frontiers in Genetics

OPINION
published: 11 October 2017
doi: 10.3389/fgene.2017.00148

The Necessity of Diploid Genome Sequencing to Unravel the Genetic Component of Complex Phenotypes

*Fernando Aleman**

Scripps Research Institute, La Jolla, CA, United States

Keywords: chromosome phasing, diploid alignment, diploid genomes, GWAS (genome-wide association study), structural variants, genetic variants, SNP association study, Diploid Manhattan Plot

Genome-Wide Association Studies (GWAS) correlate the genotype with the phenotype, identifying the genetic variants that are linked to any particular trait or disease. In 2005, a ground-breaking successful GWAS in humans associated the complement factor H gene with age-related macular degeneration (Klein et al., 2005). Since then, many successful GWAS using genotyping arrays have been published (Manolio, 2017), but due to the lowering cost of DNA sequencing, whole genome sequencing GWAS are becoming more frequent. However, the usefulness of classical GWAS has recently been questioned in a *Cell* publication (Boyle et al., 2017). The authors explain that genetic variants causing a disease should be part of a pathway connected with the etiology or prognosis of the disease, and moreover, they describe the benefits of linking GWAS with cell specific gene expression data. Still, many GWAS fail to correlate a specific genetic variant with a gene or a pathway leading to disease. This is partially due to the loose definition of how to establish an association between each genetic variant (frequently in non-coding regions) and the causal gene. In addition, the size of the effect of each genetic variant in polygenic traits and low penetrance genetic diseases is difficult to accurately establish due to confounding factors such as population stratification.

One of the main weaknesses of whole genome sequencing GWAS is the fact that for every diploid (or polyploid) organism we only obtain the "haploid genome." Due to the prevalent short-reads technology, we merge both gene copies of every chromosome into one, losing physical connections and proximity between genetic variants in homologous chromosomes. Integrating both allele sequences as if they were one hampers the elucidation of haplotype specific structural variants (SVs). Indeed, SVs are more frequent in one haplotype vs. homozygous SVs (Sudmant et al., 2015; Hehir-Kwa et al., 2016). In addition, linkage disequilibrium and genetic linkage are difficult to accurately elucidate when the homologous chromosomes are merged, which decrease the power of many gene- and pathway-based association studies (Mooney et al., 2014).

To solve this issue, there have been several studies reporting the separation of alleles into chromosomes (phased chromosomes) of several genomes, but so far, only four studies have reported *de novo* human diploid genomes (Levy et al., 2007; Cao et al., 2015; Seo et al., 2016; Weisenfeld et al., 2017). In the paper *Direct determination of diploid genome sequences,* Weisenfeld et al. recently demonstrated that an accurate and cost effective method can be routinely used with the most popular Illumina sequencing technology. However, this method has only been tested on human genomes and some difficulties may arise for other species. In fact, it is worth to mention that GWAS have been widely used in plants (Korte and Farlow, 2013; Huang and Han, 2014) where the polyploidy of some species can introduce even more noise in the final haploid sequence. Thus, the benefits of using diploid (or polyploid) genomes materialize in two ways. First, better disease/trait variant calling (since we would have the real genome without noise coming from the "mix and match" of homologous chromosomes). Still, a high number of diploid genomes would increase the statistical power for the identification of new variants causing disease or a trait. The second advantage is the potential to detect protective genetic variants which, as mentioned below, are now potentially actionable with CRISPR/Cas9 in combination with correcting the faulty variant. Other general benefits can come from closing the "missing

Edited by:
Youri I. Pavlov, University of Nebraska Medical Center, United States

Reviewed by:
Steven Andrew Roberts, Washington State University, United States

***Correspondence:**
Fernando Aleman
faleman@scripps.edu

Specialty section:
This article was submitted to Genomic Assay Technology, a section of the journal Frontiers in Genetics

Received: *18 August 2017*
Accepted: *27 September 2017*
Published: *11 October 2017*

Citation:
Aleman F (2017) The Necessity of Diploid Genome Sequencing to Unravel the Genetic Component of Complex Phenotypes. Front. Genet. 8:148. doi: 10.3389/fgene.2017.00148

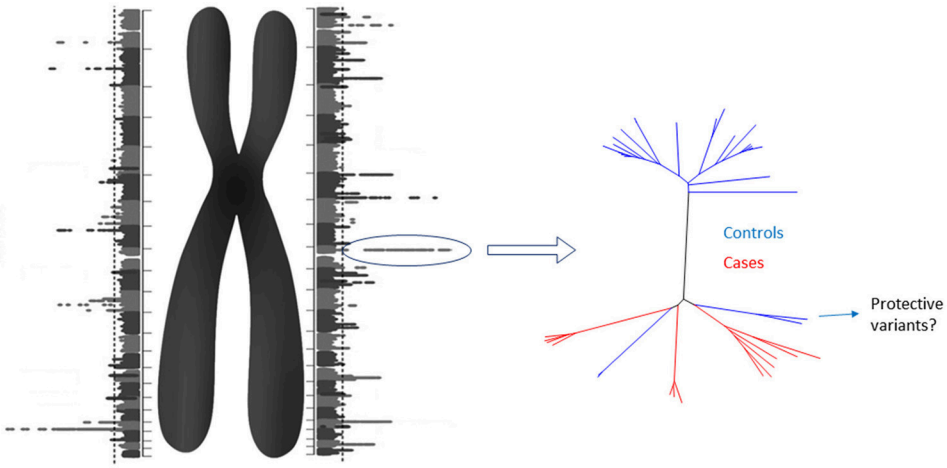

FIGURE 1 | The Diploid Manhattan Plot: To achieve personalized medicine we need to isolate individual diploid genomes and compare them with a reference genome. Each homologous chromosome can be compared with the regular "haploid reference genome." Thus, two GWAS that include the individual physical context of each genetic variant (including structural variants) will be produced. Since those DNA stretches would have a manageable size, the highest individual hits could be selected for further characterization comparing them with cases and control diploid genomes. For low penetrance diseases, protective SNPs in cis could be found. Diploid genomes have the potential to close the gap of the missing heritability problem.

heritability gap" problem, and a better quantification of penetrance. So far, we do not fully understand the reasons for incomplete penetrance of most genetic diseases. Indeed, the analysis of diploid sequences has the potential to modify how we measure penetrance, since we would be able to include in the analysis not only the genetic variant that directly cause disease, but also any other protective variant that might co-exist in *cis*. Even concepts such as *conditional full penetrance* may arise (conditional to the sequence in a different interacting locus).

Diploid genomes are not only required for understanding allele-specific expression, but also to understand the real output of each allele. Two frameshifts in a gene will have completely different outcomes if they are both in the same allele or if each frameshift occurs in a separate allele. In addition, penetrance levels should be determined based not only on one genetic variant, but also on the genetic variants occurring in close proximity that are in linkage disequilibrium. This is particularly important for genetic diseases with incomplete penetrance such as celiac disease and allele-specific diseases such as Huntington's disease. Furthermore, with a high quality diploid sequence, CRISPR/Cas technology provides a potential actionability in two different ways. Allele-specific diseases can be precisely targeted, without affecting the healthy allele (Paquet et al., 2016), and diploid genomes may enable the discovery of allele-specific protective genetic variants, which could be targeted with CRISPR to improve health. Examples of how phasing loci improve the identification of disease causing variants are still limited but increasing (Safrany et al., 2013; Sharp et al., 2016; Subramanian et al., 2017). Plant breeding programs will also benefit from phased chromosomes since many important crops are polyploid and the genetic makers for heterosis may be revealed with polyploid sequences (Chen, 2013; Minio et al., 2017).

Although the cost of obtaining a diploid genome could cease to be a problem, other challenges lie ahead. Finding a proper reference for comparison will be daunting. However, the analysis may be split in two steps. First, each homologous chromosome of a diploid genome can be compared to a reference "haploid genome" obtaining a "Diploid Manhattan Plot" (**Figure 1**). The benefits of choosing a stratified population as reference need to be elucidated yet. Then, the selected loci in individual chromosomes with higher statistical significance should be explore in detail and compared with control diploid loci. When causes and controls are used, this method would work to reveal not only the causal genetic variants, but also potential protective variants from low penetrance diseases. Finally, comprehensive graphical models will be needed along with the human resources required to analyse, interpret and provide genetic counseling.

Overall, to enable the rising field of personalized medicine, we need to unwind the whole genomic information in our diploid cells and elucidate what contributes to health and disease. The field of personalized medicine has to lead the change from "haploid" genomes to the real diploid ones since it is not only the wealthiest genomic area but also the one with a potential higher impact in the society. Therefore, it is paramount that we start a new genomic generation with a diploid revolution using the resources that have just been developed.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work with the fruitful discussions mentioned in the acknowledgments.

ACKNOWLEDGMENTS

I would like to thank Drs. Amalio Telenti and Mehis Pold for insightful comments on the manuscript, which contributed to the improvement of this piece.

REFERENCES

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169, 1177–1186. doi: 10.1016/j.cell.2017.05.038

Cao, H., Wu, H., Luo, R., Huang, S., Sun, Y., Tong, X., et al. (2015). De novo assembly of a haplotype-resolved human genome. *Nat. Biotechnol.* 33, 617–622. doi: 10.1038/nbt.3200

Chen, Z. J. (2013). Genomic and epigenetic insights into the molecular bases of heterosis. *Nat. Rev. Genet.* 14, 471–482. doi: 10.1038/nrg3503

Hehir-Kwa, J. Y., Marschall, T., Kloosterman, W. P., Francioli, L. C., Baaijens, J. A., Dijkstra, L. J., et al. (2016). A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.* 7:12989. doi: 10.1038/ncomms12989

Huang, X., and Han, B. (2014). Natural variations and genome-wide association studies in crop plants. *Annu. Rev. Plant Biol.* 65, 531–551. doi: 10.1146/annurev-arplant-050213-035715

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* 308, 385–389. doi: 10.1126/science.1109557

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9, 29–29. doi: 10.1186/1746-4811-9-29

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., et al. (2007). The diploid genome sequence of an individual human. *PLoS Biol.* 5:e254. doi: 10.1371/journal.pbio.0050254

Manolio, T. A. (2017). In Retrospect: A decade of shared genomic associations. *Nature* 546, 360–361. doi: 10.1038/546360a

Minio, A., Lin, J., Gaut, B. S., and Cantu, D. (2017). How single molecule real-time sequencing and haplotype phasing have enabled reference-grade diploid genome assembly of wine grapes. *Front. Plant Sci.* 8:826. doi: 10.3389/fpls.2017.00826

Mooney, M. A., Nigg, J. T., McWeeney, S. K., and Wilmot, B. (2014). Functional and genomic context in pathway analysis of GWAS data. *Trends Genet.* 30, 390–400. doi: 10.1016/j.tig.2014.07.004

Paquet, D., Kwart, D., Chen, A., Sproul, A., Jacob, S., Teo, S., et al. (2016). Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* 533, 125–129. doi: 10.1038/nature17664

Safrany, E., Szabo, M., Szell, M., Kemeny, L., Sumegi, K., Melegh, B. I., et al. (2013). Difference of interleukin-23 receptor gene haplotype variants in ulcerative colitis compared to Crohn's disease and psoriasis. *Inflam. Res.* 62, 195–200. doi: 10.1007/s00011-012-0566-z

Seo, J. S., Rhie, A., Kim, J., Lee, S., Sohn, M. H., Kim, C. U., et al. (2016). *De novo* assembly and phasing of a Korean human genome. *Nature* 538, 243–247. doi: 10.1038/nature20098

Sharp, K., Kretzschmar, W., Delaneau, O., and Marchini, J. (2016). Phasing for medical sequencing using rare variants and large haplotype reference panels. *Bioinformatics* 32, 1974–1980. doi: 10.1093/bioinformatics/btw065

Subramanian, L., Khan, A. A., Allu, P. K. R., Kiranmayi, M., Sahu, B. S., Sharma, S., et al. (2017). A haplotype variant of the human chromogranin A gene (CHGA) promoter increases CHGA expression and the risk for cardiometabolic disorders. *J. Biol. Chem.* 292, 13970–13985. doi: 10.1074/jbc.M117.778134

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75–81. doi: 10.1038/nature15394

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., and Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Res.* 27, (5):757–767. doi: 10.1101/gr.214874.116

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.