Planned analyses for evaluating the effect of transcriptomic data on gene content simulations

Principal Investigator: Joseph Ryan

Draft or Version Number: v.1.0

20 December 2017

LIST OF ABBREVIATIONS

<Insert text>	<Insert text>
ML	Maximum likelihood
AU	Approximately unbiased
GTR	Generalised time reversible
MCMC	Markov chain Monte Carlo

1 INTRODUCTION: BACKGROUND INFORMATION AND SCIENTIFIC RATIONALE

1.1 BACKGROUND INFORMATION

Gene content provides an independent set of characters from which phylogenetic relationships can be determined. Currently, genome data across the tree of life is limited, but transcriptome data is much more plentiful (Dunn & Ryan, 2015). The proposed set of simulation experiments test to see if transcriptomic data contribute positively to phylogenetic reconstruction despite the incomplete nature of the data (i.e., not all genes are expressed in a particular transcriptome).

1.2 RATIONALE

Several studies have used gene content data from whole-genome data to reconstruct phylogenies (Burger et al. 2011; Fang et al. 2013; Ryan et al. 2013; Pisani et al. 2015), but we are not aware of studies that consider using transcriptomic data in this context. If we can prove that the data is beneficial and determine minimum completeness levels, the amount of data that can be applied to gene content phylogenies could increase exponentially. These analyses will also shed light on the effectiveness of the two most applied methods for phylogenetic reconstruction using gene content.

1.3 OBJECTIVES

We will simulate gene presence/absence data on a tree that includes genomes and transcriptomes and test to see if the inclusion of the transcriptomic data is beneficial. We will also adjust levels of transcriptomic incompleteness to see what type of completeness is required.

2 STUDY DESIGN AND ENDPOINTS

We have written a program (gene_content_sim) that simulates gene gain and loss and also transcriptomic completeness. The program accepts a tree and a set of parameters (e.g. percent-gain, percent-loss, transcriptomic-completeness, number of characters, etc.) and produces a gene content matrix. We will use matrices generated with a range of parameters to reconstruct phylogenies using ML and Bayesian methods described in Ryan et al. (2013) as well as Bayesian methods described in Pisani et al. (2015). We will compare reconstructions with and without transcriptomic data. As our starting tree, we will use a composite tree with the following 2 topologies (Porifera,(Ctenophora,(Placozoa,(Cnidaria,Bilateria)) and (Ctenophora,(Porifera,(Placozoa,(Cnidaria,Bilateria)); each of the major lineages will come from the following individual studies: Porifera from Simion et al. 2017; Ctenophora

from Whelan et al. 2017; Cnidaria from Kayal et al. 2017; Bilateria from Cannon et al. 2016 (composite_spongesis.tre and composite ctenosis.tre in this repository).

Our simulator sets an initial probability of loss for each of the initial set of columns; by default the left most column is assigned a probability of loss = 0.1 and the rightmost column is assigned a probability of loss = 0.5 and the in-between columns are assigned an equal spread of probabilities between 0.1 and 0.5. All new genes are assigned a probability of 0.5. This effectively makes genes at the left end of the matrix less likely to be lost than those at the right end of the matrix. The intention is to reflect the reality that certain types of genes are less likely to be lost and the effect is to increase the level of homoplasy since under this model, individual genes are more likely to be lost multiple times. We implement the same logic for the removal of "non-expressed" genes in a transcriptome set. By default genes less likely to be lost are more likely to be "expressed", but we also will run with the --conserved_genes_less_likely_expressed option which does the opposite.

1. The following commandlines will generate 1000 matrices with the corresponding parameters for each of our two composite trees:

a) gene_content_sim --tree=composite_spongesis.tre --perc_loss=0.01 --perc_gain=0.01 --num_chars=23910

b) gene_content_sim --tree=composite_ctenosis.tre --perc_loss=0.01 --perc_gain=0.01 --num_chars=23910

We also run the above after removing transcriptome data:

c) gene_content_sim --tree=composite_spongesis.tre --perc_loss=0.01 --perc_gain=0.01 --num_chars=23910 --taxa_file=genome_taxa.txt

d) gene_content_sim --tree=composite_ctenosis.tre --perc_loss=0.01 --perc_gain=0.01 --num_chars=23910 --taxa_file=genome_taxa.txt

2. For each simulated gene matrix the following commandlines will generate trees using ML and Bayesian methods.

2.1 Perform a maximum-likelihood analysis used the GTR+gamma model.

a) raxmlHPC -m BINGAMMA -K GTR -p 420 -s <matrix file> -n <name>

2.2 Bayesian MCMC analyses using MrBayes methods described in Ryan et al (2013). The NEXUS block is as follows:

```
#NEXUS

begin data;

 dimensions ntax=12 nchar=898;

 format datatype=restriction interleave=no gap=-;

 matrix

 <DATA MATRIX HERE>

  ;
```

```
end;

begin mrbayes;

  lset rates=invgamma;

  mcmc;

sumt;

end;
```

2.3 Bayesian MCMC analysis using MrBayes methods described in Pisani et al. (2015).

A) applying no ascertainment bias correction

```
mb

exec [matrix file]

lset coding=noabsencesites|nosingletonpresence

lset rates=gamma

set autoclose=yes

mcmc filename=metazoa

sumt
```

B) applying corrections developed to account specifically for the removal of genes present in fewer than two taxa

```
mb

exec [matrix file]

lset coding=informative

lset rates=gamma

set autoclose=yes

mcmc filename=metazoa

sumt
```

C) applying a correction for the removal of parsimony uninformative sites

```
mb

exec [matrix file]
```

```
lset coding=all

lset rates=gamma

set autoclose=yes

mcmc filename=metazoa

sumt
```

3. Use Phyutility to prune non-genomic taxa from trees that include non-genomic taxa (i.e. trees run on datasets from 1.a and 1.b). Also prune the original tree used for the simulation. Then concatenate all 10 trees for a each dataset and also include the pruned original tree, for a total of 11 trees per dataset.

4. Run CONSEL (Shimodaira & Hasegawa 2001) on each of these 11 trees (2000 sets in total)

```
raxmlHPC -f G -m BINGAMMA --no-bfgs -z 11trees.N.tre -s [alignment_file] -n 11trees.N

seqmt --puzzle RAxML_perSiteLLs.11trees.N

makermt RAxML_perSiteLLs

consel RAxML_perSiteLLs

catpv RAxML_perSiteLLs > out.au
```

6. Scoring

For each of the 5 different phylogenetic methods we will increment “genome_only” score if the genome_only dataset was ranked higher or increment the “with_transcriptomes” score if the full tree from the full dataset is ranked higher. Whichever has the highest rank score (“genome_only” or “with_transcriptomes”) will be the most effective in terms of rank. We will also generate a p-value score where we will simply add p-values for each test. The higher the P-value score the better.

3 WORK COMPLETED SO FAR W DATES

gene_content_sim was created using a toy dataset. None of the above analyses have been run.

4 LITERATURE REFERENCES

Dunn CW, Ryan JF. The evolution of animal genomes. Current opinion in genetics & development. 2015 Dec 31;35:25-32.

Ryan JF, Pang K, Schnitzler CE, Nguyen AD, Moreland RT, Simmons DK, Koch BJ, Francis WR, Havlak P, Smith SA, Putnam NH. The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution. Science. 2013 Dec 13;342(6164):1242592.

Fang H, Oates ME, Pethica RB, Greenwood JM, Sardar AJ, Rackham OJ, Donoghue PC, Stamatakis A, de Lima Morais DA, Gough J. A daily-updated tree of (sequenced) life as a reference for genome research. Scientific Reports. 2013 Jun 18;3:srep02015.

Pisani D, Pett W, Dohrmann M, Feuda R, Rota-Stabelli O, Philippe H, Lartillot N, Wörheide G. Genomic data do not support comb jellies as the sister group to all other animals. Proceedings of the National Academy of Sciences. 2015 Dec 15;112(50):15402-7.

Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Quéinnec É, Ereskovsky A, Lapébie P. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. Current Biology. 2017 Apr 3;27(7):958-67.

Whelan NV, Kocot KM, Moroz TP, Mukherjee K, Williams P, Paulay G, Moroz LL, Halanych KM. Ctenophore relationships and their placement as the sister group to all other animals. Nature ecology & evolution. 2017 Oct 9;1(11):1737.

Kayal E, Bastian B, Pankey MS, Ohdera A, Medina M, Plachetzki DC, Collins A, Ryan JF. Comprehensive phylogenomic analyses resolve cnidarian relationships and the origins of key organismal traits. PeerJ Preprints. 2017 Aug 21.

Cannon JT, Vellutini BC, Smith J, Ronquist F, Jondelius U, Hejnol A. Xenacoelomorpha is the sister group to Nephrozoa. Nature. 2016 Feb 4;530(7588):89-93.

S. A. Berger, A. Stamatakis, R. Lucking, Morphology-based phylogenetic binning of the lichen genera Graphis and Allographa (Ascomycota: Graphidaceae) using molecular site weight calibration. *Taxon* **60**, 1450-1457 (2011).

5 PHYLOTOCOL AMENDMENT HISTORY

Version	Date	Significant Revisions