ACGT

THOUGHTS ON BIOLOGY, GENOMICS, AND THE ONGOING THREAT TO HUMANITY
FROM THE BOGUS USE OF BIOINFORMATICS ACRONYMS, BY KEITH BRADNAM

ABOUT BLOG CONTACT

Goodbye CEGMA, hello BUSCO!

May 18, 2015

Adapted from this flickr image

Some history

CEGMA (Core Eukaryotic Genes Mapping Approach) is a bioinformatics tool that I helped develop about a decade ago. It led to a paper that outlined how you could use CEGMA to find a small number of highly conserved genes in any new genome sequence that might be devoid of annotation. Once you have even a handful of genes, then you can use that subset to train a gene finder in order to annotate the rest of the genome.

It was a bit of a struggle to get the paper accepted by a journal, and so it didn't appear until 2007 (a couple of years after we had started work on the project). For CEGMA to work it needed to use a set of orthologous genes that were highly conserved across different eukaryotic species. We used the KOGs database (euKaryotic Orthologous Groups) that was first described in 2003. So by the time of our publication we were already using a dataset that was a few years old.

The original CEGMA paper did not attract much attention but we subsequently realized that the same software could be used to broadly assess how complete the 'gene space' was of any published genome assembly. To do this, we defined a subset of core genes that were the most highly conserved and which tended to be single-copy genes. The resulting 2009 paper seemed to generate a lot of interest in CEGMA and citations to the original paper have increased every year since (139 citations in 2014).

This is good news except:

1. CEGMA can be a real pain to install due to its dependency on many other tools (though we've made things easier)
2. CEGMA has been very hard to continue developing. The original developer left our group about 7 years ago and he was the principle software architect. I have struggled to keep CEGMA working and updated.
3. CEGMA continues to generate **a lot** of support email requests (that end up being dealt with by me).

We have no time or resources to devote to CEGMA but the emails keep on coming. It's easy to envisage many ways how CEGMA could be improved and extended; we submitted a grant proposal to do this but it was unsuccessful. One planned aspect of 'CEGMA v3' was to replace the reliance on the aging KOGs database. Another aspect of the new version of CEGMA would be to develop clade-specific sets of core genes. If you are interested in plant genomes, we should be able to develop a much larger set of plant-specific core genes.

And so?

Today I draw a line in the sand and say...

CEGMA is dead

CEGMA had a good life and and we shall look back with fond memories, but its time has passed. Please don't grieve (and don't send flowers), but be thankful that people will no longer have to deal with the software-dependency-headache of trying to get Genewise working on Ubuntu Linux.

But what now?

The future of CEGMA has arrived and it's called BUSCO.

- BUSCO: assessing genome assembly and annotation completeness with single-copy ortholog

This new tool (Benchmarking Universal Single-Copy Orthologs) has been developed by Felipe A. Simao, Robert Waterhouse, Panagiotis Ioannidis, Evgenia Kriventseva, and Evgeny Zdobnov. You can visit the BUSCO website, have a read of the manual, or look at the supplementary online material (this is set out like a paper...I don't think the tool has been formally published yet though). Here is the first section from that supplementary material:

> Benchmarking Universal Single-Copy Orthologs (BUSCO) sets are collections of orthologous groups with near-universally-distributed single-copy genes in each species, selected from OrthoDB root-level orthology delineations across arthropods, vertebrates, metazoans, fungi, and eukaryotes (Kriventseva, et al., 2014; Waterhouse, et al., 2013). BUSCO groups were selected from each major radiation of the species phylogeny requiring genes to be present as single-copy orthologs in at least 90% of the species; in others they may be lost or duplicated, and to ensure broad phyletic distribution they cannot all be missing from one sub-clade. The species that define each major radiation were selected to include the majority of OrthoDB species, excluding only those with unusually high numbers of missing or duplicated orthologs, while retaining representation from all major sub-clades. Their widespread presence means that any BUSCO can therefore be expected to be found as a single-copy ortholog in any newly-sequenced genome from the appropriate phylogenetic clade (Waterhouse, et al., 2011). A total of 38 arthropods (3,078 BUSCO groups), 41 vertebrates (4,425 BUSCO groups), 93 metazoans (1,008 BUSCO groups), 125 fungi (1,438 BUSCO groups), and 99 eukaryotes (431 BUSCO groups), were selected from OrthoDB to make up the initial BUSCO sets which were then filtered based on uniqueness and conservation as described below to produce the final BUSCO sets for each clade, representing 2,675 genes for arthropods, 3,023 for vertebrates, 843 for metazoans, 1,438 for fungi, and 429 for eukaryotes. For bacteria, 40 universal marker genes were selected from (Mende, et al., 2013).

BUSCO seems to do everything that we wanted to include in CEGMA v3 and it is based on OrthoDB, a resource that has generated a new set of orthologs (developed by the same authors). The online material includes a comparison of BUSCO to CEGMA, and also outlines how BUSCO can be much quicker than CEGMA (depending on what set of of orthologs you use).

DISCLAIMER: I have not installed, tested, or analyzed BUSCO in any way. I make no promises as to its performance, but they seem to have gone about things in the right way.

♥ 29 Likes

‹ Newer Older ›

Comments (5)

Newest First Subscribe via e-mail

Preview Post Comment…

James Wasmuth 2 years ago · 1 Like

I share Michael's concern. I looked at the BUSCO database a whole back. I recall that Loa loa had a better score than C. elegans.

James Wasmuth 2 years ago · 0 Likes

I share Michael Paulini's concern. I looked at BVSC database a while back. I recall seeing that Los lol had a better score than C. elegans.

Peter Cock 3 years ago · 1 Like

BUSCO looks interesting, and it is open source (GPL3), but its frustrating that http://busco.ezlab.org/ doesn't have links to a source code repository or issue tracker. If it was on GitHub/BitBucket I would be urging a PostDoc to submit a pull request rather than emailing them about white listing more species names which are supported in Augustus but not in BUSCO.

Keith Bradnam 3 years ago · 0 Likes

Hopefully, a wave of people trying BUSCO might get the developers interested in listening to feedback. Several people have commented to me about the lack of a plant specific set of orthologs as well. Room for improvements!

Michael Paulini 3 years ago · 1 Like

hmm ... seems like the C.elegans genome from release WS235 is 85-90% complete compared to 98-99% for D.melanogaster and 89-99% for H.sapiens according to the BUSCO SOM.