Genome analysis

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

Felipe A. Simão†, Robert M. Waterhouse†, Panagiotis Ioannidis, Evgenia V. Kriventseva and Evgeny M. Zdobnov*

Department of Genetic Medicine and Development, University of Geneva Medical School and Swiss Institute of Bioinformatics, rue Michel-Servet 1, 1211 Geneva, Switzerland

Maido Remm, 24.09.2015

Probleem

Peale genoomi *de novo* assambleerimist on vaja hinnata, kas genoom on kokku pandud:

a) õigesti

b) täies ulatuses

Kuidas mõõta assambleeritud genoomi?

Tüüpiline mõõdik on **N50**
(50% assambleeritud järjestusest on nii pikkades või pikemates DNA fragmentides)

Kuidas mõõta assambleeritud genoomi?

Sisulisemat infot annab geenide olemasolu kontrollimine:

* Kas mitokondri järjestus on korrektselt assambleeritud?

* Kas valku kodeerivad geenid on õigesti assambleeritud?

CEGMA

BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 9 2007, pages 1061–1067
doi:10.1093/bioinformatics/btm071

Genome analysis

CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes

Genis Parra[1], Keith Bradnam[1] and Ian Korf[1,2,*]

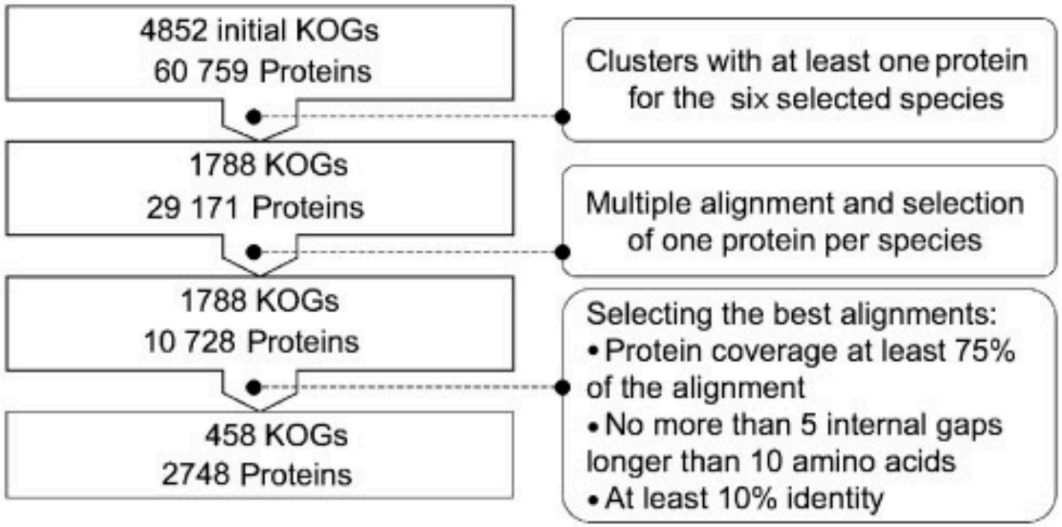
[1]UC Davis Genome Center, 451 E. Health Sciences Drive and [2]Department of Molecular and Cellular Biology, University of California Davis, Davis, CA 95616, USA

Received on December 7, 2006; revised on January 26, 2007; accepted on February 22, 2007

Advance Access publication March 1, 2007

Associate Editor: Alex Bateman

CEGMA
458 universally existing proteins



458 single copy genes/proteins that are present in:

Homo sapiens
Drosophila melanogaster
Caenorhabditis elegans
Arabidopsis thaliana
Schizosaccharomyces pombe
Saccharomyces cerevisiae

CEGMA workflow:

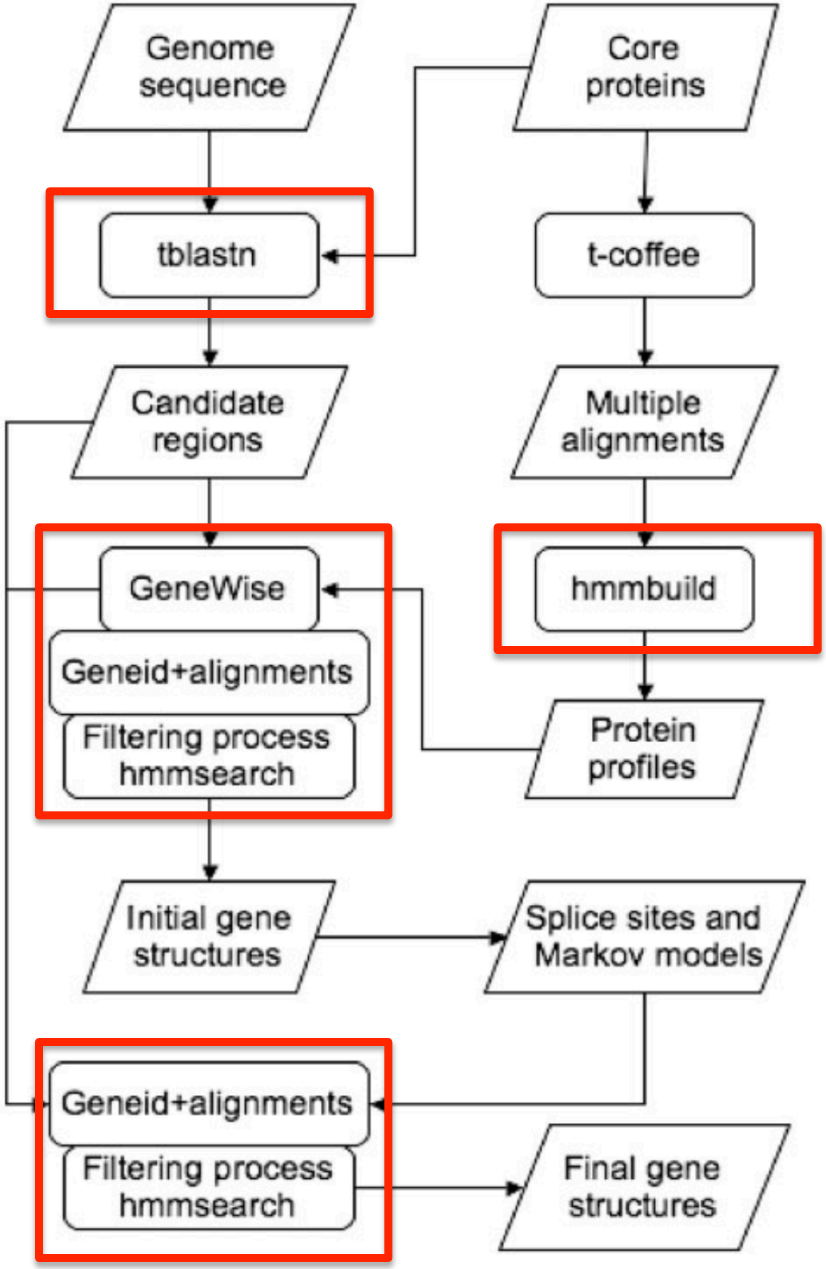
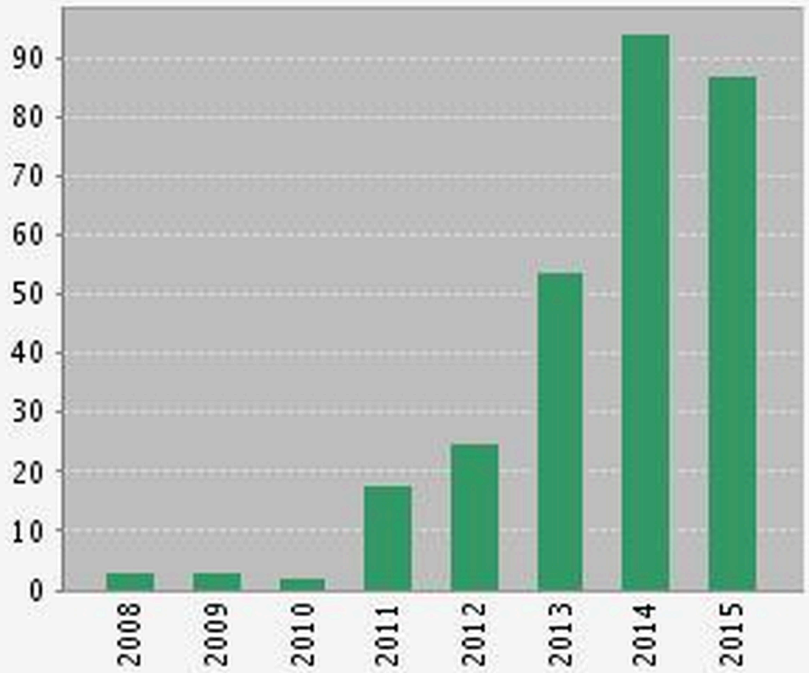


Fig. 2. Flowchart of CEGMA. The initial sources of information are the raw genomic sequence and the multiple alignment of the set of core proteins.

Citations in Each Year



Results found:	1
Sum of the Times Cited [?] :	286
Sum of Times Cited without self-citations [?] :	286
Citing Articles [?] :	286
Citing Articles without self-citations [?] :	286
Average Citations per Item [?] :	286.00
h-index [?] :	1

http://www.acgt.me/blog/2015/5/18/goodbye-cegma-hello-busco

CEGMA

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.

Goodbye CEGMA, hello BUSCO!

May 18, 2015

QUEST FOR QUALITY

"BUSCO CALIDAD"

"BUSCO QUALIDADE"

Assessing genome assembly and annotation completeness with Benchmarking Universal Single-Copy Orthologs

We used our OrthoDB database (http://www.orthodb.org/) of orthologs to define BUSCO sets for six major phylogenetic clades.

A total of 38 arthropods, 41 vertebrates, 93 metazoans, 125 fungi and 99 eukaryotes were selected from OrthoDB to make up the initial BUSCO sets.

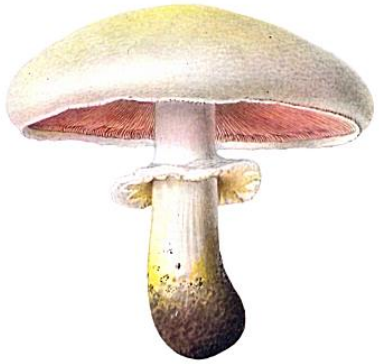
Orthologous groups with **single-copy orthologs in >90% of species** were selected.

Importantly, this threshold accommodates the fact that even well-conserved genes can be lost in some lineages, as well as allowing for incomplete gene annotations and rare gene duplications.

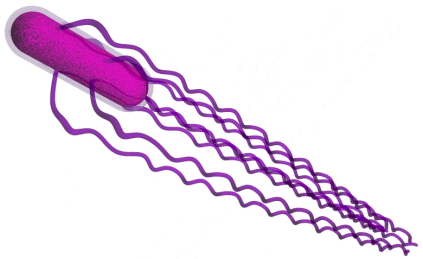
Number of genes in BUSCO sets for six major phylogenetic clades.

Vertebrate: 3023

Arthropoda: 2675



Fungi: 1438



Metazoa: 843

Eukaryotes: 429

Prokaryotes: 40

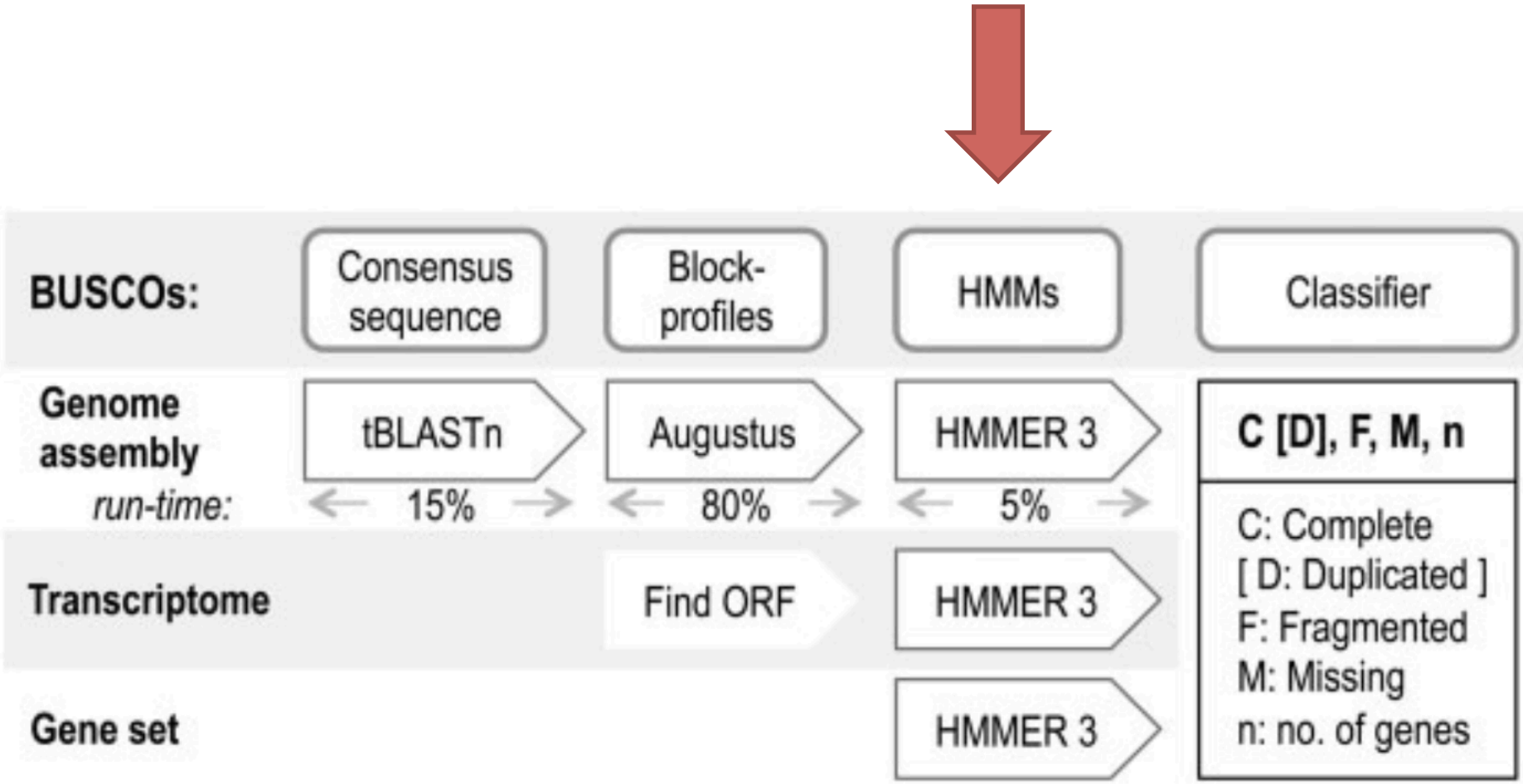


Fig. 1. BUSCO assessment workflow and relative run-times

Table 1. Assessment of fruitfly (*D. mela*,), nematode worm (*C. eleg*,), human (*H. sapi*,), owl limpet (*L. giga*,), and fungus (*A. nidu*,) genome assemblies (upper row) and gene sets (lower row) in BUSCO notation (C:complete [D:duplicated], F:fragmented, M:missing, n: gene number)

Species	Size	BUSCO notation assessment results
D. mela	139 Mbp	C:98% [D:6.4%], F:0.6%, M:0.3%, n:2 675
	13 918 genes	C:99% [D:3.7%], F:0.2%, M:0.0%, n:2 675
C. eleg	100 Mbp	C:85% [D:6.9%], F:2.8%, M:11%, n:843
	20 447 genes	C:90% [D:11%], F:1.7%, M:7.5%, n:843
H. sapi	3 381 Mbp	C:89% [D:1.5%], F:6.0%, M:4.5%, n:3 023
	20 364 genes	C:99% [D:1.7%], F:0.0%, M:0.0%, n:3 023
L. giga	359 Mbp	C:89% [D:2.3%], F:4.3%, M:5.8%, n:843
	23 349 genes	C:90% [D:13%], F:7.8%, M:2.1%, n:843
A. nidu	30 Mbp	C:98% [D:1.8%], F:0.9%, M:0.2%, n:1 438
	10 534 genes	C:95% [D:7.3%], F:3.8%, M:0.9%, n:1 438

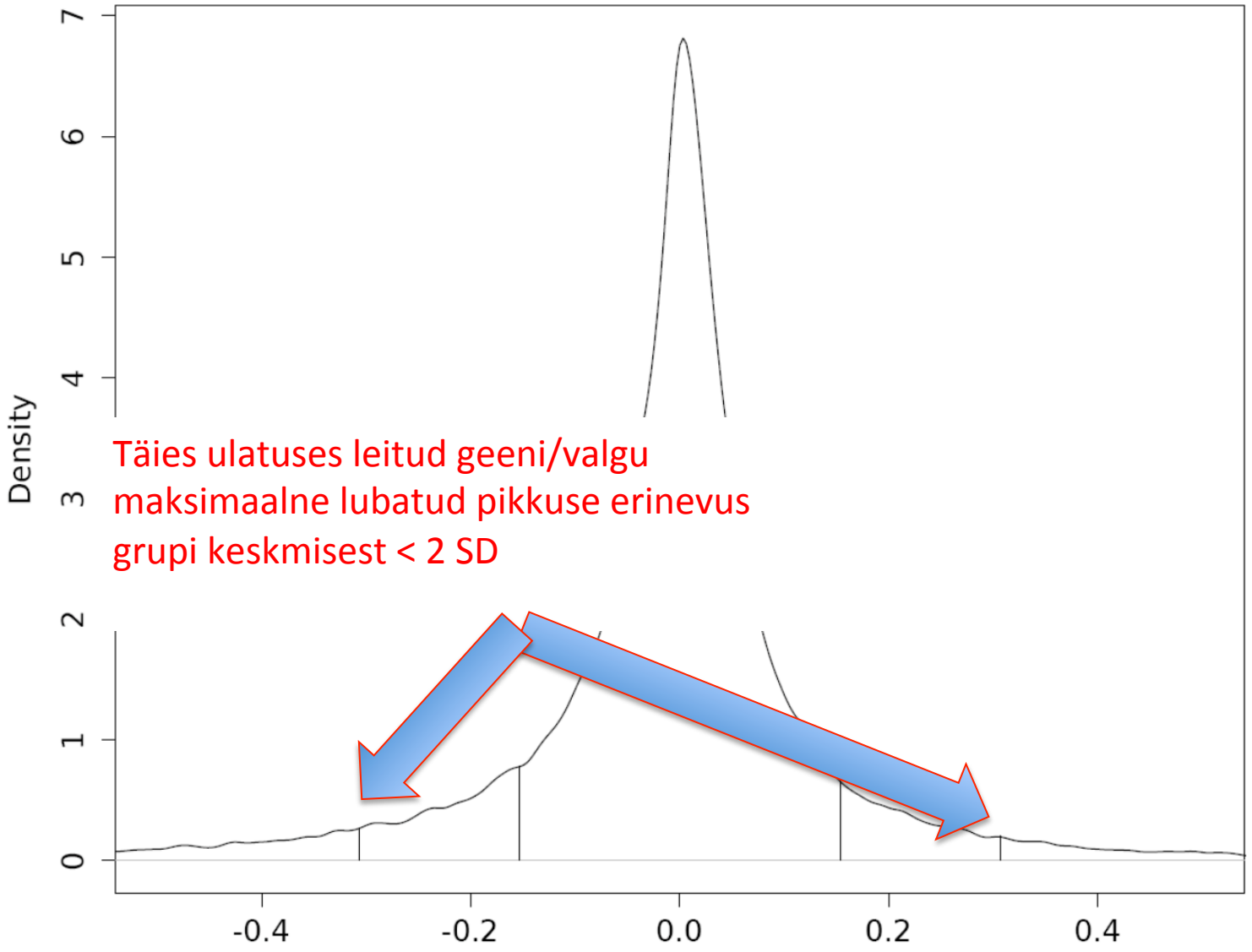
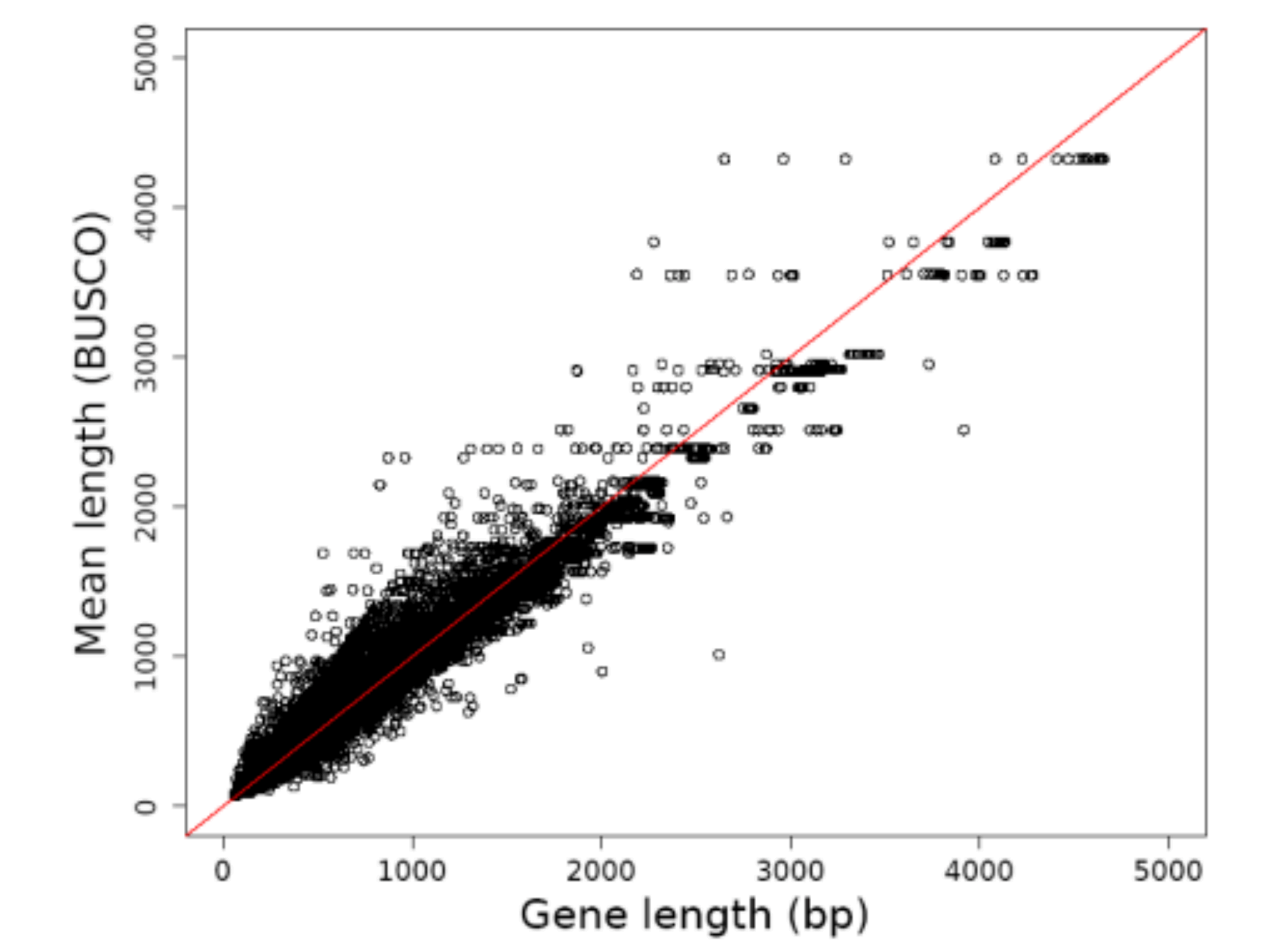
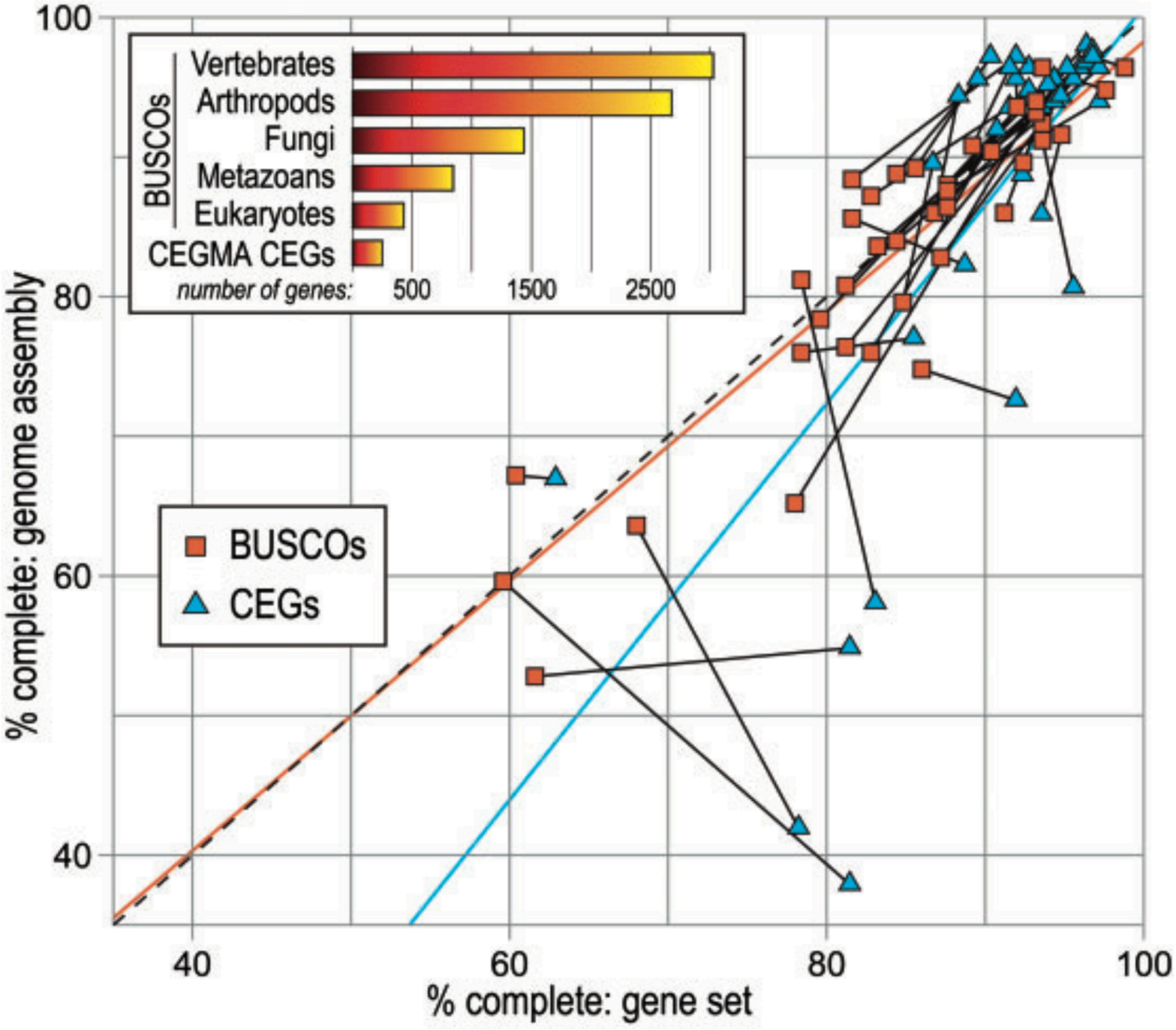


Figure S1. Distribution of the percent differences between BUSCO group member proteins and the group's mean protein length (negative = shorter than the mean, positive = longer than the mean, values of one and two standard deviations are shown with lines). Insets: spread of BUSCO group member protein lengths compared to BUSCO group mean lengths for arthropods (left) and vertebrates (right).


Mean length (BUSCO)
Gene length (bp)
0
1000
2000
3000
4000
5000


100
80
60
40
% complete: genome assembly
40
60
80
100
% complete: gene set
BUSCOs
CEGs
BUSCOs
Vertebrates
Arthropods
Fungi
Metazoans
Eukaryotes
CEGMA CEGs
number of genes:
500
1500
2500

Lineage	Species	Sample type	Identifier	N50 (Kbp)	BUSCOs assessment
		Gene set	ASM1507v1.22		C:85% [D:11%], F:12%, M:2.3%, n:843
	Ancylostoma ceylanicum	Transcriptome	GI:595744344 Unknown		C:16% [D:n.a.], F:38%, M:44%, n:843
	Aplysia californica	Transcriptome	GI:613602134 chemokine		C:88% [D:n.a.], F:8.1%, M:2.8%, n:843
		Transcriptome	GI:614063388 Gills		C:88% [D:n.a.], F:8.4%, M:3.5%, n:843
		Transcriptome	GI:606015213 Heart		C:77% [D:n.a.], F:12%, M:9.3%, n:843
		Transcriptome	GI:594457164 Salivary		C:41% [D:n.a.], F:23%, M:34%, n:843
	Apostichopus japonicus	Transcriptome	GI:638469663 Unknown		C:68% [D:n.a.], F:24%, M:6.9%, n:843
	Asterias amurensis	Transcriptome	GI:638532954 Unknown		C:59% [D:n.a.], F:28%, M:11%, n:843
	Bithynia siamensis goniomphalos	Transcriptome	GI:480970007 Unknown		C:57% [D:n.a.], F:24%, M:17%, n:843
	Evechinus chloroticus	Transcriptome	GI:559461775 Unknown		C:92% [D:n.a.], F:5.3%, M:2.6%, n:843
	Henricia sp. AR-2014	Transcriptome	GI:638872012 Unknown		C:90% [D:n.a.], F:7.9%, M:1.1%, n:843
	Patiria miniata	Transcriptome	GI:638728087 Ovary		C:88% [D:n.a.], F:10%, M:1.1%, n:843
	Patiria pectinifera	Transcriptome	GI:638651248 Unknown		C:80% [D:n.a.], F:18%, M:1.6%, n:843
	Procotyla flyviatilis	Transcriptome	GI:528026207 Unknown		C:54% [D:n.a.], F:18%, M:26%, n:843
	Lottia gigantea	Genome	GCA_00032785.1	1,870	C:89% [D:2.3%], F:4.3%, M:5.8%, n:843
		Gene set	GCA_00032785.1.22		C:90% [D:13%], F:7.8%, M:2.1%, n:843
	Nematostella vectensis	Genome	GCA_000209225.1	472	C:78% [D:3.5%], F:10%, M:10%, n:843
		Gene set	GCA_000209225.1.22		C:83% [D:15%], F:14%, M:2.8%, n:843
	Schistosoma mansoni	Genome	GCA_000237925.2	34,464	C:56% [D:4.3%], F:8.3%, M:34%, n:843
		Gene set	ASM2379v2.22		C:65% [D:7.8%], F:8.3%, M:26%, n:843
	Strongylocentrotus purpuratus	Genome	GCA_000002235.2	167	C:87% [D:6.5%], F:7.8%, M:4.9%, n:843
		Gene set	GCA_000002235.2.22		C:83% [D:19%], F:15%, M:0.7%, n:843
	Trichoplax adhaerens	Genome	GCA_000150275.1	5,978	C:81% [D:1.1%], F:7.8%, M:10%, n:843

Lineage	Species	Sample type	Identifier	N50 (Kbp)	BUSCOs assessment
Vertebrates	*Homo sapiens*	Genome	GCA_000001405.15	67,794	C:89% [D:1.5%], F:6.0%, M:4.5%, n:3023
		Gene set	GRCh37.75		C:99% [D:1.7%], F:0.0%, M:0.0%, n:3023
	Mus musculus	Genome	GCA_000001635.4	52,589	C:78% [D:3.0%], F:19%, M:2.5%, n:3023
		Gene set	GRCm38.75		C:99% [D:2.5%], F:99%, M:0.1%, n:3023
	Ornithorhyncus anatinus	Genome	GCF_000002275.2	991	C:55% [D:0.8%], F:25%, M:18%, n:3023
		Gene set	OANA5.75		C:72% [D:1.1%], F:19%, M:8.2%, n:3023
	Callithrix jacchus	Gene set	C_jacchus3.2.1.75		C:97% [D:2.9%], F:1.7%, M:0.8%, n:3023
		Transcriptome	GI:532219616 Bladder		C:76% [D:17%], F:5.5%, M:18%, n:3023
		Transcriptome	GI:532292355 hypocampus		C:79% [D:18%], F:4.5%, M:15%, n:3023
		Transcriptome	GI:532349506 Cortex		C:34% [D:7.6%], F:34%, M:64%, n:3023
		Transcriptome	GI:532452938 S. muscle		C:69% [D:13%], F:6.0%, M:24%, n:3023
		Transcriptome	GI:532524775 Cerebellum		C:76% [D:19%], F:5.1%, M:18%, n:3023
	Pan troglodytes	Gene set	CHIMP2.14.75		C:96% [D:0.5%], F:1.2%, M:1.9%, n:3023
		Transcriptome	GI:410228237adipose SC		C:75% [D:15%], F:3.8%, M:20%, n:3023
		Transcriptome	GI:410308999 Fibroblast		C:75% [D:16%], F:3.7%, M:21%, n:3023

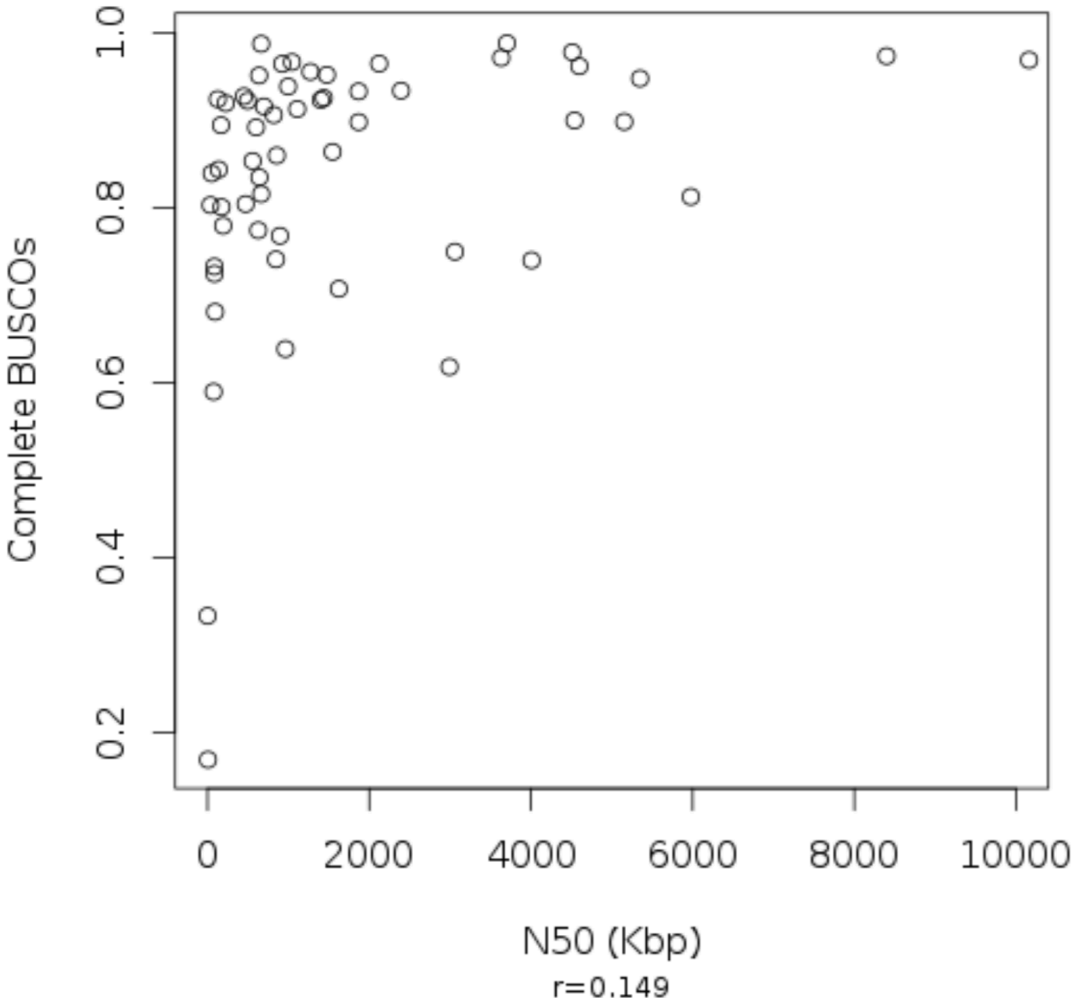
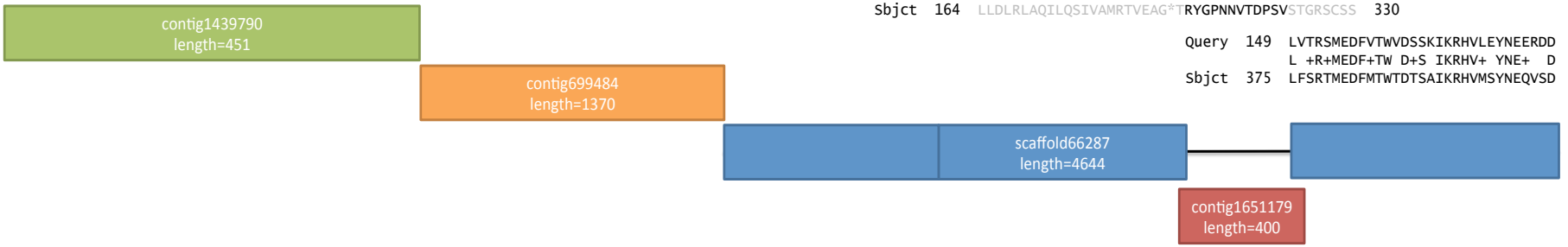


Figure S2. BUSCO completeness versus N50 contiguity. Nine outliers with N50 values above 10'000 Kbp are not shown, each of which achieve more than 90% BUSCO completeness.

>Hs8922794___KOG4655 Length=184 (IMP3, U3 small nucleolar ribonucleoprotein) Coverage: 178/184 (96.7%)
MVRKLKFHEQKLLKQVDFLNWEVTDHNLHELRVLRRYRLQRREDYTRYNQLSRAVRELARRLRDLPERDQFRVRASAALLDKLYALGLVPTRGSLELCDFVTASSFCRRRLPTVLLKLRMAQHLQAAVAFVEQGHVRVGPDVVTDPAFLVTRSMEDFVTWVDSSKIKRHVLEYNEERDDFDLEA

```
 2 VRKLKFHEQKLLKQVDFLNWEVTDHNLHE]RVLRRYRLQRREDYTRYNQLSRAVRELARR  61
   +RKLK HE KLLK+VDF+NW+ +++NL+ ++++R++R+++RED T       +VR    +
95 MRKLKHHESKLLKKVDFINWK-SENNLYFVKLMRKFRIEKREDLTL*VYCPVSVR--CCQ  265

                    Query  45    YTRYNQLSRAVRELARRLRDLPERDQFRVRASAALLDKLYALGLV  89
                           + RYN+ ++ + +LA+++++LP  D FR  A+A LL+KL A+ +V
                    Sbjct  1147  FCRYNKYTKKILDLAKKIKELPAEDPFRTVATAQLLEKL*AITVV  1013

                                                Query  84    YALGLVPTRGSLELCDFVTASSFCR  108
                                                       YA+GL+PT+ +LEL   V+A++FCR
                                                Sbjct  4270  YAMGLIPTKKNLELALKVSAAAFCR  4196

                                                        Query  93    GSLELCDFVTASSFCRRRLPTVLLKLRMAQHLQAAVAFVEQGHV  136
                                                               G +  C     + F RRRLP V++K  MAQ ++AAV FVEQG +
                                                        Sbjct  1911  G*ITCCCHCRTACF-RRRLPVVMVKSHMAQTVKAAVTFVEQGRI  1783

                                                                Query  115  LLKLRMAQHLQAAVAF--VEQGHVRVGPDVVTDPAFLVTRSMED  156
                                                                       LL LR+AQ LQ+ VA    VE G  R GP+ VTDP+     RS
                                                                Sbjct  164  LLDLRLAQILQSIVAMRTVEAG*TRYGPNNVTDPSVSTGRSCSS  330

                                                                                Query  149  LVTRSMEDFVTWVDSSKIKRHVLEYNEERDD  179
                                                                                       L +R+MEDF+TW D+S IKRHV+ YNE+  D
                                                                                Sbjct  375  LFSRTMEDFMTWTDTSAIKRHVMSYNEQVSD  283
```

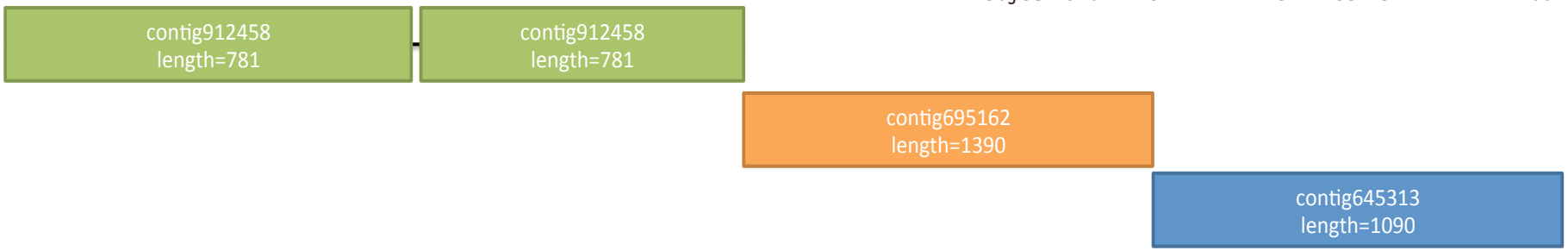


>Hs7706337___KOG3343 (COPZ1, coatomer protein complex, subunit zeta 1) Coverage: 171/177 (96.6%)
MEALILEPSLYTVKAILILDNDGDRLFAKYYDDTYPSVKEQKAFEKNIFNKTHRTDSEIALLEGLTVVYKSSIDLYFYVIGSSYENELMLMAVLNCLFDSLSQMLRKNVEKRALLENMEGLFLAVDEIVDGGVILESDPQQVVHRVALRGEDVPLTEQTVSQVLQSAKEQIKWSLLR

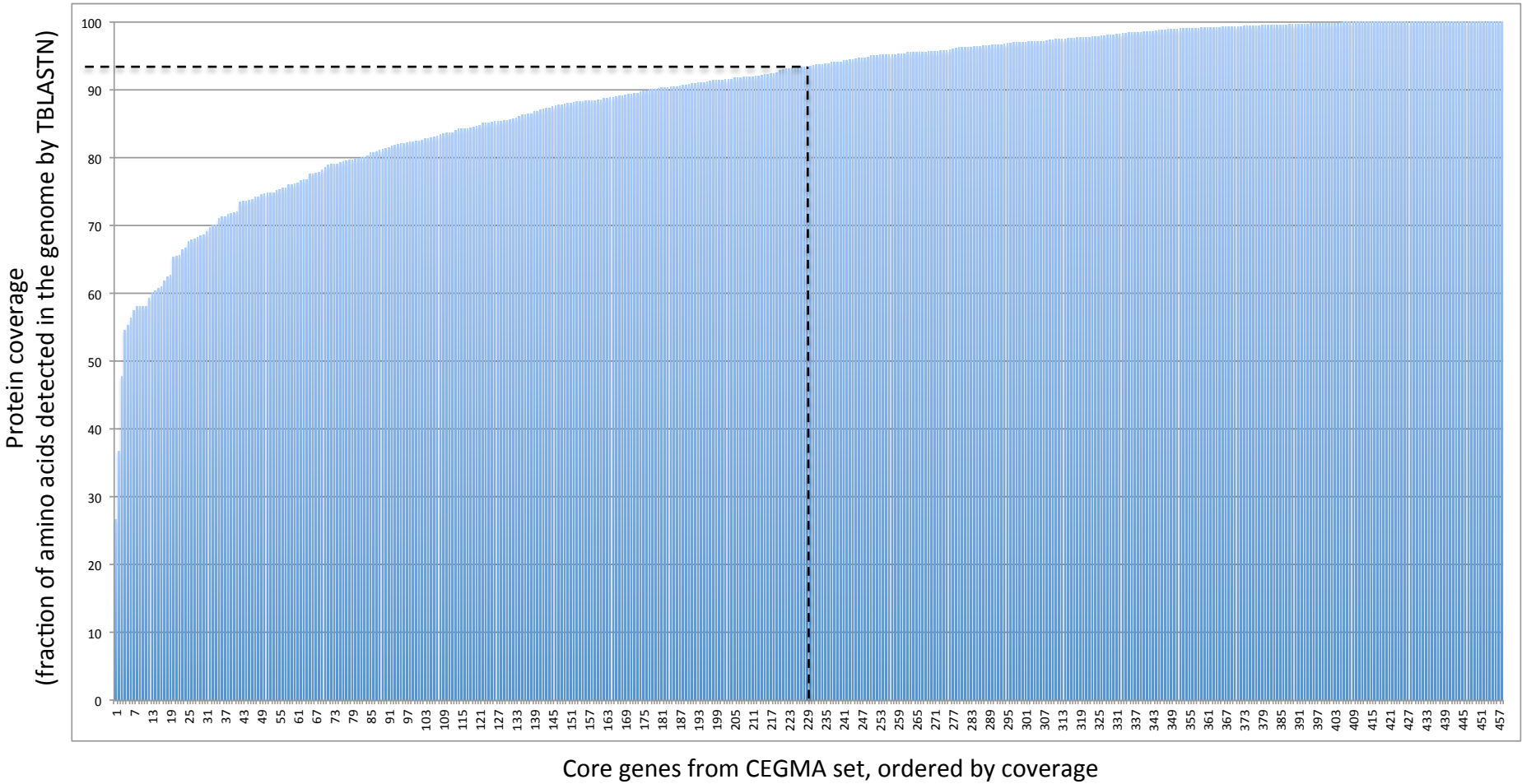
```
    7 EPSLYTVKAILILDNDGDRLFAK  29
      EP+LYT+KA+ ILDNDG+RL  K
 4040 EPTLYTIKALAILDNDGNRLLTK  4108

             Query  29    KYYDDTYPSVKEQKAFEKNIFNKTHRTDSE-----IALLEGLTVVYKSSIDLYFYVI  80
                          +YYDDT+P+VKEQKAFEKN+FNKTHR +       + L   L   + +S+D++ Y++
             Sbjct  6309  QYYDDTFPTVKEQKAFEKNLFNKTHRANGM*FNFIVCLFITL*CTW-ASLDVHLYIV  6476

                                  Query  56    DSEIALLEGLTVVYKSSIDLYFYVIGSSYENELM  89
                                               +EI + EGLT VYKS++DL+FYV+GSS ENE++
                                  Sbjct  7078  SAEIIMFEGLTCVYKSNVDLFFYVVGSSNENEVI  7179

                                                       Query  87    ELMLMAVLNCLFDSLSQMLRKNVEKRALLENMEGLFLAVDEIVDGG  132
                                                                    +L+L +VLN  +DS+SQ+LRKNVEKRALL+NM+ +FLA DEI DGG
                                                       Sbjct  8820  QLILASVLNAFYDSISQILRKNVEKRALLDNMDAVFLAADEICDGG  8957

                                                                            Query  133   VILESDPQQVVHRVALRGEDVPLTEQTVSQVL  164
                                                                                          ++LE+DP  VV +VA+R ED+PL EQTV+Q++
                                                                            Sbjct  9485  ILLEADPNAVVQKVAIRNEDIPLGEQTVAQVM  9580

                                                                                                  Query  159   TVSQVLQSAKEQIKWSLLR  177
                                                                                                               +QVLQSAKEQIKWSLL+
                                                                                                  Sbjct  1383  CATQVLQSAKEQIKWSLLK  1439
```

>Hs4507761___KOG0003 (UBA52, ubiquitin A-52 residue ribosomal protein fusion product 1) Coverage: 128/128 (100%)
MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGIIEPSLRQLAQKYNCDKMICRKCYARLHPRAVNCRKKKCGHTNNLRPKKKVK

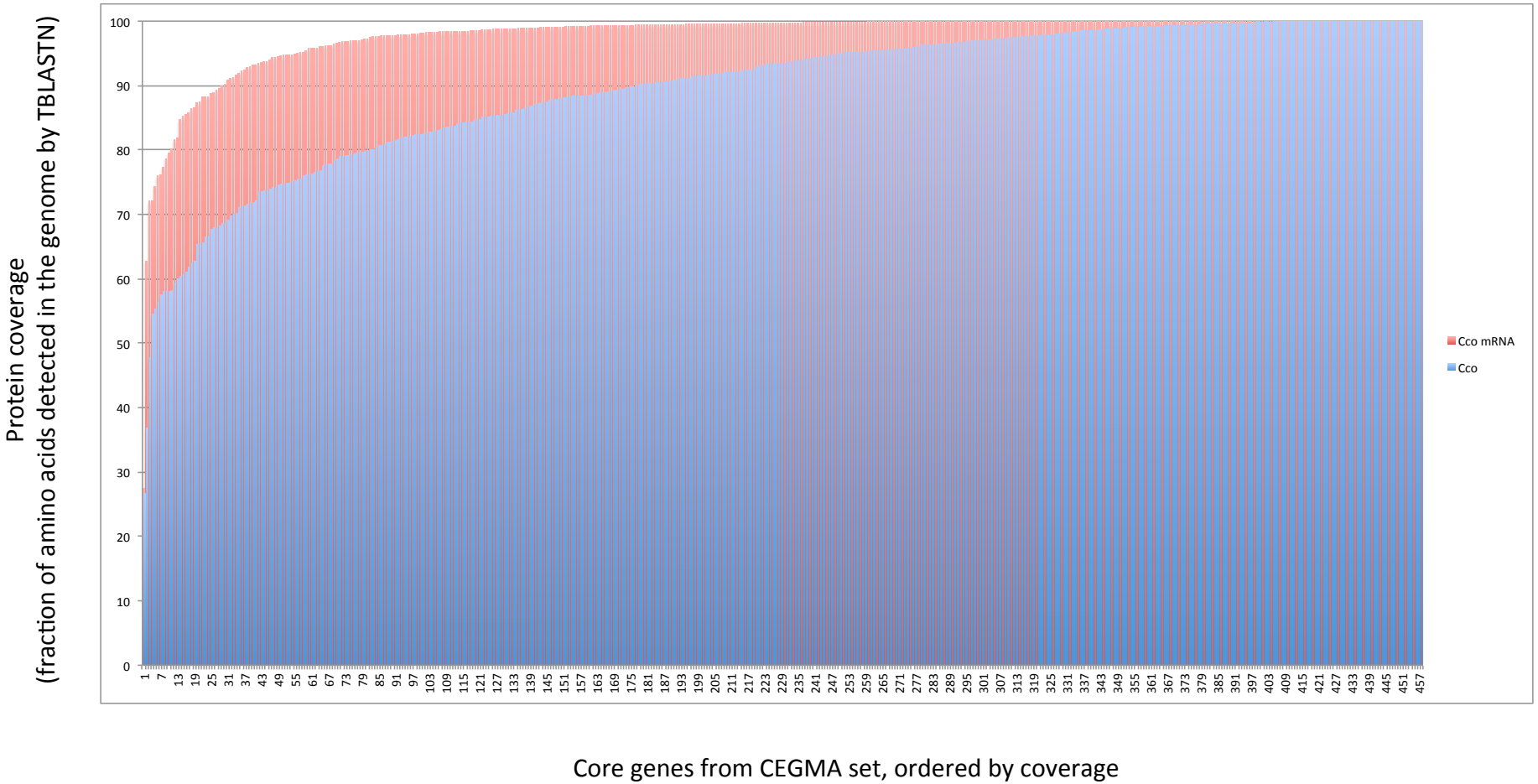
```
  1 MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEG  35
    MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEG
 11 MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEG  115

                      Query  35   GIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHL  69
                                   GIPPDQQRLIFAGKQLEDGRTLSDYNIQK   + L
                      Sbjct  603  GIPPDQQRLIFAGKQLEDGRTLSDYNIQKGERVSL  707

                                                Query  64   ESTLHLVLRLRGGIIEPSLRQLAQKYNCDKMICRK  98
                                                             ESTLHLVLRLRGGIIEPSLR LA KYNCDKMICRK
                                                Sbjct  312  ESTLHLVLRLRGGIIEPSLRILASKYNCDKMICRK  416

                                                                          Query  98    KCYARLHPRAVNCRKKKCGHTNNLRPKKKVK  128
                                                                                       +CYARLHPRA NCRK+KCGHT+N+RPKKK+K
                                                                          Sbjct  970   RCYARLHPRATNCRKRKCGHTSNIRPKKKLK  1062
```



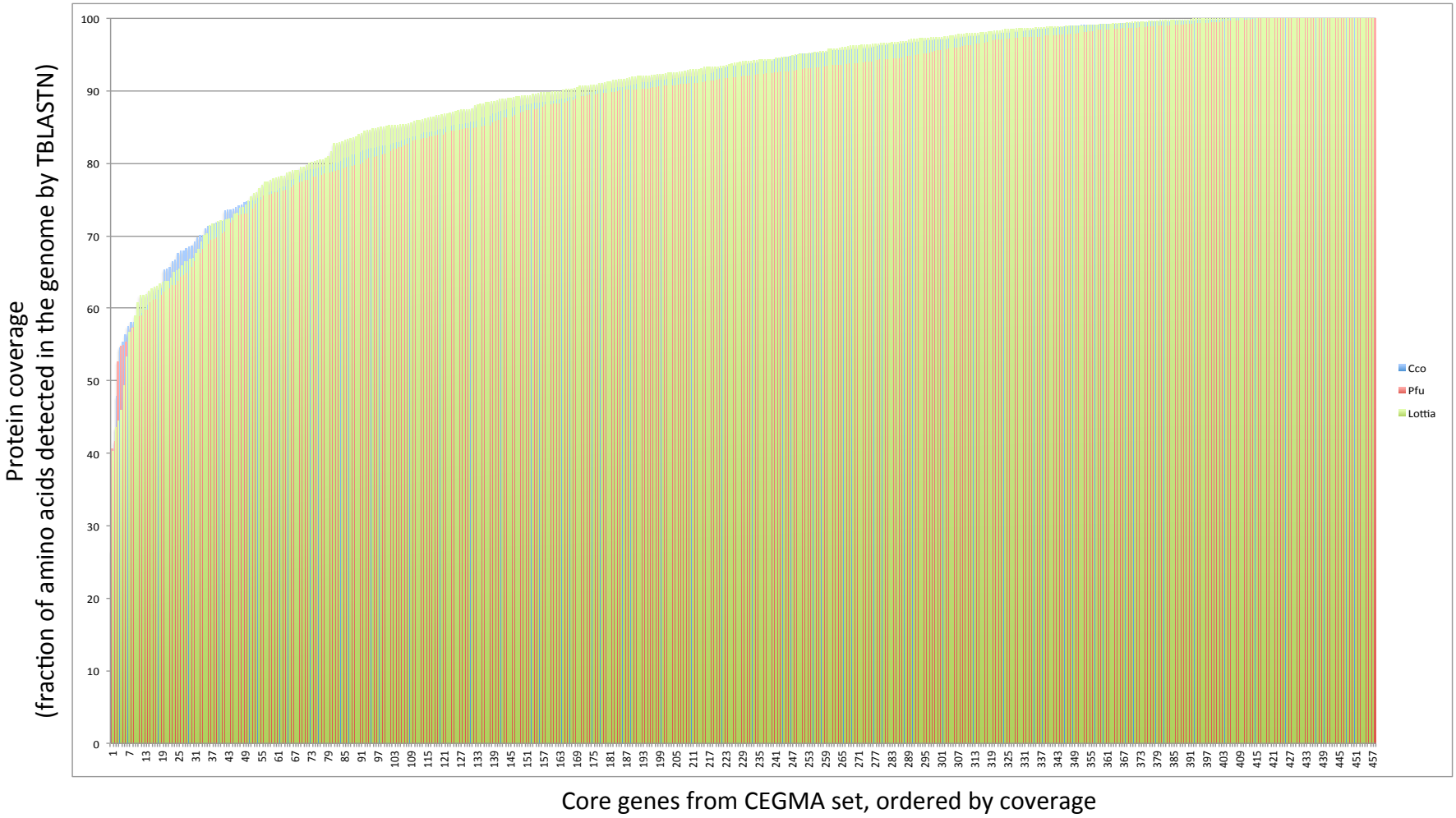
CEGMA tuumikgeenide mediaankatvus genoomis on 93% nukleotiididest



Transkriptoomis on tuumikgeenide katvus veelgi parem (99.7%)



Teiste molluskite genoomis on tuumikgeenide katvus väga sarnane



Kokkuvõte

- CEGMA 458 geeni on igati mõistlik andmestik geenisisalduise hindamisel

- Kuna CEGMA't enam ei toetata ega arendata, siis võiks testida BUSCO't

- Hea oleks vabaneda geenide ennustamise etapist, see on aeglane ja pole piisavalt robustne.