LETTER

doi:10.1038/nature09687

Comparative and demographic analysis of orang-utan genomes

Devin P. Locke[1], LaDeana W. Hillier[1], Wesley C. Warren[1], Kim C. Worley[2], Lynne V. Nazareth[2], Donna M. Muzny[2], Shiaw-Pyng Yang[1], Zhengyuan Wang[1], Asif T. Chinwalla[1], Pat Minx[1], Makedonka Mitreva[1], Lisa Cook[1], Kim D. Delehaunty[1], Catrina Fronick[1], Heather Schmidt[1], Lucinda A. Fulton[1], Robert S. Fulton[1], Joanne O. Nelson[1], Vincent Magrini[1], Craig Pohl[1], Tina A. Graves[1], Chris Markovic[1], Andy Cree[2], Huyen H. Dinh[2], Jennifer Hume[2], Christie L. Kovar[2], Gerald R. Fowler[2], Gerton Lunter[3,4], Stephen Meader[3], Andreas Heger[3], Chris P. Ponting[3], Tomas Marques-Bonet[5,6], Can Alkan[5], Lin Chen[5], Ze Cheng[5], Jeffrey M. Kidd[5], Evan E. Eichler[5,7], Simon White[8], Stephen Searle[8], Albert J. Vilella[9], Yuan Chen[9], Paul Flicek[9], Jian Ma[10]†, Brian Raney[10], Bernard Suh[10], Richard Burhans[11], Javier Herrero[9], David Haussler[10], Rui Faria[6,12], Olga Fernando[6,13], Fleur Darré[6], Domènec Farré[6], Elodie Gazave[6], Meritxell Oliva[6], Arcadi Navarro[6,14], Roberta Roberto[15], Oronzo Capozzi[15], Nicoletta Archidiacono[15], Giuliano Della Valle[16], Stefania Purgato[16], Mariano Rocchi[15], Miriam K. Konkel[17], Jerilyn A. Walker[17], Brygg Ullmer[18], Mark A. Batzer[17], Arian F. A. Smit[19], Robert Hubley[19], Claudio Casola[20], Daniel R. Schrider[20], Matthew W. Hahn[20], Victor Quesada[21], Xose S. Puente[21], Gonzalo R. Ordoñez[21], Carlos López-Otín[21], Tomas Vinar[22], Brona Brejova[22], Aakrosh Ratan[11], Robert S. Harris[11], Webb Miller[11], Carolin Kosiol[23], Heather A. Lawson[24], Vikas Taliwal[25], André L. Martins[25], Adam Siepel[25], Arindam RoyChoudhury[26], Xin Ma[25], Jeremiah Degenhardt[25], Carlos D. Bustamante[27], Ryan N. Gutenkunst[28], Thomas Mailund[29], Julien Y. Dutheil[29], Asger Hobolth[29], Mikkel H. Schierup[29], Oliver A. Ryder[30], Yuko Yoshinaga[31], Pieter J. de Jong[31], George M. Weinstock[1], Jeffrey Rogers[2], Elaine R. Mardis[1], Richard A. Gibbs[2] & Richard K. Wilson[1]

'Orang-utan' is derived from a Malay term meaning 'man of the forest' and aptly describes the southeast Asian great apes native to Sumatra and Borneo. The orang-utan species, *Pongo abelii* (Sumatran) and *Pongo pygmaeus* (Bornean), are the most phylogenetically distant great apes from humans, thereby providing an informative perspective on hominid evolution. Here we present a Sumatran orang-utan draft genome assembly and short read sequence data from five Sumatran and five Bornean orang-utan genomes. Our analyses reveal that, compared to other primates, the orang-utan genome has many unique features. Structural evolution of the orang-utan genome has proceeded much more slowly than other great apes, evidenced by fewer rearrangements, less segmental duplication, a lower rate of gene family turnover and surprisingly quiescent Alu repeats, which have played a major role in restructuring other primate genomes. We also describe a primate polymorphic neocentromere, found in both *Pongo* species, emphasizing the gradual evolution of orang-utan genome structure. Orang-utans have extremely low energy usage for a eutherian mammal[1], far lower than their hominid relatives. Adding their genome to the repertoire of sequenced primates illuminates new signals of positive selection in several pathways including glycolipid metabolism. From the population perspective, both *Pongo* species are deeply diverse; however, Sumatran individuals possess greater diversity than their Bornean counterparts, and more species-specific variation. Our estimate of Bornean/Sumatran speciation time, 400,000 years ago, is more recent than most previous studies and underscores the complexity of the orang-utan speciation process. Despite a smaller modern census population size, the Sumatran effective population size (N_e) expanded exponentially relative to the ancestral N_e after the split, while Bornean N_e declined over the same period. Overall, the resources and analyses presented here offer new opportunities in evolutionary genomics, insights into hominid biology, and an extensive database of variation for conservation efforts.

Orang-utans are the only primarily arboreal great apes, characterized by strong sexual dimorphism and delayed development of mature male features, a long lifespan (35–45 years in the wild, more than 55 years in captivity) and the longest interbirth interval among mammals (8 years on average)[2]. Orang-utans create and adeptly use tools in the wild, and while long presumed socially solitary, dense populations of Sumatran orang-utans show complex social structure and geographic variability in tool use indicative of cultural learning[3]. Both species have been subject to intense population pressure from loss of habitat, deforestation, hunting and disease. A 2004 study estimated that 7,000–7,500 Sumatran individuals and 40,000–50,000 Bornean individuals remained in the wild in fragmented subpopulations[4,5]. The International Union for Conservation of Nature lists Sumatran orang-utans as critically endangered and Bornean orang-utans as endangered.

[1]The Genome Center at Washington University, Washington University School of Medicine, 4444 Forest Park Avenue, Saint Louis, Missouri 63108, USA. [2]Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. [3]MRC Functional Genomics Unit and Department of Physiology, Anatomy and Genetics, University of Oxford, Le Gros Clark Building, South Parks Road, Oxford OX1 3QX, UK. [4]Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN, UK. [5]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA. [6]IBE, Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, PRBB, Doctor Aiguader, 88, 08003 Barcelona, Spain. [7]Howard Hughes Medical Institute, 1705 NE Pacific Street, Seattle, Washington 98195, USA. [8]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, UK. [9]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [10]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA. [11]Center for Comparative Genomics and Bioinformatics, Penn State University, University Park, Pennsylvania 16802, USA. [12]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Campus Agrário de Vairão, 4485-661 Vairão, Portugal. [13]Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Oeiras 2780-157, Portugal. [14]ICREA (Institució Catalana de Recerca i Estudis Avançats) and INB (Instituto Nacional de Bioinformática) PRBB, Doctor Aiguader, 88, 08003 Barcelona, Spain. [15]Department of Biology, University of Bari, Via Orabona 4, 70125, Bari, Italy. [16]Department of Biology, University of Bologna, Via Selmi 3, 40126, Bologna, Italy. [17]Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA. [18]Center for Computation and Technology, Department of Computer Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA. [19]Institute for Systems Biology, Seattle, Washington 98103, USA. [20]Department of Biology and School of Informatics and Computing, Indiana University, Bloomington, Indiana 47405, USA. [21]Instituto Universitario de Oncologia, Departamento de Bioquimica y Biologia Molecular, Universidad de Oviedo, Oviedo 33006, Spain. [22]Faculty of Mathematics, Physics and Informatics, Comenius University, Mlynska Dolina, Bratislava 84248, Slovakia. [23]Institut für Populations genetik, Vetmeduni Vienna, Veterinärplatz 1, A-1210 Wien, Austria. [24]Department of Anatomy and Neurobiology, Washington University School of Medicine, Saint Louis, Missouri 63110, USA. [25]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA. [26]Department of Biostatistics, Columbia University, New York, New York 10032, USA. [27]Department of Genetics, Stanford University, Stanford, California 94305, USA. [28]Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona 85721, USA. [29]Bioinformatics Research Centre, Aarhus University, DK-8000 Aarhus C, Denmark. [30]San Diego Zoo's Institute for Conservation Research, Escondido, California 92027, USA. [31]Children's Hospital Oakland Research Institute, Oakland, California 94609, USA. †Present address: Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.

Table 1 | Sumatran orang-utan assembly statistics

Total contig bases	3.09 Gb
Total contig bases >Phred Q20	3.05 Gb (98.5%)
Ordered/oriented contigs and scaffolds	3.08 Gb
Number of contigs >1 kb	410,172
N50 contig length	15.5 kb
N50 contig number	55,989
Number of scaffolds >2 kb	77,683
N50 scaffold length	739 kb
N50 scaffold number	1,031
Average read depth	5.53 ×

Q20 refers to a score of 20 on the Phred scale of base quality scores; here we present the total number of bases in the assembly with a Phred score greater than 20 (3.05 Gb, which is 98.5% of assembled bases). N50 refers to a length-weighted median such that 50% of the genome is contained in contigs or scaffolds of the indicated size or greater.

Table 2 | Number of genome rearrangements by species

Species	Rearrangements >100 kb	Rearrangements >5 kb
Orang-utan	38	861
Chimpanzee	85 (+124%)	1,095 (+27%)
Human	54 (+42%)	1,238 (44%)

The number in parentheses indicates the percentage change with respect to the orang-utan genome. Note 40 events >100 kb and 532 events >5 kb were assigned to the human-chimpanzee ancestor by ancestral reconstruction (Supplementary Information section 6).

We sequenced the genome of a female Sumatran orang-utan using a whole-genome shotgun strategy. The assembly provides 5.5-fold coverage on average across 3.08 gigabases (Gb) of ordered and oriented sequence (Table 1) (Supplementary Information section 1). Accuracy was assessed by several metrics, including comparison to 17 megabases (Mb) of finished bacterial artificial chromosome (BAC) sequences and a novel method of detecting spurious insertions and deletions (Supplementary Information section 2). Further validation resulted from orang-utan–human divergence estimates based on alignment of whole-genome shotgun reads to the human reference (Hs.35; Fig. 1, Supplementary Information section 3). We also sequenced the genomes of 10 additional unrelated wild-caught orang-utans, five Sumatran and five Bornean, using a short read sequencing platform (297 Gb of data in total; Supplementary Information section 4). The orang-utan gene set was constructed using a combination of human gene models and orang-utan complementary DNA data generated for this project (available at www.ensembl.org/Pongo_pygmaeus/Info/StatsTable; see also Supplementary Information section 5).

Among hominids, the orang-utan karyotype is the most ancestral[6], and sequencing the orang-utan genome allowed a comprehensive assessment of conservation among the wide range of rearrangement types and sequence classes involved in structural variation. We characterized orang-utan synteny breaks in detail cytogenetically in concert with an *in silico* approach that precisely tracked rearrangements between primate (human, chimpanzee, orang-utan and rhesus macaque) and other mammalian assemblies (mouse, rat and dog) (Supplementary Information section 6). Alignment-level analyses at 100 kilobase (kb) and 5 kb resolution found that the orang-utan genome underwent fewer rearrangements than the chimpanzee or human genomes, with a bias for large-scale events (>100 kb) on the chimpanzee branch (Table 2). Orang-utan large-scale rearrangements were further enriched for segmental duplications (52%) than for small-scale events (27%), suggesting that mechanisms other than non-allelic homologous recombination may have made a greater contribution to small rearrangements. Genome-wide, we estimated less segmental duplication content (3.8% total) in the orang-utan genome compared to the chimpanzee and human genomes (5%) using equivalent methods (Supplementary Information section 11). We also assessed the rate of turnover within gene families as an additional measure of genome restructuring (Supplementary Information section 12). Our analysis indicated that the human and chimpanzee lineages, as well as their shared ancestral lineage after the orang-utan split, had the highest rates of gene turnover among great apes (0.0058 events per gene per Myr)—more than twice the rate of the orang-utan and macaque lineages (0.0027)—even as the nucleotide substitution rate decreased[7]. Collectively, these data strongly suggest that structural evolution proceeded much more slowly along the orang-utan branch, in sharp contrast to the acceleration of structural variation noted for the chimpanzee and human genomes[8,9].

One structural variant that we characterized in detail was a previously described polymorphic ‘pericentric inversion’ of orang-utan chromosome 12 (ref. 10). Surprisingly, both forms of this chromosome showed no difference in marker order by fluorescence *in situ* hybridization (FISH) despite two distinct centromere positions—the hallmark of a neocentromere (Fig. 2; Supplementary Information section 8). Neocentromere function was confirmed by chromatin immunoprecipitation with antibodies to centromeric proteins CENP-A and CENP-C and subsequent oligo array hybridization (ChIP-on-chip), which narrowed the neocentromere to a ~225 kb gene-free window devoid of α satellite-related sequences. Our observations bore similarity to a recently described centromere repositioning event in the horse genome[11]; however, this is to our knowledge the first observation of such a variant among primates, with the additional complexity of polymorphism in two closely related species. Potentially related, orang-utan chromosome 12 did not show any appreciable centromeric alphoid FISH signal in comparison to other autosomes. The neocentromere most probably arose before the Bornean/Sumatran split as it is found in both species, and represents a unique opportunity to study the initial stages of centromere formation

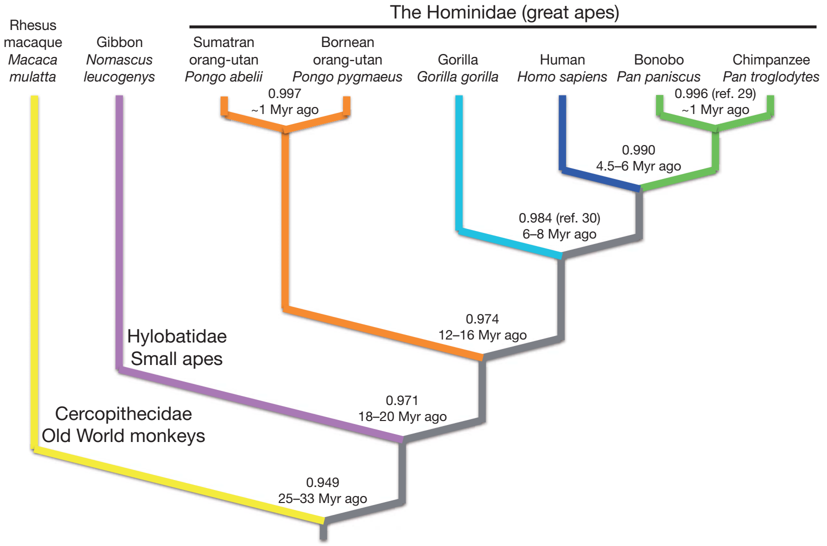

Figure 1 | Divergence among great apes, a small ape, and an Old World monkey with respect to humans. We estimated nucleotide divergence in unique gap-free sequence, indicated at each node, from the alignment of rhesus macaque (yellow), gibbon (purple), orang-utan (orange), gorilla (aqua), chimpanzee (green) and human (blue) whole genome shotgun reads to the human reference (Hs.35; Supplementary Information section 3). Note that the Bornean (*P. pygmaeus*) and Sumatran (*P. abelii*) orang-utan species showed nucleotide identity comparable to that of bonobo (*Pan paniscus*) and chimpanzee (*Pan troglodytes*). Estimates of divergence time based on sequence identity are indicated at each node, ~1 Myr implies approximately 1 Myr or less. Values taken from refs 29 and 30 where indicated.

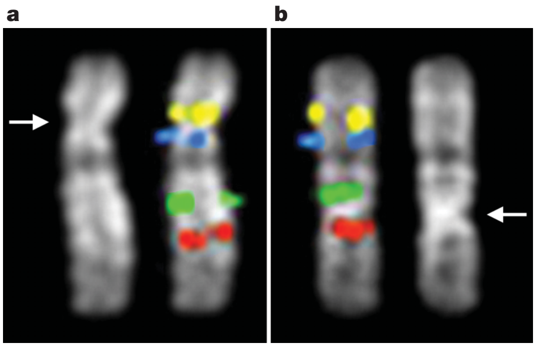

Figure 2 | The neocentromere of orang-utan chromosome 12. Note the identical order of four sequentially arranged BAC-derived FISH probes indicated in yellow, blue, green and red (given in Supplementary Information section 8) between the normal (**a**) and neocentromere-bearing (**b**) configurations of orang-utan chromosome 12, despite discordant centromere positions. The left image of **a** and the right image of **b** are DAPI-only images that show the primary constriction of both chromosomal forms, indicated by the arrows. The neocentromere recruits centromeric proteins CENP-A and CENP-C and lies within a ~225 kb gene-free and α satellite-free region. The neocentromere-bearing variant is polymorphic in both Bornean and Sumatran populations, suggesting the neocentromere arose before the Bornean/Sumatran split, yet has not been fixed in either species.

and the impact of such a large chromosomal variant on population variation and recombination.

The orang-utan genome has a comparable cadre of mobile elements to that of other primates, comprising roughly half the genome[12–14]. Orang-utan long interspersed element 1 (LINE1; L1) and SVA (SINE-R, VNTR and Alu) element expansions were expectedly broad, with roughly 5,000 and 1,800 new insertions respectively, consistent with other primates (Supplementary Information section 9). Surprisingly, Alu elements were relatively quiescent, with only ~250 recent insertions identified by computational and laboratory approaches (Fig. 3). By comparison, 5,000 human-specific and 2,300 chimpanzee-specific Alu elements were identified by similar methods. The rate of processed pseudogene formation, which like Alu insertion requires functional L1 machinery[15], was similar for the human (8.0 per Myr), chimpanzee (12.7 per Myr) and orang-utan (11.6 per Myr) lineages (Supplementary Information section 10). We identified a small number of polymorphic Alu elements exclusive to *P. abelii* (Supplementary Information section 19), indicating that Alu retroposition has been strongly limited, but not eliminated. This dramatic Alu-specific repression represents an unprecedented change in primate retrotransposition rates[16,17]. Possible explanations include L1 source mutations that lowered Alu affinity and *cis* mobilization preference[18], pressure against Alu retroposition from the *APOBEC* RNA editing family[19], or fixation of less effectively propagated Alu 'master' variants.

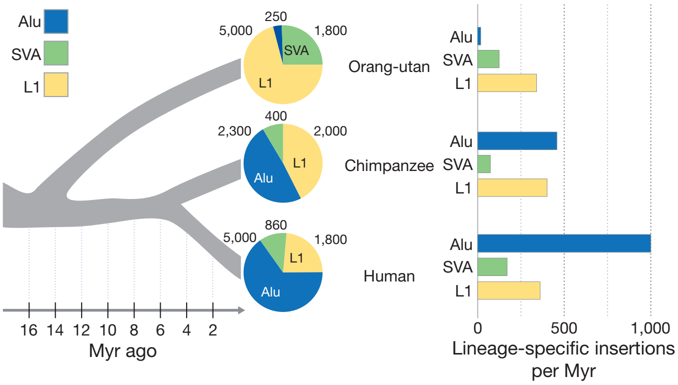

Figure 3 | Alu quiescence in the orang-utan lineage. We identified only ~250 lineage-specific Alu retroposition events in the orang-utan genome, a dramatically lower number than that of other sequenced primates, including humans. The total number of lineage-specific L1, SVA and Alu insertions is shown (pie chart) at the terminus of each branch of the phylogeny of sequenced great apes shown in grey at left, along with the rate of insertion events per element type (bar graph). Reduced Alu retroposition potentially limited the effect of a wide variety of repeat-driven mutational mechanisms in the orang-utan lineage that played a major role in restructuring other primate genomes.

It is tempting to propose a correlation between reduced Alu retroposition and the greater structural stability of the orang-utan genome. More than 10^6 Alu elements exist within primate genomes. Because of their large copy number and high sequence identity, Alu repeats play a crucial role in multiple forms of structural variation through insertion and post-insertion recombination[20]. By virtue of reduced Alu retroposition, the orang-utan lineage experienced fewer new insertions and a putative decrease in the number of regions susceptible to post-insertion Alu-mediated recombination events genome-wide, limiting the overall mobile element threat to the genome.

The unique phylogenetic position of *Pongo* species also offered the opportunity to detect signals of positive selection with increased power. We assessed positive selection in 13,872 human genes with high-confidence orthologues in the orang-utan genome, and in one or more of the chimpanzee, rhesus macaque and dog genomes, using branch-site likelihood ratio tests (Supplementary Information section 15)[14,21]. Two new Gene Ontology categories were statistically enriched for positive selection in primates: 'visual perception' and 'glycolipid metabolic processes'[22]. The enrichment for visual perception includes strong evidence from two major visual signalling proteins: arrestin (*SAG*, $P = 0.007$) and recoverin (*RCVRN*, $P = 0.008$), as well as the opsin, *OPN1SW1* ($P = 0.020$), associated with blue colour vision[23]. The enrichment for glycolipid metabolism is particularly intriguing owing to medium-to-strong evidence for positive selection (nominal $P < 0.05$) from six genes expressed in nervous tissue that cluster in the cerebroside-sulphatid region of the sphingolipid metabolism pathway (Fig. 4). This pathway

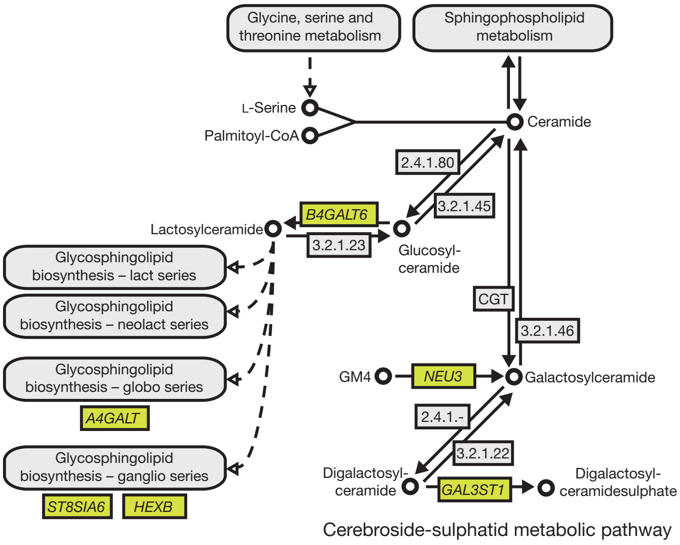
Glycine, serine and threonine metabolism
Sphingophospholipid metabolism
L-Serine
Palmitoyl-CoA
Ceramide
2.4.1.80
3.2.1.45
B4GALT6
Lactosylceramide
3.2.1.23
Glucosyl-ceramide
Glycosphingolipid biosynthesis – lact series
Glycosphingolipid biosynthesis – neolact series
CGT
3.2.1.46
Glycosphingolipid biosynthesis – globo series
A4GALT
GM4
NEU3
Galactosylceramide
2.4.1.-
3.2.1.22
Glycosphingolipid biosynthesis – ganglio series
ST8SIA6
HEXB
Digalactosyl-ceramide
GAL3ST1
Digalactosyl-ceramidesulphate
Cerebroside-sulphatid metabolic pathway

Figure 4 | Enrichment for positive selection in the cerebroside-sulphatid metabolism pathway. We identified six genes (indicated in yellow) under moderate to strong positive selection in primates ($P < 0.05$) that fall within the cerebroside-sulphatid region of the sphingolipid metabolism pathway (adapted from human KEGG pathway 00600). This pathway is associated with several human lysosomal storage disorders, such as Gaucher's disease, Sandhoff's disease, Tay-Sachs disease and metachromatic leukodystrophy. Abbreviations, annotations and connections are presented in accordance with KEGG standards: solid lines represent direct relationships between enzymes (boxes) and metabolites (circular nodes), dashed lines represent indirect relationships, arrowheads denote directionality (see http://www.genome.jp/kegg-bin/show_pathway?map00600 for further details).

is associated with human neurodegenerative diseases such as Gaucher's, Sandhoff's, Tay-Sachs, and metachromatic leukodystrophy. Variation in lipid metabolism may have affected neurological evolution among primates, and diversity of diets and life history strategies, as apes—especially orang-utans—have slower rates of reproduction and dramatically lower energy usage than other primates and mammals[1].

Ancestral orang-utan species ranged broadly across southeast Asia, including the mainland, while modern species are geographically restricted to their respective islands owing to environmental forces and human population expansion. Historically, protein markers, restriction fragment length polymorphisms, and small sets of mitochondrial and nuclear markers have been used to estimate the divergence and diversity of orang-utan species. We used short read sequencing to address this question from a genome-wide perspective. We first estimated average Bornean/Sumatran nucleotide identity genome-wide (99.68%) based on the alignment of 20-fold coverage of short read data from a Bornean individual to the Sumatran reference (Supplementary Information section 16). We then called single nucleotide polymorphisms (SNPs) from the alignment of all short read data from 10 individuals (five Bornean, including the 20-fold coverage mentioned above, and five Sumatran) (Supplementary Information section 4). We analysed each species separately using a Bayesian approach with 92% power to detect SNPs (Supplementary Information section 20). Because of relatively deep sequencing, allele frequency spectra were estimated accurately, but with an overestimation of singletons compared to other allele frequency categories of approximately 7.8% based on re-sequencing a subset of SNPs ($n = 108$) (Supplementary Information section 20). This level of error had only a marginal effect on downstream population genetic analyses (Supplementary Information section 21). Overall, 99.0% (931/940) of genotypes were accurately called within the re-sequenced subset of SNPs.

In total, we identified 13.2×10^6 putative SNPs across 1.96 Gb of the genome, or 1 SNP every 149 base pair (bp) on average. Within the Bornean and Sumatran groups we detected 6.69×10^6 (3.80×10^6 Bornean-exclusive) and 8.96×10^6 (5.19×10^6 Sumatran-exclusive) SNPs, respectively (Fig. 5). Observing 36% more SNPs among Sumatran individuals strongly supports a larger N_e. In addition, independent analysis of 85 polymorphic retroelement loci among 37 individuals (19 Sumatran, 18 Bornean) also showed more complex Sumatran population structure (Supplementary Information section 19). Using Watterson's approach[24], we estimated nucleotide diversity from the SNP data as $\theta_W = 1.21$ and $\theta_W = 1.62$ per kb for the Bornean and Sumatran species, respectively, and $\theta_W = 1.89$ per kb for the orang-utan species combined, roughly twice the diversity of modern humans[25].

The modal category of SNPs were singletons, with 2.0×10^6 and 3.7×10^6 SNPs observed as single heterozygous sites in a Bornean or Sumatran individual, consistent with the expectation that most genetic variation for an outcrossing population ought to be rare due to mutation-drift equilibrium. We observed little correlation between Bornean and Sumatran SNPs in the allele frequency spectra (that is, the 'heat' of the map is not along the diagonal as expected for populations with similar allele frequencies, but rather along the edges) (Fig. 5b). This was further supported by principal component analysis, in which PC1 corresponded to the Bornean/Sumatran population label and explained 36% of the variance (Supplementary Information section 20).

On the basis of these data, our demographic model consisted of a two-population model with divergence and potential migration, growth and difference in population size (Supplementary Information section 21). Among several models tested, we found very strong statistical support (10^5 log-likelihood units) for the most complex model, which included a split with growth and subsequent low-level migration. We estimated a relative N_e of 210% for Sumatran orang-utans relative to the ancestral and 49% for Bornean orang-utans, noting a fourfold difference for the derived populations (Fig. 5c). Assuming a mutation rate of 2.0×10^{-8} and 20 years per generation, we estimated an ancestral N_e of 17,900 and a split time of 400,000 years ago.

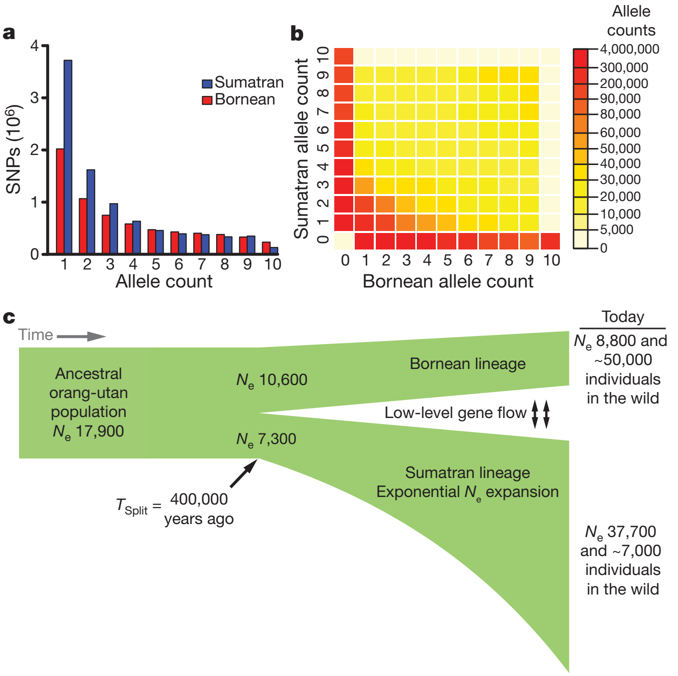

Figure 5 | Orang-utan population genetics and demographics. **a**, Site-frequency spectra for 13.2×10^6 Bornean (red) and Sumatran (blue) SNPs are shown based on the ascertainment of 10 chromosomes per species; note the enrichment of low-frequency SNPs among Sumatran individuals. **b**, The majority of SNPs were restricted to their respective island populations as the 'heat' of the two-dimensional site-frequency spectra, representing high allele counts, lay along the axes. **c**, Our demographic model estimated that the ancestral orang-utan population ($N_e = 17{,}900$) split approximately 400,000 years ago, followed by exponential expansion of Sumatran N_e and a decline of Bornean N_e, culminating in higher diversity among modern Sumatran orang-utans despite a lower census population size. The model also supported low-level gene flow (<1 individual per generation), indicated by arrows.

Parallel to the SNP-based effort, we employed a coalescent hidden Markov model (coal-HMM) approach to estimate speciation time, recombination rate and ancestral N_e from the alignment of 20-fold coverage of a Bornean individual to the Sumatran reference (Supplementary Information section 17). This method also supported a relatively recent Bornean/Sumatran speciation time (334 ± 145 kyr ago), and estimated a recombination rate of 0.95 ± 0.72 cM Mb^{-1}. We independently estimated the ancestral N_e of the autosomes (26,800 ± 6,700) and the X chromosome (20,400 ± 7,400), which was consistent with the theoretical 3/4 effective population size of X chromosomes compared to autosomes. The Bornean and Sumatran X chromosome thus diverged as expected, in contrast to the human–chimpanzee speciation process[26,27].

The orang-utan story is thus a tale of two islands with distinct evolutionary histories. Our high-resolution population studies explored the counter-intuitive nature of orang-utan diversity—greater variation among Sumatran orang-utans than their Bornean counterparts despite a smaller population size (approximately sevenfold lower by recent estimates). Further dissection of the orang-utan speciation process will require a broader survey, incorporating representatives from additional orang-utan subpopulations.

Finally, even though we found deep diversity in both Bornean and Sumatran populations, it is not clear whether this diversity will be maintained with continued habitat loss and population fragmentation. Evidence from other species suggests fragmentation is not the death knell of diversity[28], but their slow reproduction rate and arboreal lifestyle may leave orang-utan species especially vulnerable to rapid dramatic environmental change. It is our hope that the genome assembly and population variation data presented here provide a valuable resource to the community to aid the preservation of these precious species.

METHODS SUMMARY

Whole-genome sequencing was performed as described previously[12–14]. The genome assembly was constructed with a custom computational pipeline (Supplementary Information section 1). Assembly source DNA was derived from a single Sumatran female (Susie; Studbook no. 1044; ISIS no. 71), courtesy of the Gladys Porter Zoo, Brownsville, Texas. Short fragment sequencing libraries for population studies (Supplementary Information section 4) were constructed in accordance with standard Illumina protocols and sequenced on the Illumina GAIIx platform. The resulting data were processed with Illumina base-calling software and analysed using custom computational pipelines. See Supplementary Information for additional details.

Received 11 March; accepted 19 November 2010.

1. Pontzer, H., Raichlen, D. A., Shumaker, R. W., Ocobock, C. & Wich, S. A. Metabolic adaptation for low energy throughput in orangutans. *Proc. Natl Acad. Sci. USA* **107,** 14048–14052 (2010).
2. van Noordwijk, M. A. & van Schaik, C. P. Development of ecological competence in Sumatran orangutans. *Am. J. Phys. Anthropol.* **127,** 79–94 (2005).
3. van Schaik, C. P. *et al.* Orangutan cultures and the evolution of material culture. *Science* **299,** 102–105 (2003).
4. Singleton, I. *et al. Orangutan Population and Habitat Viability Assessment: Final Report* (IUCN/SSC Conservation Breeding Specialist Group, Apple Valley, 2004).
5. Meijaard, E. & Wich, S. Putting orang-utan population trends into perspective. *Curr. Biol.* **17,** R540 (2007).
6. Stanyon, R. *et al.* Primate chromosome evolution: ancestral karyotypes, marker order and neocentromeres. *Chromosome Res.* **16,** 17–39 (2008).
7. Yi, S., Ellsworth, D. L. & Li, W. H. Slow molecular clocks in Old World monkeys, apes, and humans. *Mol. Biol. Evol.* **19,** 2191–2198 (2002).
8. Hahn, M. W., Demuth, J. P. & Han, S. G. Accelerated rate of gene gain and loss in primates. *Genetics* **177,** 1941–1949 (2007).
9. Marques-Bonet, T. *et al.* A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457,** 877–881 (2009).
10. Seuanez, H., Fletcher, J., Evans, H. J. & Martin, D. E. A chromosome rearrangement in orangutan studied with Q-, C-, and G-banding techniques. *Cytogenet. Cell Genet.* **17,** 26–34 (1976).
11. Wade, C. M. *et al.* Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326,** 865–867 (2009).
12. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
13. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
14. Gibbs, R. A. *et al.* Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316,** 222–234 (2007).
15. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.* **24,** 363–367 (2000).
16. Liu, G. *et al.* Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res.* **13,** 358–368 (2003).
17. Lee, J. *et al.* Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons. *Gene* **390,** 18–27 (2007).
18. Kulpa, D. A. & Moran, J. V. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. *Nature Struct. Mol. Biol.* **13,** 655–660 (2006).
19. Bogerd, H. P. *et al.* Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. *Proc. Natl Acad. Sci. USA* **103,** 8780–8785 (2006).
20. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nature Rev. Genet.* **10,** 691–703 (2009).
21. Kosiol, C. *et al.* Patterns of positive selection in six Mammalian genomes. *PLoS Genet.* **4,** e1000144 (2008).
22. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
23. Makino, C. L. *et al.* Recoverin regulates light-dependent phosphodiesterase activity in retinal rods. *J. Gen. Physiol.* **123,** 729–741 (2004).
24. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7,** 256–276 (1975).
25. Li, W. H. & Sadler, L. A. Low nucleotide diversity in man. *Genetics* **129,** 513–523 (1991).
26. Hobolth, A., Christensen, O. F., Mailund, T. & Schierup, M. H. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet.* **3,** e7 (2007).
27. Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S. & Reich, D. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441,** 1103–1108 (2006).
28. Alcaide, M. *et al.* Population fragmentation leads to isolation by distance but not genetic impoverishment in the philopatric Lesser Kestrel: a comparison with the widespread and sympatric Eurasian Kestrel. *Heredity* **102,** 190–198 (2009).
29. Yu, N. *et al.* Low nucleotide diversity in chimpanzees and bonobos. *Genetics* **164,** 1511–1518 (2003).
30. Chen, F. C. & Li, W. H. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68,** 444–456 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The orang-utan genome project was funded by the National Human Genome Research Institute (NHGRI), including grants U54 HG003079 (R.K.W.) and U54 HG003273 (R.A.G), with further support from National Institutes of Health R01 GM59290 (M.A.B.), PO1 AG022064 (M.A.B.), HG002385 (E.E.E.) and HG002238 (W.M.), National Science Foundation DBI-0644111 (A.S. and B.B.), David and Lucile Packard Foundation (A.S., V.T. and T.V.), Cornell University Provost's Fellowship (A.L.M.), UK Medical Research Council (C.P.P., G.L., S.M. and A.H.), Marie Curie Fellowship (T.M.-B.), Ministerio de Ciencia e Innovación-Spain (MCI-Spain) and Fundación M. Botín (V.Q., X.S.P., G.R.O. and C.L.-O.), MCI-Spain BFU2006-15413-C02-01 and BFU2009-13409-C02-02 (A.N.), Spanish National Institute for Bioinformatics (INAB) and Fundação para a Ciência e a Tecnologia (Portugal), SFRH/BPD/26384/2006 (R.F.) and SFRH/BD/15856/2005 (O.F.), PRIN and CEGBA (M.R., N.A. and G.D.V.), and the Commission of the European Communities IRG-224885 (T.V.), IRG-231025 (B.B.). We thank the Gladys Porter Zoo, and S. O'Brien and S. Pääbo for use of orang-utan samples. D.P.L. thanks S. D. McGrath, A. Wollam, R. M. Abbott and L. Chemnick for technical assistance. We acknowledge all the important work that could not be cited owing to space limitations. Resources for exploring the orang-utan genome are available at UCSC (http://genome.ucsc.edu), Ensembl (http://www.ensembl.org), NCBI (http://ncbi.nlm.nih.gov) and The Genome Center at Washington University (http://genome.wustl.edu/genomes/view/pongo_abelii/).

Author Contributions D.P.L. led the project and manuscript preparation. D.P.L., A.S., T.M.-B., C.P.P., M.A.B., A.N., E.E.E., M.W.H., C.L.-O., C.D.B., J.M. and M.H.S. led the analyses. Sanger data production, assembly construction, testing and submission: L.W.H, W.C.W., S.-P.Y., Z.W., A.T.C., P.M., M.M., L.A.F., R.A.F., J.O.N., C.P., K.C.W, L.V.N., D.M.M., A.C., H.H.D., J.H., C.L.K., G.R.F. and J.R. BAC sequencing: T.A.G. 454 cDNA sequencing: V.M. and C.M. Illumina sequencing: L.C., K.D.D. and C.F. SNP validation: H.S. Indel assessment: G.L., S.M., A.H. and C.P.P. Segmental duplication, divergence and structural variation studies: T.M.-B., C.A., L.C., Z.C., J.M.K. and E.E.E. Gene models: S.W., S.S. and A.J.V. Assembly-based SNPs: Y.C. and P.F. Ancestral reconstruction and rearrangement analyses: J.M., B.R., B.S., R.B., J.H., D.H., R.S.H. and W.M. Regional variation in nucleotide divergence analyses: R.F., O.F., F.D., D.F., E.G., M.O. and A.N. Cytogenetics and neocentromere characterization: R.R., O.C., N.A., G.D.V., S.P and M.R. Repeat analyses: M.K.K., J.A.W., B.U., M.A.B., A.F.A.S. and R.H. Gene family evolution analyses: C.C., D.R.S. and M.W.H. Protease gene family studies: V.Q., X.S.P., G.R.O. and C.L.-O. Orthologue and defensin analyses: T.V., B.B., A.R. and W.M. Positive selection analyses: C.K., T.V., H.A.L., V.T., A.L.M. and A.S. Short read alignments, SNP calling and population genetics: A.R., X.M., J.D. and C.D.B. Demographic analyses: R.N.G. Coalescent-HMM analyses: T.M., J.Y.D., A.H. and M.H.S. Orang-utan samples for diversity sequencing: O.A.R. BAC library construction: Y.Y. and P.J.d.J. Principal investigators: G.M.W., E.R.M., R.A.G. and R.K.W.

Author Information The *P. abelii* whole-genome shotgun project has been deposited in DDBJ/EMBL/GenBank under the project accession ABGA00000000. The version described in this Letter is ABGA00000000.1. Assembly-based SNPs and SNPs derived from short read sequence data have been deposited in dbSNP. All short read data have been deposited into the Short Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accessions listed in Supplementary Information. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike Licence, and is freely available to all readers at www.nature.com/nature. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.P.L. (dlocke@wustl.edu) or W.C.W. (wwarren@wustl.edu).