Unveiling The Hidden Life Of A Very Visible Phytoplankton:

Use of normalized Sanger sequenced libraries and non-normalized 454 libraries to compare the transcriptomes of haploid and diploid *Emiliania huxleyi*.

Peter von Dassow[1], Hiroyuki Ogata[2], Ian Probert[1], Stéphane Audic[2], Jean-Michel Claverie[2], Patrick Wincker[3], Corinne Da Silva[3] and Colomban de Vargas[1]

[1]Evolution du Plancton et PaleOceans, Station Biologique de Roscoff, CNRS UPMC UMR7144, 29682 Roscoff, France
[2]Structural and Genomic Information, CNRS UPR2589, 13288 Marseille, France
[3]Genoscope Centre National de Sequençage, 91057 Evry, France

Funding:
Genoscope
Grants BOOM and ATIP (ANR)
Marie Curie International Incoming Post-doctoral Fellowship grant FUNSEXDEPHYND (European Commission Research Directorate General)

EPPO - Evolution du Plancton et Paleo-Oceans

Sequencage haut-debit: 2008-2012

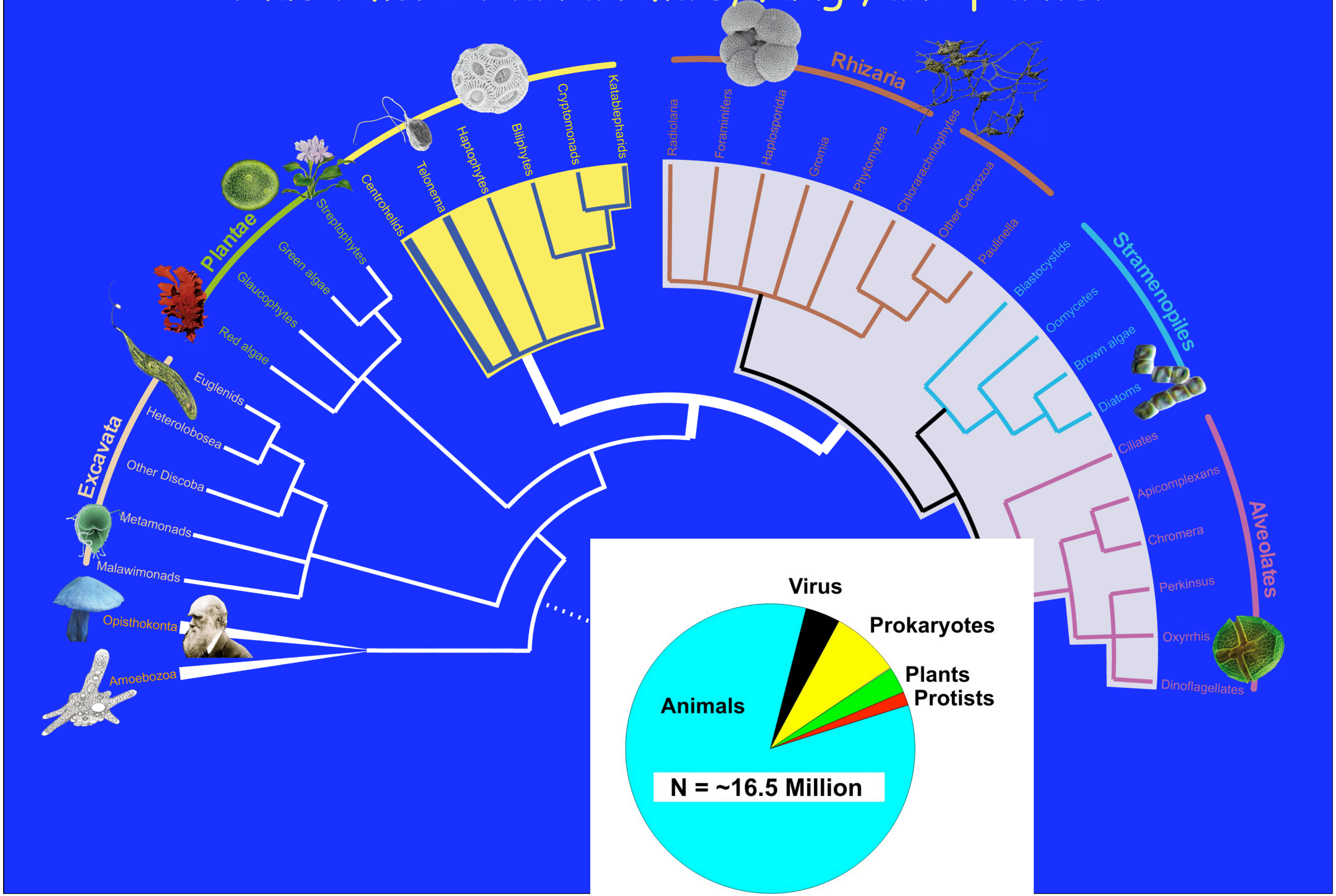
CNRS

GENOSCOPE Centre National de Séquençage

CNRS UPMC INSU Station Biologique Roscoff

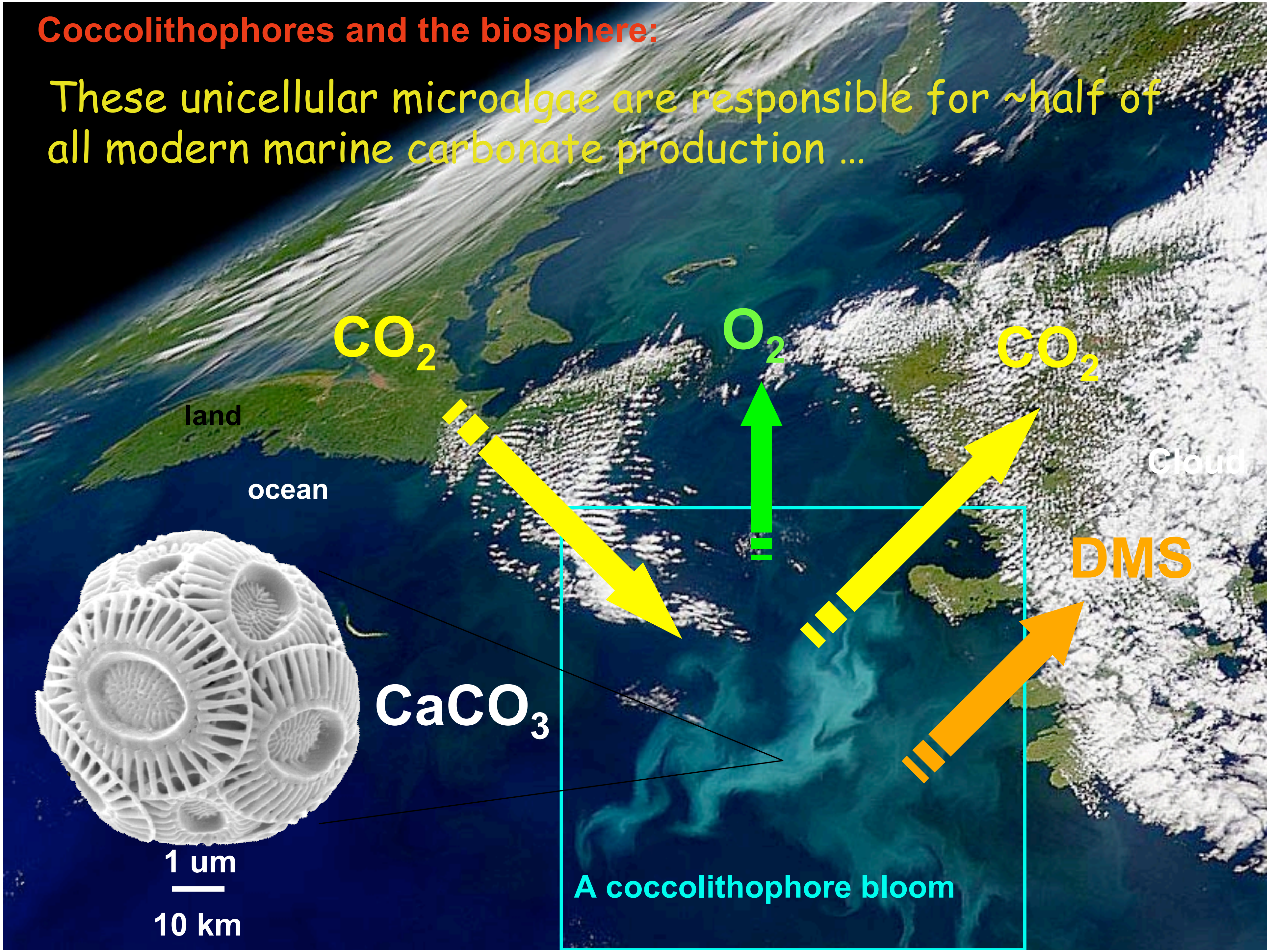
EMBL

The grand diversity of eukaryotic life:
Much more than animals, fungi, and plants!

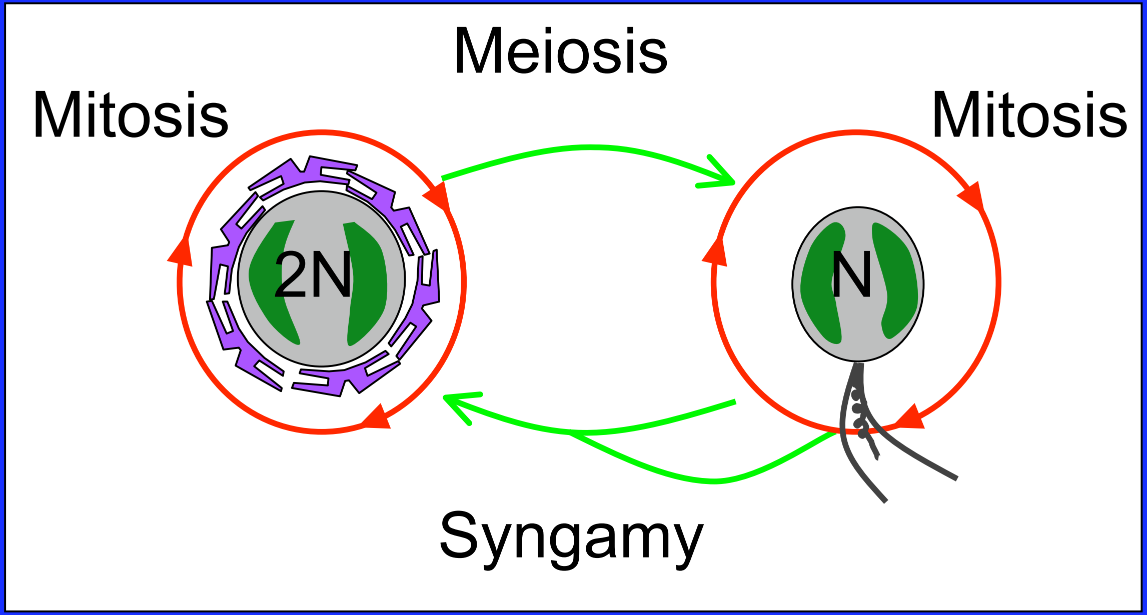


Coccolithophores and the biosphere:

These unicellular microalgae are responsible for ~half of all modern marine carbonate production …



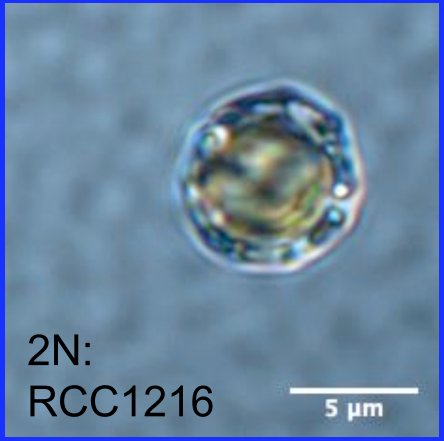
Emiliania huxleyi life cycle

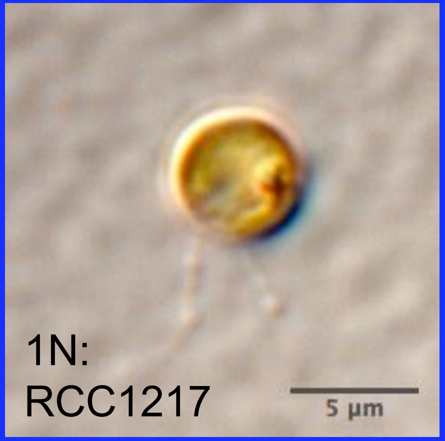


Diploid (2N):	Haploid (1N):
Non-motile	Motile
Calcified	Non-calcified
Forms massive blooms	**Doesn't bloom?**
Not photoinhibited	**Photoinhibited**
Killed by *EhVs*	**Resistant to *EhVs***

Frada et al., 2008

1N (motile) and 2N (calcified) strains with same parent





Generate Sanger and 454 EST libraries from both.

Compare to JGI's genome assembly of a different *E. huxleyi* strain, CCMP 1516

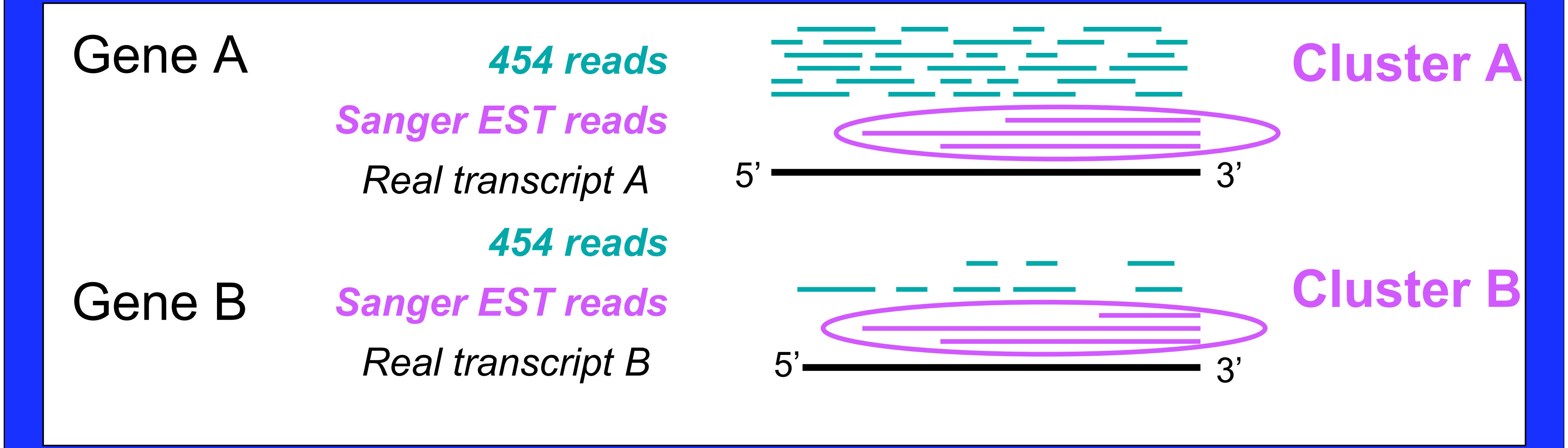
Two types of library combine sequence coverage and depth

Sanger-sequenced libraries

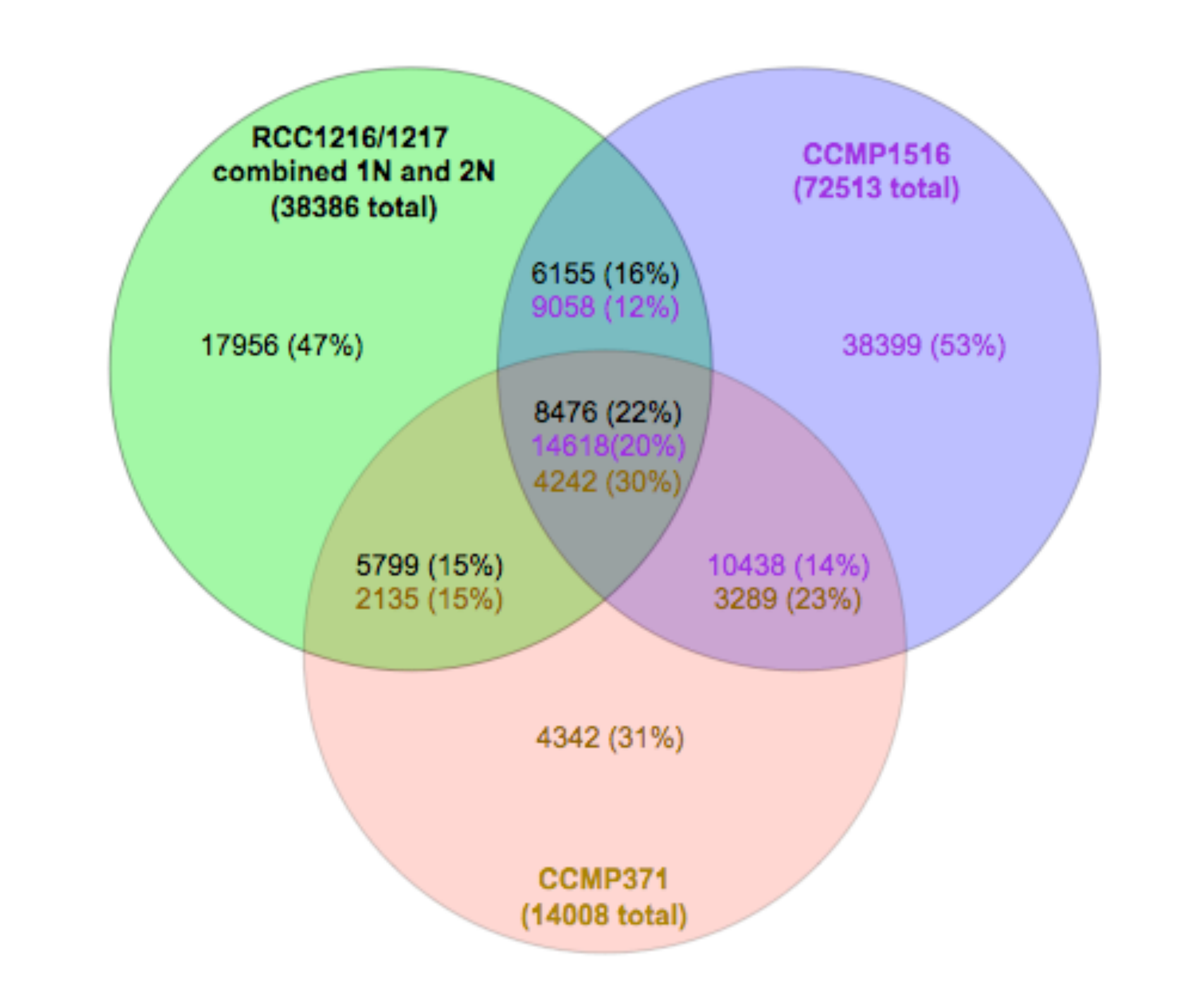
1. Oligo-dT primed cDNA
2. Normalized
3. ≈19000 longer reads

454-sequenced libraries

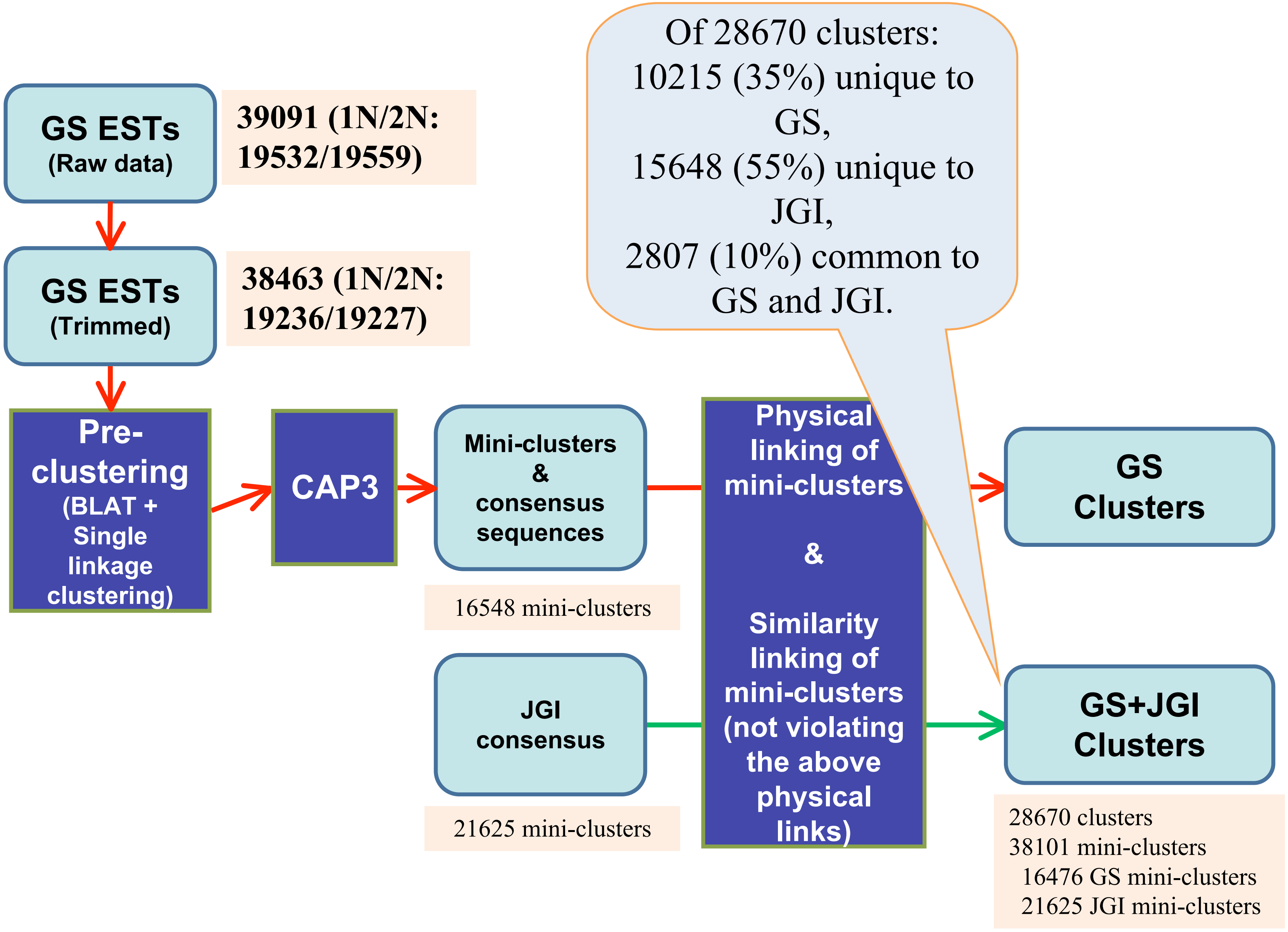
1. random primed cDNA
2. *NOT* normalized
3. ≈255000 shorter reads



Sanger ESTs reveal a large amount of new transcriptomic information for *Emiliania huxleyi*



Sanger read mapping/clustering statistics



GS ESTs
(Raw data)

39091 (1N/2N:
19532/19559)

GS ESTs
(Trimmed)

38463 (1N/2N:
19236/19227)

Pre-clustering
(BLAT +
Single
linkage
clustering)

CAP3

Mini-clusters
&
consensus
sequences

16548 mini-clusters

JGI
consensus

21625 mini-clusters

Physical
linking of
mini-clusters

&

Similarity
linking of
mini-clusters
(not violating
the above
physical
links)

GS
Clusters

GS+JGI
Clusters

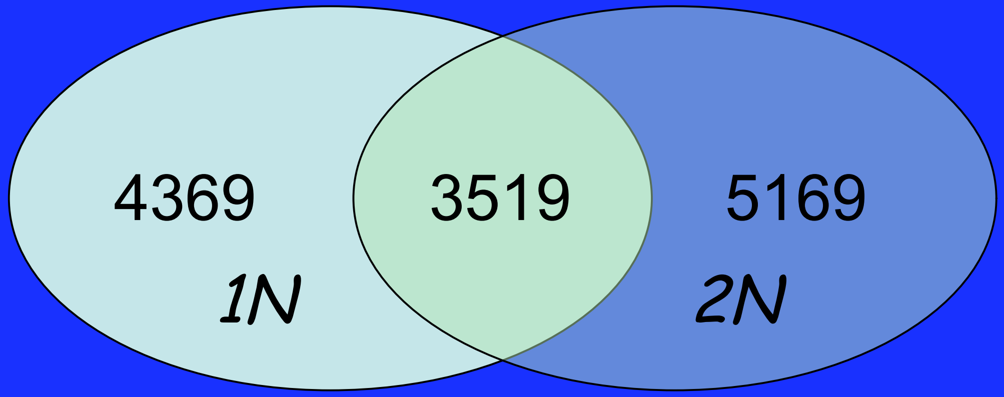
28670 clusters
38101 mini-clusters
16476 GS mini-clusters
21625 JGI mini-clusters

Of 28670 clusters:
10215 (35%) unique to GS,
15648 (55%) unique to JGI,
2807 (10%) common to GS and JGI.

Comparing alternate phases of the life cycle greatly increases transcripts detected

39000 Sanger ESTs:

13057 clusters total

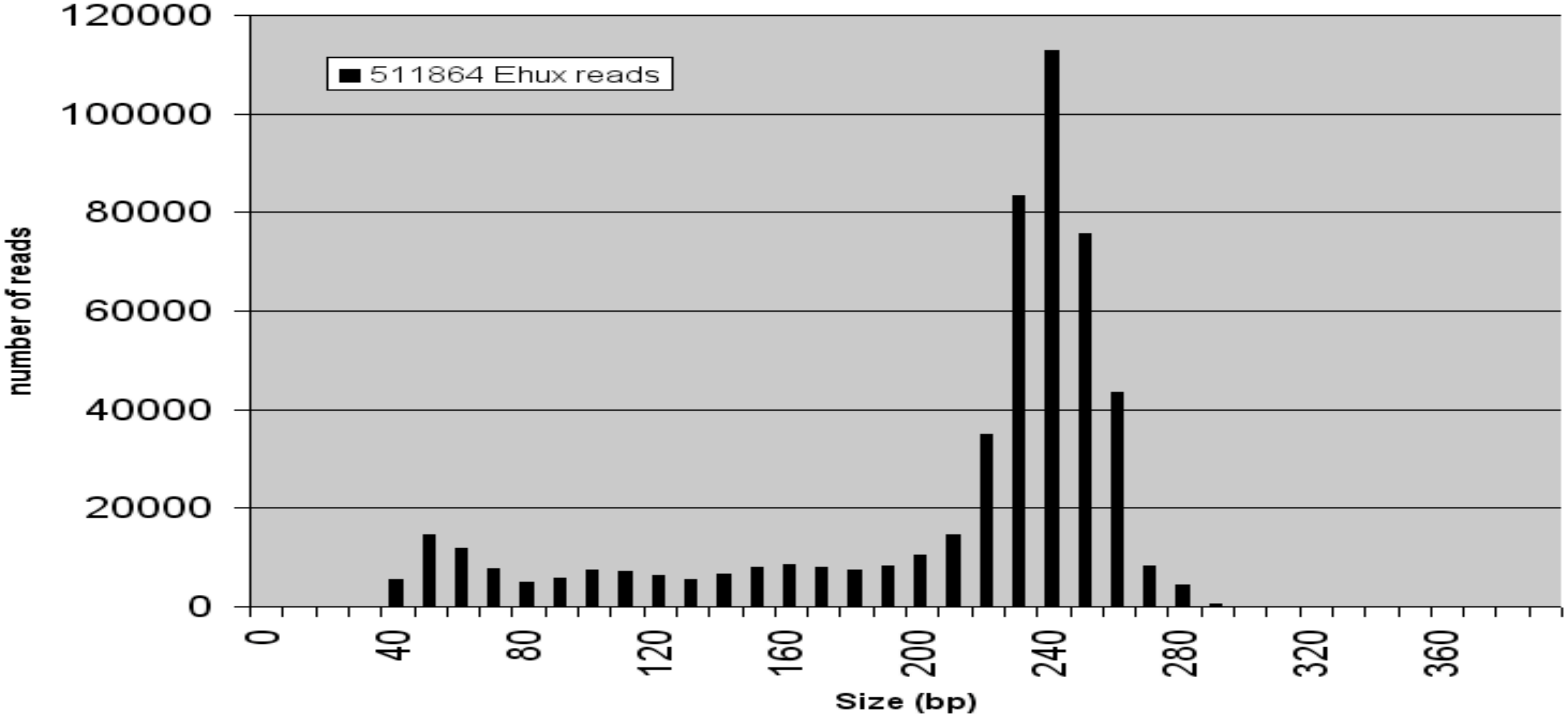


	1N	2N
# clusters	7888	8688
ML estimate of transcriptome richness	10039	11988
Chao1 estimate of transcriptome richness	12840±214 (12438,13278)	15931±289 (15385,16522)
Coverage	61.4-78.6%	54.5-72.5%

Chao Jaccard-type similarity: ≤50% of expressed genes shared!

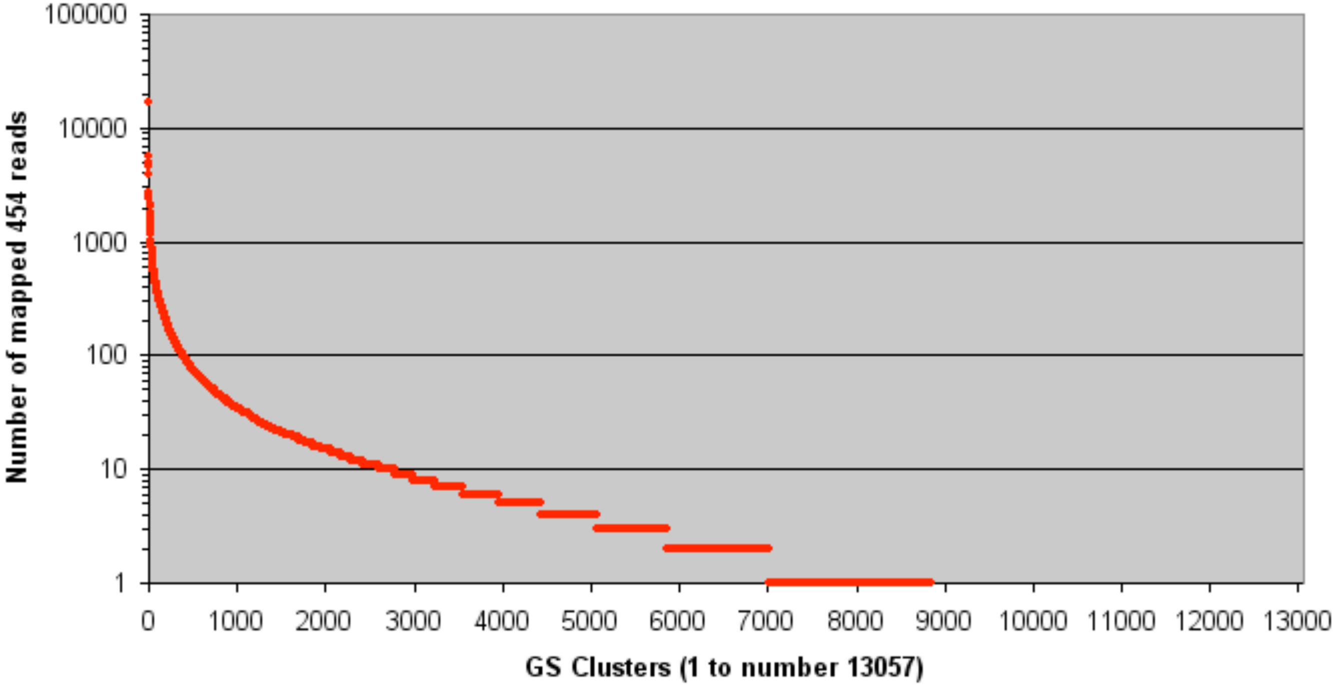
Ehux 454 read data

- 1N (file: FGFGJ1101)
  - 256484 reads
  - Average size: 216.30 (S.D.: 59.09)
  - G+C: 62.76%
- 2N (file: FG5FMAE01)
  - 255380 reads
  - Average size: 209.84 (S.D.: 62.23)
  - G+C: 62.55%



Mapping of reads on Genoscope ESTs

- (Simple criteria) BLAT, AL>=90nt, Identity>=95%, Coverage of read by alignment>=70%
- Of 511864 reads, 262023 (51%) were mapped on the Ehux mini-clusters derived from Genoscope EST data sets
- Corresponding ESTs
  - 10611 mini-clusters out of 16471 total mini-clusters (64%)
  - 8844 clusters out of 13057 total clusters (68%)

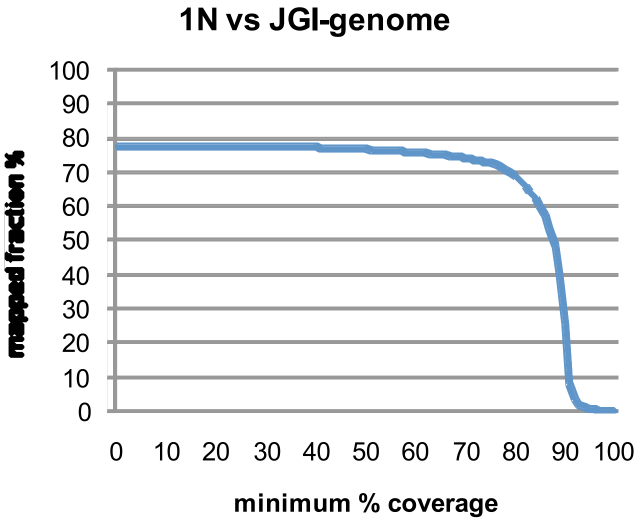


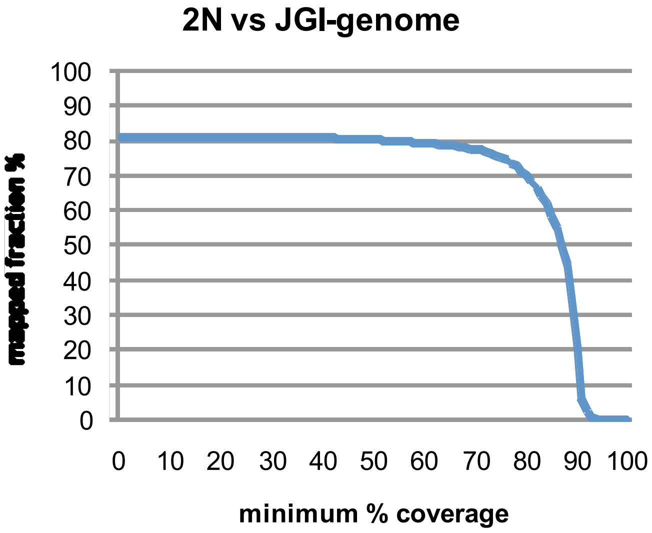
Ehux 454 reads mapping statistics
BLAT + alignment length≥90nt
by minimum %-coverage of read

1N

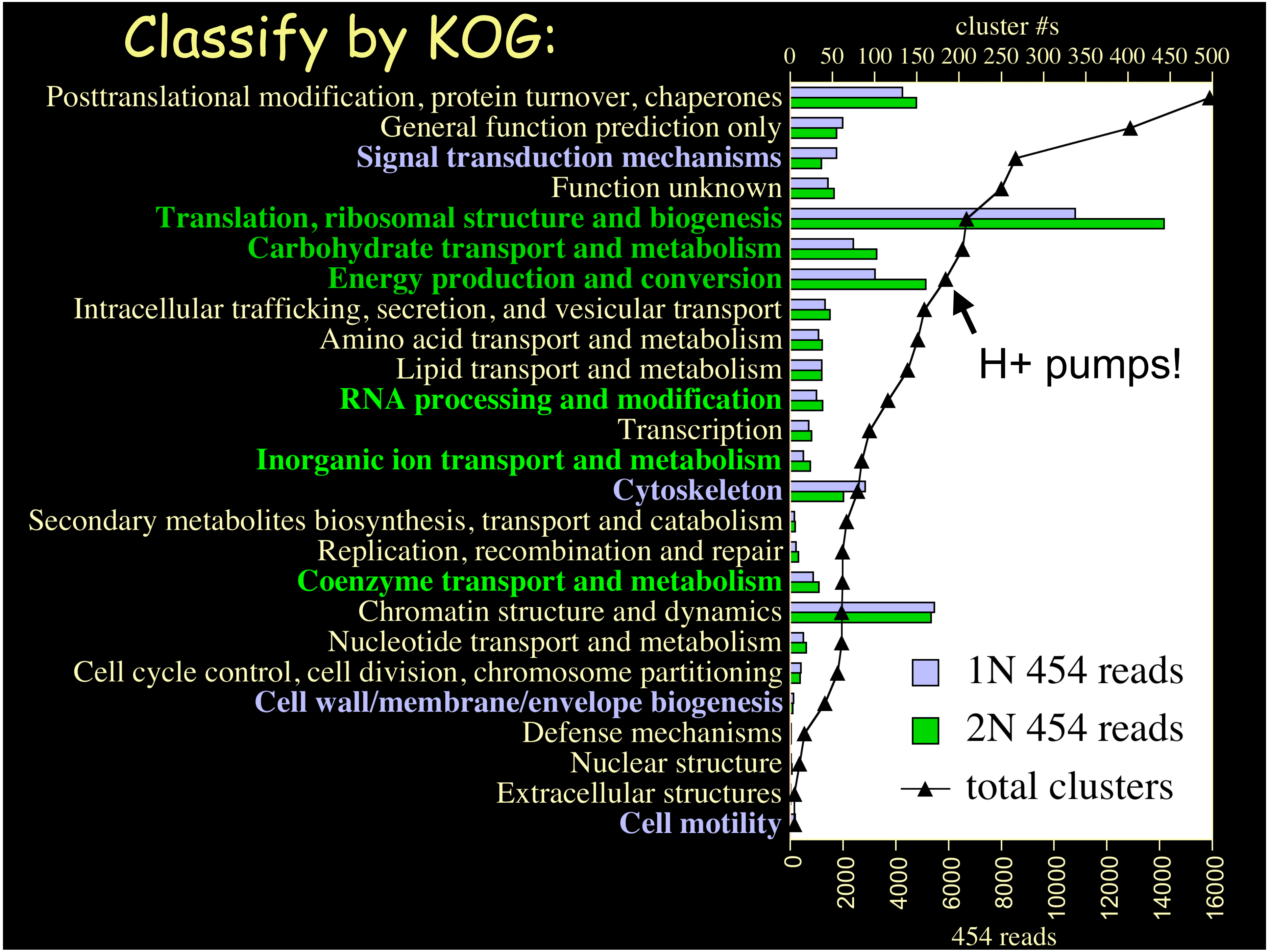
2N

Against JGI genomic sequences



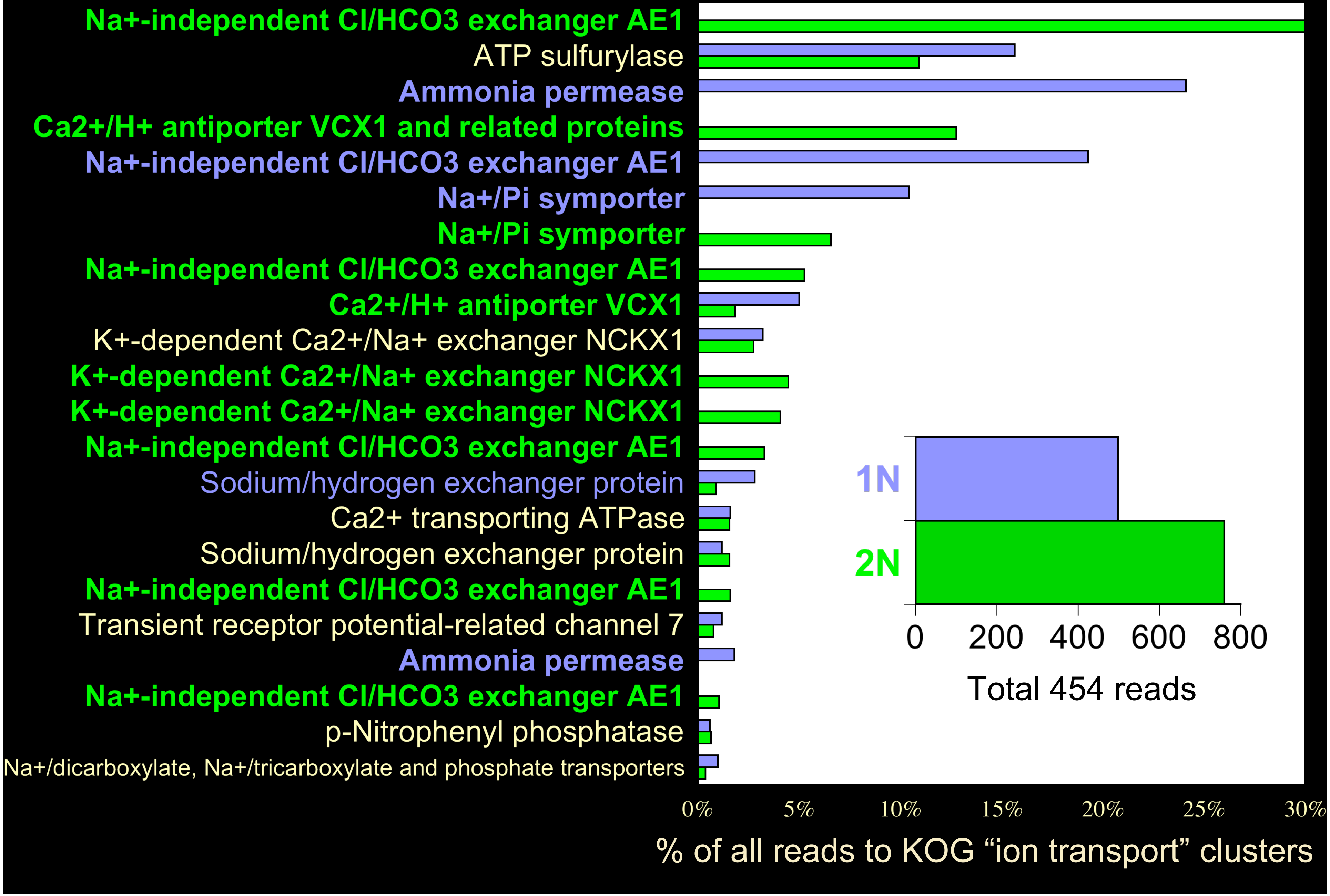


Classify by KOG:

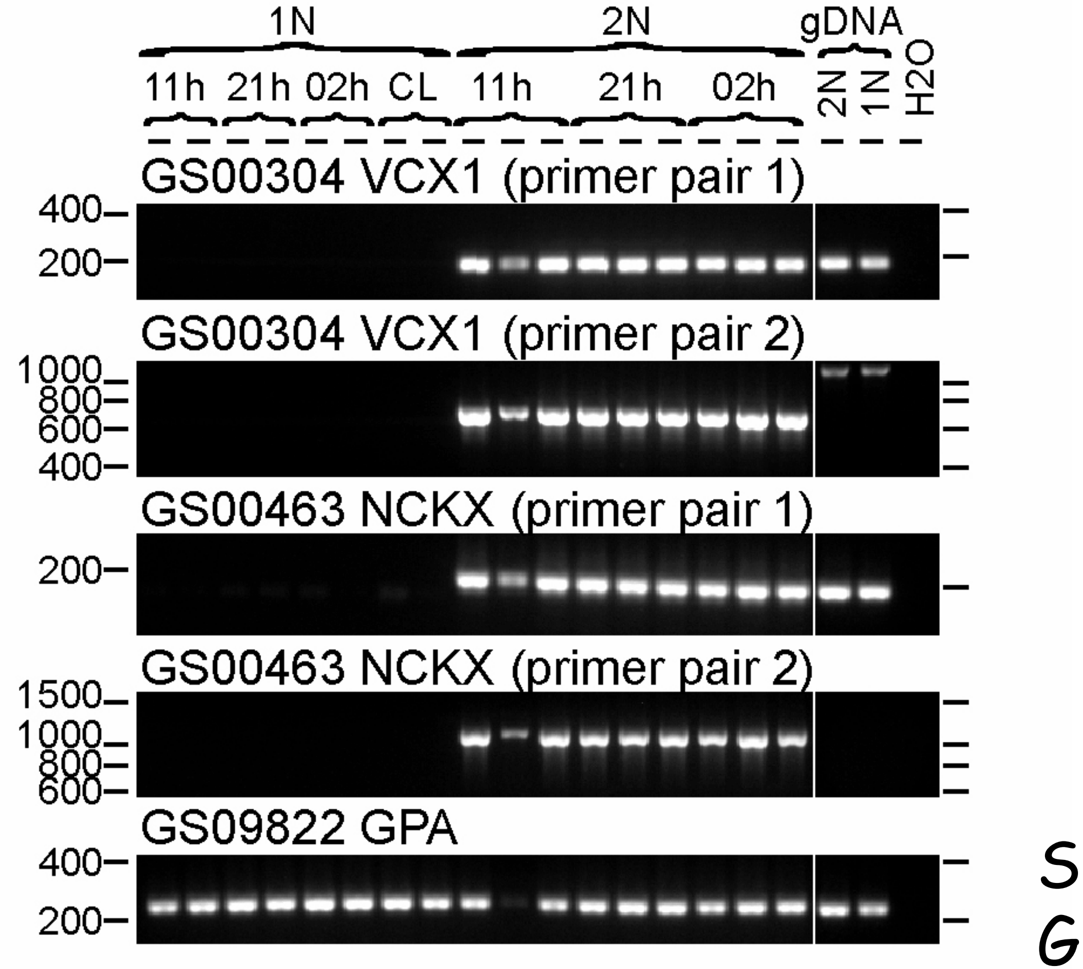


H+ pumps!

Calcifying 2N cells pump ions more



RT-PCR tests of potential calcification genes


1N
2N
gDNA
11h
21h
02h
CL
11h
21h
02h
2N
1N
H2O
GS00304 VCX1 (primer pair 1)
400
200
GS00304 VCX1 (primer pair 2)
1000
800
600
400
GS00463 NCKX (primer pair 1)
200
GS00463 NCKX (primer pair 2)
1500
1000
800
600
GS09822 GPA
400
200

Confirm highly 2N-specific expression of a VCX1 and NCKX gene

Surprise!
GPA is not 2N-specific!
9 454 reads from 1N
Only 1 454 read from 2N

Motile 1N cells: Flagella and sensor systems

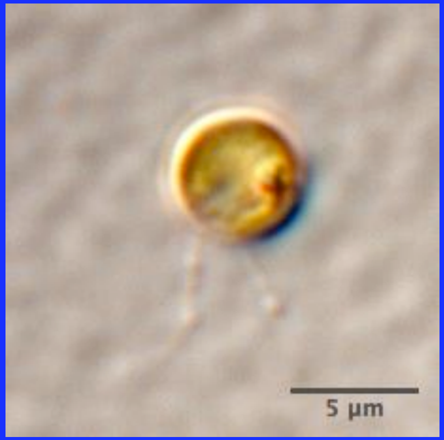
154 flagellar-related clusters

Expressed <u>only</u> in 1N cells:

86 flagellar or basal body structural elements
with no known cytoplasmic role

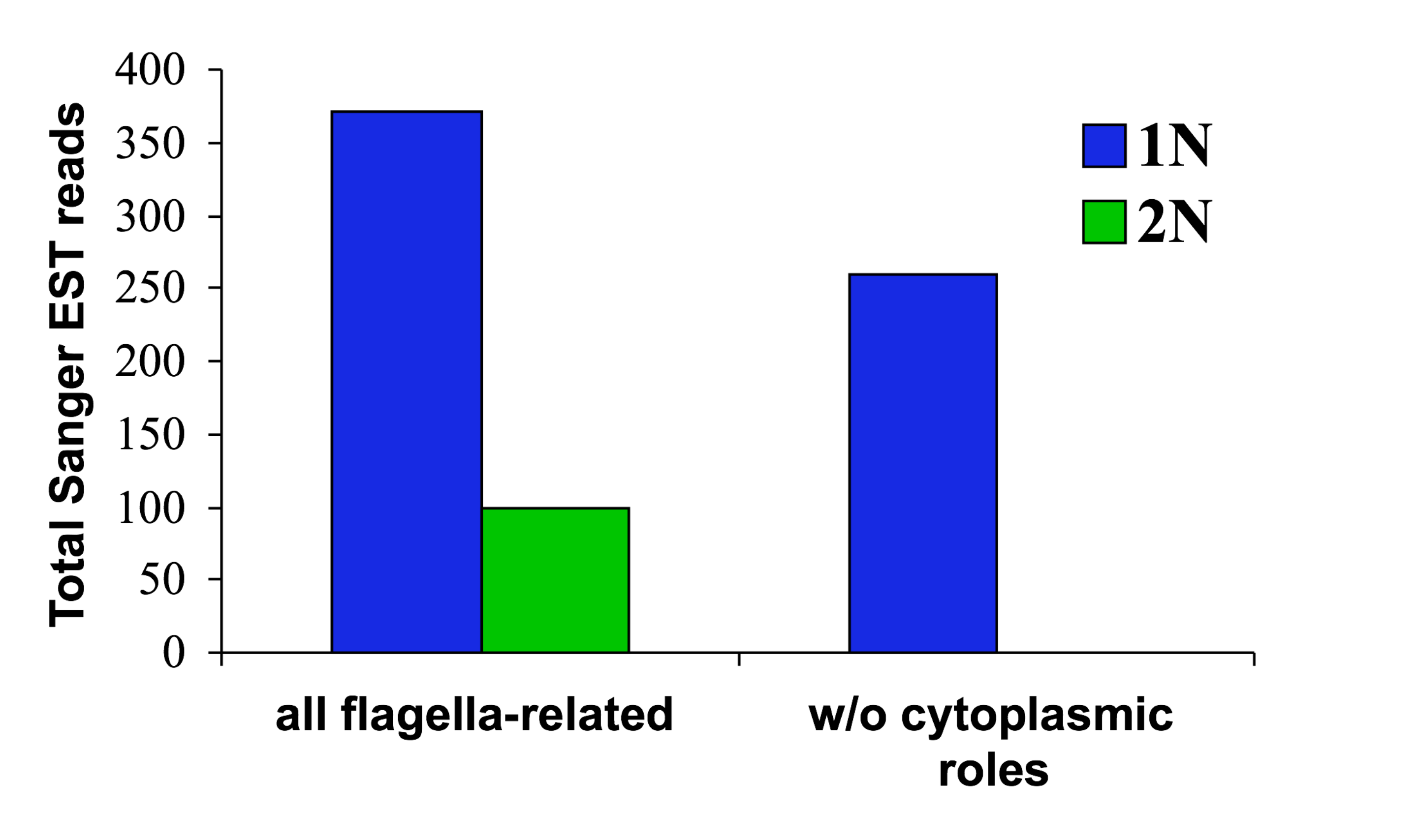
13 flagellar dynein heavy chains

1 cytoplasmic dynein heavy chains



Sensors:
Two 1N-specific phototropin-like LOV2 proteins
1N-specific cGMP protein kinase homolog

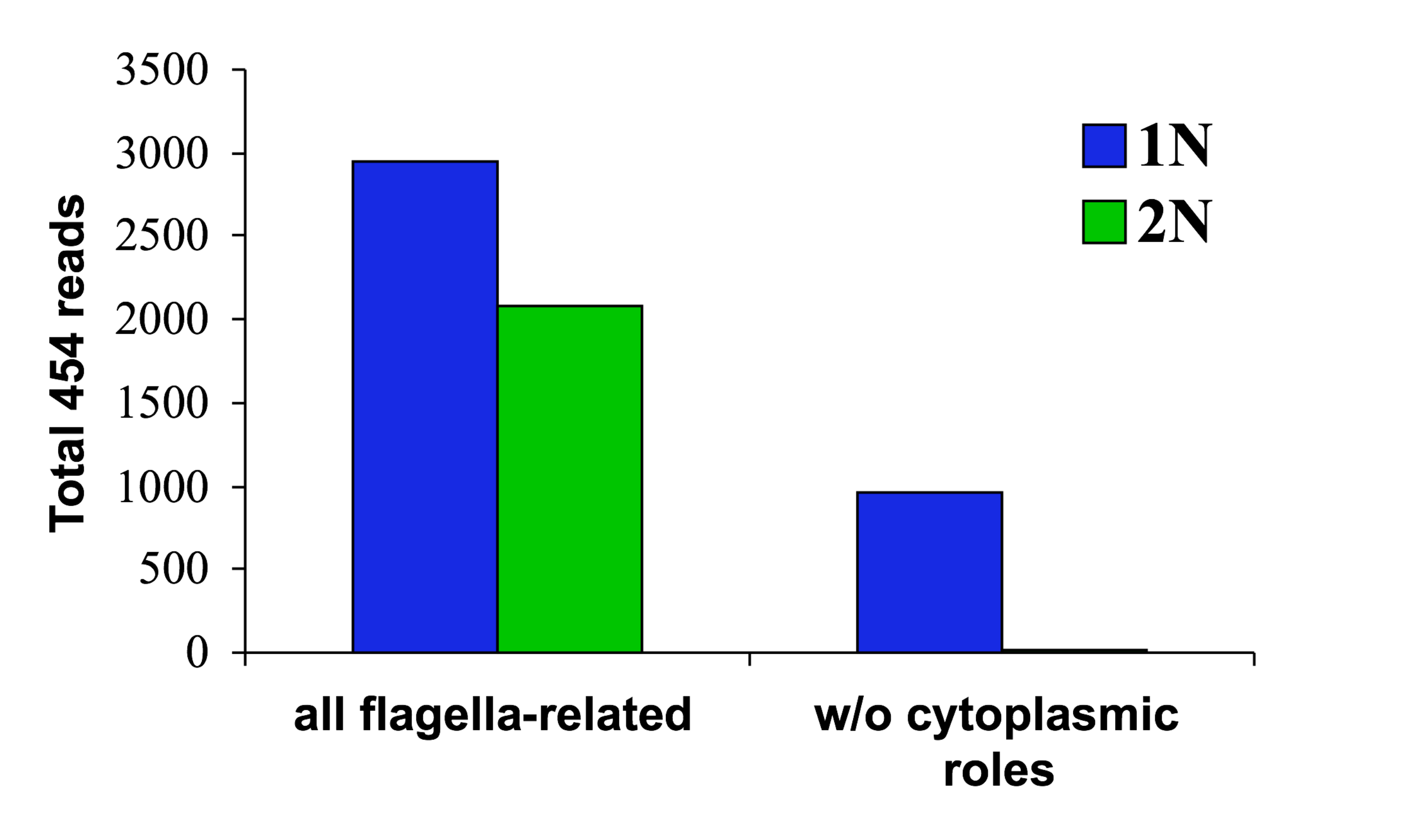
Flagellar-related transcripts are 1N-specific



156 clusters

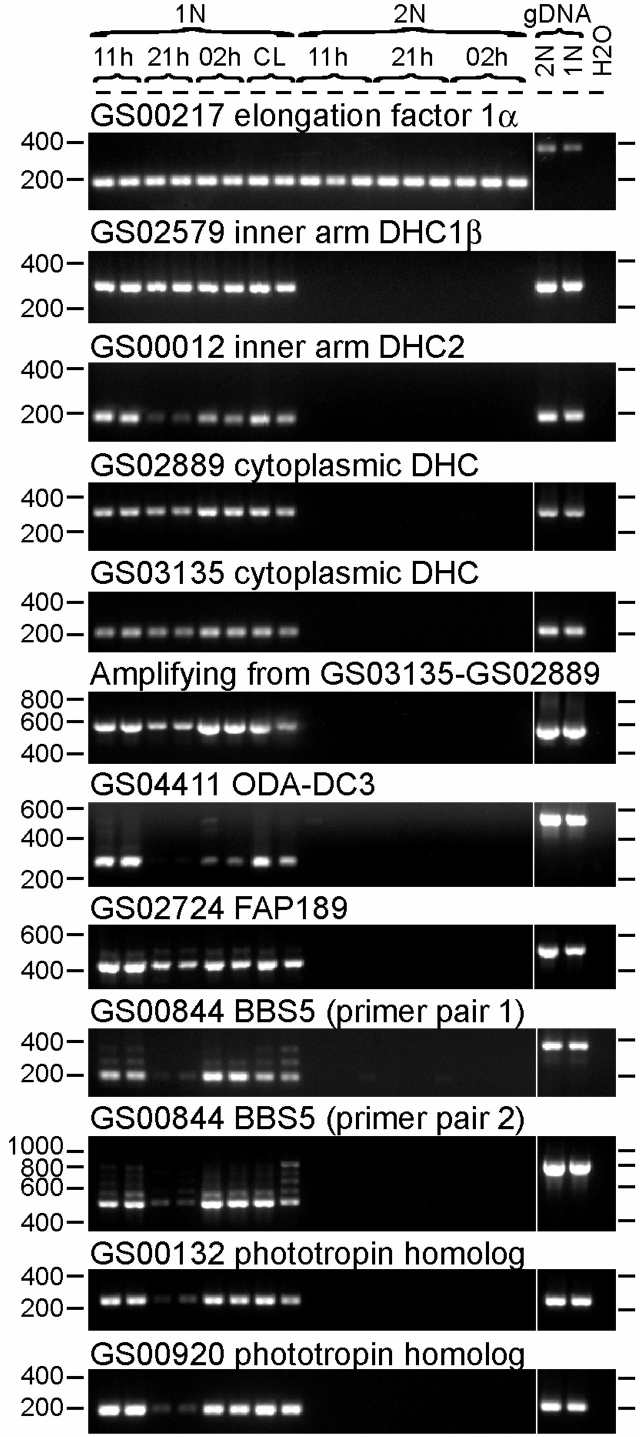
85 clusters

Flagellar-related transcripts are 1N-specific



156 clusters

85 clusters


1N
2N
gDNA
11h 21h 02h CL 11h 21h 02h
2N 1N H2O
GS00217 elongation factor 1α
400
200
GS02579 inner arm DHC1β
400
200
GS00012 inner arm DHC2
400
200
GS02889 cytoplasmic DHC
400
200
GS03135 cytoplasmic DHC
400
200
Amplifying from GS03135-GS02889
800
600
400
GS04411 ODA-DC3
600
400
200
GS02724 FAP189
600
400
GS00844 BBS5 (primer pair 1)
400
200
GS00844 BBS5 (primer pair 2)
1000
800
600
400
GS00132 phototropin homolog
400
200
GS00920 phototropin homolog
400
200

RCC1216/1217 has a lot of genes not found in the JGI whole genome sequence of CCMP 1516!

	Combined (1N+2N)	Highly 1N-specific (>10x difference $p < 10^{-5}$)	Highly 2N-specific (>10x difference $p < 10^{-5}$)
Sanger clusters	22.6%	33.3%	23.9%
454 reads	14.9%	39.0%	20.6%

Flagellar genes lost!

	RCC1216/1217	CCMP1516
Calcification	Well calcified	Poorly/non-calcified
Life cycle	Forms motile 1N cells	Does NOT form motile 1N cells
G1 DNA content	≈4x Isochrysis	≈2.9x Isochrysis

Genomic variation between *Ehux* strains

22.6% of Sanger-sequenced clusters do not map to JGI genome assembly!

47% of flagellar-associated genes do not map

5 out of 13 distinct dynein heavy chain genes do not map

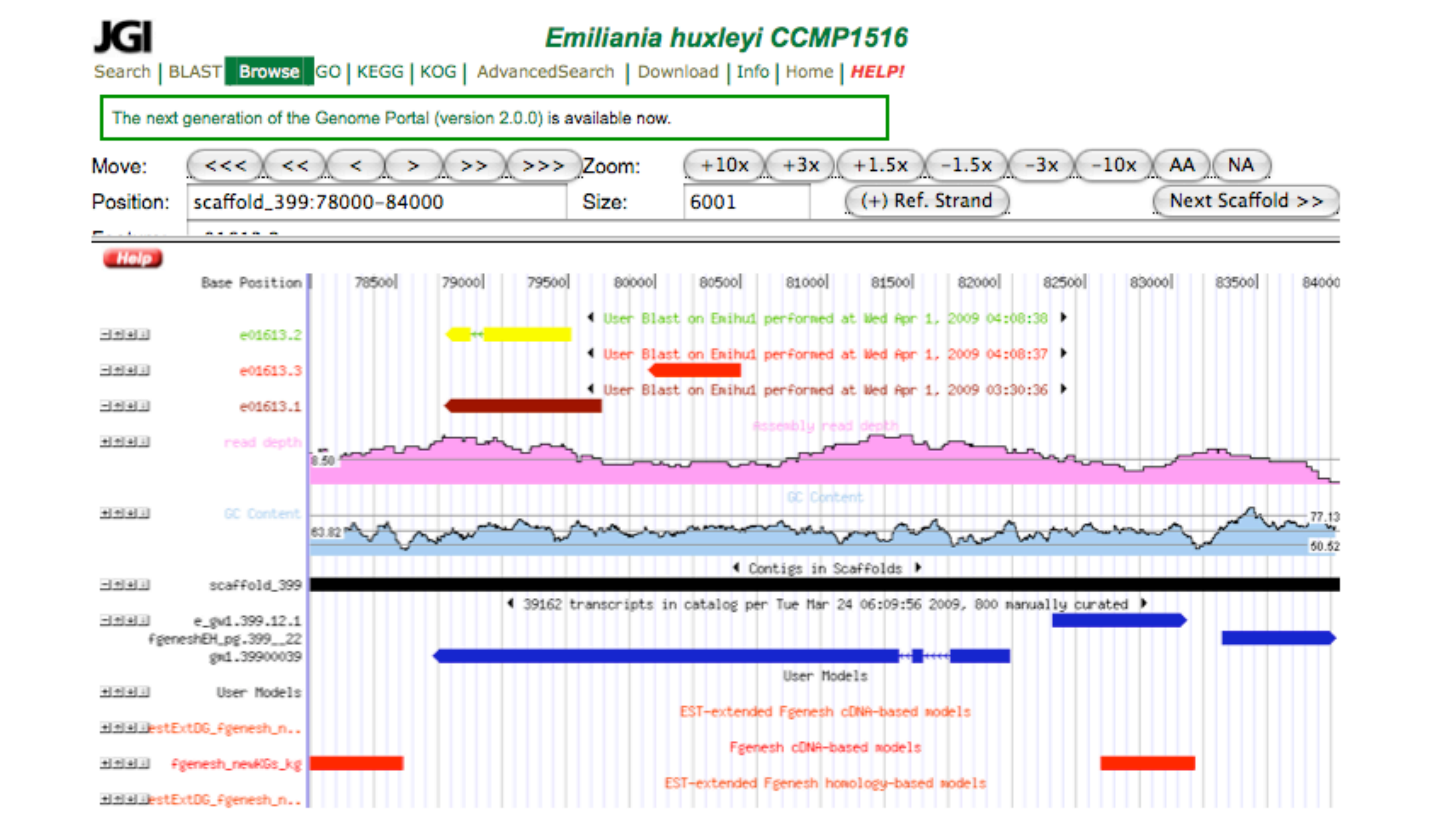
Only three loci in JGI/CCMP 1516 assembly have sufficient space to encode the ≈4000 aa dynein heavy chain genes

Several regions of fragmented homology

CCMP 1516 never observed to form flagellated cells

PCR suggests other strains may lose flagellar dyneins too!

The CCMP1516 genome assembly does not have room to encode 4000 amino acid dynein heavy chains?



The next generation of the Genome Portal (version 2.0.0) is available now.

Position: scaffold_399:78000-84000

Size: 6001

User Blast on Emihu1 performed at Wed Apr 1, 2009 04:08:38

User Blast on Emihu1 performed at Wed Apr 1, 2009 04:08:37

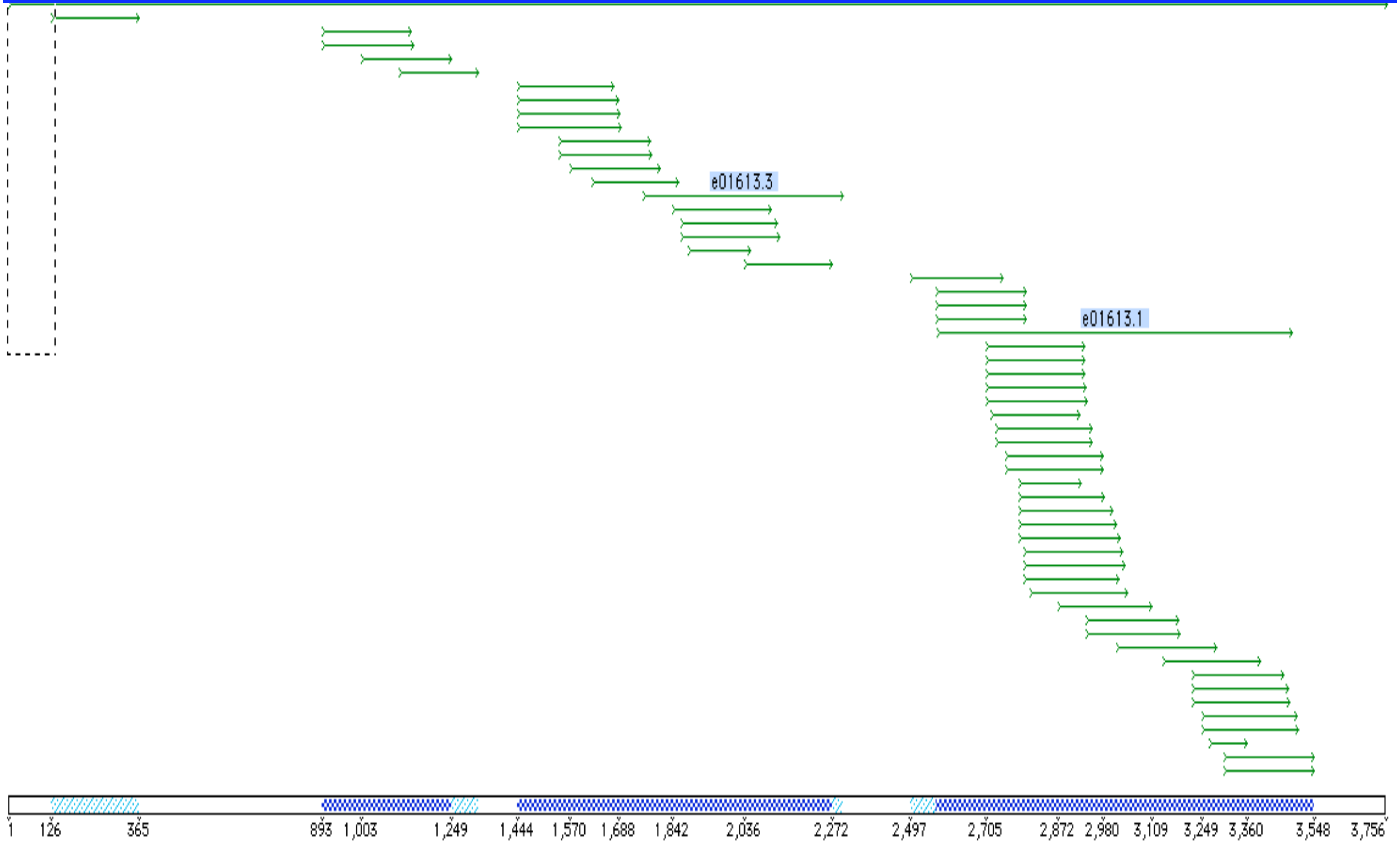
User Blast on Emihu1 performed at Wed Apr 1, 2009 03:30:36

Contigs in Scaffolds

39162 transcripts in catalog per Tue Mar 24 06:09:56 2009, 800 manually curated

Can we use 454 reads to extend transcript models when part of the gene is missing from the assembly??

454 read distribution across dynein heavy chain genes



Summary of results so far:

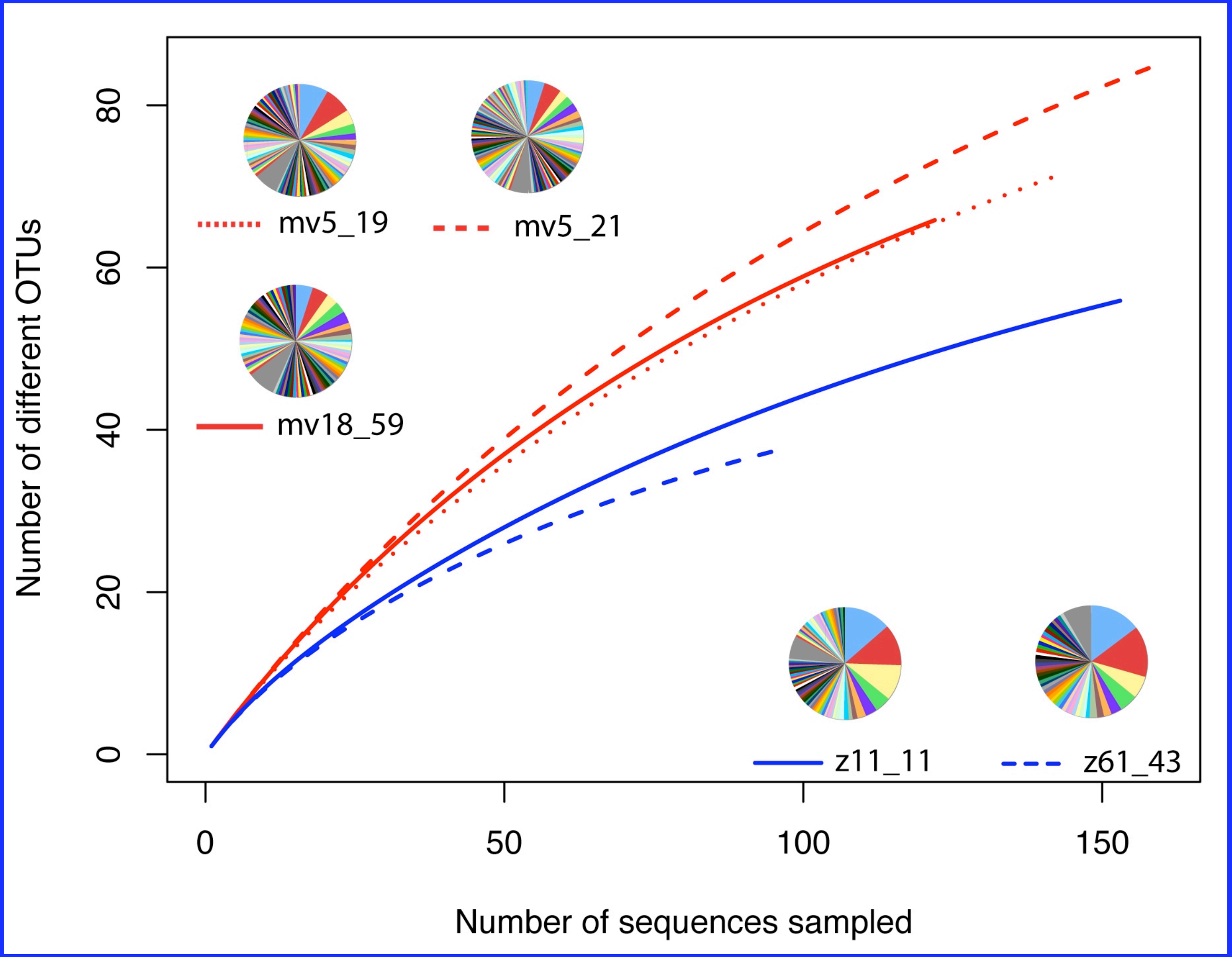
1. Combination of 454 and Sanger technology and proper biology allows deep comparison of transcriptomes
2. We discovered the JGI genome assembly is selectively missing haploid genes!
3. We have to rely much less than planned on existing genome assembly
4. Hybrid 454-Sanger transcript models might help to retrieve missing information

What we still need to do and to figure out:

1. Microarray verification (new arrays ready)
2. Finish mapping 454 reads to JGI genome
3. Create hybrid 454-Sanger clusters with and without using JGI's genome assembly
4. Global statistical description of transcriptomes and their differences based on 454:
   1. Chao1, ML, and Shannon-diversity estimates of transcriptome complexity
   2. Chao Jaccard-type estimates of transcriptome differences
   3. Account for gene or cluster length in statistical analyses of 454 data

Next use of 454 sequencing by Roscoff team: Completely sample protist diversity in the ocean!

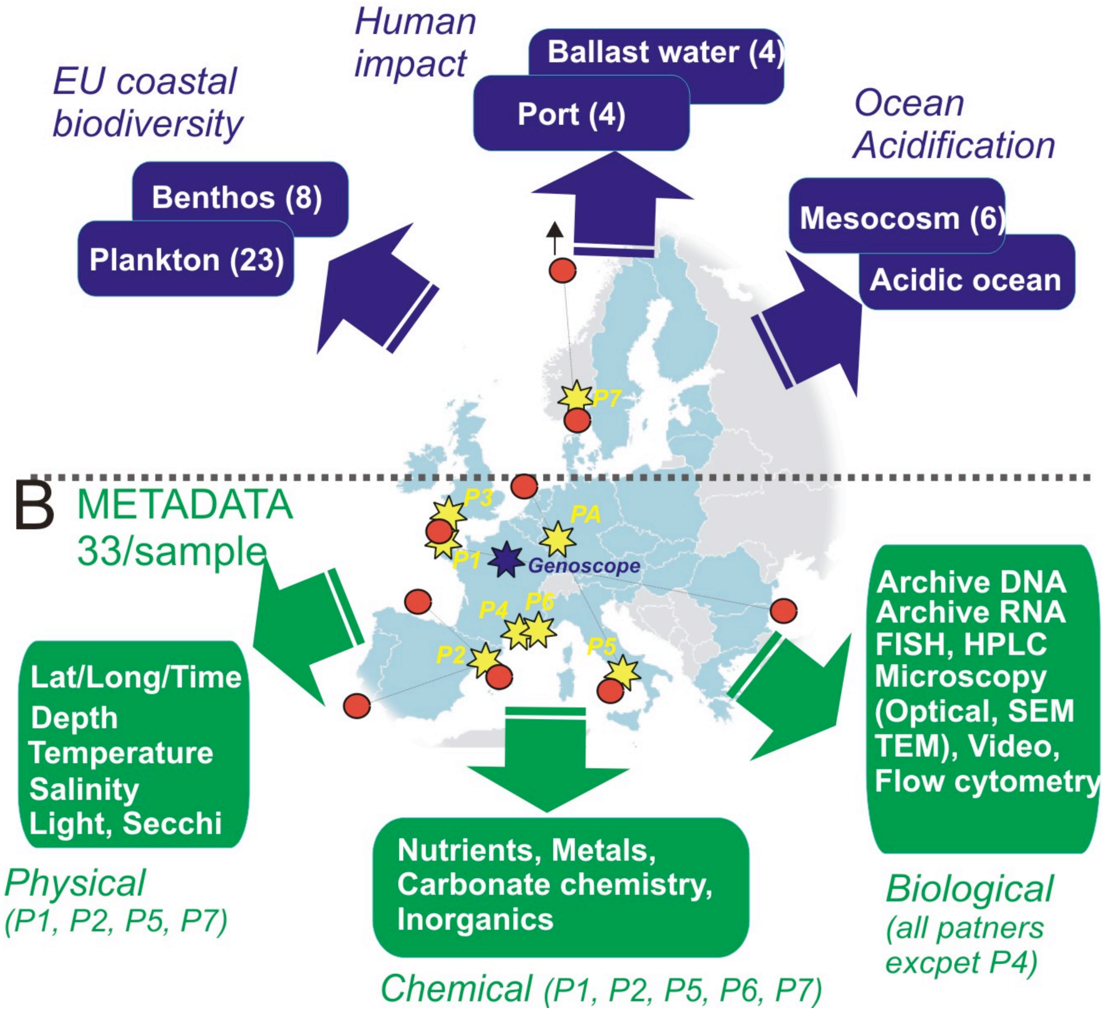
~200 28S rDNA clones per sample: far from saturation



BioMarKs

Biodiversity of Marine euKaryoteS

A 454-SEQUENCING
49 discrete samples, 45 *454*-runs



EU coastal biodiversity

Benthos (8)

Plankton (23)

Human impact

Ballast water (4)

Port (4)

Ocean Acidification

Mesocosm (6)

Acidic ocean

B METADATA
33/sample

Lat/Long/Time
Depth
Temperature
Salinity
Light, Secchi

Physical
(P1, P2, P5, P7)

Nutrients, Metals,
Carbonate chemistry,
Inorganics

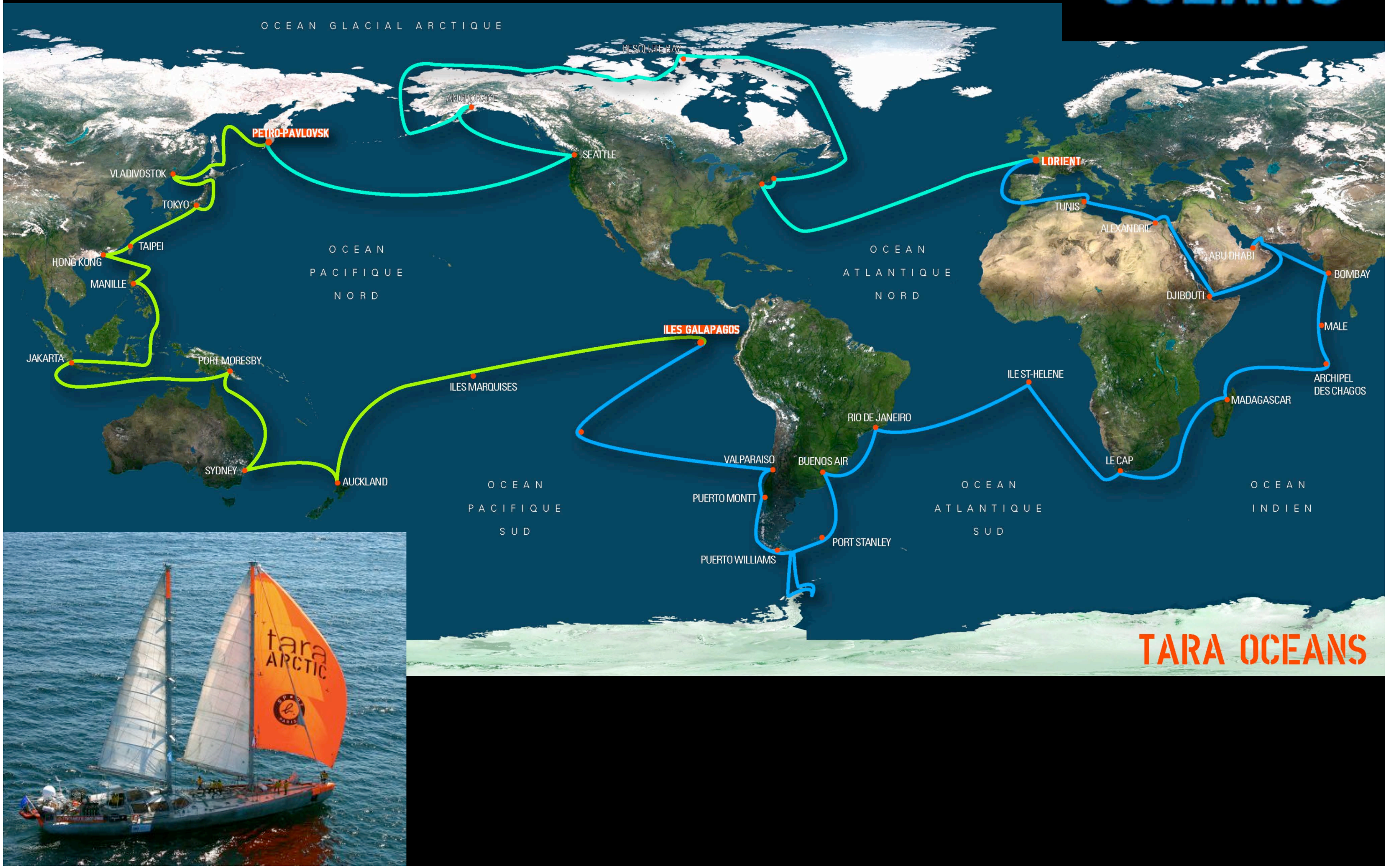
Chemical (P1, P2, P5, P6, P7)

Archive DNA
Archive RNA
FISH, HPLC
Microscopy
(Optical, SEM
TEM), Video,
Flow cytometry

Biological
(all patners
excpet P4)

POSEIDON

TARA OCEANS



Fin!

Merci!