Tempo and mode of genomic mutations unveil human evolutionary history

Yuichiro Hara*

Phyloinformatics Unit, RIKEN Center for Life Science Technologies, 2-2-3 Minatojima-minami, Chuo-ku, Kobe, Hyogo 650-0047, Japan

(Received 9 February 2015, accepted 2 August 2015)

Mutations that have occurred in human genomes provide insight into various aspects of evolutionary history such as speciation events and degrees of natural selection. Comparing genome sequences between human and great apes or among humans is a feasible approach for inferring human evolutionary history. Recent advances in high-throughput or so-called ‘next-generation’ DNA sequencing technologies have enabled the sequencing of thousands of individual human genomes, as well as a variety of reference genomes of hominids, many of which are publicly available. These sequence data can help to unveil the detailed demographic history of the lineage leading to humans as well as the explosion of modern human population size in the last several thousand years. In addition, high-throughput sequencing illustrates the tempo and mode of de novo mutations, which are producing human genetic variation at this moment. Pedigree-based human genome sequencing has shown that mutation rates vary significantly across the human genome. These studies have also provided an improved timescale of human evolution, because the mutation rate estimated from pedigree analysis is half that estimated from traditional analyses based on molecular phylogeny. Because of the dramatic reduction in sequencing cost, sequencing on-demand samples designed for specific studies is now also becoming popular. To produce data of sufficient quality to meet the requirements of the study, it is necessary to set an explicit sequencing plan that includes the choice of sample collection methods, sequencing platforms, and number of sequence reads.

Key words: De novo mutation, human evolution, next-generation sequencing, high-throughput sequencing

INTRODUCTION

The whole-genome sequences of humans and great apes have been cataloged (Chimpanzee Sequencing and Analysis Consortium, 2005; International Human Genome Sequencing Consortium, 2004; Locke et al., 2011; Prüfer et al., 2012; Scally et al., 2012). This sequence information has helped to clarify the evolutionary history of the lineage leading to modern humans in the last millions of years. Additionally, high-throughput sequencing technologies have significantly advanced in the last five years, leading to the further encouragement of these studies at the whole-genome level. While these sequencers produce very short reads (e.g., 125 bp paired-end reads with the Illumina HiSeq SBS Kit v4), they generate hundreds of gigabase pairs of sequence information in a single sequencing run. Resequencing methodology, which utilizes the characteristics of high-throughput sequencers (DePristo et al., 2011; Stratton, 2008), has revealed the whole-genome sequence variants of more than 1,000 individuals of modern humans including a considerable number of rare variants (The 1000 Genomes Project Consortium, 2012). This massive amount of sequence information provides a fine-scale picture of the demographic histories of ancient modern humans during the last tens of thousands of years (Mathieson and McVean, 2014; O’Connor et al., 2015). In addition, resequencing techniques have successfully sequenced the genomes of archaic humans and ancient modern humans whose DNAs were extracted from fossils (Fu et al., 2014; Meyer et al., 2012; Prüfer et al., 2014).

High-throughput sequencing has also clarified how mutations occurred in the human genome, directly providing a timescale for human evolutionary studies from the sequence data. Effective population size $N$ and speciation time $T$ are scaled by $\mu$, the number of mutations per nucleotide per generation, which is estimated based
on sequence comparison. The value of $\mu$ in the human genome had been estimated based on a phylogenetic approach, calculating nucleotide sequence divergence between orthologous sequences of human and chimpanzee with the fossil record as a calibration (Eyre-Walker and Keightley, 1999; Nachman and Crowell, 2000). On the other hand, high-throughput sequencing directly counts the number of mutations per nucleotide per generation by calling de novo mutations in the human germline based on a whole-genome comparison between parents and children (Campbell et al., 2012; Conrad et al., 2011; Kong et al., 2012; Michaelson et al., 2012; Roach et al., 2010).

These studies have also updated our understanding of the tempo and mode of de novo mutations, leading to an elucidation of aspects of human evolutionary history ranging from speciation between humans and great apes to mitosis and meiosis in the germline taking place at this moment. High-throughput DNA sequencing is becoming a core technology for genetic studies owing to the rapid increase in its throughput and the progress of in silico post-sequencing analysis methods (Koboldt et al., 2013). In addition, a dramatic reduction in sequencing cost facilitates on-demand high-throughput sequence production optimized to an experimental plan. Here, I introduce recent progress in our understanding of the tempo and mode of spontaneous mutations in the human and great ape genomes based on high-throughput sequencing. I then review the achievements of human evolutionary studies based on massive sequence information. Finally, I consider how we handle large-scale sequence information for human evolutionary studies.

### UPDATING OUR UNDERSTANDING OF MUTATION RATES

Analysis of the enormous amount of data yielded by high-throughput sequencers has demonstrated newly occurring mutations that were identified by comparing the genome sequences of parents and offspring. Differences between these genome sequences are formed by de novo mutations, which arose in the parents’ germlines and were inherited by their children. Based on such pedigree genome sequencing, the de novo mutation rate in the human nuclear genome was estimated as approximately $0.96–1.3 \times 10^{-8}$ per nucleotide per generation (Table 1) (Besenbacher et al., 2015; Campbell et al., 2012; Conrad et al., 2011; Kong et al., 2012; Michaelson et al., 2012; Roach et al., 2010). This mutation rate is half the

| Species and samples | Number of mutations per nucleotide per generation ($\times 10^{-9}$) | Number of mutations per nucleotide per year ($\times 10^{-9}$) | Paternal age | Maternal age | Paternal contribution (%) | Number of mutations per genome per cell division in spermatogenesis $^{*1}$ | References |
|---------------------|-------------------------------------------------------------------|-------------------------------------------------|--------------|--------------|---------------------------|-----------------------------------------------|------------|
| Human, 10 quartets  | 1                                                                 | 0.31                                           | 33.6         | 31.6         | 74                        | 0.11                                          | Kong et al. (2012) |
| Human, 78 trios     | 1.2                                                               | 0.72                                           | 29.7         | 26.5         | 76                        | 0.16                                          | Michaelson et al. (2012) |
| Human, 8 sperms from an individual | 2.9                                                               | 0.72                                           | 40.0         | –            | 100                       | 0.15                                          | Wang et al. (2012) |
| Human, an ancient individual from 45,000 years ago | 1.1–1.6                                                             | 0.44–0.63                                      | 25$^{*2}$    | 25$^{*2}$    | –                        | –                                             | Fu et al. (2014) |
| Chimpanzee, a pedigree of 9 individuals | 1.2                                                               | 0.46                                           | 24.3         | 26.3         | 89                        | 0.18                                          | Venn et al. (2014) |
| Human, whole exome  | 1.3–2.2                                                           | 0.52–0.88                                      | 32.5$^{*3}$  | 30.4$^{*3}$  | 80$^{*3}$                 | 0.27$^{*2}$                                   | Neale et al. (2012); O’Roak et al. (2012); Sanders et al. (2012) |
| Human, mitochondrial genome | 4.2                                                               | 1.67                                           | 25$^{*2}$    | 25$^{*2}$    | –                        | –                                             | Soares et al. (2009) |
| Human and chimpanzee, 18 processed pseudogenes | 2.5                                                               | 1.25                                           | 20           | 20           | 78$^{*5}$                | –                                             | Nachman and Crowell (2000) |

$^{*1}$ Number of cell divisions in spermatogenesis according to paternal age was calculated based on an equation by Crow (2000). I assumed that cell division rates of human and chimpanzee are 16 and 14 per year, respectively (Crow, 2000; Smithwick et al., 1996), and that chimpanzee sexual maturity is reached at 9 years old (Pusey, 1990). Numbers of mutations per genome per generation were calculated based on genome coverage, assuming human genome size as 3.33 Gb. For the chimpanzee pedigree genome, the number of mutations per genome per generation was calculated based on that of autosomes, assuming $\alpha = 7.8$ and chimpanzee autosomal and X chromosome sizes as 2.7 Gb and 0.16 Gb, respectively.

$^{*2}$ Assuming the generation time provided by Eyre-Walker and Keightley (1999).

$^{*3}$ O’Roak et al. (2012).

$^{*4}$ Inferred based on the mutation rate, paternal age, and paternal contribution provided by O’Roak et al. (2012).

$^{*5}$ Inferred based on $\alpha$ (Miyata et al., 1987) assuming $\alpha = 3.6$ (Nachman and Crowell, 2000).
rate estimated from a molecular phylogenetic approach using processed pseudogenes of human and chimpanzee (2.5 × 10⁻⁸ per nucleotide per generation; Table 1) (Nachman and Crowell, 2000).

Nachman and Crowell (2000) estimated the mutation rate from the equation \( k = 2\mu T/g+4\mu N \), assuming a nucleotide sequence difference between human and chimpanzee (\( k \)) of 0.0133, speciation divergence (\( T \)) time between these species at 5 million years ago, ancestral population size at speciation (\( N \)) of 10,000, and generation time (\( g \)) of 20 years. However, recently discovered ancestral hominid fossils have pushed the speciation time back (see WHOLEGENOME SEQUENCES DISENTANGLE ANCESTRAL HUMAN EVOLUTIONARY HISTORY), and population genetic analyses at the whole-genome scale have indicated a larger ancestral population size than that used in the mutation rate estimation. In addition, the nucleotide divergence (\( k = 0.0133 \)) included the insertion-deletion rate, whereas only single-nucleotide changes should be considered as nucleotide divergence for comparing the mutation rates between the phylogenetic and high-throughput sequencing approaches. These observations are likely to provide a lower mutation rate even with the phylogenetic approach. Indeed, assuming \( k = 0.012 \) after excluding indels, \( N \) ranging from 59,300 to 122,000 (Hara et al., 2012; Scally et al., 2012), \( T \) at 6 million years ago, and a generation time of 20–25 years, I calculated \( \mu \) as 1.10–1.67 × 10⁻⁹, which is comparable with the estimates based on high-throughput sequencing (Table 1).

Prior to the study by Nachman and Crowell, a similar mutation rate (2.36 × 10⁻⁸ per nucleotide per generation) was estimated based on synonymous substitutions in the major histocompatibility complex loci of human, apes and Old World monkeys (Satta et al., 1993). The discrepancy of mutation rates between this study and the pedigree genome analyses can be explained by a decrease of the mutation rate in the great ape lineage, known as hominoid slowdown (Elango et al., 2006; Kim et al., 2006; Li and Tanimura, 1987). This variation in mutation rates is likely because of the longer generation time of hominoids than that of Old World monkeys (Gage, 1998; Langergraber et al., 2012), leading to an increase in the number of replication errors as the number of cell divisions in gametogenesis increases (Li and Tanimura, 1987). A comparison of whole-genome sequences among human, great apes and rhesus macaque indicated that mutation rates in the hominid lineages were approximately half that in the lineage of the Old World monkeys (Scally and Durbin, 2012). For these reasons, it is plausible to use the mutation rate based on pedigree data for elucidating human evolution within the last millions of years.

For identifying de novo mutations based on high-throughput sequencing, candidate mutation sites are usually treated with extreme caution by employing one-by-one validation such as Sanger resequencing and hybridization capture (Conrad et al., 2011; Kong et al., 2012), even when de novo mutations have been identified with highly deep coverage. This is because the sequencing error rate is still non-negligible (e.g., 8.16 × 10⁻⁸ per base pair; Roach et al., 2010). However, discrepancies in estimated mutation rates were observed between whole-genome and exome sequence analyses. Exome sequencing of hundreds of families showed a mutation rate (1.3–2.2 × 10⁻⁸ per nucleotide per generation; Table 1) (Neale et al., 2012; O’Roak et al., 2012; Sanders et al., 2012) that was higher than rates based on whole-genome sequencing. The difference can be explained by more frequent de novo mutations in coding regions than in intergenic regions, which reflect the higher mutation rate in GC-rich coding regions than in intergenic regions (O’Roak et al., 2012; Sanders et al., 2012).

The large contribution of paternal de novo mutations (Table 1) can be explained by the large number of opportunities for replication errors to occur during cell divisions in spermatogenesis. In addition, spermatids in older individuals have experienced more cell divisions than those in younger ones, leading to more mutations caused by DNA replication errors (Crow, 2000). These observations lead to the idea that the number of de novo mutations transmitted to the children’s genomes should increase with paternal age. A project in Iceland sequenced the whole genomes of 78 trios consisting of 219 individuals, which included families with hereditary Autism Spectrum Disorders (ASD) and schizophrenia. The results showed a positive correlation between numbers of de novo mutations in the offspring genomes and paternal age: two mutations arise per year and the number of mutations doubles every 16.5 years (Kong et al., 2012). This observation explained the increased risks of ASD and schizophrenia with increasing paternal age (Kong et al., 2012). The results are also consistent with the correlation between paternal age and the birth rate of children affected with chondrodystrophia, which was noted by Weinberg more than 100 years ago (reviewed in Crow, 2000). Although paternal age did not correlate with the point mutation rate, the project revealed that the numbers of recombinations in the offspring’s genomes increased with maternal age (Kong et al., 2004).

A similar mutation rate to that in modern humans was observed in chimpanzees (Table 1): whole-genome resequencing of a chimpanzee family showed a mutation rate (0.46 × 10⁻⁸ per nucleotide per year) equivalent to that of the human genome (Venn et al., 2014). On the other hand, the paternal contribution of de novo mutations in chimpanzee (89%), as well as the paternal age effect (three mutations per year), was relatively high compared to human (Table 1). The high bias of the mutation rate toward males may be because more sperm is produced in
Comparison of extant and ancient modern human genomes enables us to infer mutation rate based on molecular phylogeny over a shorter time scale than that of human and chimpanzee. Based on the inference of additional mutations in the genomes of extant humans compared to the genome of an ancient modern human, a male from a population that lived in western Siberia 45,000 years ago, the de novo mutation rate was inferred to be $0.44-0.63 \times 10^{-8}$ per nucleotide per year (Table 1) (Fu et al., 2014). This result is consistent with those from the human pedigree genome sequencing (Table 1).

Single-cell genome sequencing, which was accomplished by whole-genome amplification, illustrated genomic variation across gametes as well as somites (Blainey and Quake, 2014). Eight single sperms, from a 40-year-old Caucasian, were collected and sequenced individually (Wang et al., 2012). The number of de novo mutations per nucleotide per generation based on the sperm genome sequencing was more than twice that calculated from human pedigree genome sequencing. This is because pedigree genome sequencing provides an averaged mutation rate of a parental diploid while sperm genome sequencing identifies mutations in a paternal haploid. In addition, the donor in the sperm genome sequencing (40 years old; Table 1) was older than the paternal donors in the pedigree sequencing experiments (up to 33.6 years old; Table 1), which is likely to have resulted in more mutations in the sperm genomes. Based on the number of paternal mutations, I calculated the number of mutations per genome per cell division in spermatogenesis and found that these values were comparable among the sperm and pedigree genome sequencing experiments (Table 1).

**HETEROGENEOUS MUTATION RATE OF NUCLEOTIDES ACROSS THE HUMAN GENOME**

The nucleotide mutation rate is not always homogeneous throughout the genome. A well-known example is the hypermutability of CpG dinucleotides, which have a 15-fold higher mutation rate than other sites (Elango et al., 2008), because of the spontaneous deamination of methylated cytosine in a CpG into thymine (Bird, 1980). Recently, non-random distribution of mutation rate across the human genome has been examined in detail based on high-throughput sequencing (Campbell et al., 2012; Kong et al., 2012; Michaelson et al., 2012). De novo mutations identified in the genomes of ten families, each composed of parents and a pair of identical twins, revealed that DNase I hypersensitivity, high GC content, predominant nucleosome occupancy, high recombination rate, and rich simple- and trinucleotide-repeats are associated with increased mutation rate (Michaelson et al., 2012).

Analyses based on large-scale sequence data have indicated that non-random variation of nucleotide mutation rates is also associated with genomic structural characters. Comparative genomics between humans and other primates revealed that neutral mutations were abundant in open chromatin (Martincorena and Luscombe, 2013) as well as highly transcribed regions (Park et al., 2012). In addition, mapping SNV distribution and replication timing to human chromosomes revealed that point mutations were abundant in genomic regions that are replicated late (Koren et al., 2012; Stamatoyanopoulos et al., 2009). One explanation for this is a shortage of time for DNA repair (Stamatoyannopoulos et al., 2009). Moreover, variation in mutation rates has been observed even between chromosomes. The high mutation rates of the mitochondrial genome and Y chromosome are well-known examples. The mutation rate of human mtDNA ($1.67 \times 10^{-8}$ per nucleotide per year; Table 1) (Soares et al., 2009) is more than ten times as high as that of the human nuclear genome, possibly because of the highly oxidative environment within the mitochondrion (Galtier et al., 2009). For the Y chromosome, an excess of cell divisions in male gametogenesis (Crow, 2000; Miyata et al., 1987) is thought to increase the mutation rate. In addition, nucleotide sequence divergence between humans and chimpanzees differs across the autosomes, implying that the mutation rate is different among the autosomes in the human and chimpanzee lineages (Hodgkinson and Eyre-Walker, 2011). This idea was supported by a coalescence analysis using the human and great ape genomes. While the averages of estimated $\mu_{HC}$ and $\mu_{HCG}$, where $T_{HC}$ and $T_{HCG}$ represent speciation times between human and chimpanzee and between the ancestor of human/chimpanzee and gorilla, respectively, varied among the human chromosomes, the averages of $\mu_{HC}$ and $\mu_{HCG}$ of the human chromosomes were highly correlated (Fig. 1A) (Hara et al., 2012). One simple explanation for this observation is that mutation rates differ across chromosomes through mechanisms that have not yet been clarified (Hara et al., 2012).

As well as point mutations, genomic structural mutation rates vary across the human genome. In the Y chromosome, for example, repeat-rich structure increases the structural mutation rate (Repping et al., 2006). Also, large-scale copy number variation (CNV) data demonstrated that CNVs are abundant in repeat motifs, recombination hotspots, tandem arrays of duplicated sequences (Campbell and Eichler, 2013), and the genomic regions that are replicated in the early and late S phases (Koren et al., 2012). In addition, ultramicro inversions, minute-scale inversions ranging from 5 to 125 bp, were identified from human-chimpanzee genome alignments, and their frequencies vary in different human chromosomes (Fig.
Genomic mutations unveil human evolution

1B) (Hara and Imanishi, 2011). These ultramicro inversions are considered to be generated by homologous recombination during meiosis (Hara and Imanishi, 2011). A negative correlation exists between inversion rates and chromosomal size, which is consistent with higher recombination rates in smaller chromosomes in human and other mammalian genomes (Jensen-Seaman et al., 2004). Whole-genome sequences disentangle ancestral human evolutionary history

Sequence comparison based on even two reference genomes of closely related species can illustrate the process of speciation and the history of ancestral populations. This is because genomes are separated into regions by a large number of recombination events during evolution, and variation in nucleotide sequence divergence across the genome accounts for different demographic histories (Osada, 2014). Based on a coalescence theory, speciation processes of the lineage leading to humans have been inferred at multiple loci using sequences of three or more species. In these studies, population size and speciation time parameters, $\theta = 4\mu N$ and $\tau = \mu T/g$, at all branching points are calculated simultaneously based on maximum likelihood, Bayesian and Hidden Markov Model frameworks. Recent progress in hominid genome sequencing has enabled us to infer human evolutionary history using the whole genomes of human and great apes (Hara et al., 2012; Locke et al., 2011; Mailund et al., 2012; Prado-Martinez et al., 2013; Scally et al., 2012; Yamamichi et al., 2012). Hara et al. (2012), for example, inferred speciation processes in the lineage leading to humans based on the MCMC framework by building whole-genome alignments among human, chimpanzee, gorilla and orangutan. In some of these studies, speciation times were scaled with the newly inferred mutation rate based on human pedigree genome resequencing (Scally and Durbin, 2012). Assuming a mutation rate of around $0.5 \times 10^{-9}$ per nucleotide per year (Conrad et al., 2011; Lynch, 2010; Roach et al., 2010), the speciation time between human and chimpanzee was estimated as around 6–7 million years ago (Hara et al., 2012; Langergraber et al., 2012; Scally et al., 2012), which was consistent with the estimated speciation time based on the fossil record (Brunet et al., 2002, 2005). On the other hand, these speciation times were older than those from previous studies that employed a higher mutation rate based on phylogenetics (Hobolt et al., 2007; Locke et al., 2011; Takahata and Satta, 1997).

Coalescence analyses have also been used for examin-
ing processes of speciation. Some studies using whole-
genome alignments of human and chimpanzee suggested
simple allopatric speciation between the species (Hara et
al., 2012; Yamamichi et al., 2012). This was incompati-
ble with an introgression hypothesis for ancestral human
and chimpanzee, proposed previously (Patterson et al.,
2006), in which introgression was observed only in the X
chromosome. However, the possibility of introgression
remains open to discussion even when the human-chim-
panzee speciation process is examined using the whole-
genome sequences. Human-chimpanzee speciation sce-
narios were tested employing a Hidden Markov Model
framework, and the results favored a model of gradual
migration of the two new species after their initial separa-
tion (Mailund et al., 2012). On the other hand, the results
of a recent comparative genomics between human and
 chimpanzee were incompatible with Patterson et al.’s intro-
gression hypothesis described above (Dutheil et al., 2015).
Detailed inference of the demographic history of human
populations is feasible owing to whole-genome sequencing
of individuals. By tracing alignments of individual
genomes from several human populations, analyses based
on the pairwise sequentially Markovian coalescent model
revealed the time to the most recent common ancestor of
each locus separated by ancestral recombination events
(Li and Durbin, 2011). This method also provided effective
population sizes and their changes over time as well as
the timing of bottlenecks (Li and Durbin, 2011). This
approach has been applied to various studies for unveil-
ing population histories in the hominoids, including anal-
ysis using archaic human genomes (Meyer et al., 2012;
Pagani et al., 2015; Prado-Martinez et al., 2013; Prufer et
al., 2014; Schiffels and Durbin, 2014).

RARE VARIANTS CLARIFY THE EVOLUTION OF
MODERN HUMANS

Rare variants in the human genome, defined as a minor
allele frequency of less than 1% in a population, often rep-
resent newly derived mutations and thus have been used
for clarifying the demographic histories of modern
humans (Coventry et al., 2010; The 1000 Genomes Project
Consortium, 2012; Gravel et al., 2011). Although sequenc-
ing a substantial number of individuals is required to verify the rareness of the variants (Keinan and Clark, 2012), large-scale sequence projects targeting thousands of individuals have enabled rare-variant analyses. The 1,000 Genomes Project provided more than 20 million novel rare mutations, which have been widely used to infer population structures between and within populations as well as the demographic history of the populations (Colonna et al., 2014; The 1,000 Genomes Project Consortium, 2012; Gravel et al., 2011; Mathieson and McVean, 2014). Similarly, target resequencing of
202 genes in more than 14,000 individuals revealed that
the degree of rare variants reflects the geography and
demography of European populations (Nelson et al.,
2012). Currently, exome sequencing is the best approach
to sequence thousands of individuals with consistent cov-
erage for certain regions across the genome (Fu et al.,
2013; Tennessen et al., 2012). Sequencing of more than
6,500 human exomes revealed that most of the deleteri-
ous mutations in the coding regions emerged 5,000–
10,000 years ago (Fu et al., 2013). This result suggested
that deleterious mutations have been spreading rapidly
in the human population due to its recent explosive
increase. Another study using these exome data
revealed the fine-scale population structure of European-
American populations (O’Connor et al., 2015).

A BRIEF GUIDE TO INFERRING HUMAN EVOLU-
TION FROM LARGE-SCALE SEQUENCE DATA

When conducting population genetics analyses using
large-scale sequence information, it is necessary to evalu-
ate the outputs in multiple ways. To identify variations
or mutations based on genomic sequence comparison,
assessments are required to see if the nucleotides aligned
in a column are truly homologous. This is because mis-
matches in alignments do not necessarily correspond to
mutations: they may be merely alignment errors. A
large number of whole-genome alignment tools have been
developed and recently benchmarked in a competition
called Alignathon (Earl et al., 2014), where substantial
differences were observed between the alignment meth-
ods using sequences artificially generated based on hom-
ind phylogeny. This competition framework is expected
to accelerate the improvement of alignment perform-
ances, leading to an expansion of alignable regions and
thus identification of more mutations in the human
genome. Post-alignment processing is also important
to obtain valid genetic variation. The first and necessary
step is to discard ambiguously aligned sites (e.g., aligned
sites near gaps) that include abundant false positives of
the variants. An optional step is to remove hypermut-
able regions such as CpG sites and apparently hypermut-
able sites that reflect genomic changes other than simple
point mutations and indels. These filtering steps are
also helpful for identifying genomic structural mutations
such as ultramicro inversions and biased gene conver-
sions (Capra et al., 2013; Hara and Imanishi, 2011).
However, it is likely that the discarded sites will include
considerable numbers of true positives and therefore that
the results obtained from such filtered alignments will be
underestimates. The de novo mutation rate based on the
human pedigree genome sequencing is expected to be the
lower limit due to the discarding of suspicious alignments
(Campbell and Eichler, 2013). When comparative anal-
yses are carried out using whole-genome alignments of
human and great apes, this mutation rate will be still
applicable. This is because ambiguously aligned sites, which can include true positive mismatches, need to be filtered from whole-genome alignments, and the filtered alignments are then modeled with the lower mutation rate rather than the true rate.

Large-scale sequence information from high-throughput sequencers is obtainable by sequencing original samples as well as by downloading data from public bio-archives. In both cases, the credibility of results relies on the validity of the sequencing strategy, from sample collection to sequencing platforms, as well as sequence quality. Sequence data based on facile experiments produce little information however the sequences are subsequently analyzed. For example, adequate numbers of individuals must be sequenced to confirm the rareness of variants (Keinan and Clark, 2012). To reduce problems caused by technical errors, all sampling for an exome study must be performed using the same target, and all libraries should be sequenced using the same sequencing platform (Clark et al., 2011; Tennessen et al., 2011). In addition, sequence coverage needs to be high to distinguish variants from sequencing errors (Clark et al., 2011; Tennessen et al., 2011).

Understanding variation of the mutation rate across the human genome leads to establishing evolutionary models for specific genomic regions as well as accurate identification of sites that are under natural selection. A mutation rate based on exome rather than whole-genome sequencing should be applied to evolutionary analyses that use coding regions because of the difference in estimated mutation rates. Some regions with extraordinarily high mutation rates such as indel hotspots may lead to an incorrect interpretation of an evolutionary event unless homoplasy is carefully considered (Kvikstad and Duret, 2014). For this reason, both tempo and mode of de novo mutations should be analyzed in detail for fine-scale evolutionary studies. Sequencing of cancer genomes is a possible candidate for such a purpose because of their high mutation rate (Pleasance et al., 2010). However, the mutation spectrum of cancer genomes is considerably different from that of the germline: SNV density in cancer genomes correlates with some features of somatic cell chromatin organization such as H3K9me3 modification levels, while SNP density in populations correlates with other features (Schuster-Bockler and Lehner, 2012). Therefore, a model system for collecting various germline mutations effectively should be established in the near future. This system will directly identify de novo mutations that occur during spermatogenesis, which can be separated from natural selection during mating and are an ultimate source in generating the diversity of life.

REFERENCES

Besenbacher, S., Liu, S., Izarzugaza, J. M., Grove, J., Belling, K., Bork-Jensen, J., Huang, S., Als, T. D., Li, S., Yadav, R., et al. (2015) Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. Nat. Commun. 6, 5969.

Bird, A. P. (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. 8, 1499–1504.

Blainey, P. C., and Quake, S. R. (2014) Dissecting genomic diversity, one cell at a time. Nat. Methods 11, 19–21.

Brunet, M., Guy, F., Pilbeam, D., Mackaye, H. T., Likius, A., Ahounta, D., Beauvilain, A., Blondel, C., Bocherens, H., Boissière, J. R., et al. (2002) A new hominid from the Upper Miocene of Chad, Central Africa. Nature 418, 145–151.

Brunet, M., Guy, F., Pilbeam, D., Lieberman, D. E., Likius, A., Mackaye, H. T., Ponce de Leon, M. S., Zollikofer, C. P., and Vignaud, P. (2005) New material of the earliest hominid from the Upper Miocene of Chad. Nature 434, 752–755.

Campbell, C. D., and Eichler, E. E. (2013) Properties and rates of germline mutations in humans. Trends Genet. 29, 575–584.

Campbell, C. D., Chong, J. X., Malig, M., Ko, A., Dumont, B. L., Han, L., Vives, L., O’Roak, B. J., Sudmant, P. H., Shendure, J., et al. (2012) Estimating the human mutation rate using autozygos in a founder population. Nat. Genet. 44, 1277–1281.

Capra, J. A., Hubisz, M. J., Kostka, D., Pollard, K. S., and Siepel, A. (2013) A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. PLoS Genet. 9, e1003684.

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69–87.

Clark, M. J., Chen, R., Lam, H. Y., Karzewski, K. J., Chen, R., Euskirchen, G., Butte, A. J., and Snyder, M. (2011) Performance comparison of exome DNA sequencing technologies. Nat. Biotechnol. 29, 906–914.

Colonna, V., Ayub, Q., Chen, Y., Pagani, L., Luisi, P., Pybus, M., Garrison, E., Xue, Y., Tyler-Smith, C., and The 1000 Genomes Project Consortium (2014) Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. Genome Biol. 15, R88.

Conrad, D. F., Keebler, J. E., DePristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C. L., Torroja, C., Garimella, K. V., et al. (2011) Variation in genome-wide mutation rates within and between human families. Nat. Genet. 43, 712–714.

Coventry, A., Bull-Otterson, L. M., Liu, X., Clark, A. G., Maxwell, T. J., Crosby, J., Hixson, J. E., Rea, T. J., Muzny, D. M., Lewis, L. R., et al. (2010) Deep resequencing reveals excess rare random variants consistent with explosive population growth. Nat. Commun. 1, 131.

Crow, J. F. (2000) The origins, patterns and implications of human spontaneous mutation. Nat. Rev. Genet. 1, 40–47.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippiakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498.

Dutheil, J. Y., Munch, K., Nam, K., Mailund, T., and Schierup, M. H. (2015) Strong selective sweeps on the X chromosome in the human-chimpanzee ancestor explain its low divergence. PLoS Genet. 11, e1005451.

Earl, D., Nguyen, N., Hickey, G., Harris, R. S., Fitzgerald, S.,
Genomic mutations unveil human evolution

297–304.

Nam, K., and Ellegren, H. (2012) Recombination drives vertebrate genome contraction. PLoS Genet. 8, e1002680.

Neale, B. M., Kou, Y., Liu, L., Ma’ayan, A., Samocha, K. E., Sabo, A., Lin, C. F., Stevens, C., Wang, L. S., Makarov, V., et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485, 242–245.

Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., St Jean, P., Verzili, C., Shen, J., Tang, Z., Bacana, S. A., Fraser, D., et al. (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337, 100–104.

O’Connor, T. D., Fu, W., NHLBI GO Exome Sequencing Project, ESP Populatin Genetics and Statistical Analysis Working Group, Turner, E., Mychaleckyj, J. C., Logsdon, B., Auer, P., Carlson, C. S., Leal, S. M., et al. (2015) Rare variation facilitates inferences of fine-scale population structure in humans. Mol. Biol. Evol. 32, 653–660.

O’Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., Levy, R., Ko, A., Lee, C., Smith, J. D., et al. (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485, 246–250.

Osaada, N. (2014) Extracting population genetics information from a diploid genome sequence. Front. Ecol. Evol. 2, 7.

Pagani, L., Schifflers, S., Gurdasani, D., Danecek, P., Scally, A., Chen, Y., Xue, Y., Haber, M., Ekon, R., Olijira, T., et al. (2015) Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. Am. J. Hum. Genet. 96, 986–991.

Park, C., Qian, W., and Zhang, J. (2012) Genomic evidence for elevated mutation rates in highly expressed genes. EMBO Rep. 13, 1123–1129.

Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S., and Reich, D. (2006) Genetic evidence for complex speciation of humans and chimpanzees. Nature 441, 1103–1108.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M. L., Ordóñez, G. R., Bignell, G. R., et al. (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463, 191–196.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O’Connor, T. D., Santpere, G., et al. (2013) Great ape genetic diversity and population history. Nature 499, 471–475.

Prüfer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J. R., Wałęsa, B., Koren, S., Sutton, G., Kodira, C., Winer, R., et al. (2012) The bonobo genome compared with the chimpanzee and human genomes. Nature 486, 527–531.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinez, A., Renaud, G., Sudmant, P. H., de Filippo, C., et al. (2014) The complete genome sequence of a Neandertal from the Altai Mountains. Nature 505, 43–49.

Pusey, A. E. (1990) Behavioural changes at adolescence in chimpanzees. Behaviour 115, 203–246.

Repping, S., van Daalen, S. K., Brown, L. G., Korver, C. M., Lange, J., Marszalek, J. D., Pyniatikova, T., van der Veen, F., Skaletsky, H., Page, D. C., et al. (2006) High mutation rates have driven extensive structural polymorphism among human Y chromosomes. Nat. Genet. 38, 463–467.

Roach, J. C., Glusman, G., Smit, A. F., Huff, C. D., Hubley, R., Shattuck, T., Rowen, L., Pani, K. P., Goodman, N., Bamshad, M., et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328, 636–639.

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., Ercan-Sencicke, A. G., DiLullo, N. M., Parikshak, N. N., Stein, J. L., et al. (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485, 237–241.

Satta, Y., O’Huigin, C., Takahata, N., and Klein, J. (1993) The synonymous substitution rate of the major histocompatibility complex loci in primates. Proc. Natl. Acad. Sci. USA 90, 7480–7484.

Scally, A., and Durbin, R. (2012) Revising the human mutation rate: implications for understanding human evolution. Nat. Rev. Genet. 13, 745–753.

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., et al. (2012) Insights into hominid evolution from the gorilla genome sequence. Nature 483, 169–175.

Schiffls, S., and Durbin, R. (2014) Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 46, 919–925.

Schuster-Bockler, B., and Lehner, B. (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature 488, 504–507.

Smithwick, E. B., Young, L. G., and Gould, K. G. (1996) Duration of spermatogenesis and relative frequency of each stage in the seminiferous epithelial cycle of the chimpanzee. Tissue Cell 28, 357–366.

Soares, F., Ermini, L., Thomson, N., Mormina, M., Rito, T., Rohl, A., Salas, A., Oppenheimer, S., Macaulay, V., and Richards, M. B. (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. Am. J. Hum. Genet. 84, 740–759.

Stamatoyannopoulos, J. A., Adzhubei, I., Thurman, R. E., Kryukov, G. V., Mirkin, S. M., and Sunyaev, S. R. (2009) Human mutation rate associated with DNA replication timing. Nat. Genet. 41, 393–395.

Stratton, M. (2008) Genome sequencing and genetic variation. Nat. Biotechnol. 26, 65–66.

Takahata, N., and Satta, Y. (1997) Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. Proc. Natl. Acad. Sci. USA 94, 4811–4815.

Tennissen, J. A., O’Connor, T. D., Bamshad, M. J., and Akey, J. M. (2011) The promise and limitations of population exomes for human evolution studies. Genome Biol. 12, 127.

Tennissen, J. A., Bigham, A. W., O’Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64–69.

The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–66.

Venn, O., Turner, I., Mathieson, I., de Groot, N., Bontrop, R., and McVean, G. (2014) Strong male bias drives germline mutation in chimpanzees. Science 344, 1272–1275.

Wang, J., Fan, H. C., Behr, B., and Quake, S. R. (2012) Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. Cell 150, 402–412.

Yamamichi, M., Gojobori, J., and Innan, H. (2012) An autosomal analysis gives no genetic evidence for complex speciation of humans and chimpanzees. Mol. Biol. Evol. 29, 145–156.