Application Note

VCF2Networks: applying Genotype Networks to Single Nucleotide Variants data

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

[1] Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Barcelona, Catalonia, Spain.
[2] Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland.
[3] The Swiss Institute of Bioinformatics, Bioinformatics, Lausanne, Switzerland.
[4] The Santa Fe Institute, Santa Fe, USA.
[5] Department de Genètica i de Microbiologia, Universitat Autonòma de Barcelona, Spain.
[+] Current Address: Division of Cancer Studies, King's College of London, UK.

Received on:

Associate Editor:

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 network 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 network analysis has rarely been applied to population genetics data. To help fill this gap, 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: Improved documentation and description of the output is given in the Supplementary Materials 1.

INTRODUCTION

Genotype networks can be used to describe the evolutionary properties of a set of genotypes associated with a qualitative and well defined 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 (Fontana and Schuster, 1998; Aguirre et al, 2013), and to metabolic networks (Matias-Rodrigues et al, 2011). 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 data sets, we developed VCF2Networks, a tool to apply genotype network analysis to next generation sequencing data. The tool permits determination of genomic regions with high robustness of a given phenotype, i.e., mutations in this region affect the phenotype little, or high potential to create novel phenotypes.

APPROACH AND IMPLEMENTATION

A genotype network 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, while two nodes are connected by an edge if their genotypes differ 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 zeroes, where a zero represents the reference allele, and a one the alternative allele.

In 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 vie DNA mutations. For example, the number of nodes and the average node degree can be interpreted as a measure of a phenotype's robustness (Payne et al, 2014; Ibáñez-Marcelo and Alarcón, 2014). The diameters of some networks, can serve as a proxy for innovative potential (Ciliberti et al, 2007). For a more complete review of genotype networks, see in (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

1 To whom correspondence should be addressed.

documentation on the output produced, and a discussion of best practices, is provided in Supplementary Materials 1.

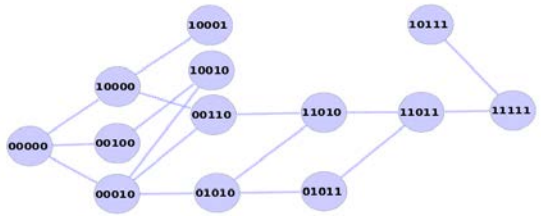

Fig. 1. A hypothetical genotype network. Nodes represent the genotypes of five Single Nucleotide Variants. Two nodes are connected if they differ only minimally (in this case, at only one of the five loci).

USE CASE A: 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 previous work (Dall'Olio GM et al 2014), we showed that genomic regions under selection tend to have more vertices, greater average degree, and average path lengths in a genotype network 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 potentially subject to selection, in the 1000 Genomes or any other data.

In the same work (Dall'Olio GM 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 about 850 individuals, a size of 11 SNVs is optimal (see Supplementary Materials 1 for a discussion on choosing the network size).

USE CASE B: 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, while heterogeneity may help them evolve new traits, such as drug resistance or new tumorigenic characteristics.

VCF2Networks can be used to analyze multiple DNA data sets 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), one can execute the following command:

```
$: vcf2networks --vcf myvcf.vcf --individuals
ind_annotations.txt --phenotype cancer_status
--network_size 5
```

Here, each column of the vcf file refers to a different tumor sample from within the same patient, and the ind_annotations.txt file annotates tumor samples instead of individuals. Since the number of samples is lower than in the 1000 genomes file, we use a network size of only 5 SNVs.

The above command will generate a whole-genome scan of all cancer samples in the data. 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.

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 for proposed in (Seemann, 2013).

ACKNOWLEDGEMENTS

We thank Tiago Carvalho, Brandon Invergo, Juan Ramón González Vallinas, and Christian Pérez-Llamas, and Juan Antonio Rodríguez from the Universitat Pompeu Fabra for feedback.

Funding: Grant BFU2010-19443 (subprogram BMC) from Ministerio Ciencia y Tecnologia (Spain). Grup de Recerca Consolidat 2009 SGR 1101, Direcció General de Recerca, Generalitat de Catalunya. FPI fellowship BES-2009-017731.

REFERENCES

Ciliberti,S. *et al.* (2007) Innovation and robustness in complex regulatory gene networks. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 13591–6.

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**, 2156–8.

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

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**, 227-35.

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.

Payne,J.L. *et al.* (2013) Robustness, Evolvability, and the Logic of Genetic Regulation. *Artif. Life*, **16**, 1–16.

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.