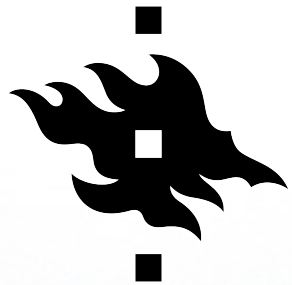
Targeted assembly of RNA-seq data using phylogenetic information

Alan Medlar and Ari Löytynoja

Institute of Biotechnology, University of Helsinki
ari.loytynoja@helsinki.fi

9 January, 2014 / AllBio, Espoo, Finland



UNIVERSITY OF HELSINKI

joint-work with Alan Medlar

UNIVERSITY OF HELSINKI

joint-work with Alan Medlar

(currently on leave)

UNIVERSITY OF HELSINKI

Outline

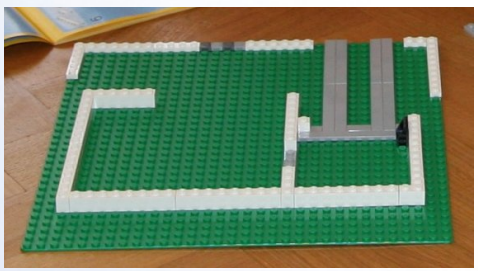
1. Why?
2. How?
3. What then?

UNIVERSITY OF HELSINKI

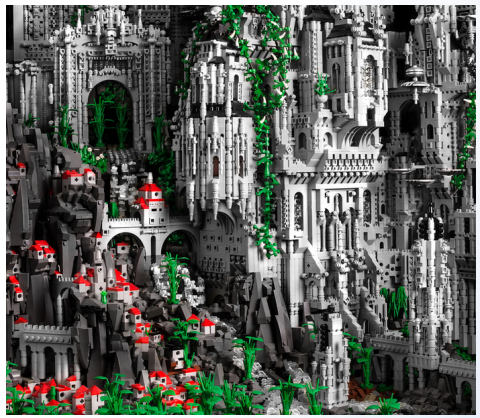
Outline

1. Why?
2. How?
3. What then?

Work in progress – few details to finish!



some work



UNIVERSITY OF HELSINKI

The role of transcriptome (1)

Quantitative RNA-seq analyses

- read mapping requires genome or transcriptome
- when not available, transcriptome can be inferred from data
- de novo transcriptome represents input data
  - sequencing coverage highly unequal
  - lowly expressed and silent transcripts not present

De novo transcriptome assembly

- inferred transcripts often fragmented
- chimeric contigs (e.g. among paralogues) problematic
- contig functions unknown
  - annotated using similarity to known sequences

The role of transcriptome (2)

RNA-seq data not always for quantitative analyses

- read data mainly from protein-coding genes
  - fraction of genome, majority(?) of information
  - complex sequence, relatively easy to assemble

low-cost alternative to whole genome sequencing

- useful for gene-centric studies
  - gene content, gene family evolution
  - phylogenetic analyses, inference of selection

Problems of de novo assembly persist

- inferred transcripts often fragmented
- chimeric contigs (e.g. among paralogues) problematic
- contig functions unknown

UNIVERSITY OF HELSINKI

The role of transcriptome (2)

RNA-seq data not always for quantitative analyses

- read data mainly from protein-coding genes
  - fraction of genome, majority(?) of information
  - complex sequence, relatively easy to assemble

  ➡ low-cost alternative to whole genome sequencing
- useful for gene-centric studies
  - gene content, gene family evolution
  - phylogenetic analyses, inference of selection

Problems of de novo assembly persist

- inferred transcripts often fragmented
- chimeric contigs (e.g. among paralogues) problematic
- contig functions unknown

UNIVERSITY OF HELSINKI

RNA-seq in phylogenetic analysis

Pioneered by Antonis Rokas et al. (2010, PNAS 107, 1476–1481)

Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life

Chris Todd Hittinger[a,b], Mark Johnston[a,b], John T. Tossberg[c], and Antonis Rokas[c,1]

UNIVERSITY OF HELSINKI

RNA-seq in phylogenetic analysis

Pioneered by Antonis Rokas et al. (2010, PNAS 107, 1476–1481)

Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life

Chris Todd Hittinger[a,b], Mark Johnston[a,b], John T. Tossberg[c], and Antonis Rokas[c,1]

Table 1. Ortholog number, contig number, and ortholog assignment accuracy for data matrices constructed from ≥100 and ≥300 bp contigs using the single-contig and supercontig strategies

	Single-contig strategy				Supercontig strategy					
	≥100-bp contigs		≥300-bp contigs		≥100-bp contigs			≥300-bp contigs		
Data set	Orthologs*	Accuracy,† %	Orthologs	Accuracy, %	Orthologs	Contigs‡	Accuracy, %	Orthologs	Contigs	Accuracy, %
100,000	0	NA	0	NA	4	3	33	0	0	NA
250,000	1	0	0	NA	50	34	79	2	2	50
500,000	11	91	0	NA	124	128	86	10	6	67
1,000,000	72	96	0	NA	226	287	85	64	36	86
2,000,000	120	94	8	75	430	550	84	148	86	86
3,000,000	173	93	29	93	630	825	82	198	146	88
4,000,000	212	93	36	92	850	1,152	82	255	183	87
5,000,000	252	93	40	95	1,054	1,449	83	302	222	86
~6,500,000	333	93	37	97	1,591	1,872	84	445	290	87
~13,000,000	553	95 (89)§	69	99 (94)§	2,661	4,118	85	725	523	86

*No. of *A. aegypti* orthologs in data matrix.
†Percentage of *A. gambiae* contigs accurately assigned to their *A. aegypti* reference transcript orthologs in the data matrix.
‡No. of *A. gambiae* contigs assigned to *A. aegypti* reference transcripts in the data matrix.
§Percentage of accurately inferred orthologs using a phylogeny-based assessment of orthology assignment.

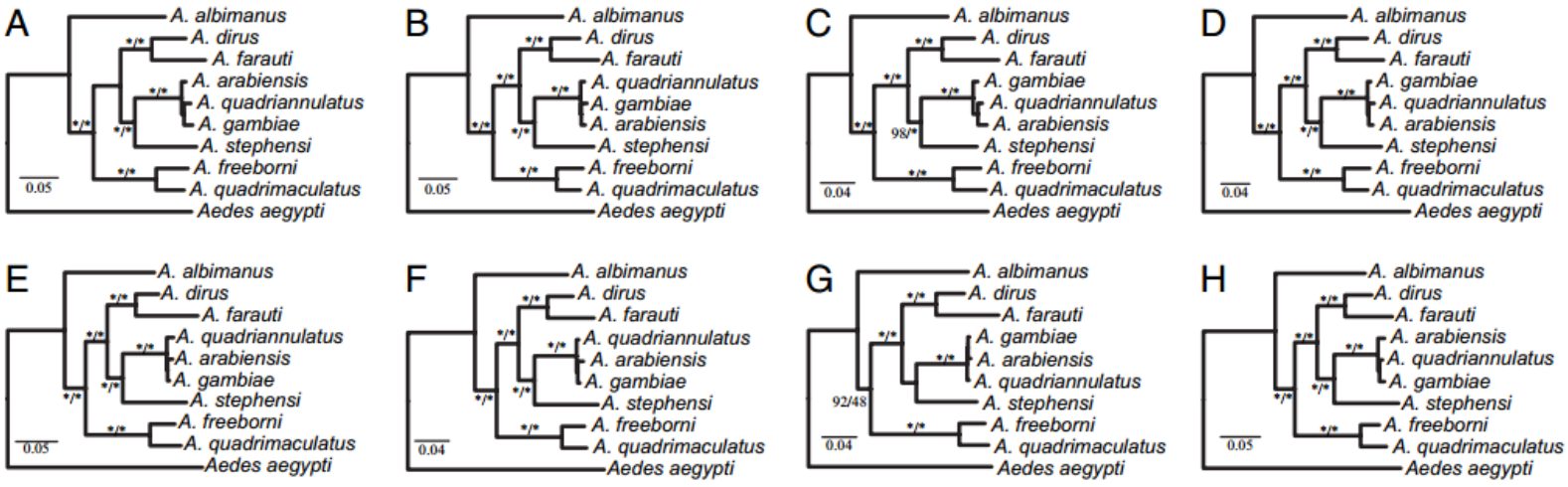
UNIVERSITY OF HELSINKI

RNA-seq in phylogenetic analysis

Pioneered by Antonis Rokas et al. (2010, PNAS 107, 1476–1481)

Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life

Chris Todd Hittinger[a,b], Mark Johnston[a,b], John T. Tossberg[c], and Antonis Rokas[c,1]



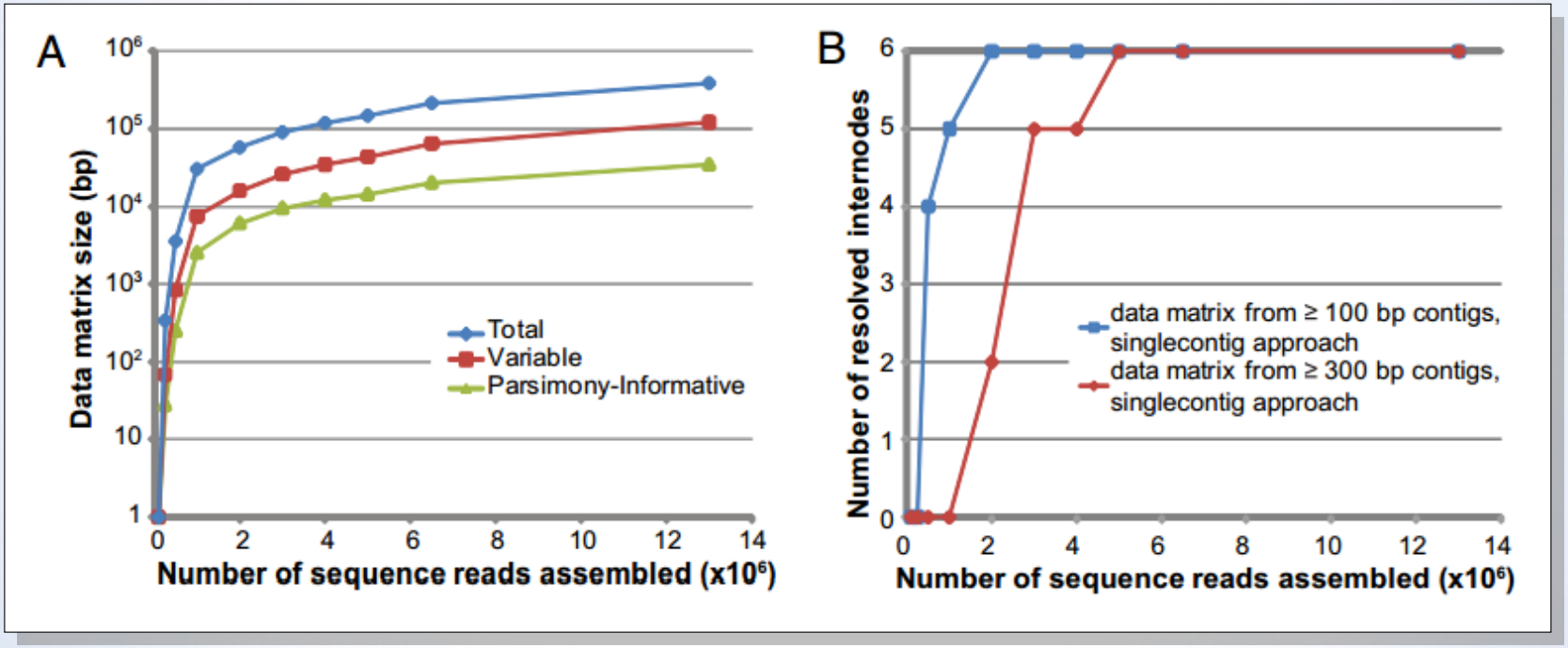
UNIVERSITY OF HELSINKI

RNA-seq in phylogenetic analysis

Pioneered by Antonis Rokas et al. (2010, PNAS 107, 1476–1481)

Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life

Chris Todd Hittinger[a,b], Mark Johnston[a,b], John T. Tossberg[c], and Antonis Rokas[c,1]



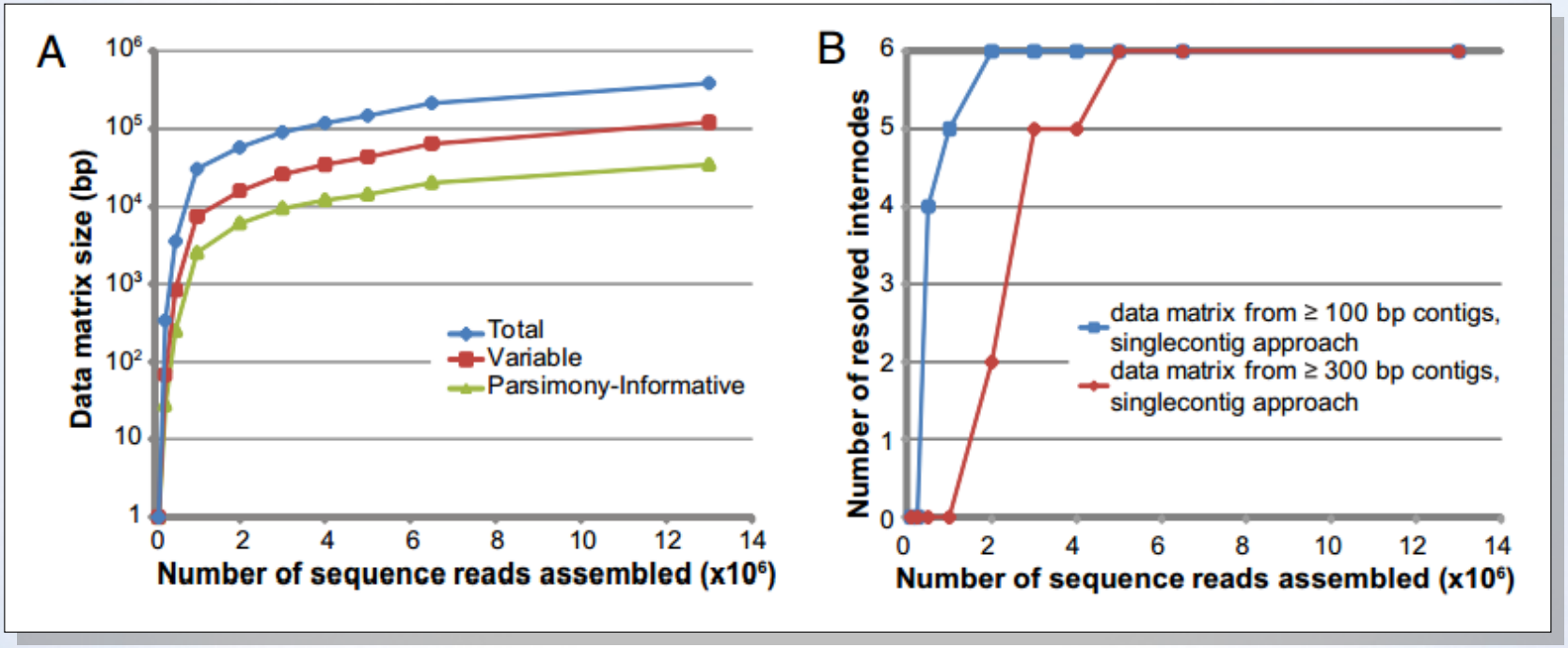
UNIVERSITY OF HELSINKI

RNA-seq in phylogenetic analysis

Pioneered by Antonis Rokas et al. (2010, PNAS 107, 1476–1481)

Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life

Chris Todd Hittinger[a,b], Mark Johnston[a,b], John T. Tossberg[c], and Antonis Rokas[c,1]



With multiplexing, sequencing costs are minimal

2008: \$50/species; 2011: \$10/species (0.5M reads)

UNIVERSITY OF HELSINKI

RNA-seq in exploratory studies

http://www.onekp.com/

The 1KP Project

Links

Home

What is the 1KP Project?

Why Sequence 1000 plants?

Transcriptomes not Genomes

The 1000 plants (oneKP or 1KP) initiative is a public-private partnership generating large scale gene sequence information for 1000 different species of plants. Major supporters include Alberta's Department of Advanced Education and Technology (AET), Silicon Valley based Musea Ventures, Beijing Genomics Institute in Shenzhen, University of Alberta, and Alberta's Informatics Circle of Research Excellence (iCORE).

Subproject Catagories

UNIVERSITY OF HELSINKI

RNA-seq in exploratory studies

http://www.onekp.com/

The 1KP Project

Links

Home

What is the 1KP Project?

Why Sequence 10 plants?

Transcriptomes nc Genomes

Why Do Only Transcriptomes?

Sequencing transcripts (*i.e.* expressed genes) is inherently cheaper than sequencing genomes, because it obviates the need to sequence the intronic and intergenic regions, which can be orders of magnitude larger. Obviously one can never get all the genes just by doing transcripts, and it is not our intention to argue for one or the other, since (a) the pros and cons were laid out two decades ago for the human genome project, and (b) BGI-Shenzhen is also sequencing *de novo* genomes like the giant panda with the new technology. The thing is once you get away from the few dozen obviously important plant species, almost none of the roughly half million plant species known to humanity has been touched by genomics at any level.

UNIVERSITY OF HELSINKI

RNA-seq in exploratory studies

FishT1K

华大基因 BGI

400-706-6615 Online Customer Service 中文 search

Home | Science | Industry | Platform | Education | About BGI

News

- BGI News
- Academic Trend
- Research Collaboration
- Training

News -> Academic Trend

3

The China National Genebank Initiates Global Collaboration to Sequence Transcriptomes of 1,000 Fish Species

Publish Date:2013-11-19

November 19, 2013, Shenzhen, China-The China National Genebank (CNGB) announces the official launch of the 1,000 Fish Transcriptome Project (Fish T1K), a groundbreaking transcriptome study designed to unveil the mysteries of the origin, evolution, and diversification of the largest group of vertebrates. Such findings could enable scientists to pursue innovative approaches and strategies to address challenges in fish breeding, disease control and prevention, seafood safety, and biodiversity conservation.

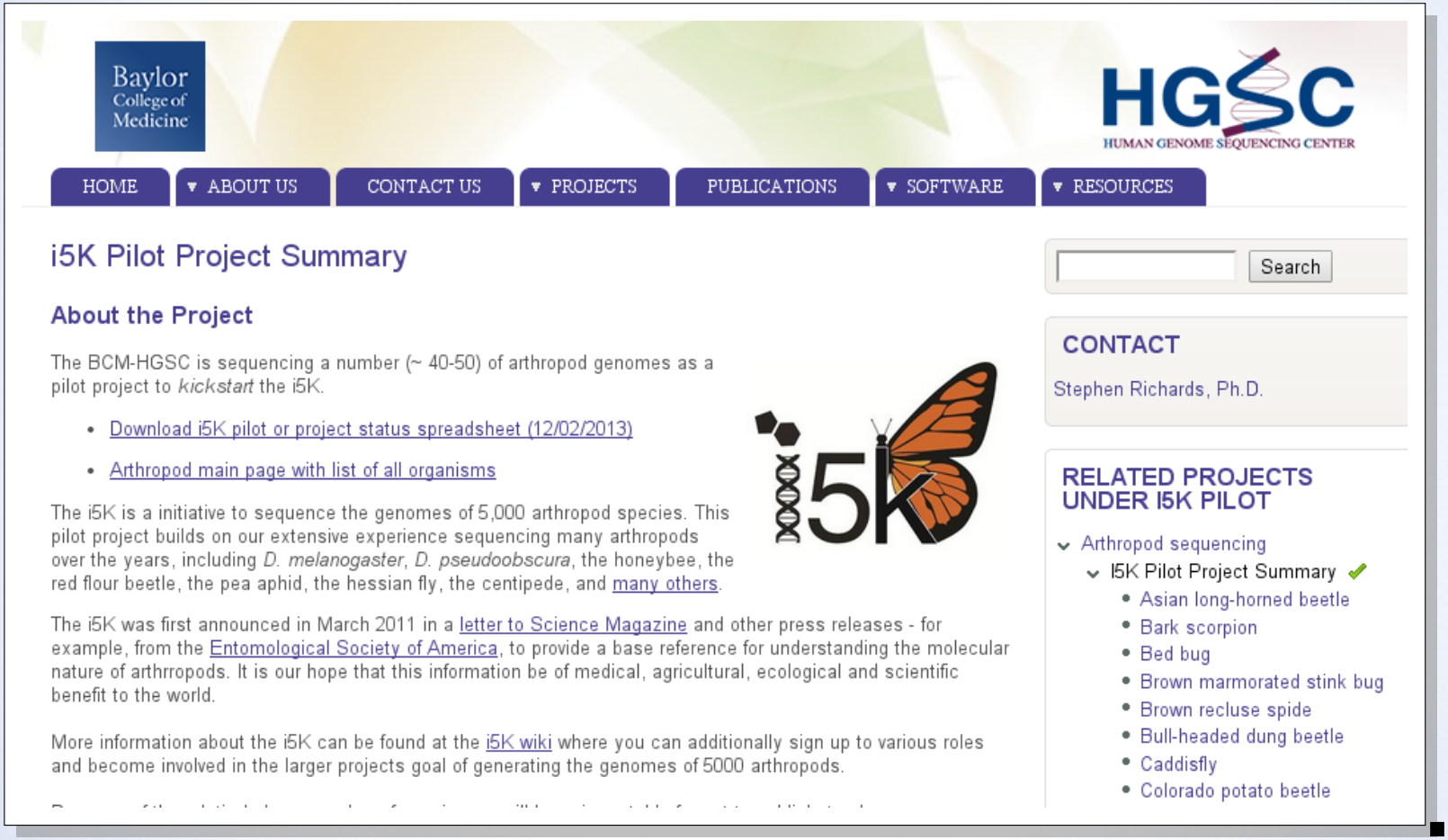
Fish T1K is unique relative to other fish transcriptome projects given its large scale and multi-disciplinary, integrative perspective. All data generated from Fish T1K will be made available publicly through CNGB, ensuring that scientists may better grasp the new developments and trends in fish research and the use of RNA-seq technology.

With over 32,000 species, fishes are the largest and most diverse group of living vertebrates. They are also an economically important group of animals. Remarkably, there are only about 10 fish genomes sequenced to date. "The lack of transcriptome data for the majority of fish species motivates us to establish a large-scale transcriptome database for fish." said Dr. Yong Zhang, Director of CNGB-Shenzhen.

UNIVERSITY OF HELSINKI

RNA-seq in exploratory studies

i5K(pilot)



UNIVERSITY OF HELSINKI

My interest in the topic

Background in methods development, sequence alignment

- a new method, PAGAN, suitable for contig analysis

Sequence analysis

Advance Access publication April 23, 2012

Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm

Ari Löytynoja[1,2,*], Albert J. Vilella[1] and Nick Goldman[1]

[1]EMBL-European Bioinformatics Institute, Hinxton, CB10 1SD, UK and [2]Institute of Biotechnology, 00014 University of Helsinki, Finland

UNIVERSITY OF HELSINKI

My interest in the topic

Background in methods development, sequence alignment

- a new method, PAGAN, suitable for contig analysis

Sequence analysis

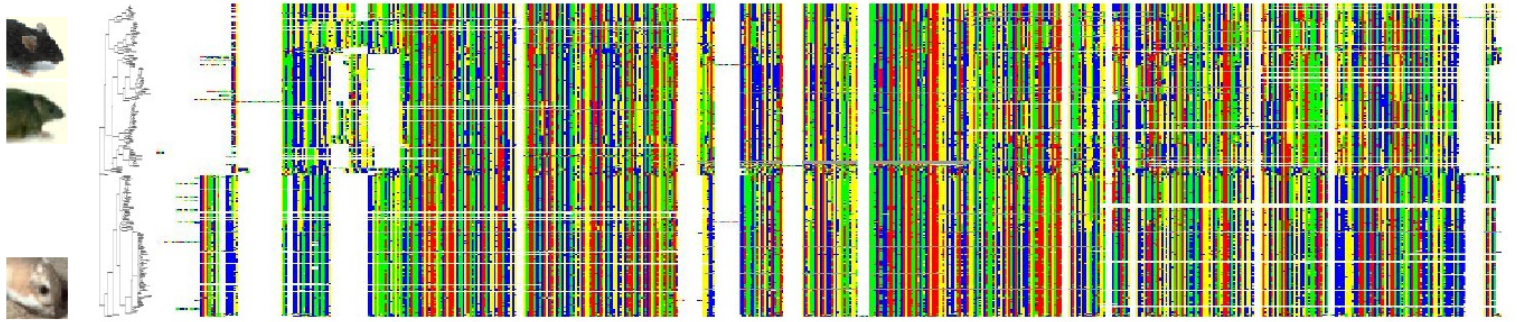
Advance Access publication April 23, 2012

Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm

Ari Löytynoja[1,2,*], Albert J. Vilella[1] and Nick Goldman[1]

[1]EMBL-European Bioinformatics Institute, Hinxton, CB10 1SD, UK and [2]Institute of Biotechnology, 00014 University of Helsinki, Finland

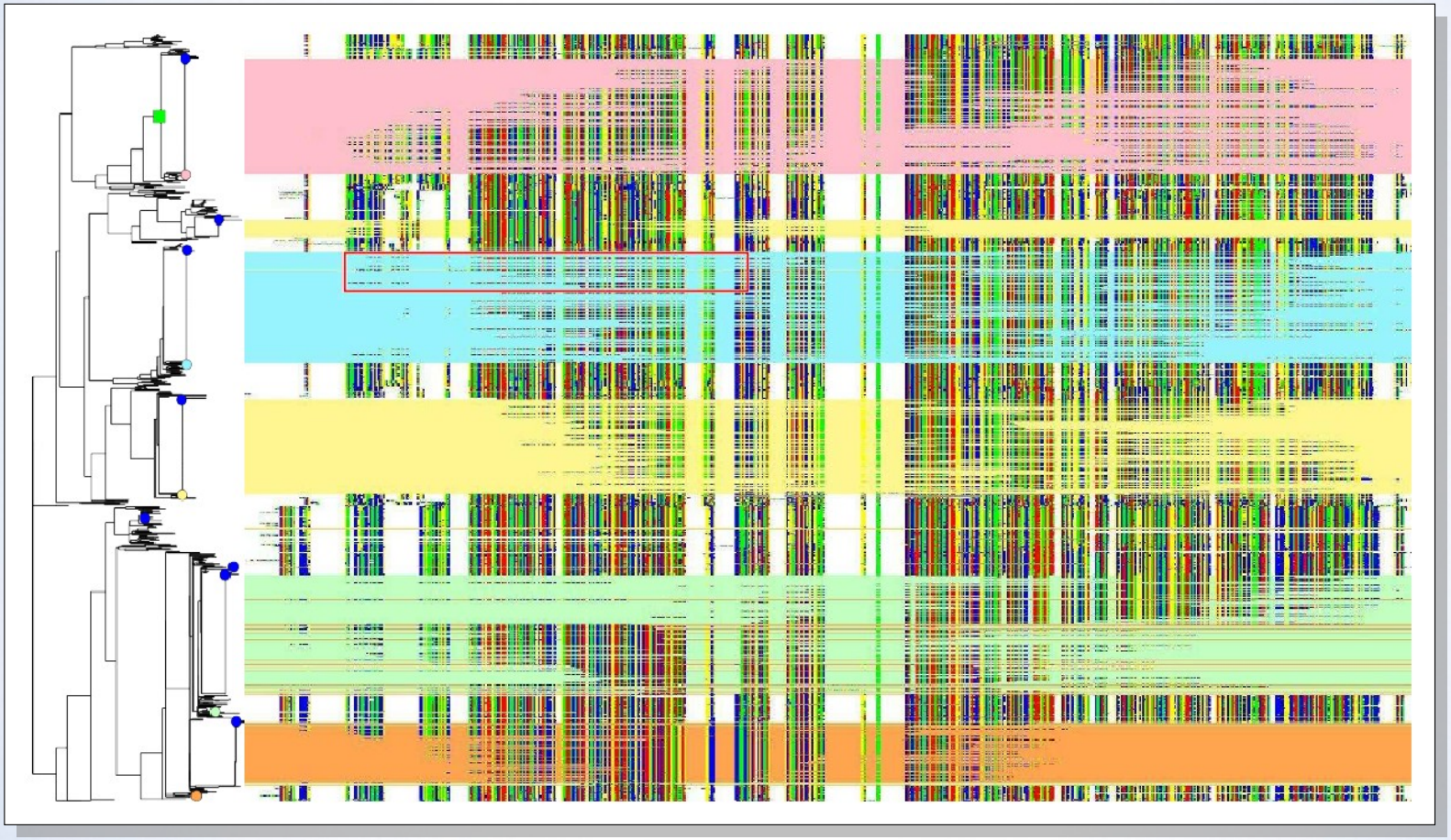
e!Ensembl



UNIVERSITY OF HELSINKI

My interest in the topic

Background in methods development, sequence alignment



- reference alignment extended with sequence fragments

UNIVERSITY OF HELSINKI

My continued interest in the topic

PAGAN works in theory, not in practice

- \+ extends gene alignments with sequence fragments
- – handles thousands of sequence fragments, not billions
- – does not provide gene alignments
- – does not assign sequence fragments to specific alignment

- \+ reads can be pre-assembled
- \+ Ensembl provides curated, open data
- \+ matching contigs with sequences is trivial

Need a tool to combine PAGAN with other tools & Ensembl

UNIVERSITY OF HELSINKI

My continued interest in the topic

PAGAN works in theory, not in practice

- \+ extends gene alignments with sequence fragments
- – handles thousands of sequence fragments, not billions
- – does not provide gene alignments
- – does not assign sequence fragments to specific alignment

- \+ reads can be pre-assembled
- \+ Ensembl provides curated, open data
- \+ matching contigs with sequences is trivial

Need a tool to combine PAGAN with other tools & Ensembl

UNIVERSITY OF HELSINKI

New approach: Glutton

Glutton analysis package

- easy-to-use, robust Python program to automate reference-based analysis of RNA-seq data

No re-inventing the wheel

- uses Ensembl, PRANK, BLAST, PAGAN, Python libraries

Aims to be the tool between Trinity and the downstream analyses

- input: assembled RNA-seq data
- output: scaffolded, annotated contigs aligned against reference sequences / alignments

Work in progress: features still incomplete or missing!

Glutton workflow

(1) get stuff from Ensembl:

```
glutton build --species Mouse --release 71
```

(2) align contigs to reference:

```
glutton align --species Mouse --contigs contigs.fasta
 --alignments mouse_alignments
```

(3) scaffold contigs:

```
glutton scaffold --contigs contigs.fasta
  --alignments mouse_alignments --scaffolds mouse_73_scaffolds.fasta
```

(4) other features coming soon!

UNIVERSITY OF HELSINKI

Constraint: protein-coding data only

We (currently) focus on protein-coding contigs only

- assume significant ORFs (Illumina or SOLiD data!)
- align nucleotide sequences as protein translations

UNIVERSITY OF HELSINKI

Constraint: protein-coding data only

We (currently) focus on protein-coding contigs only

- assume significant ORFs (Illumina or SOLiD data!)
- align nucleotide sequences as protein translations

Pipeline:

(1) all contigs (nuc) —`blastx`→ reference genes (pep)

set of contigs per reference gene

UNIVERSITY OF HELSINKI

Constraint: protein-coding data only

We (currently) focus on protein-coding contigs only

- assume significant ORFs (Illumina or SOLiD data!)
- align nucleotide sequences as protein translations

Pipeline:

(1) all contigs (nuc) —`blastx`→ reference genes (pep)

set of contigs per reference gene

(2) contig set (nuc) —`PAGAN`→ reference gene (nuc)

translated alignment of matching contigs

UNIVERSITY OF HELSINKI

Constraint: protein-coding data only

We (currently) focus on protein-coding contigs only

- assume significant ORFs (Illumina or SOLiD data!)
- align nucleotide sequences as protein translations

Pipeline:

(1) all contigs (nuc) —`blastx`→ reference genes (pep)

set of contigs per reference gene

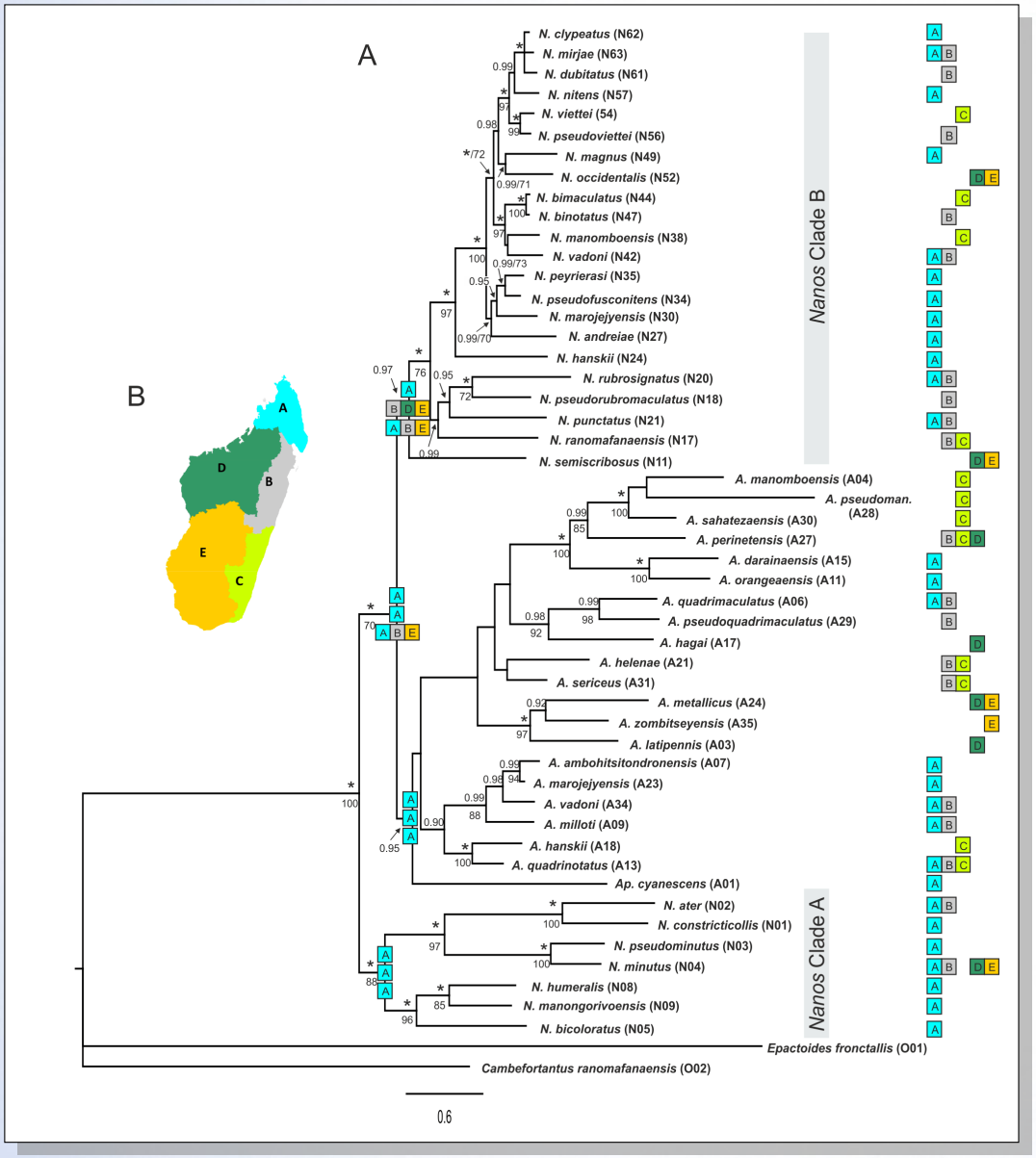
(2) contig set (nuc) —`PAGAN`→ reference gene (nuc)

translated alignment of matching contigs

(3) back-translation; scaffolding

UNIVERSITY OF HELSINKI

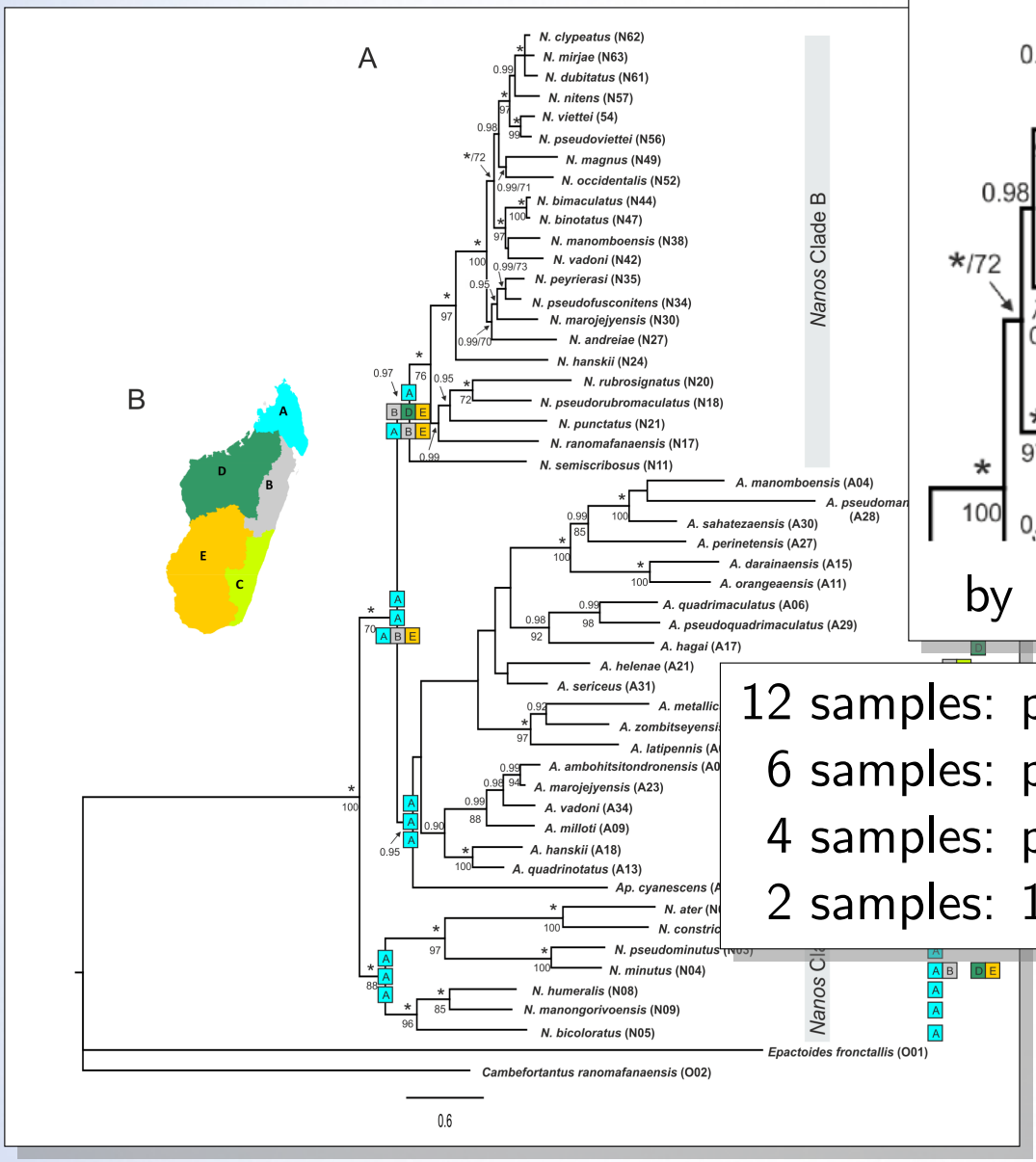
Test data: Malagasy dung beetles

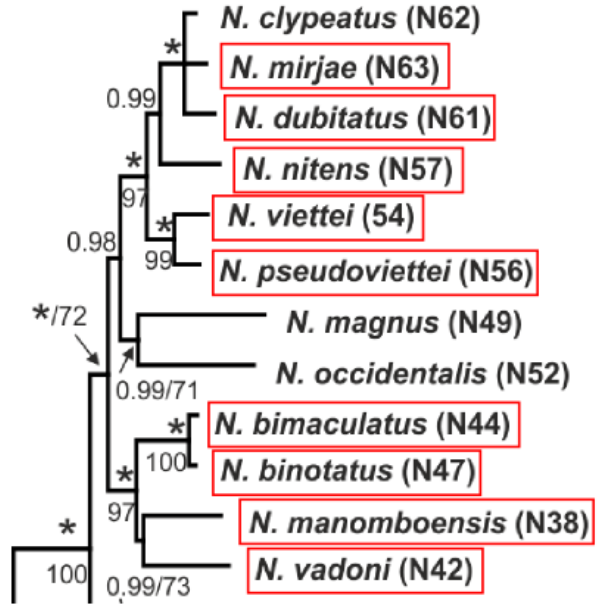


Miraldo and Hanski

UNIVERSITY OF HELSINKI

Test data: Malagas





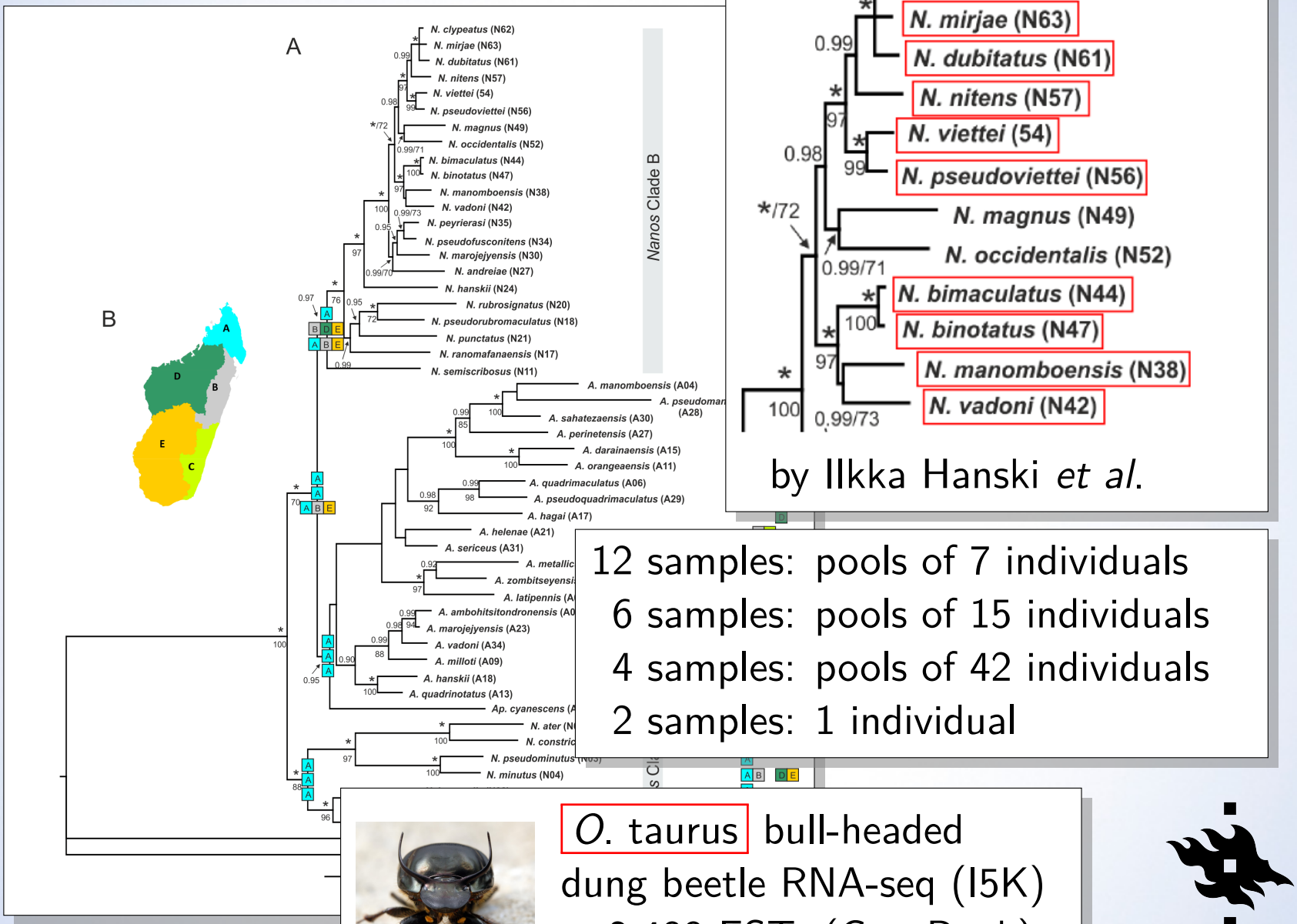
by Ilkka Hanski *et al.*

12 samples: pools of 7 individuals
6 samples: pools of 15 individuals
4 samples: pools of 42 individuals
2 samples: 1 individual

Miraldo and Hanski

UNIVERSITY OF HELSINKI

Test data: Malagas



by Ilkka Hanski *et al.*

12 samples: pools of 7 individuals
6 samples: pools of 15 individuals
4 samples: pools of 42 individuals
2 samples: 1 individual

O. taurus bull-headed dung beetle RNA-seq (I5K) + 2,488 ESTs (GeneBank)

Miraldo and Hansk

UNIVERSITY OF HELSINKI

De novo assembly

Ten data sets:

- Nanos 10–15M PE reads per sample  pooled by species
- O. taurus 212M PE reads from 3 individuals

Trinity assembly:

species	#contigs	mean length	max lengths
bimaculatus	57,839	971.4	18,666
binotatus	50,874	1,026.8	13,955
dubitatus	117,343	912.7	21,804
mamomboensis	36,561	785.3	10,215
mirjae	44,330	874.0	12,325
nitens	49,413	981.6	21,707
pseudoviettei	72,273	984.2	23,336
vadoni	114,141	865.1	25,351
viettei	107,364	906.0	18,245
taurus	252,675	1,385.6	28,119

UNIVERSITY OF HELSINKI

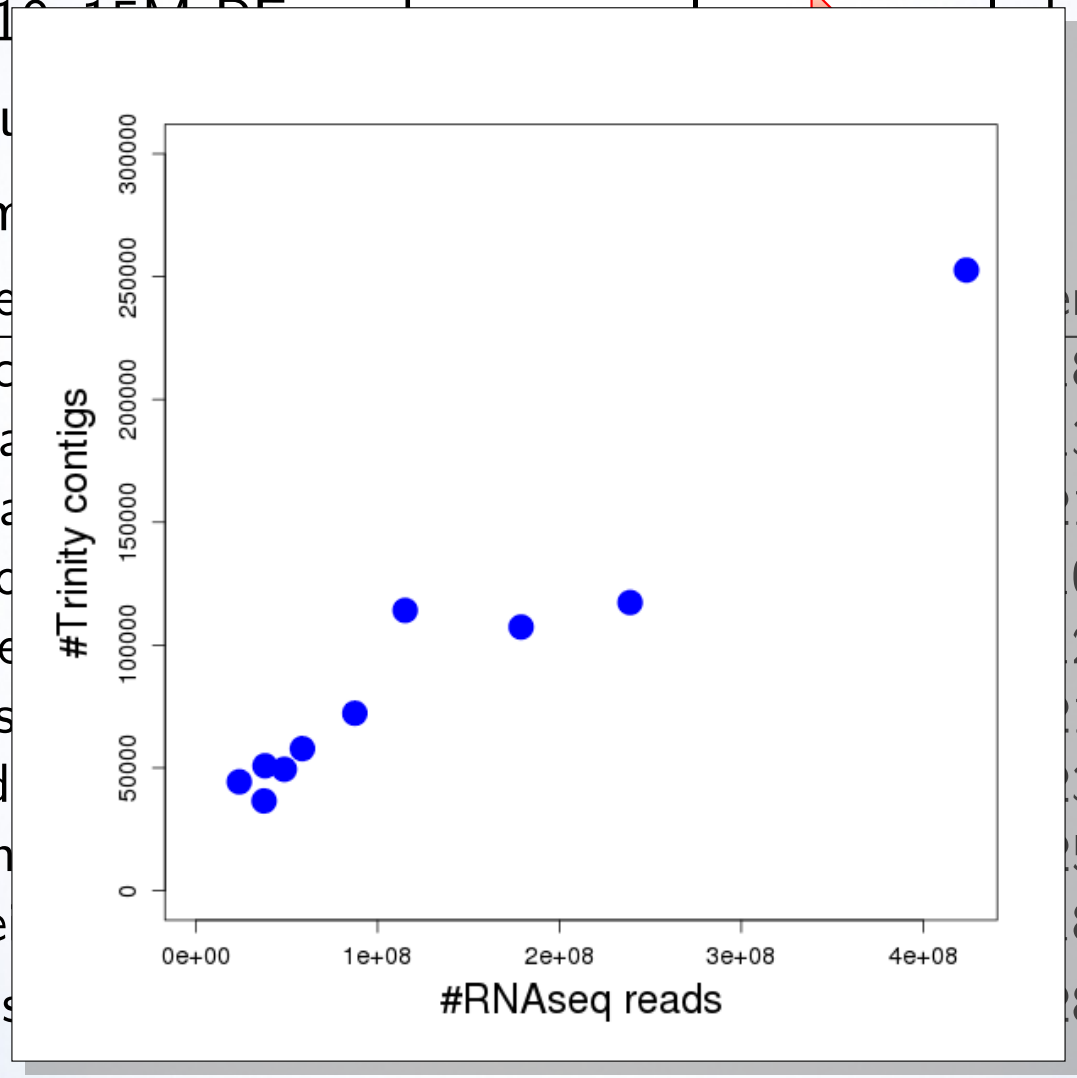
De novo assembly

Ten data sets:

- Nanos 1 [illegible] by species
- O. tauru [illegible]

Trinity assem [illegible]

specie [illegible]	ngths
bimac	8,666
binota	3,955
dubita	1,804
mamc	0,215
mirjae	2,325
nitens	1,707
pseud	3,336
vadon	5,351
viette	8,245
taurus	8,119



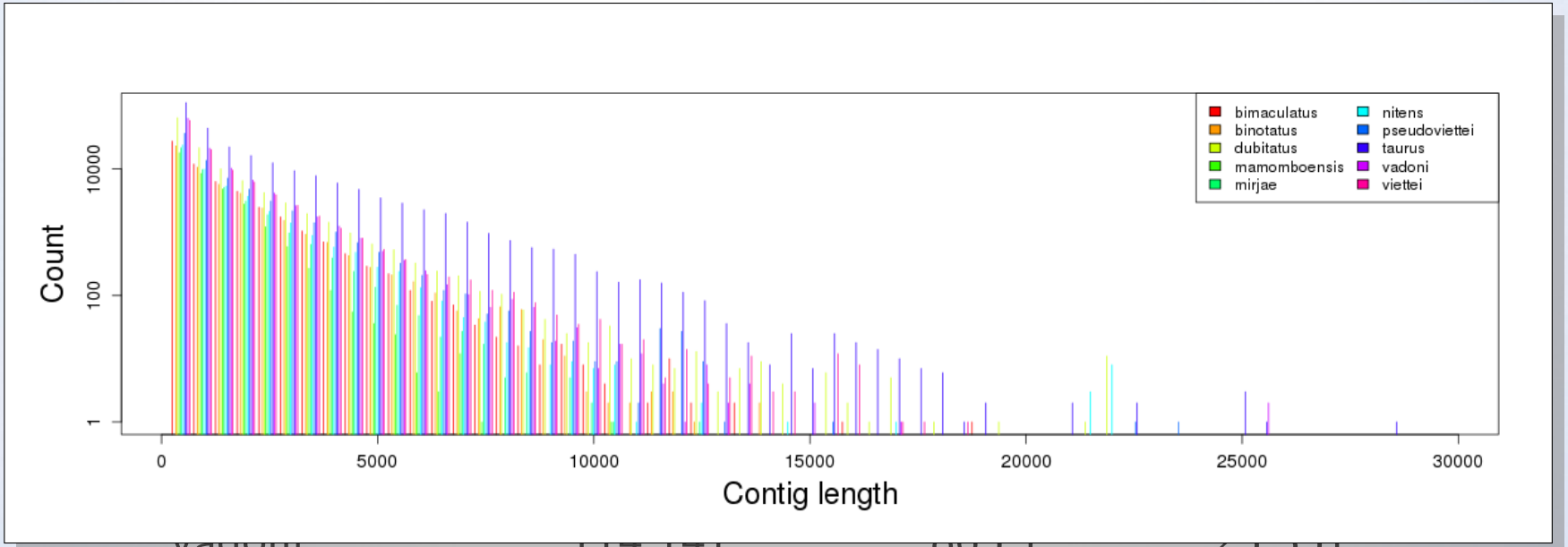
UNIVERSITY OF HELSINKI

De novo assembly

Ten data sets:

- Nanos 10–15M PE reads per sample ➡ pooled by species
- O. taurus 212M PE reads from 3 individuals

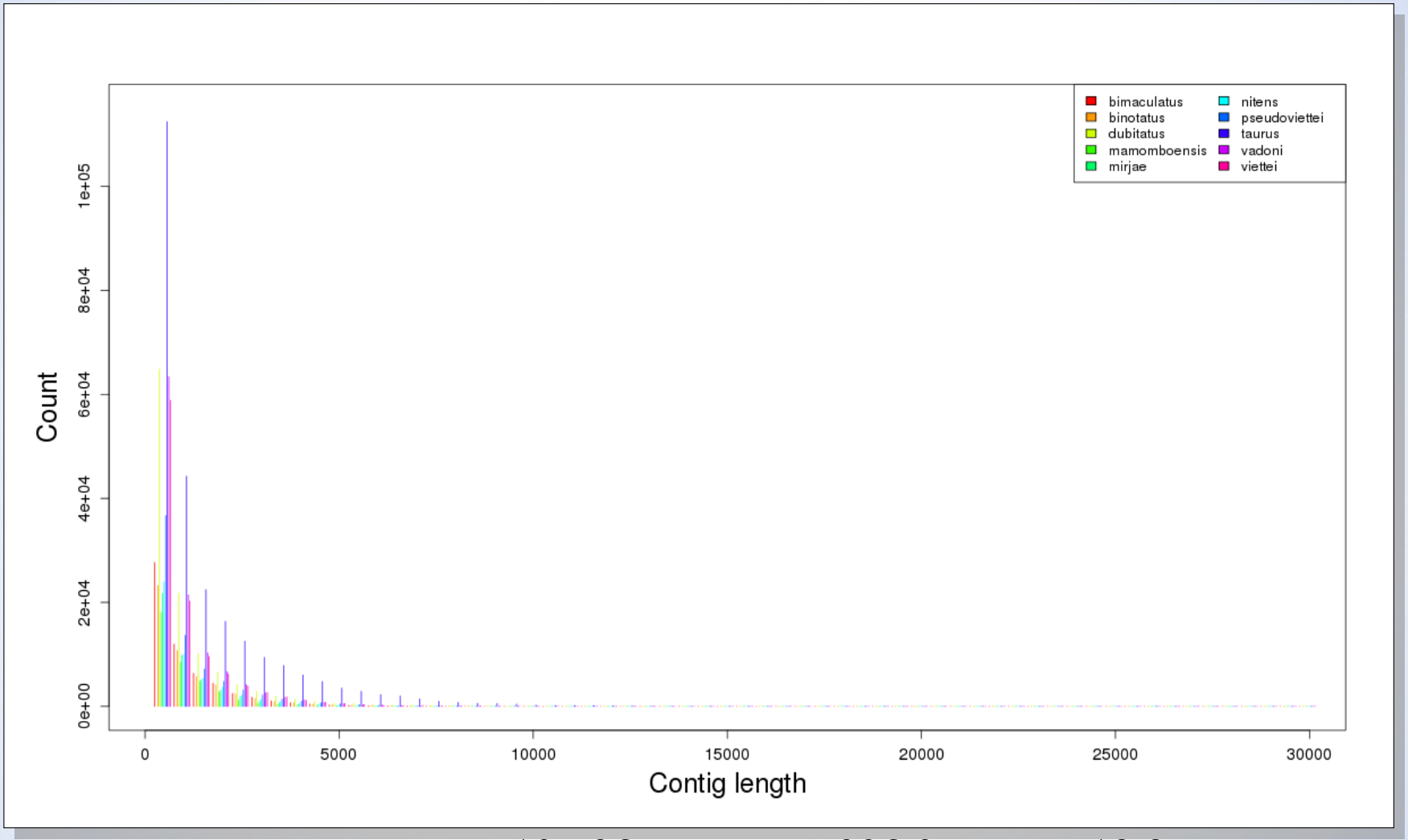
Trinity assembly:



vadoni	114,141	805.1	25,991
viettei	107,364	906.0	18,245
taurus	252,675	1,385.6	28,119

UNIVERSITY OF HELSINKI

De novo assembly



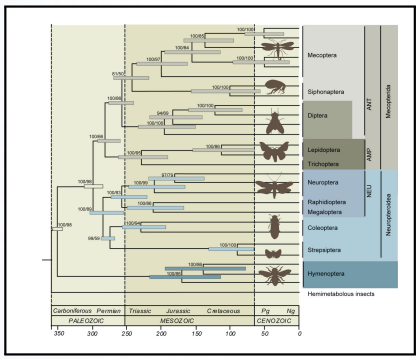
viettei	107,364	906.0	18,245
taurus	252,675	1,385.6	28,119

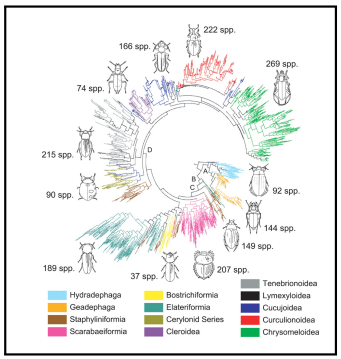
UNIVERSITY OF HELSINKI

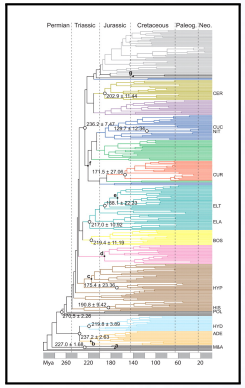
What's a suitable reference?

We use Ensembl

- Ensembl contains 75 chordates + fruitfly, nematode and yeast
- EnsemblMetazoa contains several insects, including **Tribolium castaneum**, a beetle







UNIVERSITY OF HELSINKI

What's a suitable reference?

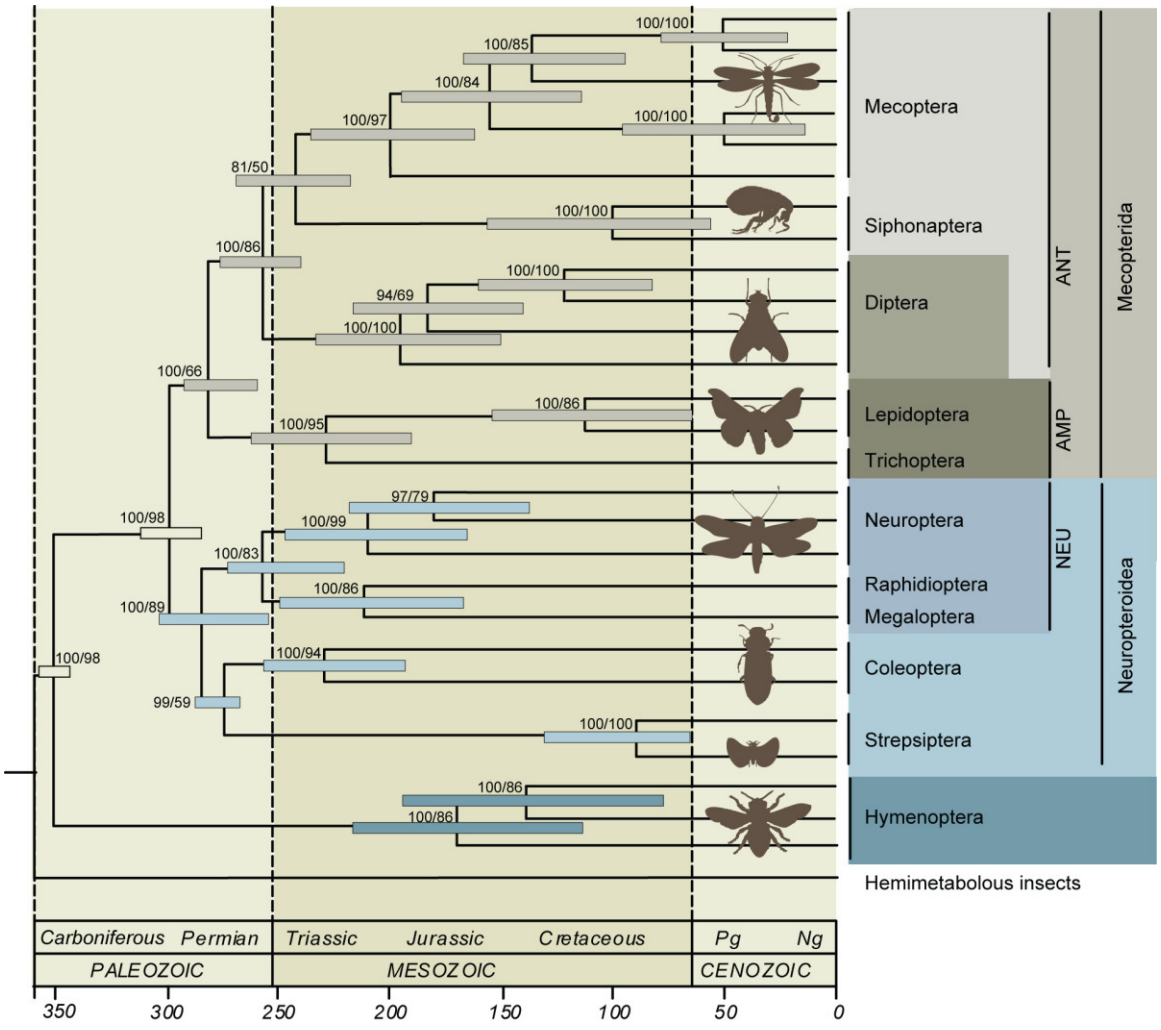
We use E

- Ense
- Ense

and yeast

Tribolium

casta



Wiegmann et al. (2009) BMC Biology.

UNIVERSITY OF HELSINKI

What's a suitable reference?

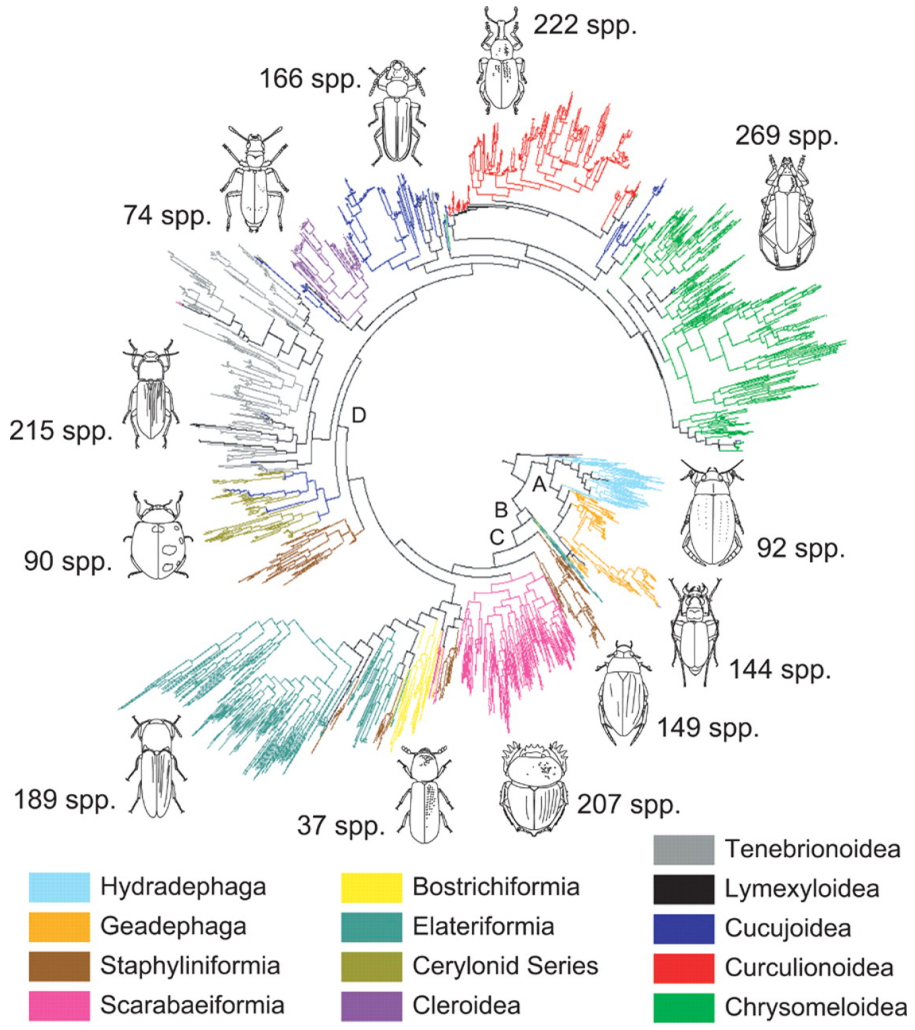
We use E

- Ense
- Ense

and yeast

Tribolium

cast



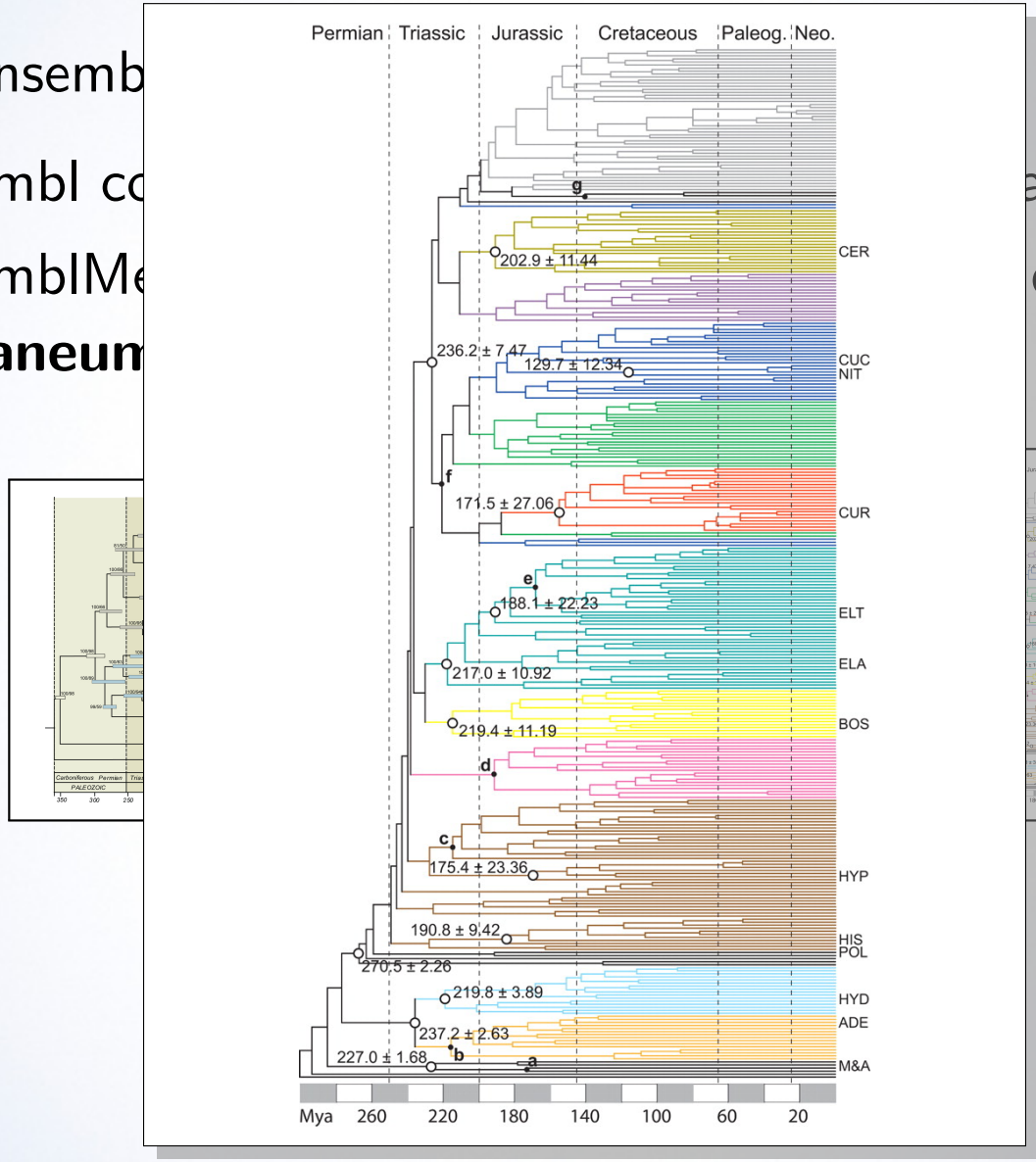
Hunt et al. (2007) Science.

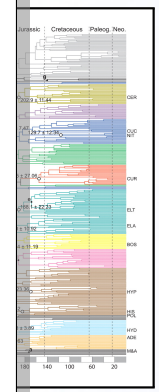
UNIVERSITY OF HELSINKI

What's a suitable reference?

We use Ensemb

- Ensembl co atode and yeast
- EnsemblMe ding **Tribolium castaneum**



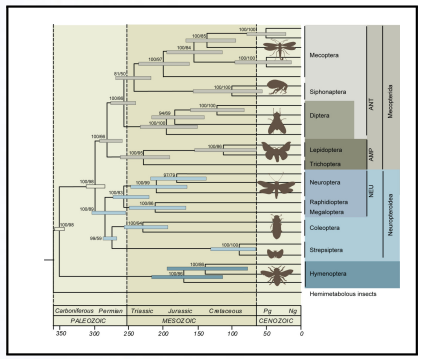


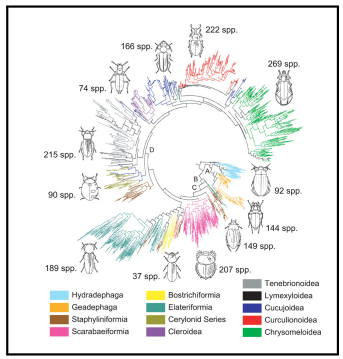
UNIVERSITY OF HELSINKI

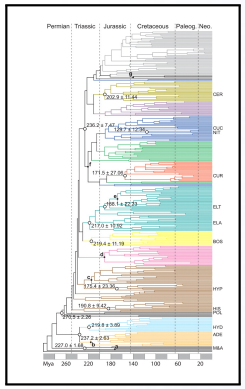
What's a suitable reference?

We use Ensembl

- Ensembl contains 75 chordates + fruitfly, nematode and yeast
- EnsemblMetazoa contains several insects, including **Tribolium castaneum**, a beetle







Our target and reference species diverged 240 mya!

UNIVERSITY OF HELSINKI

Results: BLAST hits and alignments

Data used

- reference: *T. castaneum*, 16,524 genes
- input data: ten species, 36,500 – 252,500 contigs per species

BLAST

- considering hits with identity $\geq 70\%$, length ≥ 100
- hits per species vary from from 2,750 to 3,270
- 4,818 reference genes hit, 2,060 genes hit by ten

PAGAN

- 1,720 genes have 1^{st} ranked BLAST hits from all ten species
- 1,654 genes have contigs aligned by PAGAN
- 1,623 from ten species, 19 from nine species

UNIVERSITY OF HELSINKI

Results: BLAST hits and alignments

Data used

- referenc
- input da

per species

BLAST

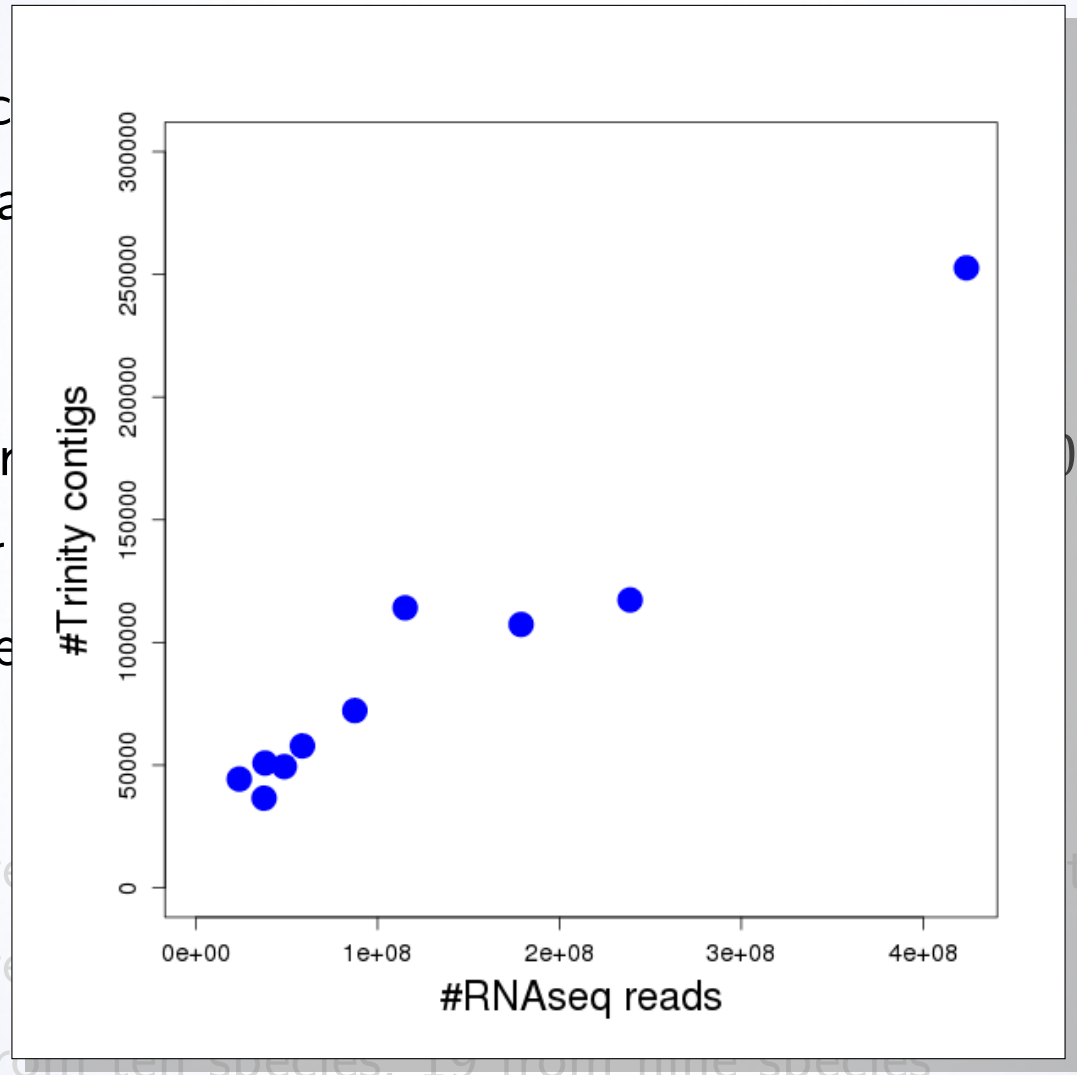
- consider
- hits per
- 4,818 re

)0

PAGAN

- 1,720 g
- 1,654 g
- 1,623 fr

ten species



UNIVERSITY OF HELSINKI

Results: BLAST hits and alignments

Data used

- referenc
- input da

per species

BLAST

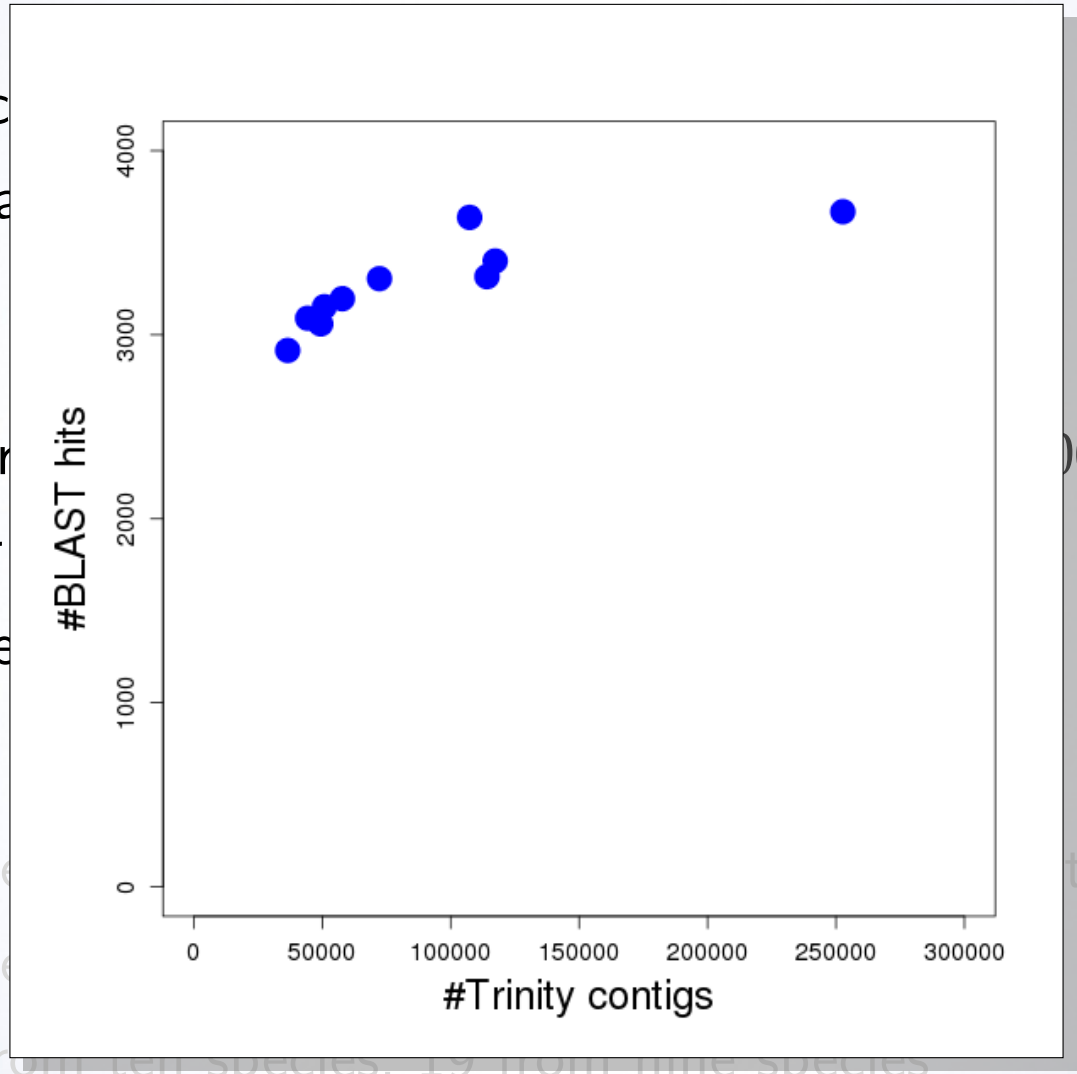
- consider
- hits per
- 4,818 re

)0

PAGAN

- 1,720 g
- 1,654 g
- 1,623 fr

ten species



UNIVERSITY OF HELSINKI

Results: BLAST hits and alignments

Data used

- reference: *T. castaneum*, 16,524 genes
- input data: ten species, 36,500 – 252,500 contigs per species

BLAST

- considering hits with identity $\geq 70\%$, length ≥ 100
- hits per species vary from from 2,750 to 3,270
- 4,818 reference genes hit, 2,060 genes hit by ten

PAGAN

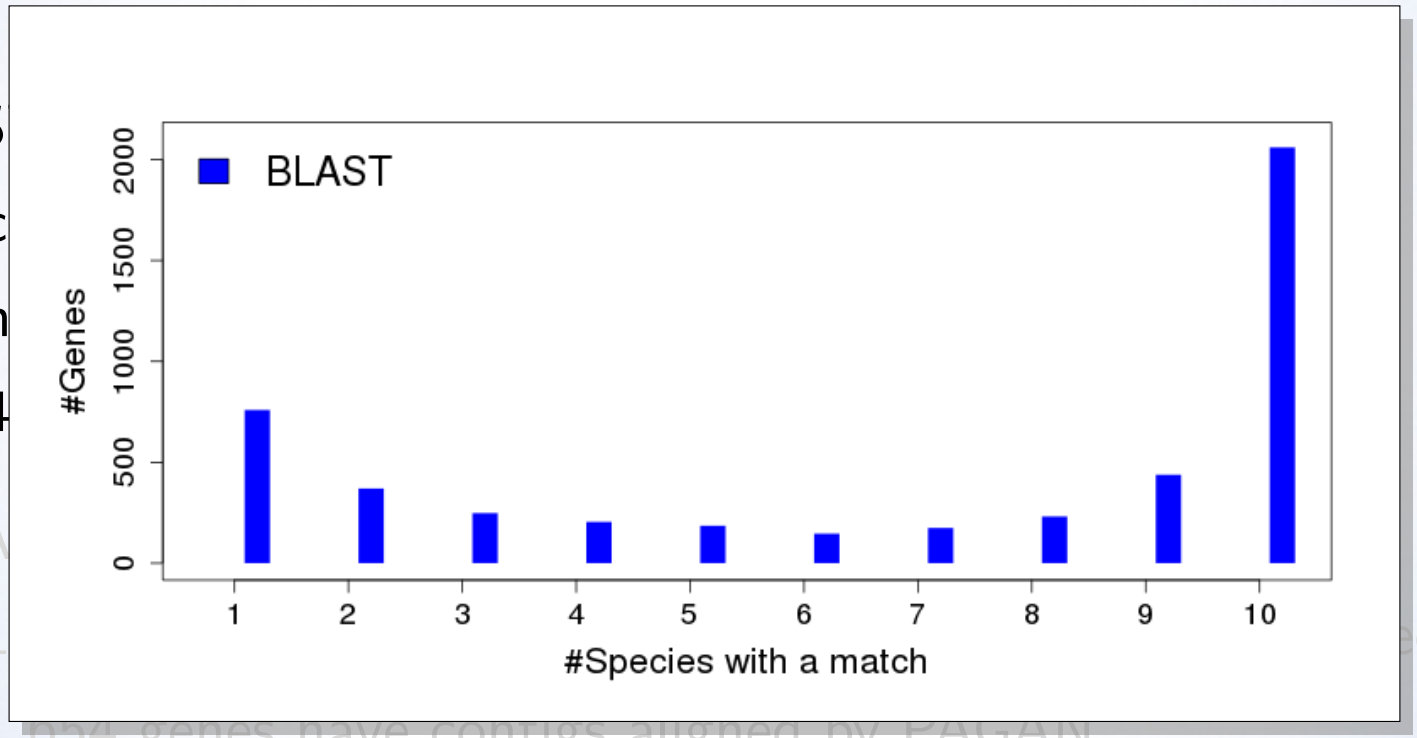
- 1,720 genes have 1^{st} ranked BLAST hits from all ten species
- 1,654 genes have contigs aligned by PAGAN
- 1,623 from ten species, 19 from nine species

UNIVERSITY OF HELSINKI

Results: BLAST hits and alignments

Data used

- reference: *T. castaneum*, 16,524 genes
- input data: ten species, 36,500 – 252,500 contigs per species



BLAS

- c
- h
- 4

PAGA

- 1
- 1,654 genes have contigs aligned by PAGAN
- 1,623 from ten species, 19 from nine species

UNIVERSITY OF HELSINKI

Results: BLAST hits and alignments

Data used

- reference: *T. castaneum*, 16,524 genes
- input data: ten species, 36,500 – 252,500 contigs per species

BLAST

- considering hits with identity $\geq 70\%$, length ≥ 100
- hits per species vary from from 2,750 to 3,270
- 4,818 reference genes hit, 2,060 genes hit by ten

PAGAN

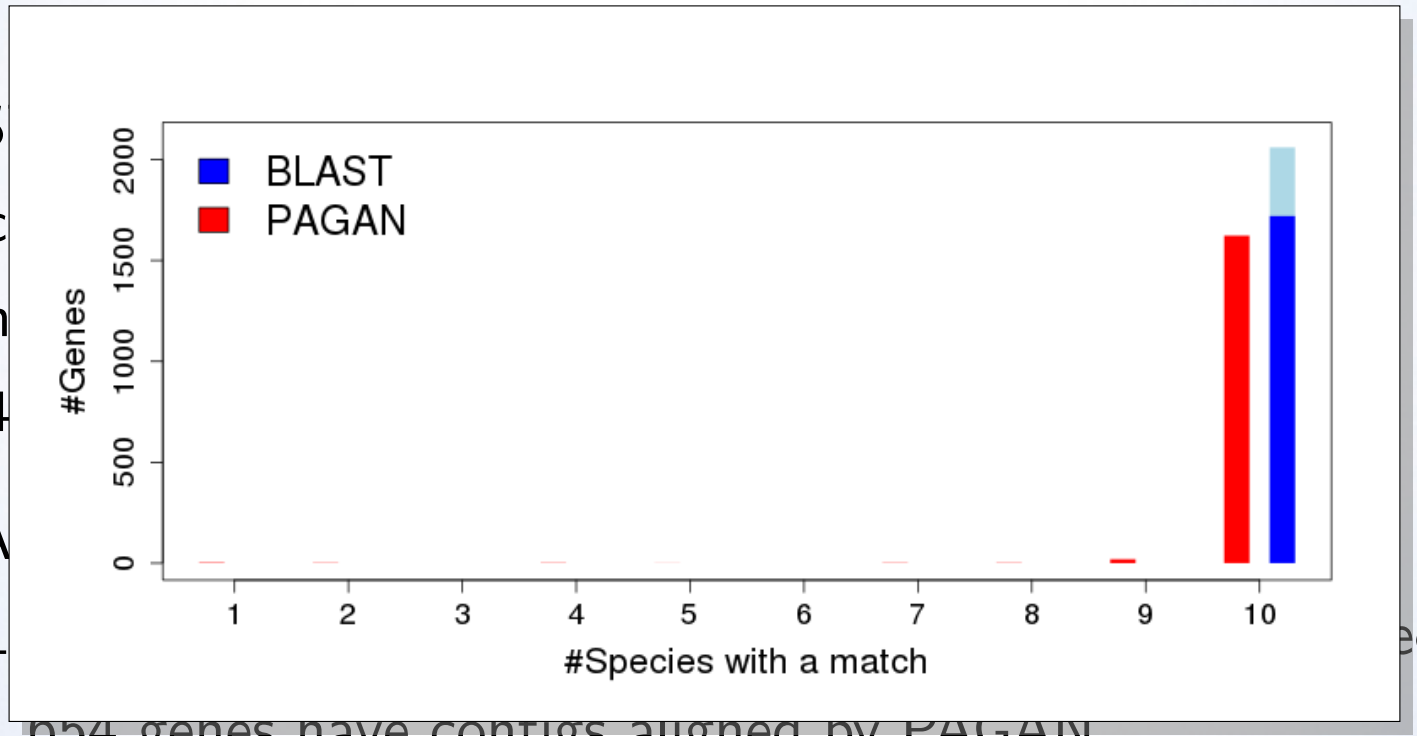
- 1,720 genes have 1^{st} ranked BLAST hits from all ten species
- 1,654 genes have contigs aligned by PAGAN
- 1,623 from ten species, 19 from nine species

UNIVERSITY OF HELSINKI

Results: BLAST hits and alignments

Data used

- reference: *T. castaneum*, 16,524 genes
- input data: ten species, 36,500 – 252,500 contigs per species



BLAS

- c
- h
- 4

PAGA

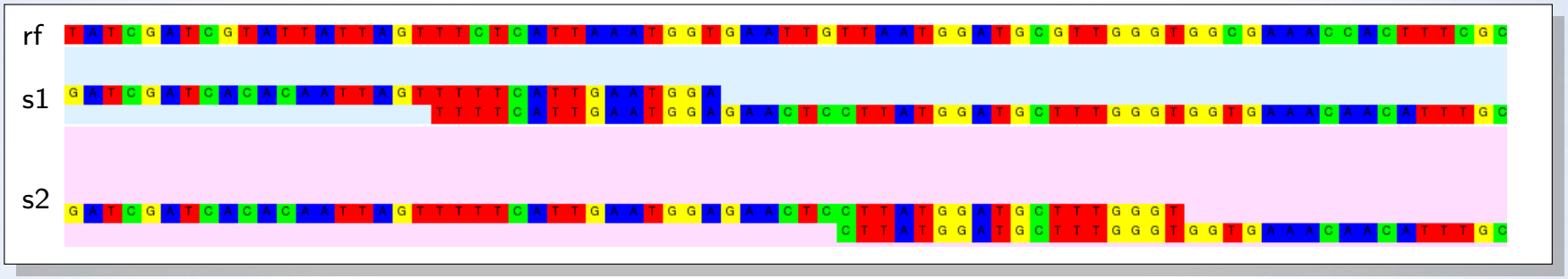
- 1 ecies
- 1,654 genes have contigs aligned by PAGAN
- 1,623 from ten species, 19 from nine species

UNIVERSITY OF HELSINKI

Results: scaffolding (1)

Matching to reference allows to connect overlapping contigs

- read overlap not always sufficient for assembly

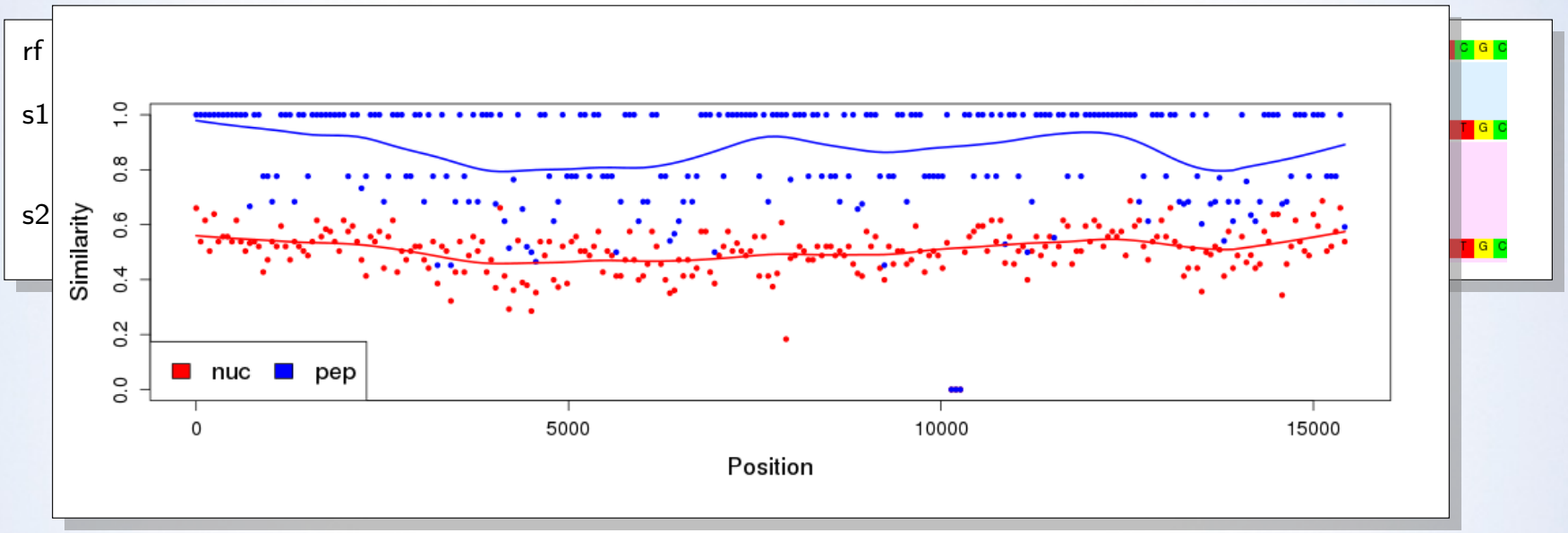


UNIVERSITY OF HELSINKI

Results: scaffolding (1)

Matching to reference allows to connect overlapping contigs

- read overlap not always sufficient for assembly

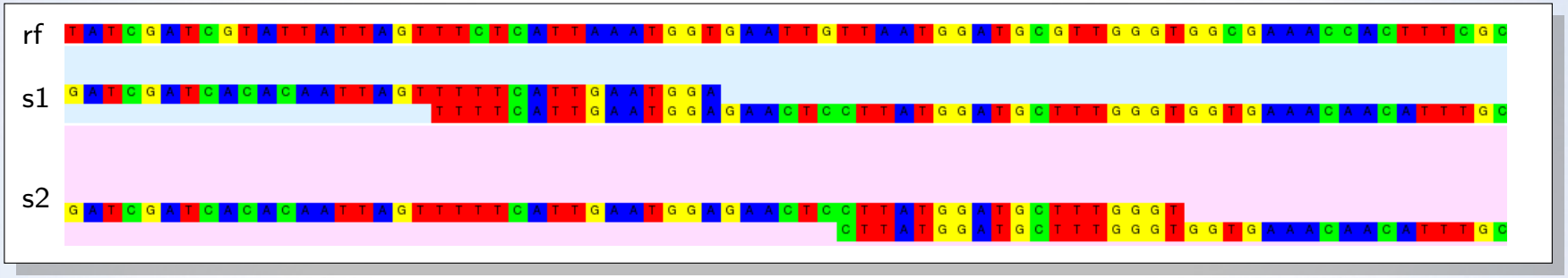


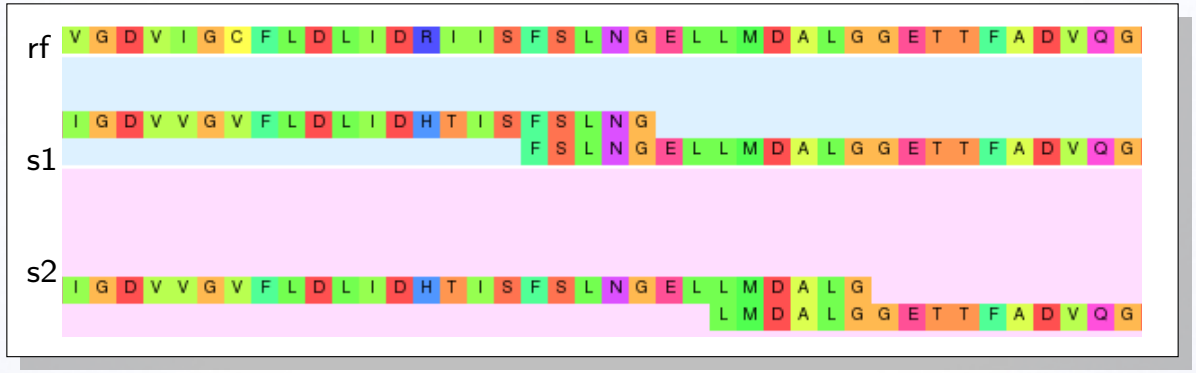
UNIVERSITY OF HELSINKI

Results: scaffolding (1)

Matching to reference allows to connect overlapping contigs

- read overlap not always sufficient for assembly



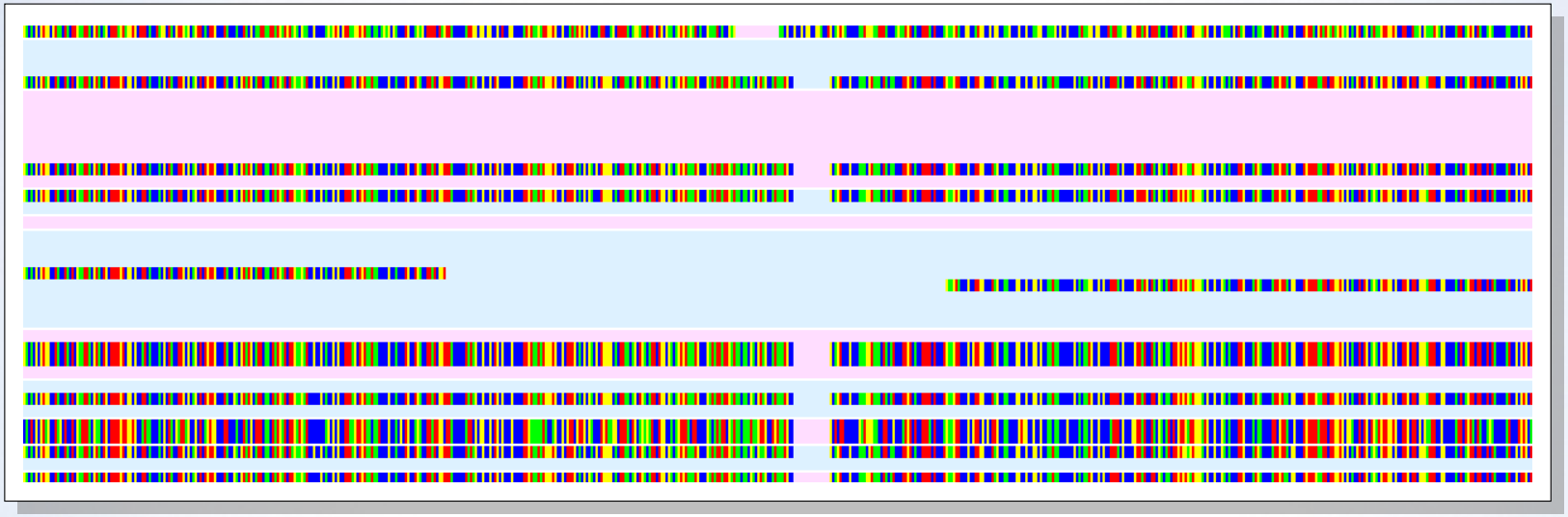


- similar approach for some non-overlapping contigs

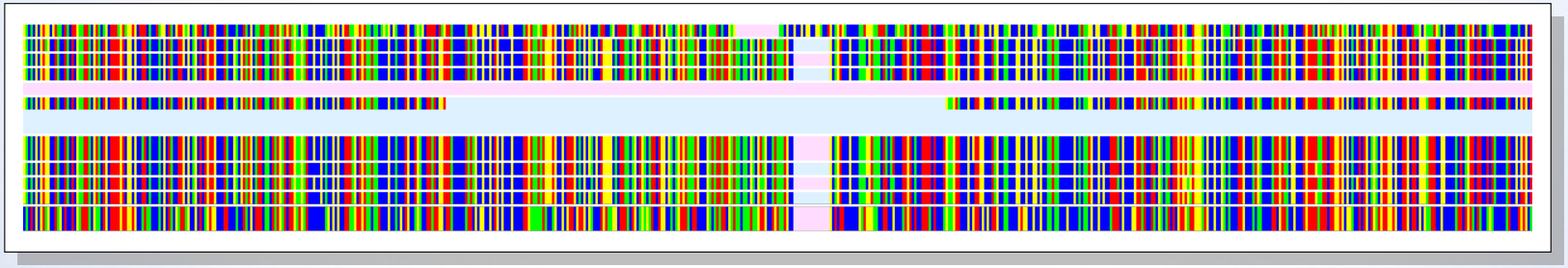
UNIVERSITY OF HELSINKI

Results: scaffolding (2)

PAGAN alignment of all homologous contigs



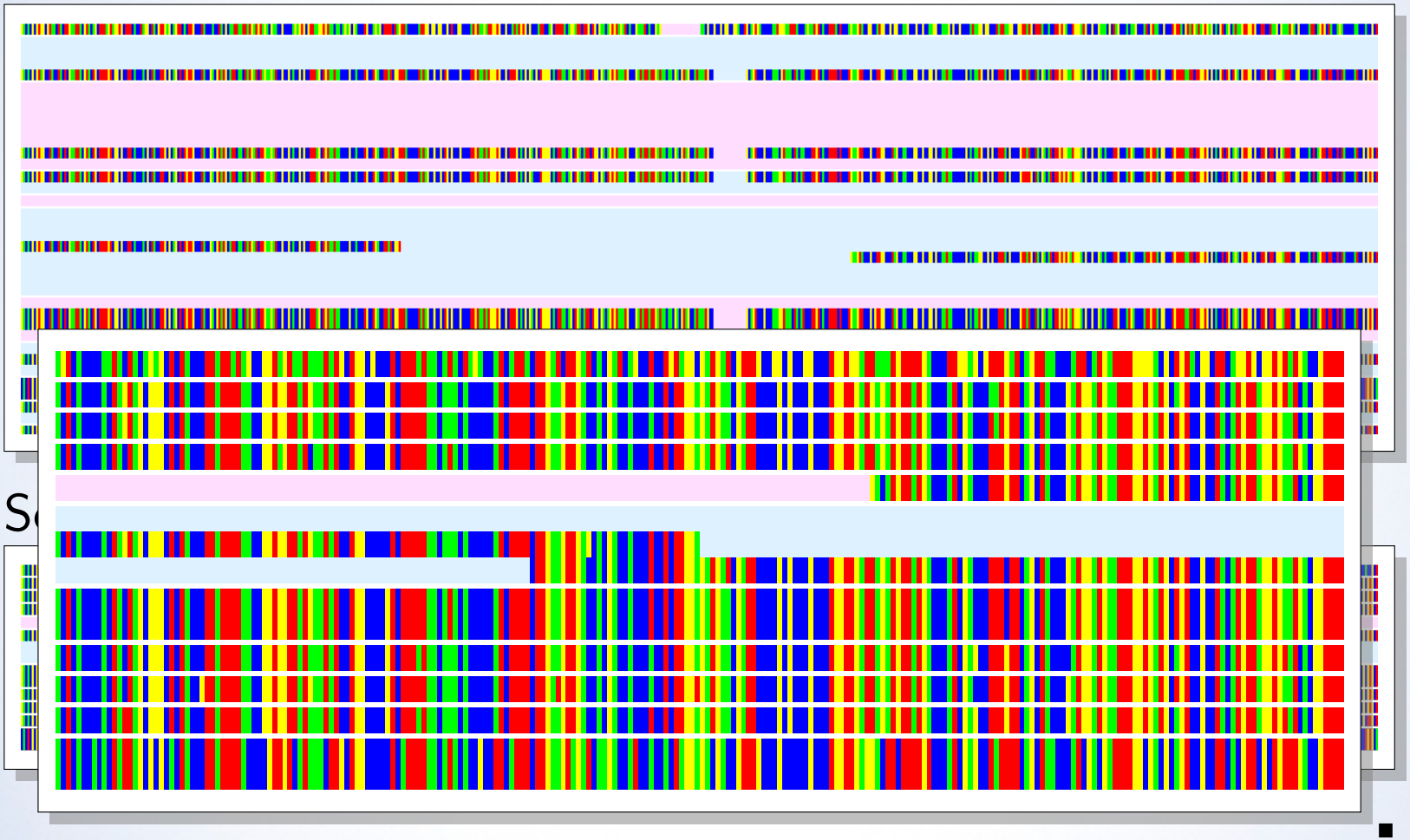
Scaffolded alignment



UNIVERSITY OF HELSINKI

Results: scaffolding (2)

PAGAN alignment of all homologous contigs



UNIVERSITY OF HELSINKI

Results: alignment coverage (1)

Sorted by coverage, hits 1–25

gene	species	contigs	length	union	mean	annotation
tcogs2:tc012671	10	26	3010	1.000	1.000	"Calcium ATPase at 60A"
tcogs2:tc008784	10	22	2540	1.000	1.000	"Elongation factor 2b"
tcogs2:tc009174	10	18	2417	1.000	0.995	"TER94"
tcogs2:tc014606	10	25	2170	1.000	0.990	"Heat shock protein 83"
tcogs2:tc005725	10	24	2379	1.000	0.989	""
tcogs2:tc015727	10	14	2133	1.000	0.980	"suppressor of forked"
tcogs2:tc014907	10	30	2054	1.000	0.979	""
tcogs2:tc008620	10	21	3114	1.000	0.977	"Na pump alpha subunit"
tcogs2:tc002825	10	11	2280	1.000	0.968	"Xeroderma pigmentosum D"
tcogs2:tc005672	10	15	3669	1.000	0.966	"SMC1"
tcogs2:tc010322	10	19	3661	1.000	0.964	""
tcogs2:tc030725	10	11	2064	1.000	0.963	"burgundy"
tcogs2:tc004395	10	15	2019	1.000	0.962	"crooked neck"
tcogs2:tc011771	10	16	3530	1.000	0.945	"RNA polymerase II 140kD subunit"
tcogs2:tc012342	10	11	2031	1.000	0.941	""
tcogs2:tc013713	10	14	2161	1.000	0.932	"Minichromosome maintenance 7"
tcogs2:tc005163	10	15	2412	1.000	0.928	""
tcogs2:tc014506	10	11	2202	1.000	0.926	"Cleavage and polyad. specif. f. 100"
tcogs2:tc005513	10	19	5730	1.000	0.919	"zipper"
tcogs2:tc010565	10	34	7101	1.000	0.908	"pre-mRNA processing factor 8"
tcogs2:tc007563	10	13	2024	1.000	0.761	"Trim9"
tcogs2:tc016126	10	11	819	0.999	0.999	""
tcogs2:tc015973	10	11	1035	0.999	0.999	""
tcogs2:tc014998	10	42	1101	0.999	0.999	""
tcogs2:tc014755	10	10	950	0.999	0.999	""

UNIVERSITY OF HELSINKI

Results: alignment coverage (2)

Sorted by coverage, hits 301–325

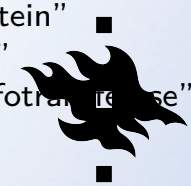
gene	species	contigs	length	union	mean	annotation
tcogs2:tc003674	10	10	449	0.998	0.988	"Aps"
tcogs2:tc000030	10	12	507	0.998	0.988	"variable nurse cells"
tcogs2:tc002742	10	10	2129	0.998	0.987	"Ced-12"
tcogs2:tc014680	10	12	2095	0.998	0.986	"Mitoch. trifunct. prot. alpha subunit"
tcogs2:tc006318	10	11	636	0.998	0.986	"insomniac"
tcogs2:tc006106	10	12	405	0.998	0.986	"Ribosomal protein L32"
tcogs2:tc013064	10	11	2259	0.998	0.985	""
tcogs2:tc007987	10	15	430	0.998	0.985	"Dim1"
tcogs2:tc006782	10	15	586	0.998	0.985	"Ribosomal protein S9"
tcogs2:tc000315	10	10	1892	0.998	0.985	""
tcogs2:tc030656	10	10	452	0.998	0.984	""
tcogs2:tc014716	10	10	512	0.998	0.984	"NC2beta"
tcogs2:tc012116	10	10	608	0.998	0.984	""
tcogs2:tc007324	10	29	661	0.998	0.984	"Ribosomal protein L10Ab"
tcogs2:tc004293	10	10	551	0.998	0.984	""
tcogs2:tc001139	10	10	551	0.998	0.984	""
tcogs2:tc001317	10	13	460	0.998	0.983	"Ribosomal protein S16"
tcogs2:tc000443	10	20	447	0.998	0.983	"effete"
tcogs2:tc008409	10	10	548	0.998	0.982	""
tcogs2:tc004789	10	12	1888	0.998	0.982	""
tcogs2:tc016093	10	10	545	0.998	0.981	""
tcogs2:tc010138	10	14	643	0.998	0.980	"Rab-protein 2"
tcogs2:tc005390	10	13	436	0.998	0.979	"lethal (1) 10Bb"
tcogs2:tc013931	10	11	477	0.998	0.978	"lethal (1) G0230"
tcogs2:tc013212	10	10	599	0.998	0.978	""

UNIVERSITY OF HELSINKI

Results: alignment coverage (3)

Sorted by coverage, hits 601–625

gene	species	contigs	length	union	mean	annotation
tcogs2:tc006466	10	12	1464	0.991	0.887	""
tcogs2:tc015505	10	11	1140	0.991	0.867	"Signal peptide peptidase-like"
tcogs2:tc013893	10	10	1529	0.990	0.990	""
tcogs2:tc013318	10	12	966	0.990	0.990	"Mov34"
tcogs2:tc010161	10	13	984	0.990	0.990	"eukar. translat. Init. Factor 2alpha"
tcogs2:tc009325	10	10	1259	0.990	0.990	"Fumarylacetoacetase"
tcogs2:tc005644	10	11	1041	0.990	0.990	""
tcogs2:tc005191	10	12	1552	0.990	0.990	""
tcogs2:tc005099	10	13	1542	0.990	0.990	"raspberry"
tcogs2:tc002299	10	10	695	0.990	0.990	""
tcogs2:tc000069	10	12	709	0.990	0.990	"Proteasome 26kD subunit"
tcogs2:tc008261	10	14	728	0.990	0.989	"Ribosomal protein S3"
tcogs2:tc010318	10	22	1885	0.990	0.987	"PAPS synthetase"
tcogs2:tc002330	10	10	953	0.990	0.981	""
tcogs2:tc012896	10	13	975	0.990	0.973	""
tcogs2:tc010335	10	11	1008	0.990	0.970	""
tcogs2:tc014743	10	15	1329	0.990	0.965	"CDP diglyceride synthetase"
tcogs2:tc001773	10	10	1316	0.990	0.950	"tetracycline resistance"
tcogs2:tc030733	10	10	1553	0.990	0.949	""
tcogs2:tc009994	10	13	1290	0.990	0.938	"DMAP1"
tcogs2:tc010587	10	12	1273	0.990	0.934	""
tcogs2:tc014954	10	11	1548	0.990	0.916	"epithelial membrane protein"
tcogs2:tc002610	10	14	14962	0.990	0.902	"Ryanodine receptor 44F"
tcogs2:tc014815	10	10	1271	0.990	0.899	"Heparan sulfate 6-O-sulfotransferase"
tcogs2:tc008190	10	10	2591	0.990	0.896	"Nat1"



UNIVERSITY OF HELSINKI

Results: alignment coverage (4)

Sorted by coverage, hits 901–925

gene	species	contigs	length	union	mean	annotation
tcogs2:tc004935	10	10	1364	0.979	0.968	""
tcogs2:tc000412	10	15	1470	0.979	0.966	"Abelson interacting protein"
tcogs2:tc006185	10	21	610	0.979	0.962	""
tcogs2:tc001089	10	10	1652	0.979	0.955	"Helicase"
tcogs2:tc014294	10	12	3339	0.979	0.950	"dre4"
tcogs2:tc015941	10	14	1208	0.979	0.948	"homer"
tcogs2:tc010363	10	11	2700	0.979	0.947	""
tcogs2:tc002951	10	14	761	0.979	0.937	"Ribosomal protein L7"
tcogs2:tc000524	10	14	631	0.979	0.936	""
tcogs2:tc009202	10	10	2150	0.979	0.930	""
tcogs2:tc005928	10	11	888	0.979	0.930	""
tcogs2:tc002088	10	13	2337	0.979	0.926	"Minichromosome maintenance 3"
tcogs2:tc011785	10	10	1043	0.979	0.874	"fringe"
tcogs2:tc005583	10	19	1025	0.979	0.863	"S-adenosylmethionine decarboxylase"
tcogs2:tc015300	10	13	3346	0.979	0.799	"Integrator 2"
tcogs2:tc011404	10	10	2012	0.979	0.764	"pole hole"
tcogs2:tc007336	10	13	1442	0.979	0.736	"CoRest"
tcogs2:tc012453	10	16	1336	0.979	0.713	"C-terminal Binding Protein"
tcogs2:tc006305	10	14	2172	0.979	0.668	"Rapgap1"
tcogs2:tc013786	10	10	461	0.978	0.978	""
tcogs2:tc016225	10	13	1275	0.978	0.975	"ade5"
tcogs2:tc015633	10	10	599	0.978	0.973	"CHOp24"
tcogs2:tc009472	10	10	1844	0.978	0.973	"NOP2-Sun domain family
tcogs2:tc004425	10	45	1703	0.978	0.971	"Heat shock protein cognate
tcogs2:tc006024	10	12	1845	0.978	0.968	"sluggish A"

UNIVERSITY OF HELSINKI

Results: alignment coverage (5)

Sorted by coverage, hits 1,201–1,225

gene	species	contigs	length	union	mean	annotation
tcogs2:tc007466	10	10	2015	0.959	0.953	""
tcogs2:tc002547	10	47	1265	0.959	0.946	""
tcogs2:tc007911	10	11	900	0.959	0.941	""
tcogs2:tc000687	10	11	1482	0.959	0.928	"genderblind"
tcogs2:tc011210	10	15	1717	0.959	0.923	"absent
tcogs2:tc012273	10	14	1430	0.959	0.915	"twins"
tcogs2:tc009295	10	10	1271	0.959	0.898	""
tcogs2:tc004839	10	14	2280	0.959	0.874	""
tcogs2:tc002125	10	11	765	0.959	0.872	""
tcogs2:tc015882	10	12	1710	0.959	0.803	""
tcogs2:tc003318	10	19	2002	0.959	0.767	"Cbl"
tcogs2:tc008121	10	15	1744	0.958	0.953	"Myb-interacting protein 130"
tcogs2:tc010401	10	11	1158	0.958	0.944	""
tcogs2:tc002496	10	12	1653	0.958	0.914	"pale"
tcogs2:tc016079	10	13	1747	0.958	0.910	""
tcogs2:tc016315	10	10	1313	0.958	0.879	""
tcogs2:tc004563	10	20	3239	0.958	0.828	""
tcogs2:tc008130	10	10	1265	0.957	0.957	""
tcogs2:tc002441	10	12	651	0.957	0.953	"Fkbp13"
tcogs2:tc011321	10	10	1610	0.957	0.951	"prolyl-4-hydroxylase-alpha EFB"
tcogs2:tc011541	10	15	1777	0.957	0.921	"lethal (2) k01209"
tcogs2:tc002602	10	14	2539	0.957	0.895	""
tcogs2:tc014175	10	10	647	0.957	0.877	""
tcogs2:tc011068	10	10	2747	0.957	0.807	"papillote"
tcogs2:tc000797	10	53	3759	0.957	0.725	"retinal degeneration A"

UNIVERSITY OF HELSINKI

Results: alignment coverage (6)

Sorted by coverage, hits 1,501–1,525

gene	species	contigs	length	union	mean	annotation
tcogs2:tc000321	10	33	1646	0.878	0.770	""
tcogs2:tc007820	10	15	3150	0.878	0.699	""
tcogs2:tc013939	10	13	2418	0.877	0.814	""
tcogs2:tc011089	10	10	1583	0.876	0.859	"alpha Mannosidase I"
tcogs2:tc014113	10	28	1455	0.874	0.854	"Eukaryotic initiation factor 4a"
tcogs2:tc004773	10	10	671	0.873	0.855	"Dihydropteridine reductase"
tcogs2:tc006614	10	15	1623	0.871	0.787	""
tcogs2:tc010200	10	73	5794	0.870	0.790	"lethal (1) G0196"
tcogs2:tc005777	10	36	1421	0.866	0.859	"discs overgrown"
tcogs2:tc007831	10	11	1676	0.866	0.803	"anterior open"
tcogs2:tc009631	10	11	1335	0.864	0.848	"Vacuolar H[+] ATPase 44kD C sub"
tcogs2:tc005924	10	40	6878	0.864	0.842	"Myosin heavy chain"
tcogs2:tc004390	10	23	980	0.864	0.811	""
tcogs2:tc004825	10	19	1841	0.864	0.632	""
tcogs2:tc004567	10	11	1698	0.863	0.804	""
tcogs2:tc011298	10	30	1347	0.860	0.830	"tropomodulin"
tcogs2:tc006408	10	25	1055	0.860	0.820	"Porin2"
tcogs2:tc011682	10	12	817	0.859	0.815	""
tcogs2:tc015819	10	12	594	0.857	0.812	"anti-silencing factor 1"
tcogs2:tc005709	10	11	1089	0.856	0.807	"gustavus"
tcogs2:tc012861	10	10	884	0.856	0.750	""
tcogs2:tc006203	10	10	2801	0.855	0.814	""
tcogs2:tc003020	10	11	1805	0.854	0.844	"delta-coatomer protein"
tcogs2:tc012381	10	10	1085	0.853	0.788	""
tcogs2:tc010869	10	20	2849	0.853	0.762	"Furin 1"

UNIVERSITY OF HELSINKI

Results: alignment coverage (7)

Sorted by coverage, hits 1,629–1,654

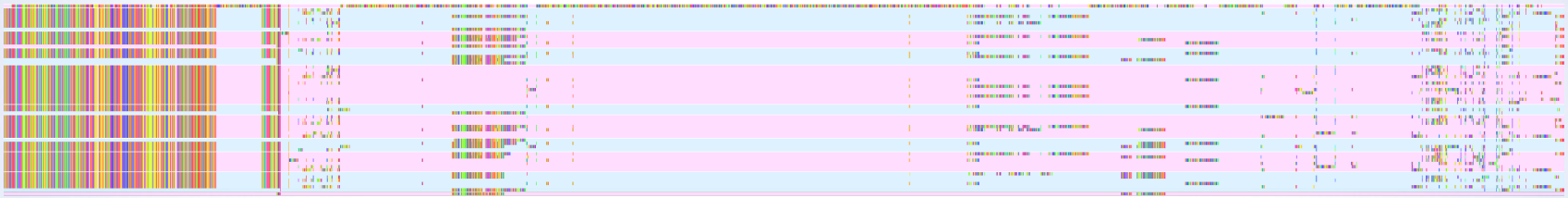
gene	species	contigs	length	union	mean	annotation
tcogs2:tc004856	10	21	1336	0.564	0.504	""
tcogs2:tc009085	10	12	2826	0.564	0.433	""
tcogs2:tc015202	10	10	1015	0.563	0.502	""
tcogs2:tc008782	10	12	3192	0.553	0.494	"cactin"
tcogs2:tc009488	10	17	1207	0.550	0.498	"tungus"
tcogs2:tc004497	10	28	4105	0.549	0.538	"hu li tai shao"
tcogs2:tc015307	10	15	2297	0.527	0.514	"Major Facilit. Superfam. Transp. 18"
tcogs2:tc008507	10	13	4980	0.525	0.263	"toothrin"
tcogs2:tc013252	10	14	2585	0.516	0.486	""
tcogs2:tc007694	10	32	6859	0.507	0.504	""
tcogs2:tc014370	10	19	2385	0.501	0.497	"Syntrophin-like 1"
tcogs2:tc015417	10	18	2201	0.488	0.463	""
tcogs2:tc004728	10	29	3980	0.465	0.396	"Na[+]/H[+] hydrogen exchanger 3"
tcogs2:tc016023	10	11	1680	0.455	0.407	"suppressor of white-apricot"
tcogs2:tc008240	10	47	1637	0.429	0.401	"wings up A"
tcogs2:tc000855	1	1	1538	0.416	0.416	"discs large 1"
tcogs2:tc000606	10	15	15650	0.411	0.252	"Lost PHDs of trr"
tcogs2:tc011562	10	12	2385	0.362	0.362	"Dipeptidase C"
tcogs2:tc013587	10	10	5647	0.362	0.348	"Cullin-4"
tcogs2:tc003097	10	56	5545	0.361	0.297	"longitudinals lacking"
tcogs2:tc014344	10	16	2839	0.347	0.344	""
tcogs2:tc008947	10	11	1619	0.335	0.308	"mitoch. ribosomal protein L13"
tcogs2:tc013260	10	12	5157	0.252	0.237	""
tcogs2:tc013018	10	13	3247	0.167	0.161	""
tcogs2:tc015896	10	10	2147	0.149	0.148	"fau"

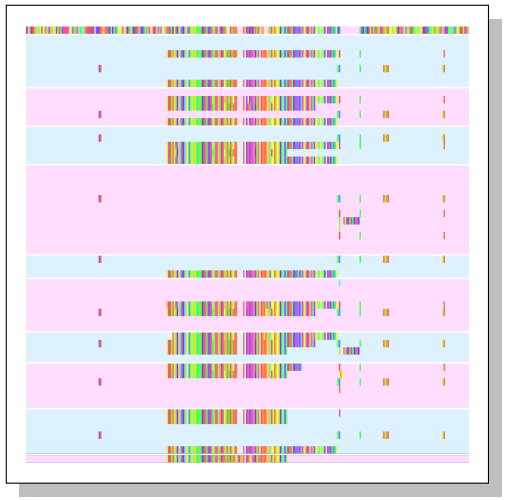
UNIVERSITY OF HELSINKI

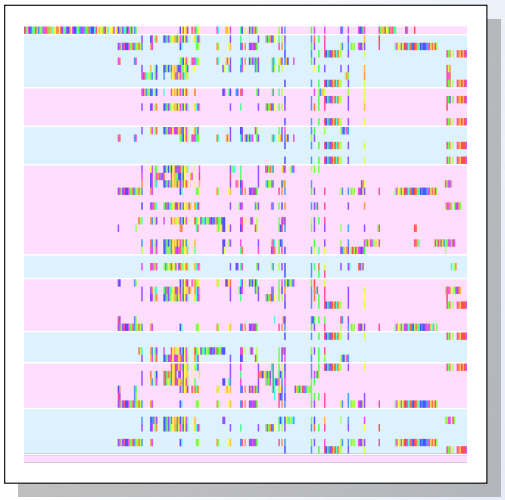
Low coverage: incomplete or erroneous?

Sorted by coverage, hit 1,649

gene	species	contigs	length	union	mean	annotation
tcogs2:tc003097	10	56	5545	0.361	0.297	"longitudinals lacking"







- reference may be erroneous, too!

UNIVERSITY OF HELSINKI

High coverage: not always easy!

Sorted by coverage, hit 24

gene	species	contigs	length	union	mean	annotation
tcogs2:tc014998	10	42	1101	0.999	0.999	””

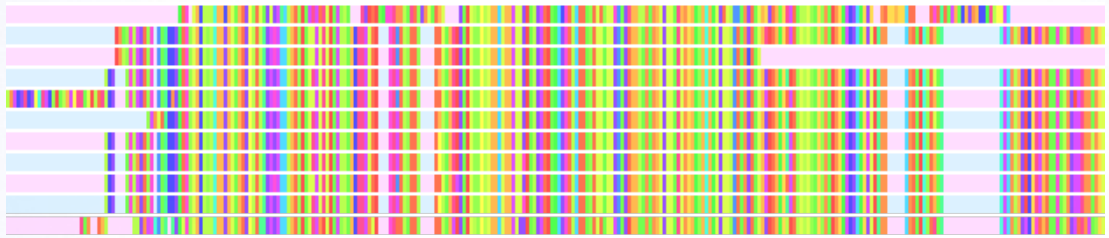
- high expression ➡ alternative transcripts, polymorphism

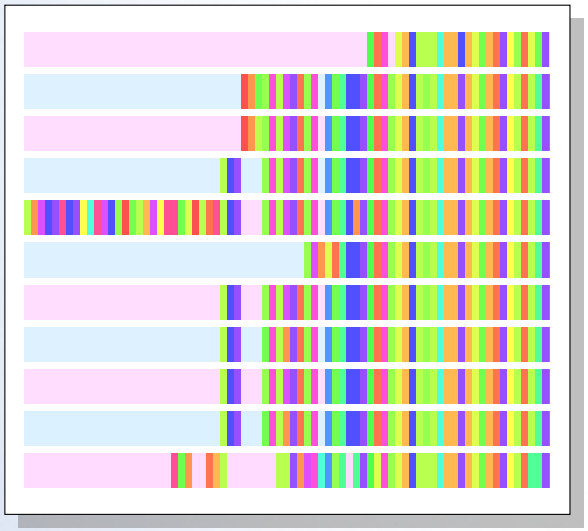
UNIVERSITY OF HELSINKI

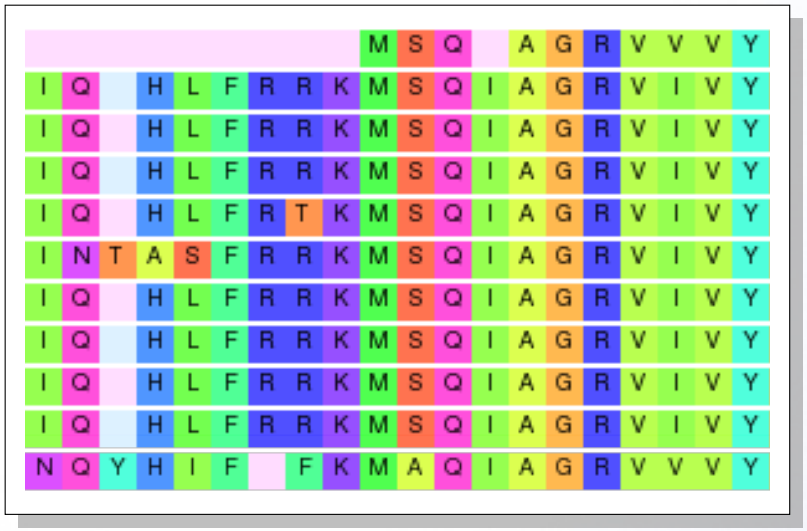
Reference: job done?

Sorted by coverage, hit 1,506

gene	species	contigs	length	union	mean	annotation
tcogs2:tc004773	10	10	671	0.873	0.855	"Dihydropteridine reductase"







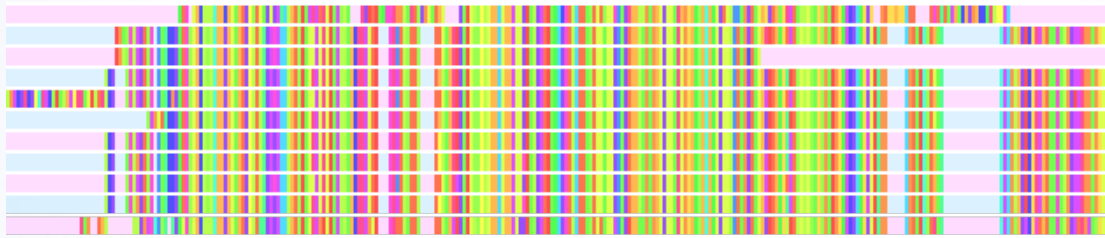
- proteins lack UTRs, exons/splicing may be different!

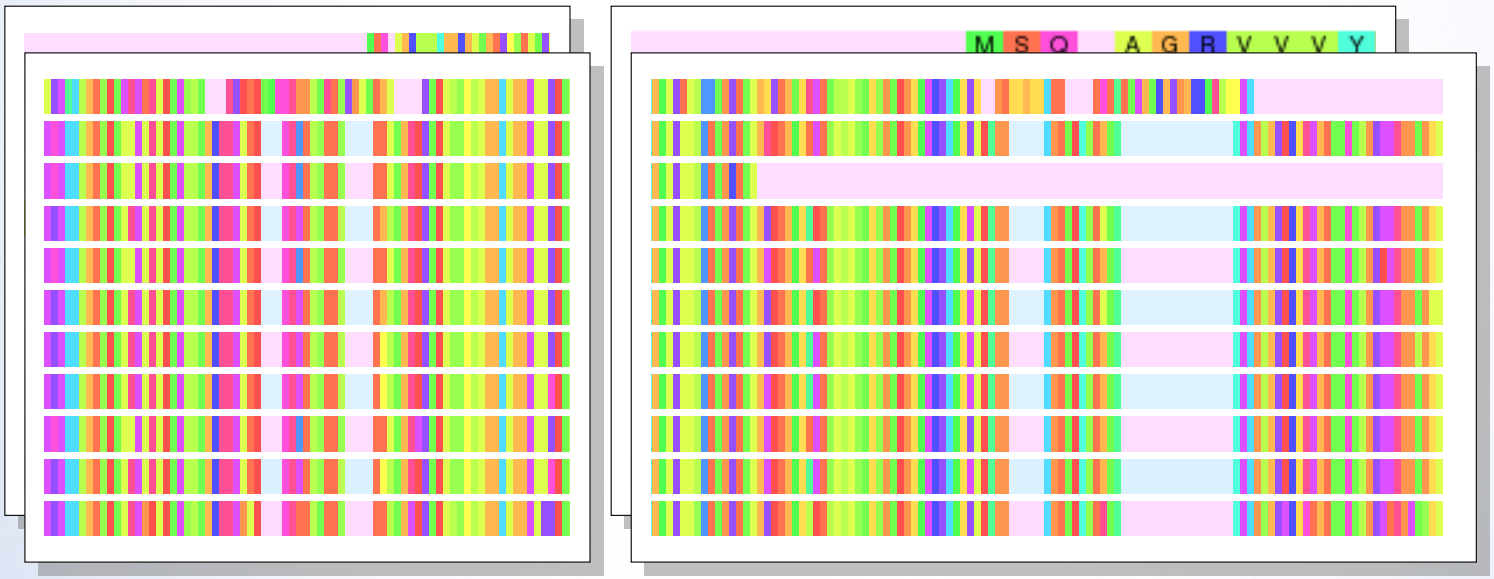
UNIVERSITY OF HELSINKI

Reference: job done?

Sorted by coverage, hit 1,506

gene	species	contigs	length	union	mean	annotation
tcogs2:tc004773	10	10	671	0.873	0.855	"Dihydropteridine reductase"





- proteins lack UTRs, exons/splicing may be different!

UNIVERSITY OF HELSINKI

Conclusions by now

Learned from dung beetles

- 480my divergence to reference species seems acceptable
- $>$1,500 gene alignments ($\sim$ 10%) with data from all species
- many data sets with a candidate annotation from fruitfly

We have reason to believe

- using fruitfly as reference would be possible and useful
- lower thresholds and heuristics would capture more genes
- however, a majority of genes may already been captured

It seems self-evident

- close, well annotated reference would make things lot easier

UNIVERSITY OF HELSINKI

Things to do next

Tuning Glutton v.1.0

- more aggressive scaffolding, at least as an option
- iterative search with consensus nucleotide reference
  - capture UTRs, alternative splicing
- multiple reference species with phylogenetic modelling

For Glutton v.2.0

- fully phylogenetic approach: extension of existing analyses
- when data available, incorporate gene splicing
- for close species, full analysis in DNA space

Software available in a few months time. Stay tuned!

Funding:

MARIE CURIE ACTIONS

BIOCENTRUM HELSINKI

BF Biocenter Finland

UNIVERSITY OF HELSINKI

http://en.wikipedia.org/wiki/Gluttony

was@bi

Beta version available at

http://wasabi.biocenter.helsinki.fi

Graphical interface for evolutionary sequence alignment and sharing of alignment data

- runs inside a regular web browser
  - ⇨ platform independent, no installation required
- locally installed application coming soon

Andres Veidenberg

UNIVERSITY OF HELSINKI