frontiers in
ECOLOGY AND EVOLUTION

OPINION ARTICLE
published: 02 April 2014
doi: 10.3389/fevo.2014.00007

# Extracting population genetics information from a diploid genome sequence

***Naoki Osada*** [1,2]*

[1] Department of Population Genetics, National Institute of Genetics, Mishima, Japan
[2] Department of Genetics, The Graduate University for Advanced Studies (SOKENDAI), Mishima, Japan
*Correspondence: nosada@nig.ac.jp

***Edited by:***
James J. Cai, Texas A&M University, USA

***Reviewed by:***
Tina Hu, Princeton University, USA
Gerton Lunter, Wellcome Trust Centre for Human Genetics, UK

**Keywords: polymorphism, nucleotide diversity, sequencing, non-model organism, genome sequence**

Due to advances in sequencing technologies, large-scale genomic research has become feasible for many biologists who study organisms that are not traditionally used as model organisms. Many genomes from populations of non-model organisms have been sequenced using these new technologies, providing novel insights into the underlying mechanisms and patterns of evolution of particular traits (e.g., Ellegren et al., 2012; Jones et al., 2012; Martin et al., 2013). However, many biologists studying non-model organisms, particularly those with large genomes, have not yet entered the era of population genomics because of costs limit. Therefore, generally genome sequencing projects, in which a genome from single individual is sequenced as a reference genome, and population genomics projects, in which complete genomes of multiple individuals are sequenced, are to be in different regimes for many researchers. Because some biologists still misunderstand that population genetic information is obtained only with "population" samples, important population genetics information from a small number of individuals are often ignored and not described in literatures.

However, population genetics theory has predicted that a selection of population genetics statistics could be estimated without studying a large number of individuals when many genetically independent loci were investigated. In the framework of massively parallel sequencing, single nucleotide polymorphisms (SNPs) can be identified by mapping many short-read sequences to reference or *de novo* assembled genomes; heterozygous SNPs between two chromosomes represent the genetic diversity of a population, unless strong population structure (e.g., inbreeding) exists.

For example, an estimation of nucleotide diversity ($\pi$) could be inferred from a single genome sequence of a diploid individual. By definition, $\pi$ is the average number of nucleotide differences between random samples of two alleles from a population. If only two alleles from one locus are examined, there could be a large stochastic variance for the estimator of $\pi$. However, genome sequences are the result of many recombination events in the past; therefore, any given genomic sequence is a sample of many different genomic loci that have different histories (genealogies). Therefore, the variance of $\pi$ would be fairly small when it is estimated using a whole genome sequence, except for very small and/or rarely recombined genomes (Pluzhnikov and Donnelly, 1996; Felsenstein, 2006). The exome study of multiple human individuals showed that the number of protein-coding heterozygous SNPs within individuals is fairly constant among individuals in the same population group (Ng et al., 2009).

One limitation for this sort of analysis is the quality of data for genome sequences and read numbers. The rate of heterozygous SNPs is highly dependent on the coverage depth (Bentley et al., 2008). Deficiencies of coverage will bias the estimate toward lower values. Another problem may be the distance between the mapping sample and the reference genome. When the genetic divergence between a sample and a reference is relatively large, reads from non-reference alleles are less plausible to be mapped on the reference genome, leading to underestimation of $\pi$. In addition, when we identify SNPs using *de novo* assembled genomes, care must be taken that genomes are not separately assembled into two haploid genomes, which could occur when genetic diversity within a population is very high. In this case, heterozygous SNPs tend to be lost in the resulting diverged contigs.

Despite the above limitation, such information will aid to understanding of much genetic variation exists in the population, how ecological factors affect genetic diversity among many types of organisms, and how the numbers of segregating non-synonymous and synonymous mutations relate to effective population sizes (Akashi et al., 2012; Lanfear et al., 2014). Recently, an alternative transcriptome-level approach to estimate population genetics parameters without sequencing genomes of multiple individuals, providing a cost-effective option, has been also proposed and implemented (Gayral et al., 2013; Loire et al., 2013). Regardless of the method used, accumulation of such population genetics data would be very important for answering many evolutionary questions, and the presentation of population genetics statistics is desirable for future genome-wide studies.

Li and Durbin recently developed the PSMC method, a pairwise version of the sequentially Markovian coalescent (McVean, 2009), to infer past demography using a single genome sequence (Li and Durbin, 2011). The method

significantly enhanced research for exploring an important aspect of demography using a single diploid genome sequence, and its use has been widely reported (e.g., Higashino et al., 2012; Miller et al., 2012; Prado-Martinez et al., 2013; Zhao et al., 2013). However, it should be noted that the method is effective only when the assembled chromosomes are sufficiently long with given recombination rate; the method is not suitable for estimating very recent changes in population size (Keinan and Clark, 2012; Sheehan et al., 2013).

More recently, Sheehan et al. (2013) developed an efficient implementation of sequentially Markovian coalescent for use with multiple individuals. Currently, the densest sampling in natural populations is achieved in humans. Many novel methods that is applicable to genome-wide polymorphism data have been developed and utilized to analyze human data, such as Approximate Bayesian Computation (ABC) methods (e.g., Beaumont et al., 2002) and their derivatives (Nakagome et al., 2013), and composite-likelihood methods using site frequency spectrum (Gutenkunst et al., 2009; Excoffier et al., 2013) or identity by state tract length (Harris and Nielsen, 2013). It is anticipated that these approaches will become widely used in future genome-wide population studies in non-model organisms.

In addition to the estimation of demography, although sampling bias may seriously affect some estimators of population genetics parameters in the presence of inbreeding and population structure, some analysis may be robust against the bias. For example, it has been shown that genetic diversity within populations decreases near functional regions of the genome owing to natural selection in mammals and *Drosophila* (selective sweep or background selection; Hernandez et al., 2011; Sattath et al., 2011; Halligan et al., 2013). Although this pattern was initially identified using the genome sequences of multiple individuals, we could observe a similar trend using a single diploid genome. Osada et al. (2013), by re-analyzing the data of Yan et al. (2011) showed that when the diversity level was normalized by divergence level, the SNP density in non-coding regions between two different chromosomes from a cynomolgus monkey (*Macaca fascicularis*) declined to approximately 90% near annotated exons, and that this of reduction is slightly stronger on X chromosomes than on autosomes. Although statistical power to detect such patterns is plausibly weaker than that of a multi-individual analysis, it is interesting to see whether the observed patterns in *Drosophila* and mammals are universal among different types of diploid organisms. Needless to say, an analysis with a small number of samples should be considered a starting point, as it would not capture all important aspects of natural populations, such as complex demography and population structure. Nevertheless, such an analysis could provide novel insight into the evolution of genomes in a wider range of taxa before we enter the true population genomics era.

## ACKNOWLEDGMENTS

This study was supported by KAKENHI Grant Numbers 22687021 and 23113008.

## REFERENCES

Akashi, H., Osada, N., and Ohta, T. (2012). Weak selection and protein evolution. *Genetics* 192, 15–31. doi: 10.1534/genetics.112.140178

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* 162, 2025–2035.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59. doi: 10.1038/nature07517

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backstrom, N., Kawakami, T., et al. (2012). The genomic landscape of species divergence in Ficedula flycatchers. *Nature* 491, 756–760. doi: 10.1038/nature11584

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., and Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genet.* 9:e1003905. doi: 10.1371/journal.pgen.1003905

Felsenstein, J. (2006). Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Mol. Biol. Evol.* 23, 691–700. doi: 10.1093/molbev/msj079

Gayral, P., Melo-Ferreira, J., Glémin, S., Bierne, N., Carneiro, M., Nabholz, B., et al. (2013). Reference-free population genomics from next-generation transcriptome data and the vertebrate–invertebrate gap. *PLoS Genet.* 9:e1003457. doi: 10.1371/journal.pgen.1003457

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5:e1000695. doi: 10.1371/journal.pgen.1000695

Halligan, D. L., Kousathanas, A., Ness, R. W., Harr, B., Eöry, L., Keane, T. M., et al. (2013). Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. *PLoS Genet.* 9:e1003995. doi: 10.1371/journal.pgen.1003995

Harris, K., and Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet.* 9:e1003521. doi: 10.1371/journal.pgen.1003521

Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G., et al. (2011). Classic selective sweeps were rare in recent human evolution. *Science* 331, 920–924. doi: 10.1126/science.1198878

Higashino, A., Sakate, R., Kameoka, Y., Takahashi, I., Hirata, M., Tanuma, R., et al. (2012). Whole-genome sequencing and analysis of the Malaysian cynomolgus macaque (*Macaca fascicularis*) genome. *Genome Biol.* 13:R58. doi: 10.1186/gb-2012-13-7-r58

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., et al. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484, 55–61. doi: 10.1038/nature10944

Keinan, A., and Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336, 740–743. doi: 10.1126/science.1217283

Lanfear, R., Kokko, H., and Eyre-Walker, A. (2014). Population size and the rate of evolution. *Trends Ecol. Evol.* 29, 33–41. doi: 10.1016/j.tree.2013.09.009

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475, 493–496. doi: 10.1038/nature10231

Loire, E., Chiari, Y., Bernard, A., Cahais, V., Romiguier, J., Nabholz, B., et al. (2013). Population genomics of the endangered giant Galapagos tortoise. *Genome Biol.* 14:R136. doi: 10.1186/gb-2013-14-12-r136

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., et al. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res.* 23, 1817–1828. doi: 10.1101/gr.159426.113

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genet.* 5:e1000686. doi: 10.1371/journal.pgen.1000686

Miller, W., Schuster, S. C., Welch, A. J., Ratan, A., Bedoya-Reina, O. C., Zhao, F., et al. (2012). Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proc. Natl. Acad. Sci. U.S.A.* 109, E2382–E2390. doi: 10.1073/pnas.1210506109

Nakagome, S., Fukumizu, K., and Mano, S. (2013). Kernel approximate Bayesian computation in population genetic inferences. *Stat. Appl. Genet. Mol. Biol.* 12, 667–678. doi: 10.1515/sagmb-2012-0050

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461, 272–276. doi: 10.1038/nature08250

Osada, N., Nakagome, S., Mano, S., Kameoka, Y., Takahashi, I., and Terao, K. (2013). Finding the factors of reduced genetic diversity on X

chromosomes of Macaca fascicularis: male-driven evolution, demography, and natural selection. *Genetics* 195, 1027–1035. doi: 10.1534/genetics.113.156703

Pluzhnikov, A., and Donnelly, P. (1996). Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* 144, 1247–1262.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., et al. (2013). Great ape genetic diversity and population history. *Nature* 499, 471–475. doi: 10.1038/nature12228

Sattath, S., Elyashiv, E., Kolodny, O., Rinott, Y., and Sella, G. (2011). Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*. *PLoS Genet.* 7:e1001302. doi: 10.1371/journal.pgen.1001302

Sheehan, S., Harris, K., and Song, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics* 194, 647–662. doi: 10.1534/genetics.112.149096

Yan, G., Zhang, G., Fang, X., Zhang, Y., Li, C., Ling, F., et al. (2011). Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nat. Biotech.* 29, 1019–1023. doi: 10.1038/nbt.1992

Zhao, S., Zheng, P., Dong, S., Zhan, X., Wu, Q., Guo, X., et al. (2013). Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nat. Genet.* 45, 67–71. doi: 10.1038/ng.2494

*Received: 31 January 2014; accepted: 16 March 2014; published online: 02 April 2014.*

*Citation: Osada N (2014) Extracting population genetics information from a diploid genome sequence. Front. Ecol. Evol.* ***2****:7. doi: 10.3389/fevo.2014.00007*

*This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution.*