Applications for Haplotype Assembly Problems

Minzhen Yi

Department of Automation, Tsinghua University

Mentor: Sorin Istrail

Department of Computer Science, Brown University

Abstract

Haplotype is a set of single-nucleotide polymorphisms which is also known as DNA-sequence variations at specific nucleotide sites, or polymorphic sites on a single chromatid that are associated statistically. It is thought that identifying these statistical associations and alleles of a specific haplotype sequence can facilitate identifying all other such polymorphic sites that are nearby on the chromosome. Consequently, improving algorithmic strategies for haplotype assembly which can determine haplotypes of individuals computationally is meaningful. Nowadays, one of the most efficient algorithm that has been proved is HapCompass[1]. With the help of this tool, specific sequence haplotypes of diploid or multiplied species can be obtained and the exact number of haplotypes in an aneuploid cell or organism can be computed. Based on these detailed information of haplotypes, a better interpretation of SNPs can be gained. As variations in the DNA sequences can affect how a species contract diseases and respond to pathogens, chemicals, drugs, vaccines and other agents, this improvement in the comprehension of haplotypes will contribute to many significative applications like corp and livestock breeding, genome-wide association studies, disease diagnosis and disease cures. This essay starts with the introduction of the description and computational research methods of haplotypes, and then will pay attention on three

significant applications of haplotypes under the help of HapCompass. At last, some original ideas about the haplotype application considering the progress in the high-throughput sequencing technology and some potential research areas combined with sociology and public health are addressed for discussion and reflection.

Haplotype

A haplotype can describe a pair of genes inherited together from one parent on one chromosome, or it can describe all of the genes on a chromosome that were inherited together from a single parent. This group of genes is inherited together because of genetic linkage, or because they are close to each other on the same chromosome.

Haplotype has a few definitions, but in this paper it exclusively indicates a collective of SNPs. SNP as the abbreviation of Single Nucleotide Polymorphism is a DNA sequence variation occurring commonly within a population in which a single nucleotide in the genome differs among members of a biological species or paired chromosomes. These genetic variations between individuals are something exploited in DNA fingerprinting, which is used in forensic science. As a representative example, these genetic variations underlie differences in the susceptibility in diseases. The severity of illness and the mechanism how each person's body responds to treatments are also manifestations of genetic variations. By examining haplotypes, scientists can identify patterns of genetic variation that are associated with health and disease states. For instance, if a haplotype is associated with a certain disease, then scientists can examine stretches of DNA near the SNP cluster to try to identify the gene or genes responsible for causing the disease.

Haplotype study has a profound influence on research in disease cure, species breeding and many other practical applications, which involves medical science, pharmacy, and biology. Whereas, how to determine accurate haplotypes for long chromosomes particularly on restricted optimizations has ever been an intractable problem due to modern high-throughput sequencing technologies. A novel algorithm — HapCompass was presented in 2012 for haplotype assembly of densely sequenced human genome data, which was furthermore extended to polyploid genomes and aneuploid genomes in the subsequent several years[2]. With this technical improvement, more work can be done by means of haplotype assembly in the research of cancer, congenital disease prevention and corp and livestock breeding. This paper will focus on these three most influential research fields of haplotype study application, and analyze how HapCompass play an essential role in them.

Cancer Cure

Until now, there have been many previous studies which confirm that the risk of cancer is highly relevant to some variations at nucleotide sites on DNA sequence. A recent research which reveals the role of haplotypes and the relationship between haplotypes and susceptibility to lung cancer reaches the conclusion that CA/CA homozygote showed an increased risk of lung cancer with borderline significance in discovery and statistical significance in replication, compared with combined genotypes (CG/CG + CG/TG) even among never-smokers[3]. Through pair-wise haplotype analysis, they identified that CHRNA5 rs588765-rs16969968 was the most significant haplotype associated with lung cancer risk. This is a consequential discovery for lung cancer detection, considering lung cancer is the number one cause of cancer deaths in both men and women in the U.S. and worldwide while the general prognosis of lung cancer is poor because doctors tend not to find the disease until it is at an advanced stage. Five-year survival for early stage lung cancer is 40%-50%, while for advanced and inoperable stage is only 1%-5%. If this genomic detection works, the detection stage of lung cancer can be much earlier by analyzing haplotypes of an uncertain patient and calculating the statistical probability of having lung cancer by examining the significance of replication borderline, which can remarkably decrease the morality of lung cancer.

Although the gene therapy may still have a long way to go, it doesn't mean the research of haplotypes is meaningless at the present stage. Actually a latest interesting research reveals that haplotype analysis can be used to evaluate the traditional therapy by comparing proportion of survival with two different alleles[4]. With this in mind, a patient can be treated in an theoretically appropriate therapy, which will probably prolong his life according to his genotype. This assumption is verified by this report that the most frequent TP53 haplotype which is recognized to be previously associated with an increased CRC(colorectal cancer) and pancreatic cancer risks is also related to an increased BC(breast cancer) susceptibility. From this, a preliminary hypotheses can be supposed that some certain haplotype is highly relevant to the susceptibility of cancer, and the detection and modification of this maybe the key point of cancer cure.

Apart from gene cure, here is another area in identifying T cell epitopes that the research of haplotypes may contributes to cancer cure[5]. It is extensively known that the main mechanism in immune system is histocompatibility antigen and transplantation antigen. Although there are up to more than 20 kinds of antigen system relevant with reject reaction, MHC(major histocompatibility complex) is mainly responsible for binding to peptide fragments derived from pathogens and displaying them on the cell surface for recognition by the appropriate T-cells. The set of alleles that is present in each chromosome is called the MHC haplotype. The MHC genes are so highly polymorphic that no two individuals have exactly the same set of MHC molecules in a mixed population. As a result, the presence of many different alleles in MHC genes ensures there will always be an individual with a specific MHC molecule able to load the correct peptide to recognize a specific microbe. The significance of MHC gene alleles makes the research about MHC haplotypes extraordinarily important. By using computational methods such as artificial neural network to predict peptide binding to MHC, it can be beneficial to describing and predicting the specificity of antigen processing. With a more complete knowledge of this, we are more likely to cure cancer by targeted therapy.

Congenital desease

There are many congenital diseases that have been proved to be caused by chromosomal problems, among which Down Syndrome is typical and hence analyzed as an example in this paper. Down Syndrome which is also known as trisomy 21 is a genetic disorder caused by the presence of all or part of a third copy of chromosome 21. It is typically associated with physical growth delays, characteristic facial features, and mild to moderate intellectual disability. Down Syndrome is one of the most common chromosome abnormalities in humans, occurring in about per 1000 babies born each year[6]. In 2013 it resulted in 36,000 deaths down from 43,000 deaths in 1990. The perniciousness of Down Syndrome is not only physical and intellectual disabilities, poor immune function, but also a increase risk of other health problems, which includes congenital heart disease, leukemia, thyroid disorders and mental disorders. Owning to these severe harm of Down Syndrome, it is necessary and momentous to detect the disease before birth.

At present, diagnostic tests utilized to detect Down Syndrome are mainly ultrasound imaging, blood test and screening, which contain combined test, quad screen, integrated screen and cell-free fetal DNA. For ultrasound imaging, findings indicate that fetus have increased risks when seen at 14 to 24 weeks of gestation include a small or no nasal bone, large ventricles, nuchal fold thickness, and an abnormal right subclavian artery compared to healthy ones[7]. And for blood tests, the detection correctness rate is approximately 60% - 70% in the second trimester during gestational period. The deficiencies in the above methods evidently lie in late detection time and high probability for false detection, which can be entirely overcome by haplotype assembly test. Screening test results are more confident, while amniocentesis and chorionic villus increase the risk of miscarriage[8], also the risk of limb problems is increased in the offspring owing to this procedure[9].

With the algorithmic strategy — HapCompass that makes haplotype assembly easily possible, and thanks to the feasibility that genome-wide aneuploidy detection by maternal plasma DNA sequencing[10], a new detection mean that determines the number of fetus' 21 chromosome during pregnancy can be effective. After extracting the pregnant maternal plasma, fetus haplotype phases can be inferred, and it will determines exactly whether the fetus has Down Syndrome in an early stage.

Livestock Breeding

Advances in genome technology have presented us an opportunity to regard the historical and genetic processes crucial to rapid phenotypic evolution under domestication in a new angle. A previous research which conducted an extensive genome-wide survey of more than 48,000 SNPs in dogs and their wild progenitors — the grey wolf found out how genetic diversity differs in these two species along with time, and discovered correspondence between genetic and phenotypic or functional breed groupings[11]. It indicated that it is feasible to understand how species evolve and differ from their progenitors genetically through analyzing SNPs of these species. Also, the correspondence between genetic and phenotypic breed groupings will help us have a better know of how epigenetic features are related to genome, and make it possible to breed livestock superior to the present species.

The significance of haplotype analysis in livestock breeding does not only lie in breeding optimization, but also benefits conserving livestock biodiversity resources in the future[12]. With the use of molecular genetics, we can better comprehend evolutionary and demographic history of human, animals and plants we have domesticated, especially in how these events take place and identification of the wild ancestors of modern livestocks and the nature of livestock expansion in past millennia. It reveals human history and how we have shaped such extraordinary biological diversity in a relatively short period of time, which may potentially be important for the management and conservation of today's animal and genetic resources.

Further Supposition

Considering the rapid progress in high-throughput sequencing technology, chromosome conformation capture (3C) technology has developed into many 3C-based methods, such as 4C(chromosome conformation capture-on-chip), 5C(chromosome conformation capture carbon copy), HiC, and ChIA-PET[13]. ChIP-seq is a method used to analyze protein interaction with DNA, which combines chromatin immunoprecipitation with massively parallel DNA sequencing to identify the binding sites of DNA associated proteins. It is used primarily to determine how transcription factors and other chromatin-associated proteins influence phenotype-affecting mechanisms.

Lots of previous work have proved that different phenotypes among different people are relevant to their different haplotypes, which is the inevitable result reflecting that phenotypes are interrelated with chromatin structure. As a result, it is logical to deduct that more work can be done in studying the relationship between phenotype and genotype by combining the ChIP-seq data and haplotype data. By recognizing the pattern between ChIP-seq data and haplotype assembly data, or by drawing a map from haplotype data to ChIP-seq data using computational methods from machine learning, we can have a better view of how people differ from each other not only from protein function respect or just from chromatin structure respect but also from combining them together.

Moreover, haplotype assembly can also be used to examine other social issues statistically, as it can provide all the gene information differences among people of different places and time. Here is an interesting paper about exploring population size changes using SNP frequency spectra published on Nature Genetics in 2015[14]. This gives us another hint about the application for SNPs. SNPs, which reflect all the gene variation among different people are just like the coding of human characters and behaviors, and are impliedly responsible for many social phenomena. Consequently, compared to previous study problems in sociology, anthropology and epidemiology with traditional methods, more understanding of this can be gained from genetics.

Also, here is another research direction about SNPs I consider as attractive and promising. As a matter of fact, among overall gene alleles in all the people, there are only 5% of them different among different individuals which are SNPs. There is a mysterious problem which appeals to me a lot is that how these different alleles decide phenotypes. Although many studies have been conducted to understand how chromatin mediate protein function, how SNPs in different areas regulate protein expression are still unknown. We propose that if SNP analysis are divided into different categories as synonymous SNPs in coding area, non-synonymous SNPs in coding area and SNPs in non-coding area, more facts about mechanism of gene regulation can be discovered. Furthermore, this work should be done in different generations or different races of human in order to get more comprehensive and detailed information of mankind.

Reference

[1] Aguiar D., Istrail S. HapCompass: A fast cycle basis algorithm for accurate haplotype assembly of sequence data. *Journal of Computational Biology*, 2012 June, **19**(6): 577-90.

[2] Aguiar D., Istrail S. Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics*, 2013, vol. 29, no. 13, pp. i352-i360.

[3] Ji X., Gui J., Han Y., Brennan P., Li Y., McKay J., Caporaso N., Bertazzi PA., Landi MT., Amos CI. The Role of Haplotype in 15q25.1 Locus in Lung Cancer Risk: Results of Scanning chromosome 15. *Carcinogenesis.* 2015, doi: 10.1093/carcin/bgv118.

[4] Vymetalkova V., Soucek P., Kunicka T., Jiraskova K, etc. Genotype and Haplotype Analyses of *TP53* Gene in Breast Cancer Patients: Association with Risk and Clinical Outcomes. *PLOS*, July 30, 2015, doi: 10.1371/journal.pone.0134463.

[5] Lauemøller SL., Kesmir C., Corbet SL., etc. Identifying cytotoxic T cell epitopes from genomic and proteomic information: “The human MHC project”. IMMU993481 P348 28-06-01 15:34:30.

[6] Malt EA., Dahl RC., Haugsand TM., Ulvestad IH., Emilsen NM., Hansen B., Cardenas YE., Skold RO., Thorsen AT., Davidsen EM. Health and disease in adults with Down syndrome. *Tidsskrift for den Norske laegeforening : tidsskrift for praktisk medicin.* 2013 Feb *5*, ny raekke **133** (3): 290–4.

[7] Agathokleous M., Chaveeva P., Poon LC., Kosinski P., Nicolaides KH. Meta-analysis of second-trimester markers for trisomy 21. *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology.* 2013 Mar, **41** (3): 247–61.

[8] Tabor A., Alfirevic Z. Update on procedure-related risks for prenatal diagnosis techniques. *Fetal diagnosis and therapy.* 2010, **27** (1): 1–7.

[9] Choi H., Van Riper M., Thoyre S (Mar–Apr 2012). "Decision making following a prenatal diagnosis of Down syndrome: an integrative review.". *Journal of midwifery & women's health* **57** (2): 156–64.

[10] Bianchi Diana W., Platt Lawrence D., Goldberg James D., Abuhamad Alfred Z., Sehnert Amy J., Rava Richard P. Genome-Wide Fetal Aneuploidy Detection by Maternal Plasma DNA Sequencing. *Obstet Gynecol.* 2012 Oct, 120(4):957.

[11] Bridgett M. vonHoldt, John P. Pollinger, Kirk E. Lohmueller, Eunjung Han, Heidi G. Parker, Pascale Quignon, Jeremiah D. Degenhardt, Adam R. Boyko, Dent A. Earl, Adam Auton, Andy Reynolds, Kasia Bryc, Abra Brisbin, James C. Knowles, Dana S. Mosher, Tyrone C. Spady, Abdel Elkahloun, Eli Geffen, Malgorzata Pilot, Wlodzimierz Jedrzejewski, Claudia Greco, Ettore Randi, Danika Bannasch, Alan Wilton, Jeremy Shearman, Marco Musiani, Michelle Cargill, Paul G. Jones, Zuwei Qian, Wei Huang, Zhao-Li Ding, Ya-ping Zhang, Carlos D. Bustamante, Elaine A. Ostrander, John Novembre & Robert K. Wayne. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* **464**, 898-902 (8 April 2010) | doi: 10.1038/nature08837.

[12] Bruford MW., Bradley DG., Luikart G. DNA markers reveal the complexity of livestock domestication. *Nature Reviews Genetics* **4**, 900-910 (November 2003) | doi:10.1038/nrg1203.

[13] Wit ED., Laat WD. A decade of 3C technologies: insights into nuclear organization. *Genes & Dev.* 2012. 26: 11-24 doi: 10.1101/gad.179804.111.

[14] Liu X., Fu Y. Exploring population size changes using SNP frequency spectra. *Nature Genetics* **47**, 555–559 (2015) doi:10.1038/ng.3254.