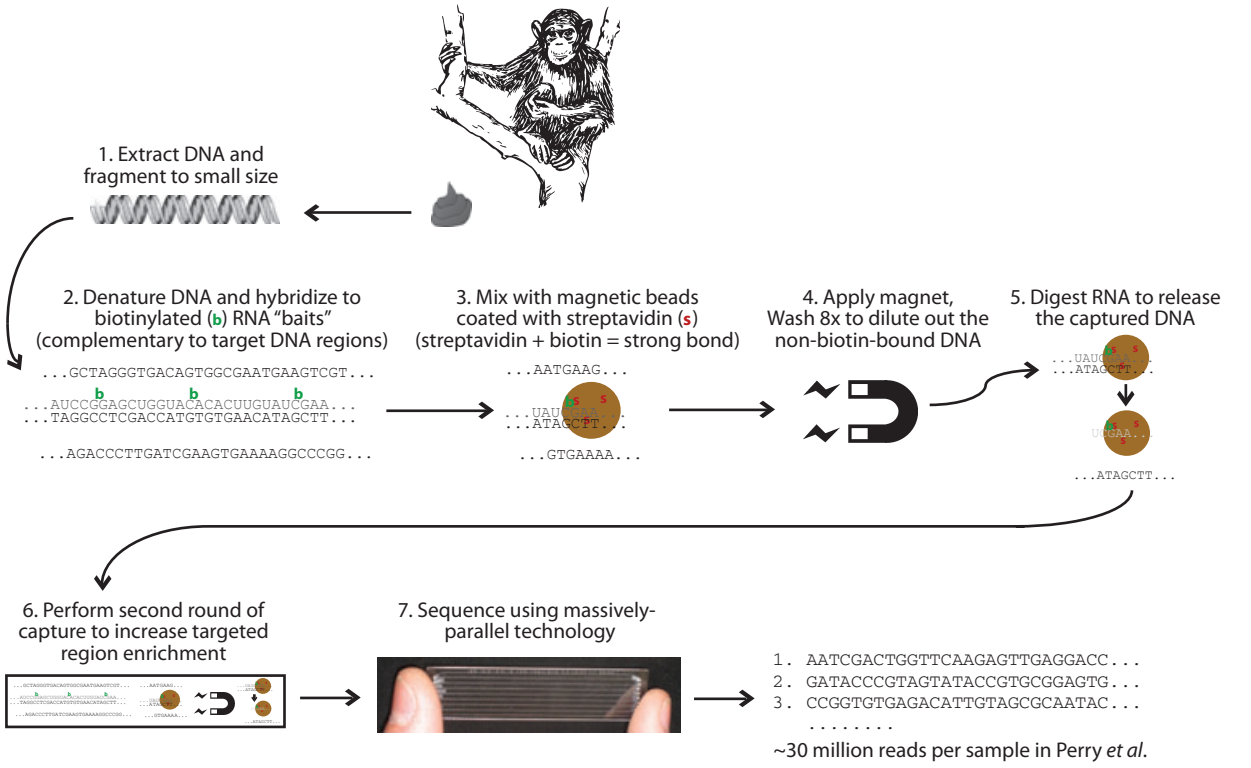

Fig. 2 Perry *et al.* collected matched blood and faecal samples from captive chimpanzees, extracted the DNA, sheared it, enriched the sheared DNA for small fragments on gels and amplified the DNA using primers that match ligated linkers (step 1). Subsequently, the desired size-selected DNA fragments (endogenous DNA and targeted loci) are selected with a series of steps (2–5). The target fragments are bound to magnetic molecular complexes made up of biotin, modified streptavidin and the bait sequence. These can then be separated from undesirable unbound DNA—from bacteria and food— or endogenous DNA that is irrelevant to the study. Perry *et al.* repeated steps 2–5 twice, adjusting the stringency of the washing steps. After these steps, second-generation technology was used to collect the sequencing data (step 7). Figure credit: George H. Perry, University of Chicago.

cies with a sequenced genome, could also qualify for noninvasive genome sampling. Modified protocols would need to be developed to study such species and would likely target only slowly evolving genomic loci, i.e. those loci for which useful baits can be designed based on the available genome sequence of a species closely related to the study species with no genome sequence published.

An even larger number of bioinformatic steps follow the molecular work (Perry *et al.* 2010). The obtained sequences need to be assembled and aligned to the orthologous region of the genome, sequence coverage must be determined, and any discrepancies between the collected and the published sequences must be clarified. This requires the application of certain criteria to filter the data. For example, SNPs should only be scored if they have been obtained from regions with good coverage. Perry *et al.* opted to set this cut-off at 20 × coverage per locus, which equates to 10 × for each DNA strand. Considering such coverage per strand can reduce the risk of SNP calling error introduced by rare, strand-specific biases in next-generation sequencing.

Moreover, each allele of a heterozygous chimp should, theoretically, be present in 50% of all reads covering a locus. In homozygous chimps, only one allele should be present. However, the stochastic processes of sequencing and different affinities of baits for individual alleles (step 2 in Fig. 2) can skew allele frequencies. Perry *et al.* chose to label chimps as genuinely heterozygous when one allele accounts for less than 80% of the reads at that locus. The plots in Fig. 1A in Perry *et al.* depicting the proportion of loci in which the common nucleotide has a particular frequency reveal a pattern that enables quality control. One peak represents loci that correspond to the 50/50 ratio expected for heterozygous animals; the other peak represents homozygous animals. Anything in between could be false heterozygotes or homozygotes. Whether the 80/20 cut-offs used to call heterozygotes chosen by Perry *et al.* worked well for this study in particular or whether it could be generally useful during other noninvasive genome sampling studies (possibly employing different next-generation sequencing platforms) remains to be seen. However, in this study, the application of this cut-off value resulted in impressively low numbers of questionable SNPs, such as a mere two of >400 000 sites studied on the X chromosome of a male chimp.

Moreover, the results obtained from faeces and from blood correspond very well. There were comparable levels of genetic diversity in matched faecal and blood samples and comparable levels of variation across faecal samples and published data on diversity in chimpanzees. The authors also employed a clever trick based on the analysis of the X chromosome of male chimps to validate this method. The male X is hemizygous, and thus, while males can differ from the published genome sequence of the X, all reads covering a locus of a male chimp on the X should be identical. Any variations thus must be false-positive SNPs. Perry *et al.* encountered very few (<0.001%) such false positives. Finally, the authors used PCR-based sequencing at some loci to compare SNP accuracy by both methods. A phylogenetic 'test' provided the reassuring result that matched blood and faecal samples cluster together.

Technical issues associated with noninvasive genetic sampling have been covered thoroughly in the literature (Taberlet *et al.* 1996, 1999) and in the context of studies of chimps specifically (Bradley *et al.* 2001; Morin *et al.* 2001; Vigilant 2002; Knapp 2005). Variations on some of these biases will likely persist even after technologies have changed. The most obvious limitation of this approach is that DNA isolated from faeces is generally of poor quality. Some species in fact are notoriously difficult to study by noninvasive genetic sampling. River otters, for example, have frustrated scientists in this regard. Faeces are simply not well preserved in semi-aquatic habitats. Any effects seen previously at individual loci, such as allelic dropout, can now potentially occur at thousands of loci. However, the results by Perry *et al.* are encouraging in this regard. Nuclear mitochondrial insertions (Numts) and duplicated genes may cause confusion as to whether SNPs are genuine or the result of comparisons between nonorthologous loci. Therefore, the bioinformatics leading to the selection of baits should be taken seriously. This will be more of a problem for poorly annotated genomes than for well-annotated genomes. Furthermore, carnivore faeces would contain DNA from prey, which could present difficulties; for example, chimpanzees are known to hunt monkeys, so highly conserved genes may contaminate enriched DNA even after capture with baits. A series of other issues surely merit contemplation. However, in all, the noninvasive genome sampling approach first implemented by Perry *et al.* will likely be considered reliable, and thus, in principle available to pursue many potentially interesting conservation genomic research avenues in primates (Vigilant & Guschanski 2009) or in general (Hedrick 2001; Kohn *et al.* 2006).

Which applications merit the cost and efforts required? Elusive animals may already be studied with a suite of 'old-fashioned' noninvasive genetic sampling approaches, and many may argue that these meet the needs of conservation biologists. Thus, it seems unlikely that the genome-scale approach will be used in the near term to sequence through hundreds of faecal samples collected in the wild simply to increase the scale of conservation projects.

Perry *et al.* rightfully emphasize that the noninvasive genome sampling of animal faeces might also find applications in molecular ecological and population genetic studies in general; not only in conservation. In these research areas, a single reference genome sequence is an asset for several obvious reasons, but is of limited value for illustrating the variation within and between populations. Thus, the most obvious application of noninvasive genome sampling is to achieve more comprehensive geographic sampling of genetic variation of rare, elusive and protected species.

The benefits of having genome sequences from multiple humans have been reviewed. Having multiple sequences

for the chimp, which has served as an important outgroup in the analysis of human genetic variation, may be similarly advantageous; noninvasive genome sampling could be used to this end. Furthermore, the information currently available on human variation could help us understand the population genetic and demographic history of chimps once this broader sampling has been achieved. If it would indeed be desirable to have multiple chimp sequences, should these be generated from faeces rather than blood? Which chimps should be chosen? To the James Watsons and Craig Venters of the chimp world: step out of the shadows of the deep African forests for sequencing!

There are a number of minor yet notable differences among chimp cultures, such as the use of tools. Perhaps, this indeed is a good place to start looking for the population genetic consequences associated with the origin of culture and new skills (Langergraber *et al.* 2010). Another issue that might merit genome sequencing concerns the possible variations of HIV susceptibility and disease course among wild chimpanzees. And then there are these scary, giant, lion-eating chimps first mentioned by the late Shelly Williams, a primatologist—are these really just a subspecies of chimp as deduced from ordinary noninvasive genotyping (c.f. Young & Bennet 2006)? Could noninvasive genome sampling reveal whether they are genetically unique? Or were they merely invented by local parents inhabiting villages in the Bili Forest, Democratic Republic of the Congo, to scare their children into behaving?

Ideas generate technologies, and in turn, as the genome-sequencing era has shown, technologies can drive ideas. Now that noninvasive genome sampling technology is available, we are free to come up with ideas for its use, but let us be thoughtful—sequencing a critically endangered species' genome does not save it from extinction in the wild.

References

Bradley BJ, Chambers KE, Vigilant L (2001) Accurate DNA-based sex identification of apes using non-invasive samples. *Conservation Genetics*, **2**, 179–181.

Constable JL, Ashley MV, Goodall J, Pusey AE (2001) Noninvasive paternity assignment in Gombe chimpanzees. *Molecular Ecology*, **10**, 1279–1300.

Gerloff U, Hartung B, Fruth B, Hohmann G, Tautz D (1999) Intracommunity relationships, dispersal pattern and paternity success in a wild living community of Bonobos (*Pan paniscus*) determined from DNA analysis of faecal samples. *Proceedings. Biological sciences/The Royal Society*, **266**, 1189–1195.

Haussler D, O'Brien SJ, Ryder OA *et al.* (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, **100**, 659–674.

Hedrick PW (2001) Conservation genetics: where are we now? *Trends in Ecology & Evolution*, **16**, 629–636.

Keele BF, Van Heuverswyn F, Li YY *et al.* (2006) Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science*, **313**, 523–526.

Knapp LA (2005) Facts, faeces and setting standards for the study of MHC genes using noninvasive samples. *Molecular Ecology*, **14**, 1597–1599.

Kohn MH, Wayne RK (1997) Facts from feces revisited. *Trends in Ecology and Evolution*, **6**, 223–227.

Kohn MH, Murphy WJ, Ostrander EA, Wayne RK (2006) Genomics and conservation genetics. *Trends in Ecology & Evolution*, **21**, 629–637.

Langergraber KE, Boesch C, Inoue E *et al.* (2010) Genetic and 'cultural' similarity in wild chimpanzees. *Proceedings. Biological sciences/The Royal Society*, doi: 10.1098/rspb.2010.1112.

Liolios K, Chen IM, Mavromatis K *et al.* (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, **38**, D346–D354.

McGrew WC, Ensminger AL, Marchant LF, Pruetz JD, Vigilant L (2004) Genotyping aids field study of unhabituated wild chimpanzees. *American Journal of Primatology*, **63**, 87–93.

Morin PA, Chambers KE, Boesch C, Vigilant L (2001) Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglodytes* verus). *Molecular Ecology*, **10**, 1835–1844.

Perry G, Marioni J, Melsted P, Gilad Y (2010) Genomic-scale capture and sequencing of endogenous DNA from feces. *Molecular Ecology*, **19**, 5332–5344.

Sharp M, Hahn BH (2010) The evolution of HIV-1 and the origin of AIDS. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **365**, 2487–2494.

Taberlet P, Griffin S, Goossens B *et al.* (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, **24**, 3189–3194.

Taberlet P, Waits LP, Luikart G (1999) Noninvasive genetic sampling: look before you leap. *Trends in Ecology & Evolution*, **14**, 323–327.

Vigilant L (2002) Technical challenges in the microsatellite genotyping of a wild chimpanzee population using feces. *Evolutionary Anthropology*, **11**, 162–165.

Vigilant L, Guschanski K (2009) Using genetics to understand the dynamics of wild primate populations. *Primates*, **50**, 105–120.

Waits LP, Paetkau D (2005) Noninvasive genetic sampling tools for wildlife biologists: a review of applications and recommendations for accurate data collection. *The Journal of Wildlife Management*, **69**, 1419–1433.

Young E, Bennet A (2006) DNA tests solve mystery of giant apes. *New Scientist*, **2558**, 32.

Michael H. Kohn is studying evolutionary genomics, conservation genetics, and medical genetics. The detection of adaptations at the level of genes is one main area of interest that is studied in rat populations resistant to anticoagulant rodenticides (warfarin).

---

doi: 10.1111/j.1365-294X.2010.04889.x