Chapter 19

Population Genomics in the Great Apes

David Castellano and Kasper Munch

Abstract

The great apes play an important role as model organisms. They are our closest living relatives, allowing us to identify the genetic basis of phenotypic traits that we think of as characteristically human. However, the most significant asset of great apes as model organisms is that they share with humans most of their genetic makeup. This means that we can extend our vast knowledge of the human genome, its genes, and the associated phenotypes to these species. Comparative genomic studies of humans and apes thus reveal how very similar genomes react when exposed to different population genetic regimes. In this way, each species represents a natural experiment, where a genome highly similar to the human one, is differently exposed to the evolutionary forces of demography, population structure, selection, recombination, and admixture/hybridization. The initial sequencing of reference genomes for chimpanzee, orangutan, gorilla, the bonobo, each provided new insights and a second generation of sequencing projects has provided diversity data for all the great apes. In this chapter, we will outline some of the findings that population genomic analysis of great apes has provided, and how comparative studies have helped us understand how the fundamental forces in evolution have contributed to shaping the genomes and the genetic diversity of the great apes.

Key words Population genomics, Great apes, Incomplete lineage sorting, Demography, Distribution of fitness effects, Recombination, X chromosome, Selective sweeps

1 Species Trees and Incomplete Lineage Sorting

The sequencing of all the great ape genomes [1–6] has allowed us to paint a detailed picture of the species relationship between humans and their closest relatives. By joint analysis of full genomes from pairs of species, coalescent hidden Markov models (CoalHMM) (*see* Chapter 8) can efficiently model both sequence divergence and recombination by approximating the full ancestral recombination graph as a Markov process along the genome. The states in these hidden Markov models represent different gene trees separated by recombination events. Such models can jointly estimate both the time of reproductive isolation (the time of speciation) and the size of the ancestral population that gave rise to the two species. Figure 1 provides an overview of the estimated split

Julien Y. Dutheil (ed.), *Statistical Population Genomics*, Methods in Molecular Biology, vol. 2090,
https://doi.org/10.1007/978-1-0716-0199-0_19, © The Author(s) 2020

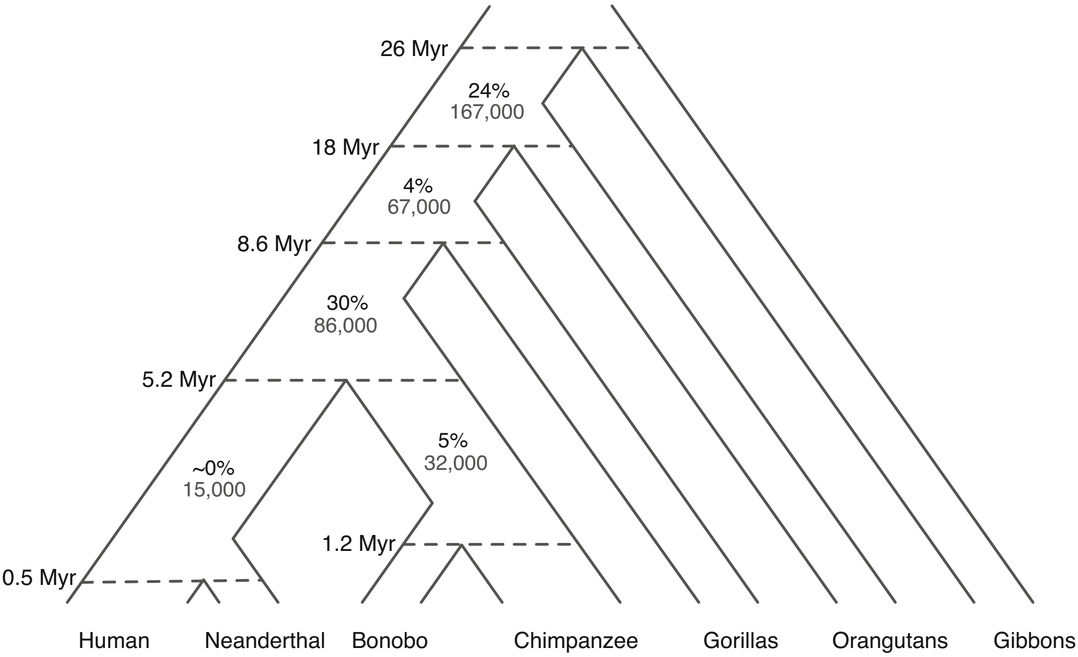

Fig. 1 Species tree of humans and the great apes. Dashed lines represent speciation events with estimated dates. Gray numbers show estimated sizes of ancestral populations, and percentages show the estimated amount of incomplete lineage sorting (*see* below) between two descendant species and their immediate outgroup (e.g., 30% ILS between human, chimpanzee, and gorilla). (The figure is adapted from Mailund et al. [7])

times and ancestral population sizes. The estimated split times are computed assuming that the mutation rate across the species tree has remained constant at the rate of 0.6×10^{-9} per year as observed in humans. However, speciation dates produced using a constant mutation rate does not square with the physical dating of ancestral fossil species. To reconcile DNA and fossil evidence, it has been proposed that the yearly mutation rate has slowed down across in the great ape lineage [8], possibly resulting from the development of larger body sizes and longer generation times.

The time to the most recent common ancestor of sequences sampled from two species lies much further into the past than the time when the species split apart. For this reason, a common ancestor of two lineages from separate species may not be found in the population ancestral to the two species, but even further into the past, in a population ancestral to additional species. When sampling more than two species, this allows for the possibility that lineages from other than the most closely related species find a common ancestor before those most closely related. This is especially true for the relationship between human, chimpanzee, and gorilla. Between the speciation events separating human and chimpanzee and that separating human and gorilla, a lot of ancestral polymorphism was conserved in the large ancestral population. The more rapid the succession of speciation events, and the larger the ancestral population between them is, the more ancestral polymorphism will be conserved. The implication is that individual gene trees along the alignment of these three species will not always group the same two species as the species tree does. The phenomenon is called incomplete lineage sorting (ILS) because the lineages of individual gene trees are not completely sorted according to species (*see* Chapter 1 for further details).

One coalescent hidden Markov model compares three closely related species and exploits information from sequence divergence and ILS to estimate the time of the two speciation events as well as the size of the ancestral population [9, 10]. From this model, it is also possible to extract the proportion of discordant gene trees with a topology different from the species tree. Applying this method to the human, chimpanzee, and gorilla showed that for ~15% percent of the genome, humans are more closely related to gorillas than chimpanzees, and for another ~15%, chimpanzees and gorillas are more closely related to each other than to humans [3] (*see* Fig. 1). The same model has been applied to alignments of bonobo, chimpanzee, and human, and showed that ~5% of the genome is subject to ILS [4]. Because the proportion of ILS is determined by ancestral population size and the time between speciation events, we can compute the estimated proportions of ILS for trios of species where these parameters have been estimated by other means. Between human, gorilla, and orangutan it is expected to be ~4%, and for human, orangutan, and gibbon it is expected to be ~24% [7]. The great apes thus also showcase how misleading phylogenies built from individual genes may be since a phylogeny built from long regions of a recombining sequence will not represent the population genetic processes that distribute individual lineages among species.

2 Gene Flow and Demography

Most coalescent hidden Markov models assume that speciation is instantaneous and that the initial split of two populations is not followed by gene flow between the diverging populations. Other coalescent hidden Markov models account for the possibility that such gene flow has occurred [11]. Among species splits in great ape evolution, most have involved a period of gene flow before consolidation of the populations as separate species [11]. The divergence of the orangutan from the human–chimp–bonobo–gorilla ancestor involved several hundred thousand years of gene flow. The speciation of humans and chimpanzees-bonobo most likely also included an extended period of gene flow. Only the speciation separating the bonobo from the chimpanzees seem to be a clear example of an abrupt and permanent split, possibly produced as the Congo River provided a physical barrier between the populations.

An alternative way to estimate gene flow between separating populations is using methods such as MSMC [12] (*see* Chapter 7) that can estimate the relative rate of cross-coalescence between populations from the present and into the past. This approach measures the proportion of gene pairs that find common ancestry between two sampled populations rather than within them. Inspecting a curve of this relative cross-coalescence rate can help identify both the time of speciation and whether this was a clean split rather than a protracted period of reduced gene flow. MSMC as well as a similar method modeling only one diploid sample (PSMC [13]) also estimate the historical effective population size of a species. Such methods have been used to identify how the great apes have responded to environmental changes and show that great apes have experienced a decline in their effective population size across the last few hundred thousand years [5]. Comparing the curves of historical effective population sizes may also reveal when the species split apart. Across the time in the past where two species share an ancestor their historical population sizes will be the same, but at the time this ancestral population split into two, the size of these two populations will be free to follow different trajectories through time and will reveal the species split as a separation of the curves of historical population sizes. Along with methods such as approximate Bayesian computation, PSMC has helped describe the relationship between chimpanzee subspecies [5] showing that eastern and central chimpanzees are most closely related, forming a group separate from Nigeria–Cameroon and western chimpanzees.

3 Selection

One of the most intriguing questions in great ape evolution is how the adaptive evolution of particular genes has contributed to shaping phenotypes in present-day species. A study comparing a large number of orthologous genes addressed adaptive evolution along the branches of humans and chimpanzees by comparing the rate of evolution at synonymous sites (sites where a mutation will not change the encoded protein) with nonsynonymous sites (sites where mutation replaces an amino acid) [14]. Many of the identified genes were involved in sensory perception and immune defenses, but the genes showing the strongest evidence of positive selection were genes involved in tumor suppression and apoptosis, and genes involved in spermatogenesis [15]. Another way to identify selection in primate genomes is by measuring the patterns of genetic diversity along the genomes. Slightly deleterious variants will reduce genetic diversity in a genomic region around the deleterious variant, a process called background selection (*see*

Charlesworth [16] for a review), in effect reducing the local effective population size. Positive selection also removes variation in a region around it, leaving a signature in local genetic variation that can be distinguished from that of background selection if the positive selection is strong enough and occurred recently. When a new variant is subject to strong positive selection, variation in the flanking regions is depleted because linked variants are carried to fixation along with the selected variant. This is called a selective sweep because it sweeps variation in a region around the selected variant and produces a wide genomic region where all individuals from a species share a recent common ancestor [17]. The size of the swept region depends on the strength of selection, the size of the population and the rate of genetic recombination. Several methods have been developed to detect sweeps from information in population samples such as the site frequency spectrum, linkage disequilibrium and population differentiation [18] (*see* Chapter 5). Due to the relatively small sample sizes available in great apes, no striking examples of recent sweeps on great ape autosomes have been reported (but see Sect. 5 for strong selective sweeps on the X chromosome). However, thanks to the McDonald and Kreitman test framework there are many estimates of the proportion of beneficial nonsynonymous substitutions (α) across primates (*see* Chapter 1 for a formal definition of α). Genome-wide estimates in humans and nonhuman primates are very low, $\alpha < 10$–20% [1, 19–23], but α can be as high as 50% for some particular genes like immune genes, testis genes, or virus interacting protein genes [24, 25].

It is still debated if positive or negative selection is more prominent in shaping diversity along great ape genomes, and we are still trying to figure out whether selective sweeps are mainly due to new mutations [17] or selection on standing variation [26], and which are more important for adaptation and the surrounding patterns of DNA diversity. One argument to suggest that sweeps from new mutations contribute significantly to variation in diversity is that great apes with larger population sizes show more dramatic reductions in diversity near genes [27]. This dependence of population size is consistent with the action of positive selection rather than negative selection and suggests that new beneficial mutations leading to sweeps arise more often in species with a larger number of individuals subject to mutation. Identification of selective sweeps, from depressions in diversity or distortions of the site frequency spectrum, is limited to the recent past, where a sample of individuals is expected to be represented by many ancestors. An alternative method to quantify the impact of sweeps on longer timescales is to identify extended regions devoid of incomplete lineage sorting. A sweep in an ancestral species will induce common ancestry for all lineages in a wide region around the selected variant and thus precludes the possibility of incomplete lineage sorting in the region. By identifying and comparing such regions in both the ancestor to human and chimpanzee and the ancestor to human and orangutan, it was possible to show that the human–chimpanzee ancestor experienced a higher frequency of strong sweeps than the human–orangutan ancestor [28].

Addressing the forces of positive and negative selection in the great apes, we need to know what proportion of new mutations are advantageous, neutral, or deleterious and whether these proportions differ across these species. The distribution of fitness effects (DFE) describes the proportions of new mutations that are effectively neutral and new mutations that are under selection [29, 30] (*see* Chapter 1). The DFE further distinguishes between advantageous mutations, which increase the fitness of the organism, and deleterious mutations, which impair survival or fertility. Several methods are available to infer this continuum of selective effects from DNA sequence data [19–21, 31–34]. Initial studies in humans with modest sample sizes found ~25% of effectively neutral nonsynonymous mutations ($-1 > 2Ns < 1$), ~15% of weakly deleterious nonsynonymous mutations ($-10 > 2Ns \leq -1$) and ~60% of moderately to strongly deleterious nonsynonymous mutations ($2Ns \leq -10$) [19–21, 31–34]. A recent study with a large sample size was able to further refine the estimate of new nonsynonymous mutations which are strongly deleterious ($2Ns \leq -100$) to 14–22% and the proportion of weakly deleterious mutations ($-10 > 2Ns \leq -1$) to 25–33% [32]. The DFE for new nonsynonymous mutations is quite similar across great apes despite the differences in the species long-term Ne [22, 35]. This similarity may be explained by the highly leptokurtic DFE of these species, which predicts that substantial changes in Ne will only have a modest impact on the selective effects of mutations. Nonetheless, very different methods and assumptions have been invoked to estimate the DFE across species, and even the shape of the DFE is still a contentious issue. There is very limited knowledge about the DFE of new noncoding mutations, and all we know relies on measures of DNA conservation across mammals and primates. Thus, for noncoding DNA we are only able to say which proportion of new mutations are effectively neutral ($-1 > 2Ns < 1$) and effectively selected against ($2Ns \leq -1$). These rough conservation scores show that only 2–5% of point mutations at noncoding sites might be under purifying selection in humans and the rest of primates [36–40].

Balancing selection is another mode of selection that differs from directional selection in that it does not drive selected variants toward fixation or extinction (*see* Chapter 1 for a definition of balancing selection). Instead, it maintains genetic variation by stabilizing alleles at intermediate frequencies. There are several methods to detect loci under recent and/or long-term balancing selection [22, 41, 42]. A recent study in great apes has confirmed that immune genes are enriched in signals of balancing selection, and it has found that genes involved in the formation of the skin are also under balancing selection [22]. Some of these polymorphisms maintained by balancing selection are even shared between humans and chimpanzees; the most prominent example is the major histocompatibility complex (MHC).

4 Recombination

The rate of recombination varies along the genome. The local recombination rate in each part of the genome can be estimated from patterns of linkage disequilibrium (LD) [43]. It can also be inferred from individually called recombination events by comparing many parent and offspring genomes [44] or by examining genomes of individuals with mixed ancestry [45]. The landscape of varying recombination rate across the genome is referred to as a recombination map. For humans, recombination maps have been produced by all three approaches. Among the great apes, detailed recombination maps only exist for bonobo, chimpanzee, and gorilla. These are produced using the same LD-based method used in humans, allowing direct comparison of recombination maps across species. In all four species, recombination rate varies on a large scale (millions of bases), and this variation is associated with the size of chromosomes, the chromosomal position, the sequence GC content, the gene density, and several other factors [46]. At the fine scale (thousands of bases) recombination rate is determined by the location of the so-called recombination hotspots where about 60% of recombinations occur despite that these hotspots constitute only ~6% of the genome [47]. The location of hotspots is determined by the affinity of the PRDM9 protein for certain DNA motifs present at hotspots. This affinity is encoded in a zinc-finger array whose DNA contacting residues are under strong positive selection. It is now clear that biased gene conversion favors alleles that disrupt hotspots. This depletion of hotspot motifs may result in selection for PRMD9 variants recognizing alternative motifs, producing a turnover of hotspot locations [48]. A comparative analysis of recombination maps of the four species [49, 50] showed that recombination rate on a megabase scale is highly conserved across species, but that the location recombination hotspots are completely different. Only a few hotspots are shared even between chimpanzee subspecies, revealing that turnover of hotspot locations commence at short evolutionary timescales [50].

Comparative studies of recombination have less power than comparative studies of genome sequences: whereas sequence change can be assigned to individual species branches using standard models of molecular evolution, change in recombination rate has so far only been observed as differences between pairs of species. This is because the differences between two species cannot be resolved into the change that occurred in each species without knowledge of recombination rates in the species common ancestor. Fortunately, it is now also possible to construct recombination maps for ancestral species if enough incomplete lineage sorting is present [2]. This approach takes advantage of the fact that gene trees with different topologies must be separated by a recombination event. When sequences are sampled from three different species, the majority of recombination events separating gene trees with different topology will occur in the species ancestral to the two most closely related species. This approach has been used to produce a recombination map of the ancestor of human and chimpanzee [3]. By resolving the differences between humans and chimpanzees into the changes that occurred in each species since their divergence it was shown that recombination rate had evolved more rapidly in humans than in chimpanzees and that striking changes in recombination rate had resulted from a genomic inversion and a chromosome fusion in the human lineage.

5 The X Chromosomes of Great Apes

The unique mode of inheritance of X chromosomes exposes them to population genetic process that differs from that of the autosomes. In a simple population genetic model, the effective population size of the X chromosome will be 3/4 that of the autosomal one. However, this ratio is influenced by many factors such as a difference in generation time and reproductive variance between the sexes, or a stronger propensity of one sex to migrate between subpopulations. More recently it has been suggested that linked selection on the X chromosome in the form of selective sweeps may contribute significantly to a reduced X–autosome ratio. Analysis of diversity along the X chromosomes of the great apes identified extreme selective sweeps in the form of wide regions with strongly reduced diversity and a higher proportion of singleton polymorphisms [51]. The swept regions overlap partially between species, suggesting some amount of recurrent positive selection on the same genes. A separate study exploiting patterns of ILS to measure the cumulative effect of sweeps in the human–chimpanzee ancestor, identified a set of wider regions, spanning the regions identified in extant great ape species [52]. This suggests that regions of the X chromosome are subject to recurrent very strong positive selection. Since these extreme sweeps are only observed on the X chromosomes, it is possible that this is the result of selection of "selfish genes." Such selfish genes, catering only for the preferential transmission of X or Y chromosomes into viable sperm are potentially subject to a particular kind of positive selection called meiotic drive. Even modest transmission distortions will provide selective advantages strong enough to explain the magnitude of these sweeps.

6 Conclusion

The examples of insights provided above represent only the first glimpses of the evolutionary history we share with the great apes as well as the evolution that is private to each species. As genetic diversity across the ranges of each great ape is assayed in more detail, we will get a much deeper understanding of how diverse population genetic processes have shaped genomes very similar to our own.

References

1. Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87
2. Locke DP et al (2011) Comparative and demographic analysis of orang-utan genomes. Nature 469:529–533
3. Scally A et al (2012) Insights into hominid evolution from the gorilla genome sequence. Nature 483:169–175
4. Prüfer K et al (2012) The bonobo genome compared with the chimpanzee and human genomes. Nature 486:527–531
5. Prado-Martinez J et al (2013) Great ape genetic diversity and population history. Nature 499:471–475
6. Kronenberg ZN et al (2018) High-resolution comparative analysis of great ape genomes. Science 360:pii: eaar6343
7. Mailund T, Munch K, Schierup MH (2014) Lineage sorting in apes. Annu Rev Genet 48:519–535
8. Scally A, Durbin R (2012) Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet 13:745–753
9. Dutheil JY et al (2009) Ancestral population genomics: the coalescent hidden Markov model approach. Genetics 183:259–274
10. Hobolth A, Christensen OF, Mailund T, Schierup MH (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet 3:e7
11. Mailund T et al (2012) A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. PLoS Genet 8:e1003125
12. Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. *Nat Genet* 46:919–925
13. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475:493–496
14. Kosiol C et al (2008) Patterns of positive selection in six Mammalian genomes. PLoS Genet 4:e1000144
15. Nielsen R et al (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3:e170
16. Charlesworth B (2013) Background selection 20 years on: the Wilhelmine E. Key 2012 invitational lecture. J Hered 104:161–171
17. Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23:23–35
18. Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39:197–218
19. Boyko AR et al (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4:e1000083
20. Eyre-Walker A, Keightley PD (2009) Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol 26:2097–2108
21. Galtier N (2016) Adaptive protein evolution in animals and the effective population size hypothesis. PLoS Genet 12:e1005774
22. Cagan A et al (2016) Natural selection in the great apes. Mol Biol Evol 33:3268–3283
23. McManus KF et al (2015) Inference of gorilla demographic and selective history from whole-genome sequence data. Mol Biol Evol 32:600–612
24. Enard D, Cai L, Gwennap C, Petrov DA (2016) Viruses are a dominant driver of protein adaptation in mammals. elife 5:pii: e12469

25. Enard D, Messer PW, Petrov DA (2014) Genome-wide signals of positive selection in human evolution. Genome Res 24:885–895
26. Przeworski M, Coop G, Wall JD (2005) The signature of positive selection on standing genetic variation. Evolution 59:2312–2323
27. Nam K et al (2017) Evidence that the rate of strong selective sweeps increases with population size in the great apes. Proc Natl Acad Sci U S A 114:1613–1618
28. Munch K, Nam K, Schierup MH, Mailund T (2016) Selective sweeps across twenty millions years of primate evolution. Mol Biol Evol 33:3065–3074
29. Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. Nat Rev Genet 8:610–618
30. Keightley PD, Eyre-Walker A (2010) What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? Philos Trans R Soc Lond Ser B Biol Sci 365:1187–1193
31. Keightley PD, Eyre-Walker A (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177:2251–2261
32. Kim BY, Huber CD, Lohmueller KE (2017) Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. Genetics 206:345–361
33. Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD (2011) A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics 189:1427–1437
34. Tataru P, Mollion M, Glémin S, Bataillon T (2017) Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. Genetics 207:1103–1119
35. Bataillon T et al (2015) Inference of purifying and positive selection in three subspecies of chimpanzees (Pan troglodytes) from exome sequencing. Genome Biol Evol 7:1122–1132
36. Cooper GM et al (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15:901–913
37. Davydov EV et al (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6:e1001025
38. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20:110–121
39. Siepel A et al (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034–1050
40. Lindblad-Toh K et al (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478:476–482
41. Andrés AM et al (2009) Targets of balancing selection in the human genome. Mol Biol Evol 26:2755–2764
42. Leffler EM et al (2013) Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science 339:1578–1582
43. McVean GAT et al (2004) The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584
44. Kong A et al (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247
45. Hinch AG et al (2011) The landscape of recombination in African Americans. Nature 476:170–175
46. Spencer CCA et al (2006) The influence of recombination on human genetic diversity. PLoS Genet 2:e148
47. Frazer KA et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861
48. Coop G, Myers SR (2007) Live hot, die young: transmission distortion in recombination hotspots. PLoS Genet 3:e35
49. Auton A et al (2012) A fine-scale chimpanzee genetic map from population sequencing. Science 336:193–198
50. Stevison LS et al (2016) The time scale of recombination rate evolution in great apes. Mol Biol Evol 33:928–945
51. Nam K et al (2015) Extreme selective sweeps independently targeted the X chromosomes of the great apes. Proc Natl Acad Sci U S A 112:6413–6418
52. Dutheil JY, Munch K, Nam K, Mailund T, Schierup MH (2015) Strong selective sweeps on the X chromosome in the human-chimpanzee ancestor explain its low divergence. PLoS Genet 11:e1005451