SCIENCE CHINA
Life Sciences

•INSIGHT•

https://doi.org/10.1007/s11427-020-1709-6

DNA sequencing: the key to unveiling genome

Suhui Chen & Xuehui Huang*

Shanghai Key Laboratory of Plant Molecular Sciences, College of Life Sciences, Shanghai Normal University, Shanghai 200234, China

Received April 13, 2020; accepted April 24, 2020; published online May 20, 2020

Citation: Chen, S., and Huang, X. (2020). DNA sequencing: the key to unveiling genome. Sci China Life Sci 63, https://doi.org/10.1007/s11427-020-1709-6

The genome, containing total genetic material in the organism, i.e., DNA, and RNA for some viruses, encodes the information needed for all life activity. Besides the DNA in cell nucleus, mitochondrial DNA and chloroplast DNA are also important components of the genome. Using high-throughput sequencing, a tremendous amount of genomic data has been obtained. Currently, 1,704 archaeal, 26,075 bacterial, 16,837 viral, and 4,688 eukaryotic genomes have been sequenced and submitted to the GenBank database (https://www.ncbi.nlm.nih.gov/genome). These abundant sequences have greatly accelerated basic research, in areas such as gene function, genomic diversity and structure, and even life origins and evolution. This review summarizes current knowledge of genome structure and genomic evolution, and advanced sequencing technologies.

Complexity and diversity of the genome. Genomes act as information storage systems, likened to electronic storage systems, that record variations among species. Typically, the number of genes required for function of an organism increases with its complexity. *Nasuia deltocephalinicola*, a kind of endosymbiotic bacteria, has only 137 coding genes and displays the smallest genome to the best of current knowledge (Bennett and Moran, 2013). In contrast, gene numbers of mammals can reach or exceed 25,000. In prokaryotes and small eukaryotes, a positive correlation exists between genome size and gene number (Hou and Lin, 2009). However, the ratio of genome size to gene number is not necessarily constant in eukaryotes, which is known as the C-value paradox. For instance, dinoflagellates are a large group of marine algae, and their nuclear genome varies from 1 to 270 Gb. The main cause of variations in genome size is the proportion of repetitive sequences (Ren et al., 2018).

Repetitive sequences range in size from several bases (simple sequence repeats, e.g., "ATATATAT") to millions of bases (large transposable elements), and account for over half of the human genome. Repeated sequences are categorized into moderately repetitive sequences and highly repetitive sequences based on copy numbers. Interspersed repeats (SINE and LINE) and partly tandem repeats (microsatellites and minisatellites) are two groups of moderately repetitive sequences. *Alu* repeat elements, the most abundant human SINE, compose approximately 11% of the human genome (Batzer and Deininger, 2002). *Alu* elements transpose within the human genome and are responsible for chromosome rearrangement during evolution. Satellite sequences are highly repeated and significant DNA components of heterochromatin. These repeats are mainly located in pericentromeric and telomeric regions of chromosomes to ensure the formation and maintenance of heterochromatin.

Mysteries of genome evolution. Genomic evolution involves small and large-scale dynamic changes in genomes over time. Evolutionary events are detected by comprehensive analysis of genome sequences at different levels of evolutionary hierarchies. Based on numerous reports, duplications of whole genomes (polyploidization) or segments, inversions, deletions, transposable element (TE)-mediated insertions and excisions all play significant roles in genomic evolution (Platt II et al., 2018).

Polyploidy is frequently observed in plants, and the frequency of polyploidy reaches 95% in ferns. Estimated polyploid frequency is ~50% in angiosperms (Grant, 1975);

*Corresponding author (email: xhhuang@shnu.edu.cn)

polyploidy is relatively rare in animals. Polyploidy is an ancient and recurrent process. An ancient genome doubling event occurred in an ancestor of modern grasses, and new polyploid populations are still forming in Goatsbeard (Soltis et al., 2015). Given the frequent occurrence of polyploidy, how genomes are maintained at a size far less than expected is intriguing. An explanation is the loss of genes after establishment of polyploidy. A study in maize demonstrated that approximately half of duplicated genes present after the first genome doubling in a progenitor species are now absent (Messing et al., 2004). The removal of duplicate genes is not random. Analysis of polyploidy in *Arabidopsis* demonstrates that classes of genes involved in transcription and signal transduction are preferentially retained compared to genes involved in DNA repair. Moreover, divergence following duplication events likely fuel evolutionary adaptation. More than half of *Arabidopsis* genes acquired different expression patterns, and 62% of recently duplicated genes showed functional diversification after polyploidy, based on a thorough analysis of evolutionary history of the genus (Makino et al., 2010). Alterations in DNA methylation, histone modification and chromatin structure, as well as small RNAs are responsible for gene silencing after the establishment of polyploidy. Although the creation of a newly polyploid species is likely a rare event, polyploidy does increase genetic diversity and alter gene expression. These processes may enhance phenotypic variability and adaptation to complex environment conditions compared to adaptability of diploid progenitors (Soltis et al., 2015).

TEs also play an important role in genome evolution. TEs are mobile genetic elements and are widespread in eukaryotic genomes. They account for at least half of the human genome and more than 80% of the maize genome. In previous investigations, TEs were regarded as "selfish DNA" that did not contribute to development or function. With new sequencing technologies and bioinformatic methods, the essential roles of TE have been elucidated. Different TE accumulation occurs among organisms, and TE expansion is considered a primary mechanism for synthesis of new nuclear DNA in eutherian (placental mammals) and avian evolution. TEs counter their expansion by unequal homologous and illegitimate recombination. Differences in the efficiency of these processes may determine variation in genome size among species (Bennetzen and Wang, 2014). Besides acting as drivers of genome size, TEs are also responsible for alterations in gene coding and gene expression. Mobile TEs may insert nearby or in a gene body, causing mutations in protein-coding sequences, introns, or promoter sequences. As a result, changes in amino acid sequences or expression patterns are encountered. TE also transposes existing promoters or enhancers. Amplification and redistribution of transcription binding sites (TFBSs) may be formed by TEs that endow genomes with transcriptional plasticity. A recent investigation showed that large amounts of TFBSs originated from TE in human genomes (Kellner and Makałowski, 2019). Moreover, accumulation of TE in heterochromatin, reinforces their heterochromatic states as TE accompanied with epigenetic silencing.

As an important genomic variation for microevolution (e.g., intraspecies genetic variation), structure variations (SV) exist in various forms, including inversions, translocations, different repeated numbers of microsatellites, and copy number variations (CNVs). CNVs display variable numbers of copies for large DNA segments, ranging from 1 kilobase to several megabases. CNVs may occupy 13% of the human genome. Presence/absence variants (PAVs) are an extreme form of CNVs, where sequences are present in one genome but absent in another. CNVs and PAVs are associated with stress adaptation. In soybean research, over 800 genes involved in biotic stress affected by CNVs and PAVs were discovered (McHale et al., 2012). SVs may occur from recombination, replication and DNA break repair errors, and polyploidy accompanies SVs in plants. SVs play an indispensable role in creating genome diversity, leading to variants in DNA and DNA rearrangements. Disruption of gene function and chromatin structure and alteration of gene expression are also part of SV activity. A huge number of SVs are in telomeric regions.

Selective pressures also affect genomic evolution. These pressures can be divided into two categories: "exaptation" and "adaptation". The term of exaptation is defined as "features that now enhance fitness but were not built by natural selection for their current role". In contrast, adaptation is a naturally selected feature. A well-documented example of exaptation is the legs and skin of tetrapods acted as wings in bats (Brosius, 2019). The concept of exaptation can assist in achieving a better understanding of evolutionary events that occurred in genomes. For instance, a novel module might be evolved from an existing gene out of neutrally altered sequences (*de novo)* origination. Whether the modification is functional is serendipitous and whether it is beneficial is unknown. TEs and SVs evolve neutrally, which supports the exaptation at the molecular level. Also, a strong positive selection for removing dispensable genes is predicted by the streamlining theory for prokaryotes (Koonin and Wolf, 2008). Many genes may be retained in support of environmental adaptation.

Updated Sequence technology accelerates genome research. DNA sequencing originated in 1977 as "Sanger sequencing". Subsequently, the Human Genome Project, the world's largest collaborative biological project to date, was completed after 13 years and a cost of almost 3 billion dollars. This project produced the first complete code of the human genome on a large scale. Sanger sequencing, which is slow and costly, was gradually replaced by next-generation sequencing (NGS). NGS has greatly accelerated genome

investigation, and genomes for many organisms have been acquired, as mentioned above (Zhang et al., 2019). NGS is relatively inexpensive and supports high-throughput for genome characterization. Nevertheless, NGS is not without drawbacks. The biggest issue is associated with short reads (100–400 bp) that lead to misalignment, mis-assembly and frequent assembly gaps. Further, repetitive sequences are widely distributed and many of them are much longer than common reads in NGS. *De novo* assembly using NGS data is, thus, more difficult. Meanwhile, although single-nucleotide variations (SNVs) and short indels are easily detected, large fragments such as structural variations (SVs) and long mRNA transcripts are still challenging (Ranz and Clifton, 2019).

The second problem is amplification bias. NGS relies on PCR to enhance signals, and regions with extreme GC% are inefficiently amplified. As a result, these regions are poorly covered. In the human genome, over 160 euchromatic gaps are unfilled, and the majority are repeats with high GC content (Genovese et al., 2013). Therefore, new technologies were needed to fill-in genomic gaps.

In 2011, third-generation sequencing (TGS) was created by Pacific Biosciences (PacBio), termed "single-molecule real-time" (SMRT) sequencing, taking advantage of the full capabilities of DNA polymerase and utilizing fluorescently labeled nucleotides. Oxford Nanopore Technology (ONT) released nanopore sequencing in 2014. SMRT and ONT, two long-read sequencing platforms, could detect epigenetic modifications such as methylation and show excellent performance at analysis of repeats, SVs, haplotype phasing, and long transcripts. Bias is reduced because no PCR is included in these technologies. Maximum read length is not limited in ONT. In the literature, 1 Mb reads have been obtained. This technology is cheaper, portable, less time-consuming, and can also identify RNA and single-molecule proteins (Kolmogorov et al., 2017). TGS can overcome weaknesses that exist in NGS, but its high error rate (10%–25%) cannot be avoided. Delicate extraction of long and intact genomic DNA segments is also required. SMRT can attain high accuracy (>99.8%) using circular consensus sequencing (CCS). In this process, a ligated circular DNA template is read multiple times by sacrificing read length due to limited polymerase capacity (Wenger et al., 2019). The ONT 1D library appears to reduce the error rate to about 3%, where both strands of a DNA molecule are sequenced successively. More coverages with a full-length 1D library are required to achieve a high level of accuracy.

Conclusion and perspectives. An increasing number of genome sequences have been published since the advent of NGS. While TGS have improved sequence analysis of some complex genomic regions, significant room for improvement, including the high error rate of ONT and the high cost of SMRT. Affordable and high-accuracy TGS further augment genomic research. Also, the significance and function of complex genomic regions require more study. The issue of integration of current databases is also crucial. The genome of individual organisms is insufficient for analysis of genomic diversity and evolution among species, and broader range of sequences is needed to identify associations of selective pressures and genetic variation. Application of pan genomic information (Zhao et al., 2018) will support efforts to decipher the origination, organization, functionality, and evolution of genomes at nucleotide and structural levels. Currently, existing technologies, such as single reference genome, limit further development and more focus on pan-genome study are acquired. Updated algorithms enabling *de novo* assembly of short reads and efficient representation of genome graphs are also an urgent need.

Despite the leap from genome sequencing to gene annotation, a greater challenge is to understand how genome sequences ultimately affect phenotype and disease and to precisely predict variant effects. Comprehensive analyses of cell types and human populations along with improvements in experimental methods and algorithms will be required to build accurate computational prediction models and directly assess the consequences of genetic variants. Large and high-resolution TGS projects will be crucial complements to cataloging of variants and identifying coding and non-coding regions associated with diseases. Such developments will help make molecular diagnosis and genome-guided disease treatment a reality.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

References

Batzer, M.A., and Deininger, P.L. (2002). Alu repeats and human genomic diversity. Nat Rev Genet 3, 370–379.

Bennett, G.M., and Moran, N.A. (2013). Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a phloem-feeding insect. Genome Biol Evol 5, 1675–1688.

Bennetzen, J.L., and Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol 65, 505–530.

Genovese, G., Handsaker, R.E., Li, H., Altemose, N., Lindgren, A.M., Chambert, K., Pasaniuc, B., Price, A.L., Reich, D., Morton, C.C., et al. (2013). Using population admixture to help complete maps of the human genome. Nat Genet 45, 406–414.

Grant, V. (1975). Genetics of Flowering Plants (New York: Columbia University Press).

Hou, Y., and Lin, S. (2009). Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. PLoS ONE 4, e6978.

Brosius, J. (2019). Exaptation at the molecular genetic level. Sci China Life Sci 62, 437–452.

Kellner, M., and Makałowski, W. (2019). Transposable elements significantly contributed to the core promoters in the human genome. Sci China Life Sci 62, 489–497.

Kolmogorov, M., Kennedy, E., Dong, Z., Timp, G., and Pevzner, P.A. (2017). Single-molecule protein identification by sub-nanopore sensors. PLoS Comput Biol 13, e1005356.

Koonin, E.V., and Wolf, Y.I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res 36, 6688–6719.

Makino, T., Knowles, D.G., and McLysaght, A. (2010). Functional divergence of duplicated genes. In Evolution after Gene Duplication. (Wiley-Blackwell). pp. 23–30.

McHale, L.K., Haun, W.J., Xu, W.W., Bhaskar, P.B., Anderson, J.E., Hyten, D.L., Gerhardt, D.J., Jeddeloh, J.A., and Stupar, R.M. (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. Plant Physiol 159, 1295–1308.

Messing, J., Bharti, A.K., Karlowski, W.M., Gundlach, H., Kim, H.R., Yu, Y., Wei, F., Fuks, G., Soderlund, C.A., Mayer, K.F.X., et al. (2004). Sequence composition and genome organization of maize. Proc Natl Acad Sci USA 101, 14349–14354.

Platt II, R.N., Vandewege, M.W., and Ray, D.A. (2018). Mammalian transposable elements and their impacts on genome evolution. Chromosome Res 26, 25–43.

Ranz, J., and Clifton, B. (2019). Characterization and evolutionary dynamics of complex regions in eukaryotic genomes. Sci China Life Sci 62, 467–488.

Ren, L., Huang, W., Cannon, E.K.S., Bertioli, D.J., and Cannon, S.B. (2018). A mechanism for genome size reduction following genomic rearrangements. Front Genet 9, 454.

Soltis, P.S., Marchant, D.B., Van de Peer, Y., and Soltis, D.E. (2015). Polyploidy and genome evolution in plants. Curr Opin Genet Dev 35, 119–125.

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37, 1155–1162.

Zhang, W., Gao, Y., Long, M., and Shen, B. (2019). Origination and evolution of orphan genes and *de novo* genes in the genome of *Caenorhabditis elegans*. Sci China Life Sci 62, 579–593.

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., et al. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet 50, 278–284.