Identification of lineage-specific gene family expansions in a database of gene families

CURRENT STATUS: POSTED

Matthew Berriman
Berriman Lab Group, Sanger Institute
mb4@sanger.ac.uk*Corresponding Author*

Diogo Ribeiro
Berriman Lab Group, Sanger Institute

Avril Coghlan
Berriman Lab Group, Sanger Institute

Bhavana Harsha
Berriman Lab Group, Sanger Institute

DOI:
10.1038/protex.2018.057

SUBJECT AREAS
Computational biology and bioinformatics *Biotechnology*

KEYWORDS
Gene families, gene family expansions, Compara, lineage-specific expansions, Ensembl

Abstract

Gene families specific to, or with significantly changed membership in, particular lineages compared to outgroups may reflect important lineage-specific changes in biology. Here we describe a computational protocol to identify gene families that vary greatly in gene count across a species tree. This protocol uses three different metrics to capture aspects of this variability, and calculates them for each family in an in-house database of gene families (e.g. built using the Ensembl Compara pipeline). One metric (C_v) identifies families that vary a lot in gene count across the species tree, and the other two (E_{max}, Z_{max}) identify families that have an elevated gene count in a certain clade of the species tree. Our protocol controls for differences in gene counts due to fragmented assemblies.

Introduction

Gene families specific to, or with significantly changed membership in, particular lineages compared to outgroups may reflect important lineage-specific changes in biology. Here we describe a computational protocol to identify gene families that vary greatly in gene count across a species tree. This protocol can be used to analyse families from an in-house database of gene families (e.g. built using the Ensembl Compara pipeline, Vilella *et al* 2009).

To identify gene families that vary greatly in gene count across a species tree, we use three metrics to capture aspects of this variability. However, to control for differences in gene counts due to fragmented assemblies (e.g. genes split across multiple contigs), we use summed protein length per species in a family as a proxy for gene counts in these metrics. The first metric, coefficient of variation (C_v), simply measures the variability in summed protein length (per species) in a family. The other two metrics, Z-score (Z_{max}) and enrichment coefficient (E_{max}), reflect whether there is a gene family expansion in a particular group of species, relative to the rest of the species tree. For this, species are classified into a set of non-overlapping species groups of interest that ideally correspond to monophyletic clades of the species tree (e.g. nematodes, flatworms, and arthropods in a species tree of animals).

The three metrics are defined as follows:

(i) coefficient of variation,

See figure in Figures section.

where the numerator (s) is the standard deviation in the summed protein length per species in the family, and the denominator is the mean of the summed protein length per species in the family.

(ii) maximum Z-score,

See figure in Figures section.

where T is the set of non-overlapping species groups, c is a species group in the set T. For each species group c in the set T, we calculate Z as the difference between the mean summed protein length (per species) in that species group c and the overall mean (the numerator), divided by the standard deviation in summed protein length per species among the species outside group c (the denominator). Z_{max} is the maximum of these Z values over all tested species groups. Note that the standard deviation used here is for the species outside species group c, so that it is not affected by any gene family expansion that occurred in the family within the species group c.

(iii) maximum enrichment coefficient,

See figure in Figures section.

That is, for each species group in the set T, we calculate E as the ratio between the mean of the summed protein length (per species) in that species group and the mean summed protein length outside that species group. E_{max} is the maximum of these E values over all tested species groups.

High values of the three metrics may be used to detect families with different patterns of variability. High values of Z_{max} and E_{max} may indicate a gene family expansion in a particular clade of the species tree, but high values of C_v may highlight cases that they miss, such as a gene family that has independently expanded in different clades of the species tree. In addition, a gene family that has expanded in a particular clade of the species tree can only have high values of Z_{max} and E_{max} if the gene family also includes genes from species outside this clade, whereas high values of C_v can be used to detect large variability within a single clade even if the family lacks members from outside that clade.

The protocol has two steps. In Step A, the set of all input gene families is filtered to remove likely

transposable element families, assuming that these are not of interest to most users, since expansions of transposable element gene families are relatively unlikely to contribute to lineage-specific changes in biology.

In Step B, we define a set of non-overlapping species groups, which are ideally monophyletic clades of the species tree (e.g. nematodes, flatworms), in which we wish to identify lineage-specific expansions (e.g. nematode-specific expansions, flatworm-specific expansions). In Step B, we also define a subset of the species as having high-quality assemblies. Then, the three metrics C_v, E_{max} and Z_{max} are calculated for each family in an in-house database of gene families. Families with high values of each metric are good candidates for families with lineage-specific expansions, and can be manually examined in follow-up analyses.

Our protocol controls for differences in gene counts due to fragmented assemblies, firstly by using summed protein length per species in a family as a proxy for gene count, and also by calculating the three metrics based only on those species that have high-quality assemblies (as defined in Step B).

A related approach for finding lineage-specific gene family expansions is to identify Pfam (Finn *et al* 2014) domains that are enriched in a certain species group compared to the rest of the species tree, and then to identify the expanded gene families containing that domain (e.g. see Tsai *et al* 2013). In contrast, our current protocol does not require that the genes in an expanded family have any predicted protein domains, so can detect expansions in uncharacterised or relatively poorly characterised gene families.

A more similar approach to the current protocol was used to identify *Spirometra* lineage-specific expansions by Bennett *et al* 2014, by ranking gene families by the ratio of the total cumulative length of encoded *Spirometra erinaceieuropaei* genes in a family to the cumulative length of the corresponding *Echinococcus multilocularis* genes in the same family. This is a very similar idea to E_{max}, but E_{max} for a family is extended to be a ratio between the mean summed protein length in a species group, compared to the mean summed protein length outside that species group. This should be a more accurate metric for detecting expansions because it can include information from more

species (e.g. several *Spirometra*-lineage species and several *Echinococcus*-lineage species).

Reagents

Ensembl Compara pipeline software

Pfam

InterProScan

Equipment

Computer cluster.

Procedure

Step A: filtering gene sets and gene families for transposable element genes

The input to Step A is a gene prediction set (and the corresponding predicted protein sequences) for each species of interest; and a database of gene families containing all the genes from all those species, e.g. built using the Compara pipeline (Vilella *et al* 2009).

1. Interproscan 5 (Jones *et al* 2014) is used to identify predicted proteins with Pfam (Finn *et al* 2014) domains associated with transposable elements (using a list from Foth *et al* 2014): PF12762, integrase; PF03221, DNA-binding; PF03184, endonuclease; PF00078, reverse transcriptase; PF03564, DUF1759; PF05380, Pao retrotransposon peptidase; PF10551, transposase; PF00077, retroviral aspartyl protease; PF13456, reverse transcriptase-like; PF00665, integrase; PF14227, Gag-polypeptide of LTR copia-type; PF03732, retrotransposon gag protein; PF01541, GIY-YIG catalytic domain; PF00680, RNA dependent RNA polymerase; PF07727, reverse transcriptase; PF13961, DUF4219; PF01359, transposase; PF08284, retroviral aspartyl protease; PF13976, GAG-pre-integrase; PF14223, Gag-polypeptide of LTR copia-type; and PF14244, Gag-polypeptide of LTR copia-type.
2. Gene families are identified that have at least one transposon Pfam domain assigned to at least one member. We consider a family as ‘transposon-related’ if ≥20% of its genes (with or without any Pfam annotation) have a transposon-associated Pfam domain, and if ≥80% of the genes with at least one Pfam annotation have a transposon-associated Pfam domain. The identified gene families are subsequently excluded from further analyses of gene family expansions.

The output from Step A is the set of gene families from which likely transposable element gene

families have been removed.

Step B: calculating metrics for identifying highly variable gene families

The input to Step B is the set of non-transposable element gene families output from Step A above, and a species tree for the full set of species.

1. The set of all species is classified into a set of non-overlapping subsets. We refer to these subsets as ‘species groups’, and they are the subsets in which we wish to identify gene family expansions. For example, if the full list of all our species is {‘A’, ‘B’, ‘C’, ‘D’, ‘E’}, the set of non-overlapping subsets could be {{‘A’, ‘C’, ‘D’}, {‘B’, ‘E’}}. Ideally each species group corresponds to a monophyletic clade of the species tree.
2. If the set of all species includes some species with relatively high-quality assemblies and some species with much more fragmented assemblies, then a subset of species that have high-quality assemblies, and span the major clades of the species tree, are designated ‘tier 1’ species. The tier 1 species should ideally be those with relatively contiguous assemblies (e.g. N50/scaffold-count > 5) and relatively complete proteomes (e.g. CEGMA (Parra *et al* 2007) partial score > 85%).
3. The metrics C_v, Z_{max} and E_{max} are calculated for each of the gene families. To increase the reliability of these metrics, only the tier 1 species are considered, since these species have the highest quality assemblies, and therefore most complete proteomes and fewest artefactual gene splits and merges. When calculating the means and standard deviations that contribute to these metrics, the ‘tier 1’ species that lack any genes (in a family) are taken into account, as follows. First, to calculate the mean and standard deviation for a whole family (e.g. for C_v), we identify the node of the species tree corresponding to the root of the gene tree for the family. Then, for each tier 1 descendent species that is present on the species tree but not in the current gene family, we count its summed protein length as zero when calculating the mean and standard deviation. Similarly, when calculating the means and standard deviations for a particular clade (for Z_{max} or E_{max}), the node of the species tree corresponding to the root of the

clade is identified, and all tier 1 species that descend from that node of the species tree are taken into account.

The code for calculating metrics is tied to our in-house infrastructure, but is available for perusal at tinyurl.com/comparaFamiliesAnalysis-py

The output from Step B is the C_v, Z_{max} and E_{max} values for each of the families.

Troubleshooting

Because at least two genes are needed to calculate the C_v, Z_{max} and E_{max} metrics for a family, and only tier 1 species are considered when calculating them, these metrics are undefined for families containing less than two genes from tier 1 species. Furthermore, in very small families, the estimates of C_v, Z_{max} and E_{max} are very noisy (because the underlying estimates of mean and standard deviation are noisy), so we recommend discarding families having few (e.g. <10 genes) from tier 1 species.

It is possible that some of the families with relatively high values of C_v, Z_{max} or E_{max} may turn out to be undetected transposable element genes; or families that have gene counts that are equally as high (or nearly as high) in outgroup species as in the species group of interest. Expansions of transposable element gene families, or expansions specific to outgroup species, are unlikely to be interesting to the user, so we suggest that users manually examine the families with high values of C_v, Z_{max} or E_{max} to discover and possibly discard such cases.

In Step B, species are classified into a set of non-overlapping species groups of interest in which we want to find lineage-specific expansions (e.g. nematodes, flatworms). These ideally correspond to monophyletic clades of the species tree. In order to detect lineage-specific expansions that occurred in narrower phylogenetic groups, the user may want to run the protocol again with narrower species group definitions (e.g. nematode clades I, II, III, IV, V, and cestode flatworms, trematode flatworms).

Anticipated Results

The output from the protocol is the C_v, Z_{max} and E_{max} values for each of the families. In order to filter the list of families to identify those families that have the most striking patterns of variability, we suggest taking the union of the top 500 highest-scoring families according each of the metrics (C_v,

Z_{max} and E_{max}), and then plotting phylogenetic trees for each of these families, and examining the plots by eye. The cutoff of 500 suggested here is arbitrary and users should investigate the effect of varying this cutoff.

References

Bennett, H.M. *et al.* The genome of the sparganosis tapeworm *Spirometra erinaceieuropaei* isolated from the biopsy of a migrating brain lesion. *Genome Biol.* **15**, 510 (2014).

Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).

Finn, R.D. *et al.* Pfam: the protein families database. *Nucleic acids research* **42**, D222-230 (2014).

Foth, B.J. *et al.* Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction. *Nat Genet.* **46**, 693-700 (2014).

Parra, G. *et al.* CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067 (2007).

Tsai, I.J. *et al.* The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* **496**, 57-63 (2013).

Vilella, A.J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327-335 (2009).

Acknowledgements

We would like to thank the WTSI Pathogen Informatics team, especially Jacqueline Keane.

Figures

$$C_v = \frac{s}{\bar{x}}$$

Figure 1

Definition of Cv

$$Z_{max} = \max_{c \in T} \left(\frac{\bar{x}_{i,i \in c} - \bar{x}}{s} \right)$$

Figure 2

Definition of Zmax

$$E_{max} = \max_{c \in T} \left(\frac{\bar{x}_{i,i \in c}}{\bar{x}_{i,i \notin c}} \right)$$

Figure 3

Definition of Emax