High resolution linkage disequilibrium structure of functional elements within exonic and intronic regions

Extended Abstract

Alejandra Vergara-Lope*
Next Generation Computational Modelling
University of Southampton, UK
nvlg1e15@soton.ac.uk

Andrew Collins†
Genetic Epidemiology and Genomic Informatics
University of Southampton, UK
A.R.Collins@soton.ac.uk

ABSTRACT

Patterns of linkage disequilibrium (LD) at fine-scale may improve filtering of Next Generation Sequencing variant lists to determinate true disease. LD patterns reflect the combined impacts of recombination, natural selection, genetic drift and mutation to understanding the genome function. Within this novel research, the potential of LD maps of the autosomal genome at a very fine scale down to the exonic and intronic level is being assessed for providing novel insights into the impact of recombination and selection on genome structure and function.

CCS CONCEPTS

• **Computing methodologies** → Linkage Disequilibrium; • **Applied Computing** → High resolution of LD;

KEYWORDS

High-resolution linkage disequilibrium, Next Generation Sequencing, Genes Filtering

ACM Reference Format:
Alejandra Vergara-Lope and Andrew Collins. 2018. High resolution linkage disequilibrium structure of functional elements within exonic and intronic regions: Extended Abstract. In *Proceedings of ACM Celebration of Women in Computing womENcourage 2018*. ACM, Belgrade Serbia, 2 pages.

1 INTRODUCTION

Next Generation Sequencing (NGS) of deoxyribonucleic acid (DNA) from patients with diseases is revolutionising medical research stimulating rapid transition towards personalised treatment [1]. Understanding disease gene characteristics is a pressing need to help interpret voluminous NGS data. The integration of genomic properties such as recombination, natural selection, genetic drift and mutation can improve filtering of NGS variant lists to determinate true disease candidates [2]. The pattern of LD at high resolution level across the genome represents the outcome of these processes and enable much more comprehensive analysis of LD structure and genome function. The aim of this study is to to improve the variant filtering of NGS by analysing LD patterns at fine-scale for ranking plausible disease-causing candidates.

*Genetic Epidemiology and Bioinformatics Research Group
†Head of the Genetic Epidemiology and Bioinformatics Research Group

ACM Celebration of Women in Computing womENcourage 2018, October 3-5, 2018, Belgrade, Serbia
2018.

2 METHODS

Single nucleotide polymorphism (SNP) data were taken from whole genome sequences from 454 individuals from the Wellderly study. Filtering of the SNP data excluded SNPs showing deviation from Hardy-Weinberg equilibrium (p-value of <0.001) and rare SNPs with a minor allele frequency of <0.01 [3]. LD maps were generated in linkage disequilibrium units (LDUs) for the autosomal chromosomes 1-22 based on the Malecot-Morton model [4]. The extent of linkage disequilibrium as kilobases/LDU was determined exons and introns within genes.

3 RESULTS AND DISCUSSION

The results show that LD intensity (extent of LD) is lower for small chromosomes compared to larger chromosomes; these results can be attributed to elevated recombination rates of smaller chromosomes [4]. The strong relationship between the extent of LD and chromosome recombination rate in centimorgans has significant and dominant effect on these patterns (Figure 1) exceeding effects due to selection and mutation [5]. The overall difference in the extent of LD between exons and introns is small with more intense LD in exons. Preliminary studies have concluded that the strength on LD in genes containing disease variation is intermediate between 'housekeeping' genes (very strong LD, intolerant to recombination/mutation) and genes involved in (for example) sensory perception where weak LD reflecting a high haplotype diversity may be adaptive [6]. Understanding patterns of LD and the interaction between recombination and selection throughout the whole genome is fundamental to identifying disease candidate genes in the context of NGS for more powerful filtering strategies.

ACKNOWLEDGMENTS

The authors acknowledge support from the University of Southampton EPSRC Centre for Doctoral Training in Next Generation Computational Modelling (CDT NGCM).

REFERENCES

[1] Eleanor G. Seaby, Reuben J. Pengelly, and Sarah Ennis. Exome sequencing explained: a practical guide to its clinical application. *Briefings in Functional Genomics*, 15(5):374–384, 2016.
[2] Reuben J. Pengelly, Alejandra Vergara-Lope, Dareen Alyousfi, M. Reza Jabalameli, and Andrew Collins. Understanding the disease genome: gene essentiality and the interplay of selection, recombination and mutation. *Briefings in Bioinformatics*, page bbx110, 2017.
[3] Reuben J. Pengelly, William Tapper, Jane Gibson, Marcin Knut, Rick Tearle, Andrew Collins, and Sarah Ennis. Whole genome sequences are required to fully resolve the linkage disequilibrium structure of human populations. *BMC Genomics*, 16(1):666, 2015.

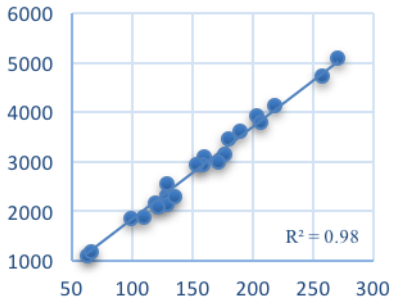

Figure 1: LDU lengths of chromosomes vs genetic length in cM.

[4] Swetlana Berger, Martin Schlather, Gustavo de los Campos, Steffen Weigend, Rudolf Preisinger, Malena Erbe, and Henner Simianer. A scale-corrected comparison of linkage disequilibrium levels between genic and non-genic regions. *PLOS ONE*, 10:1–19, 10 2015.

[5] William. Tapper, Andrew. Collins, Jane. Gibson, Nikolas. Maniatis, Sarah. Ennis, and Newton E. Morton. A map of the human genome in linkage disequilibrium units. 102(33):11835–11839, 2005.

[6] Andrew Collins. The genomic and functional characteristics of disease genes. *Briefings in Bioinformatics*, 16(1):16–23, 2015.