frontiers in GENETICS

GENERAL COMMENTARY
published: 14 October 2013
doi: 10.3389/fgene.2013.00206

The importance and application of the ancestral recombination graph

Miguel Arenas*

Centre for Molecular Biology "Severo Ochoa," Consejo Superior de Investigaciones Científicas, Universidad Autónoma de Madrid, Madrid, Spain
**Correspondence: marenas@cbm.uam.es*

Edited by:
Jeffrey Jensen, École Polytechnique Fédérale de Lausanne, Switzerland

Reviewed by:
Matthieu Foll, Ecole Polytechnique Fédérale de Lausanne, Switzerland
Gregory B. Ewing, École polytechnique fédérale de Lausanne, Switzerland

Keywords: ancestral recombination graph, recombination network, recombination breakpoints, ancestral material, recombinant fragment, phylogenetic bias

A commentary on

GraphML specializations to codify ancestral recombinant graphs
by McGill, J. R., Walkup, E. A., and Kuhner, M. K. (2013). Front. Genet. 4:146. doi: 10.3389/fgene.2013.00146

One of the most important evolutionary forces is recombination, it increases genetic diversity and promotes adaptation through exchange of genetic material and where existent mutations are shuffled. Knowledge about recombination is, for example, fundamental to understand genome structure (Reich et al., 2001), phenotypic diversity (Zhang et al., 2002), and diverse genetic diseases (Daly et al., 2001). Indeed, recombination should be considered to properly study molecular evolution and perform phylogenetic inferences (e.g., Schierup and Hein, 2000; Anisimova et al., 2003; Arenas and Posada, 2010c). The recombination evolutionary history is commonly represented by the ancestral recombination graph (ARG) (Griffiths and Marjoram, 1997), an illustrative example is shown in **Figure 1**. Counterintuitively, ARGs have not been widely used, perhaps as a consequence of the difficulties to infer explicit ARGs and the complexity of the ARG representation. The aim of this general commentary is to describe the importance and application of the ARG.

WORKING WITH THE ARG

A key aspect of ARG software is the ARG format. Notice that the description of an ARG is not straightforward because it is not a simple phylogenetic network and therefore, it cannot be represented by a classic network format [e.g., Extended Newick; see Arenas et al. (2010)]. Any ARG format must include not only the evolutionary history, but also the regions with ancestral material (material that reaches the sample) for each node (see **Figure 1**). Only a few ARG formats have been developed so far (Buendia and Narasimhan, 2006; Vos et al., 2012) and there is a need for a standard format that would allow an easy communication among different ARG software. In this regard, recently McGill et al. (2013) presented the *ArgML* format that is based on XML syntax and can include detailed information about the ARG (e.g., recombination breakpoints and ancestral material). These features suggest *ArgML* as a successful candidate to the standard ARG format.

Several methods and computer tools exist to simulate and infer ARGs.

From the simulation perspective, to my knowledge, only two programs can output a simulated ARG, namely, *Serial NetEvolve*

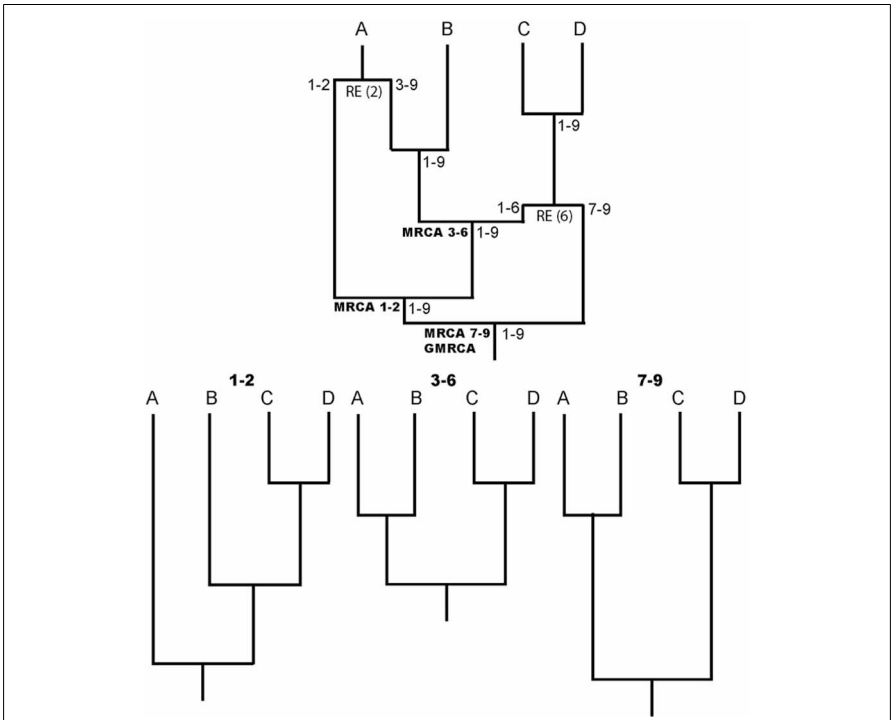

FIGURE 1 | Illustrative example of an ARG. RE indicates recombination events. Numbers in nodes indicate intervals of ancestral material. Note that each recombinant fragment (1–2, 3–6, and 7–9) has its own most recent common ancestor (MRCA), all of which finally coalesce into a grand MRCA (GMRCA). At the bottom, the trees corresponding to each recombinant fragment are shown.

(Buendia and Narasimhan, 2006) and *NetRecodon* (Arenas and Posada, 2010a). Most of simulation tools only simulate a tree for each recombinant fragment, but these trees can be then combined to generate the ARG by using tools like *CombineTrees* (see Woolley et al., 2008).

On the other hand, the inference of ARGs is complex and computationally extensive (see Rasmussen and Siepel, 2013 and references therein). For this reason, a common procedure to infer an ARG consists of the detection of recombination breakpoints (see Martin et al., 2011) followed by a phylogenetic tree reconstruction for each recombinant fragment and finally, a combination of all the reconstructed trees. Actually, the *IRiS* tool (Javed et al., 2011) has automated this whole procedure described above. However, in recent years, new methods to directly infer ARGs are emerging. For example, *ACG* (O'Fallon, 2013) that is based on a Bayesian Markov chain Monte Carlo (MCMC) procedure to compute the full likelihood of the ARG, or *ARGweaver* (Rasmussen and Siepel, 2013) that is based on hidden Markov models (HMM) and can infer ARGs from genome-wide data.

APPLYING THE ARG

ARGs can be applied for a variety of purposes, some examples are described below.

1. Detailed visualization and understanding of relationships among lineages, timing of recombination events and genetic exchange (e.g., Utro et al., 2012).
2. The ARG can be used in a variety of population genetics inferences related with demographics, population divergence times, migration and selection. Note that these population genetics scenarios may affect the length and shape of the ARG (Hudson and Kaplan, 1988; Griffiths and Marjoram, 1997; Rasmussen and Siepel, 2013). For example, it is known that variable population sizes can alter the branch lengths, in particular coalescences are more likely to occur when the population is small. Therefore, the ARG can be used to find correlations between recombination rate and population size over time (e.g., Birkner et al., 2013).
3. Recombination can bias phylogenetic tree reconstruction (e.g., Schierup and Hein, 2000; Posada and Crandall, 2002) and derived inferences (Anisimova et al., 2003; Arenas and Posada, 2010a,b,c). In order to avoid such a bias, trees embedded in the ARG can be used to correctly perform phylogenetic estimations (e.g., ancestral sequence reconstruction or molecular adaptation) accounting for recombination (e.g., Perez-Losada et al., 2009, 2011).
4. Molecular evolution can be simulated over an ARG (e.g., Buendia and Narasimhan, 2006) and the simulated genetic data can be applied for hypothesis testing, to evaluate analytical tools (see the reviews Arenas, 2012, 2013) or to estimate evolutionary parameters by using approaches like approximate Bayesian computation (ABC) (Wilson et al., 2009; Lopes et al., in press).

CONCLUSIONS

The ARG is indispensable to study evolutionary scenarios where recombination has occurred. With the development of next-generation sequencing (NGS) technologies there is a growing number of genomes at our disposal, many of which could have evolved under recombination (e.g., Utro et al., 2012; Rasmussen and Siepel, 2013). As a consequence, the importance and application of ARGs is expected to increase over the next years.

In order to make progress in the use of ARGs, a key aspect is the design of a standard format to represent the ARG. In this regard, McGill et al. (2013) presented a format based on XML syntax that can easily be used to store and communicate ARGs.

ACKNOWLEDGMENTS

I want to thank the Editor of *Frontiers in Evolutionary and Population Genetics*, for the invitation to contribute with this general commentary about the article by McGill et al. (2013). I also want to thank the two reviewers for their thoughtful comments. I thank the Spanish Government for the "Juan de la Cierva" fellowship, JCI-2011-10452.

REFERENCES

Anisimova, M., Nielsen, R., and Yang, Z. (2003). Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164, 1229–1236.

Arenas, M. (2012). Simulation of molecular data under diverse evolutionary scenarios. *PLoS Comput. Biol.* 8:e1002495. doi: 10.1371/journal.pcbi.1002495

Arenas, M. (2013). Computer programs and methodologies for the simulation of DNA sequence data with recombination. *Front. Genet.* 4:9. doi: 10.3389/fgene.2013.00009

Arenas, M., Patricio, M., Posada, D., and Valiente, G. (2010). Characterization of phylogenetic networks with NetTest. *BMC Bioinformatics* 11:268. doi: 10.1186/1471-2105-11-268

Arenas, M., and Posada, D. (2010a). Coalescent simulation of intracodon recombination. *Genetics* 184, 429–437. doi: 10.1534/genetics.109.109736

Arenas, M., and Posada, D. (2010b). Computational design of centralized HIV-1 genes. *Curr. HIV Res.* 8, 613–621. doi: 10.2174/157016210794088263

Arenas, M., and Posada, D. (2010c). The effect of recombination on the reconstruction of ancestral sequences. *Genetics* 184, 1133–1139. doi: 10.1534/genetics.109.113423

Birkner, M., Blath, J., and Eldon, B. (2013). An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics* 193, 255–290. doi: 10.1534/genetics.112.144329

Buendia, P., and Narasimhan, G. (2006). Serial NetEvolve: a flexible utility for generating serially-sampled sequences along a tree or recombinant network. *Bioinformatics* 22, 2313–2314. doi: 10.1093/bioinformatics/btl387

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nat. Genet.* 29, 229–232. doi: 10.1038/ng1001-229

Griffiths, R. C., and Marjoram, P. (1997). "An ancestral recombination graph," in *Progress in Population Genetics and Human Evolution*, eds P. Donelly and S. Tavaré (Berlin: Springer-Verlag), 257–270. doi: 10.1007/978-1-4757-2609-1_16

Hudson, R. R., and Kaplan, N. L. (1988). The coalescent process in models with selection and recombination. *Genetics* 120, 831–840.

Javed, A., Pybus, M., Mele, M., Utro, F., Bertranpetit, J., Calafell, F., et al. (2011). IRiS: construction of ARG networks at genomic scales. *Bioinformatics* 27, 2448–2450. doi: 10.1093/bioinformatics/btr423

Lopes, J. S., Arenas, M., Posada, D., and Beaumont, M. A. (in press). Coestimation of recombination, substitution and molecular adaptation rates by approximate Bayesian computation. *Heredity*.

Martin, D. P., Lemey, P., and Posada, D. (2011). Analysing recombination in nucleotide sequences. *Mol. Ecol. Resour.* 11, 943–955. doi: 10.1111/j.1755-0998.2011.03026.x

McGill, J. R., Walkup, E. A., and Kuhner, M. K. (2013). GraphML specializations to codify ancestral recombinant graphs. *Front. Genet.* 4:146. doi: 10.3389/fgene.2013.00146

O'Fallon, B. D. (2013). ACG: rapid inference of population history from recombining nucleotide sequences. *BMC Bioinformatics* 14:40. doi: 10.1186/1471-2105-14-40

Perez-Losada, M., Jobes, D. V., Sinangil, F., Crandall, K. A., Arenas, M., Posada, D., et al. (2011).

Phylodynamics of HIV-1 from a phase III AIDS vaccine trial in Bangkok, Thailand. *PLoS ONE* 6:e16902. doi: 10.1371/journal.pone.0016902

Perez-Losada, M., Posada, D., Arenas, M., Jobes, D. V., Sinangil, F., Berman, P. W., et al. (2009). Ethnic differences in the adaptation rate of HIV gp120 from a vaccine trial. *Retrovirology* 6, 67. doi: 10.1186/1742-4690-6-67

Posada, D., and Crandall, K. A. (2002). The effect of recombination on the accuracy of phylogeny estimation. *J. Mol. Evol.* 54, 396–402.

Rasmussen, M. D., and Siepel, A. (2013). Genome-wide inference of ancestral recombination graphs. *arXiv* 1306.5110v2.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., et al. (2001). Linkage disequilibrium in the human genome. *Nature* 411, 199–204. doi: 10.1038/35075590

Schierup, M. H., and Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics* 156, 879–891.

Utro, F., Cornejo, O. E., Livingstone, D., Motamayor, J. C., and Parida, L. (2012). ARG-based genome-wide analysis of cacao cultivars. *BMC Bioinformatics* 13(Suppl. 19):S17. doi: 10.1186/1471-2105-13-S19-S17

Vos, R. A., Balhoff, J. P., Caravas, J. A., Holder, M. T., Lapp, H., Maddison, W. P., et al. (2012). NeXML: rich, extensible, and verifiable representation of comparative data and metadata. *Syst. Biol.* 61, 675–689. doi: 10.1093/sysbio/sys025

Wilson, D. J., Gabriel, E., Leatherbarrow, A. J., Cheesbrough, J., Gee, S., Bolton, E., et al. (2009). Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. *Mol. Biol. Evol.* 26, 385–397. doi: 10.1093/molbev/msn264

Woolley, S. M., Posada, D., and Crandall, K. A. (2008). A comparison of phylogenetic network methods using computer simulation. *PLoS ONE* 3:e1913. doi: 10.1371/journal.pone.0001913

Zhang, Y. X., Perry, K., Vinci, V. A., Powell, K., Stemmer, W. P., and del Cardayre, S. B. (2002). Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature* 415, 644–646. doi: 10.1038/415644a

Received: 28 August 2013; accepted: 24 September 2013; published online: 14 October 2013.

*Citation: Arenas M (2013) The importance and application of the ancestral recombination graph. Front. Genet. **4**:206. doi: 10.3389/fgene.2013.00206*

This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics.