BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 3 2015, pages 438–439
doi:10.1093/bioinformatics/btu650

Genetics and population analysis

Advance Access publication October 4, 2014

VCF2Networks: applying genotype networks to single-nucleotide variants data

Giovanni Marco Dall'Olio[1,*], Ali R. Vahdati[2], Jaume Bertranpetit[1], Andreas Wagner[2,3,4] and Hafid Laayouni[1,5]

[1]Department of Experimental and Health Sciences, Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), 08003 Barcelona, Catalonia, Spain, [2]Institute of Evolutionary Biology and Environmental Studies/Swiss Institute of Bioinformatics, University of Zurich, 8057 Zurich, Switzerland, [3]SIB, CIG Quartier Sorge, bâtiment Génopode 1015 Lausanne, Switzerland, [4]The Santa Fe Institute, 1399 Hyde Parke Road, 87501 Santa Fe, New Mexico, USA and [5]Departament de Genetica i de Microbiologia, Grup de Biologia Evolutiva (GBE), Universitat Autonoma de Barcelona, 08913 Bellaterra, Barcelona

Associate Editor: Gunnar Ratsch

ABSTRACT

Summary: A wealth of large-scale genome sequencing projects opens the doors to new approaches to study the relationship between genotype and phenotype. One such opportunity is the possibility to apply genotype networks analysis to population genetics data. Genotype networks are a representation of the set of genotypes associated with a single phenotype, and they allow one to estimate properties such as the robustness of the phenotype to mutations, and the ability of its associated genotypes to evolve new adaptations. So far, though, genotype networks analysis has rarely been applied to population genetics data. To help fill this gap, here we present VCF2Networks, a tool to determine and study genotype network structure from single-nucleotide variant data.

Availability and implementation: VCF2Networks is available at https://bitbucket.org/dalloliogm/vcf2networks.

Contact: giovanni.dallolio@kcl.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 19, 2014; revised on June 26, 2014; accepted on September 29, 2014

1 INTRODUCTION

Genotype networks can be used to describe the evolutionary properties of a set of genotypes associated with a qualitative phenotype. They are derived from genotype–phenotype maps, and have been used in a wide range of systems, from genetic circuits (Espinosa-Soto *et al.*, 2011), to RNA folding (Aguirre *et al.*, 2011; Fontana and Schuster, 1998), and to metabolic networks (Matias Rodrigues and Wagner 2009). In these cases, genotype networks were used to predict the robustness of a phenotype to mutations, and the potential of the underlying genotypes to evolve new and innovative traits.

There have so far been few applications of methods based on genotype–phenotype maps to empirical data (de Visser and Krug, 2014), even though the advent of new sequencing technologies provides large datasets of genotype data associated with phenotypes. To take advantage of such datasets, we developed VCF2Networks, a tool to apply genotype networks analysis to next-generation sequencing data. The tool permits the determination of genomic regions with high robustness of a given phenotype, i.e. mutations in this region affect the phenotype little or have high potential to create novel phenotypes.

2 APPROACH AND IMPLEMENTATION

A genotype networks is a graph of all the genotypes associated with a given phenotype. Each node of the graph represents an individual's genotype at a fixed number of loci, whereas two nodes are connected by an edge if their genotype differs at only one locus. In particular, in VCF2Networks, the relevant genotypes are single-nucleotide variants (SNVs) at multiple loci. Figure 1 shows an example of a hypothetical genotype network of five SNVs. Each node represents the genotypes of the five loci, encoded as strings of ones and zeros, where a zero represents the reference allele, and a one the alternative allele.

In the previous literature, some properties of a phenotype's genotype network have been associated with phenotypic robustness and the potential of the underlying genotypes to bring forth new phenotypes via DNA mutations. For example, the number of nodes and the average node degree can be interpreted as a measure of a phenotype's robustness (Ibáñez-Marcelo and Alarcón, 2014), while the diameters of some networks can serve as proxy for innovative potentials (Ciliberti *et al.*, 2007). For a more complete review of genotype networks, see Wagner, 2011.

VCF2Networks parses genotype files in the variant call format (VCF) (Danecek *et al.*, 2011), and produces tabular output containing properties such as the number of nodes, average degree and diameter for each of the genotype networks generated. More documentation on the output produced, and a discussion of best practices, is provided in the Supplementary Materials S1.

*To whom correspondence should be addressed.

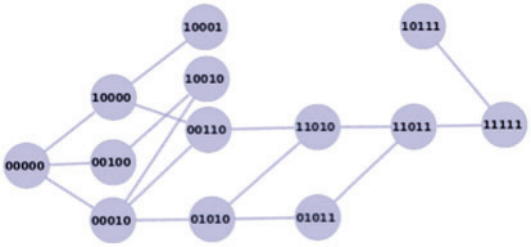

Fig. 1. A hypothetical genotype network. Nodes represent the genotypes of five SNVs. Two nodes are connected if they differ only minimally (in this case, at only one of the five loci)

3 USE CASE 1: IDENTIFYING REGIONS UNDER SELECTION IN HUMAN POPULATIONS

The tutorial of VCF2Networks uses example data from the 1000 Genomes Project. In this case, we do not have real phenotypes, but we can compare the genotype networks of different human populations, as exemplified by the command:

> $: vcf2networks –vcf 1000genomesdata.vcf –individuals ind_annotations.txt –phenotype population –network_size 11

The above command will parse the genotype data from a VCF file (–vcf 1000genomesdata.vcf), and split it into windows of 11 adjacent SNVs (–network_size 11). It will also read the phenotype of each individual from the file ind_annotations.txt, and compute a genotype network according to the phenotype 'population' defined in the same file.

In a previous work (Dall'Olio *et al.*, 2014), we showed that genomic regions under selection tend to have more vertices, greater average degree and greater average path length than regions evolving neutrally. This is in agreement with theoretical models, showing that high robustness facilitates the ability to innovate and adapt (Ibáñez-Marcelo and Alarcón, 2014). Thus, the above command can be used, in combination with other methods, to identify regions under selection in the 1000 Genomes or any other data.

In the same work (Dall'Olio *et al.*, 2014), we also derived some guidelines to calculate genotype networks from human population genetics data. Most importantly, the chosen size of the network should take into account the number of samples available. For example, we showed that for a sample size of ~850 individuals, a size of 11 SNVs is optimal (see Supplementary Materials S1 for a discussion on choosing the network size).

4 USE CASE 2: ROBUSTNESS OF CANCER PHENOTYPES

It has been proposed that cancer phenotypes are characterized by high genetic robustness and high genetic heterogeneity (Kitano, 2004; Tian *et al.*, 2010). Genetic robustness allows cancers to survive higher mutation rates, whereas genetic heterogeneity may help them evolve new traits, such as drug resistance or new tumorigenic characteristics. VCF2Networks can be used to analyze multiple DNA datasets coming from the same cancer patient, and identify regions with potentially high robustness and evolvability. For example, defining the phenotype as a binary trait (tumor or normal), and running the following line:

> $: vcf2networks –vcf myvcf.vcf –individuals ind_annotations.txt –phenotype cancer_status –network_size 5

The above command will generate a whole-genome scan of all cancer samples in the data. As the number of samples is lower than in 1000 genomes file, we use a network size of only five SNVs. Regions showing high robustness (high average degree) and high evolvability (high average path length and diameter) may be important in the evolution of the cancer phenotype.

5 AVAILABILITY

VCF2Networks is available from the Python Package Index, and can be installed through the python setuptools utilities (easy_install vcf2networks). The home page of the project is https://bitbucket.org/dalloliogm/vcf2networks/. VCF2Networks follows the best practices as proposed in Seemann, 2013.

ACKNOWLEDGEMENTS

The authors thank Tiago Carvalho, Brandon Invergo and Christian Pérez-Llamas for feedback.

Funding: This study has been possible thanks the grant BFU2013-43726-P awarded by Ministerio de Economía y Competitividad (Spain) and with the support of Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2014 SGR 866). GMD was supported by a FPI fellowship BES-2009-017731. AW was supported by the Swiss National Science Foundation and by the URPP Evolutionary Biology at the University of Zurich.

Conflict of interest: none declared.

REFERENCES

Aguirre,J. *et al.* (2011) Topological structure of the space of phenotypes: the case of RNA neutral networks. *PLoS One*, **6**, e26324.

Ciliberti,S. *et al.* (2007) Innovation and robustness in complex regulatory gene networks. *Proc. Natl Acad. Sci. USA*, **104**, 13591–13596.

Dall'Olio,G.M. *et al.* (2014) Human genome variation and the concept of genotype networks. *PLoS One*, **9**, e99424.

Danecek,P. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

de Visser,J.A. and Krug,J. (2014) Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.*, **15**, 480–490.

Espinosa-Soto,C. *et al.* (2011) Phenotypic plasticity can facilitate adaptive evolution in gene regulatory circuits. *BMC Evol. Biol.*, **11**, 5.

Fontana,W. and Schuster,P. (1998) Shaping space: the possible and the attainable in RNA genotype-phenotype mapping. *J. Theor. Biol.*, **194**, 491–515.

Kitano,H. (2004) Cancer as a robust system: implications for anticancer therapy. *Nat. Rev. Cancer*, **4**, 227–235.

Ibáñez-Marcelo,E. and Alarcón,T. (2014) The topology of robustness and evolvability in evolutionary systems with genotype-phenotype map. *J. Theor. Biol.*, **30**, 144–162.

Matias Rodrigues,J.F. and Wagner,A. (2009) Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comput. Biol.*, **5**, e1000613.

Seemann,T. (2013) Ten recommendations for creating usable bioinformatics command line software. *Gigascience*, **2**, 15.

Tian,T. *et al.* (2010) The origins of cancer and evolvability. *Integr. Biol.*, **3**, 17–30.

Wagner,A. (2011) *The Origins of Evolutionary Innovations*. Oxford University Press, Oxford (UK).