DE LA RECHERCHE À L'INDUSTRIE

Colloque EPGV

Eucaryote genome annotation at the Genoscope

Marion Dubarry

4th October

Centre National de Séquençage

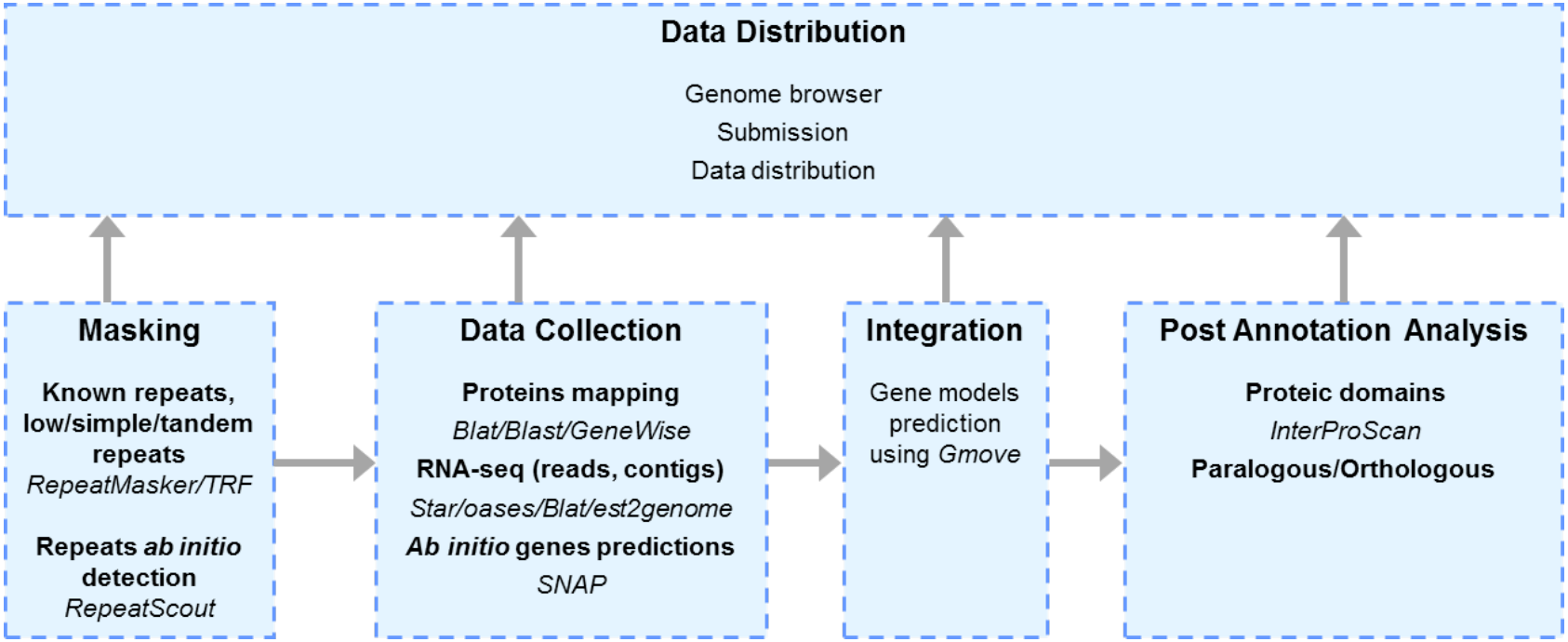
GENOSCOPE

www.cea.fr

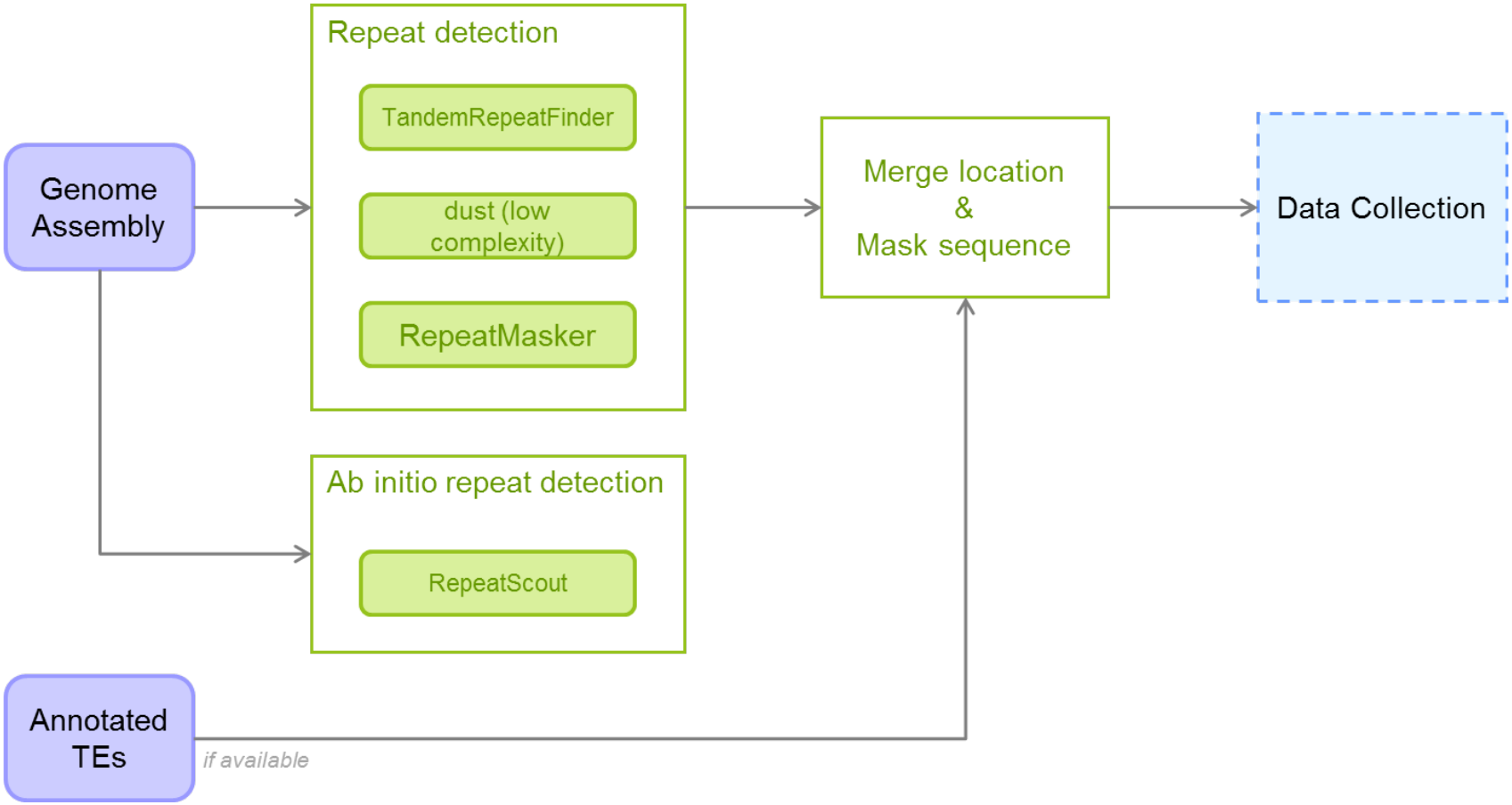
Genoscope

- French sequencing center, created in 1997 and part of the CEA / Institut de Génomique since 2007
- Provide high-throughput sequencing data to the Academic community, and carry out in-house genomic projects
- Focus on biodiversity : *de novo* sequencing and metagenomics projects
- Coordination of large sequencing projects like Tara Oceans

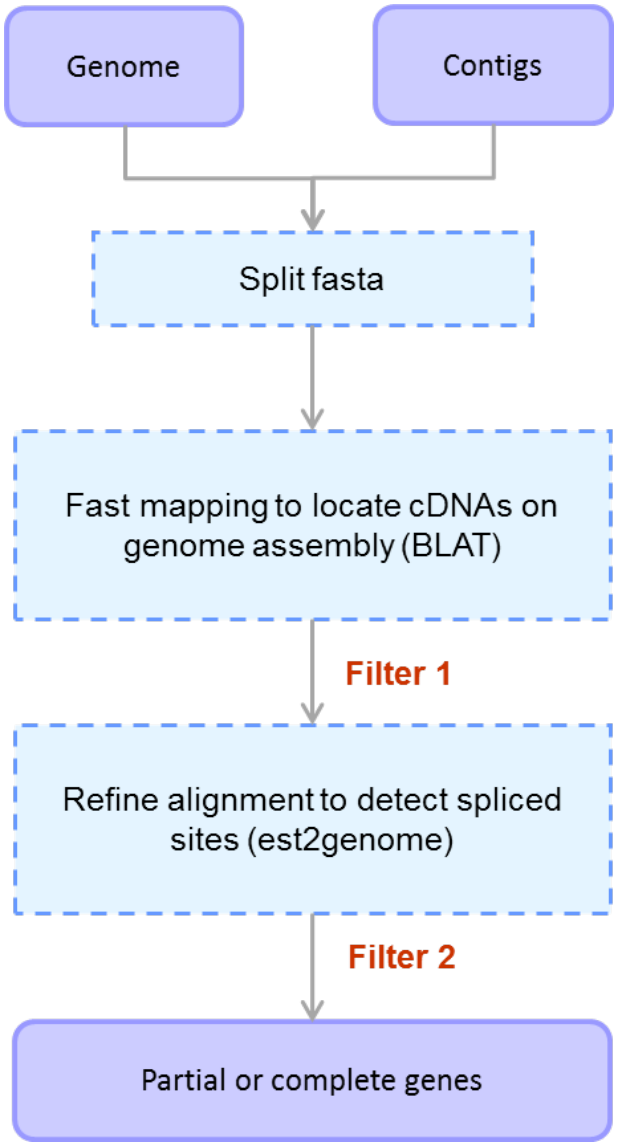
Annotation pipeline at the Genoscope



Masking
annotation pipeline

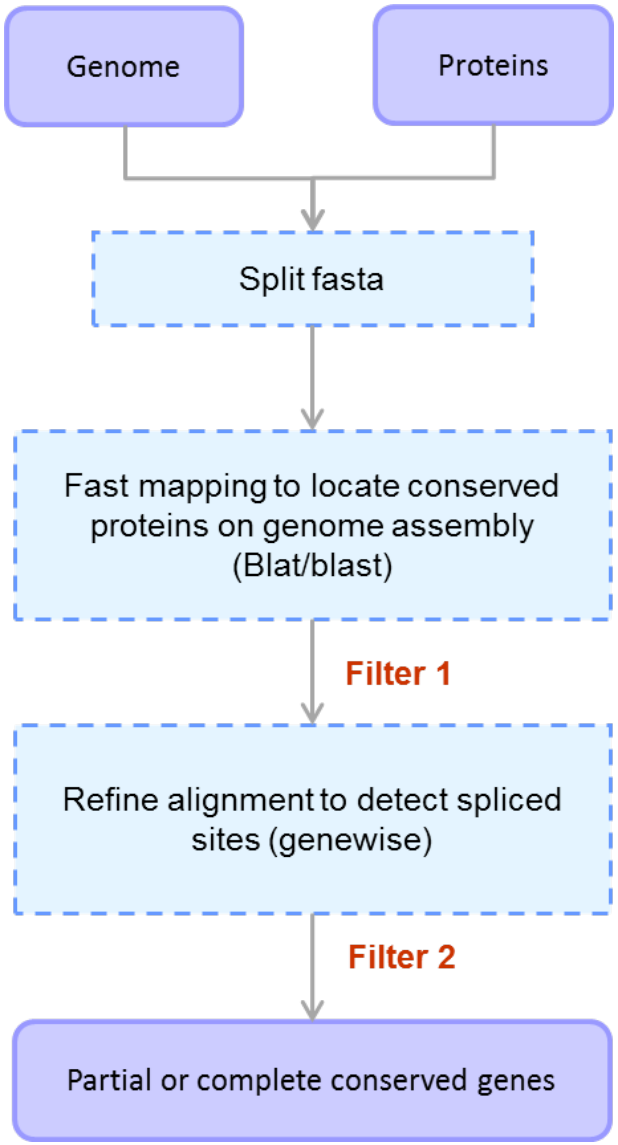


RNA mapping
annotation pipeline



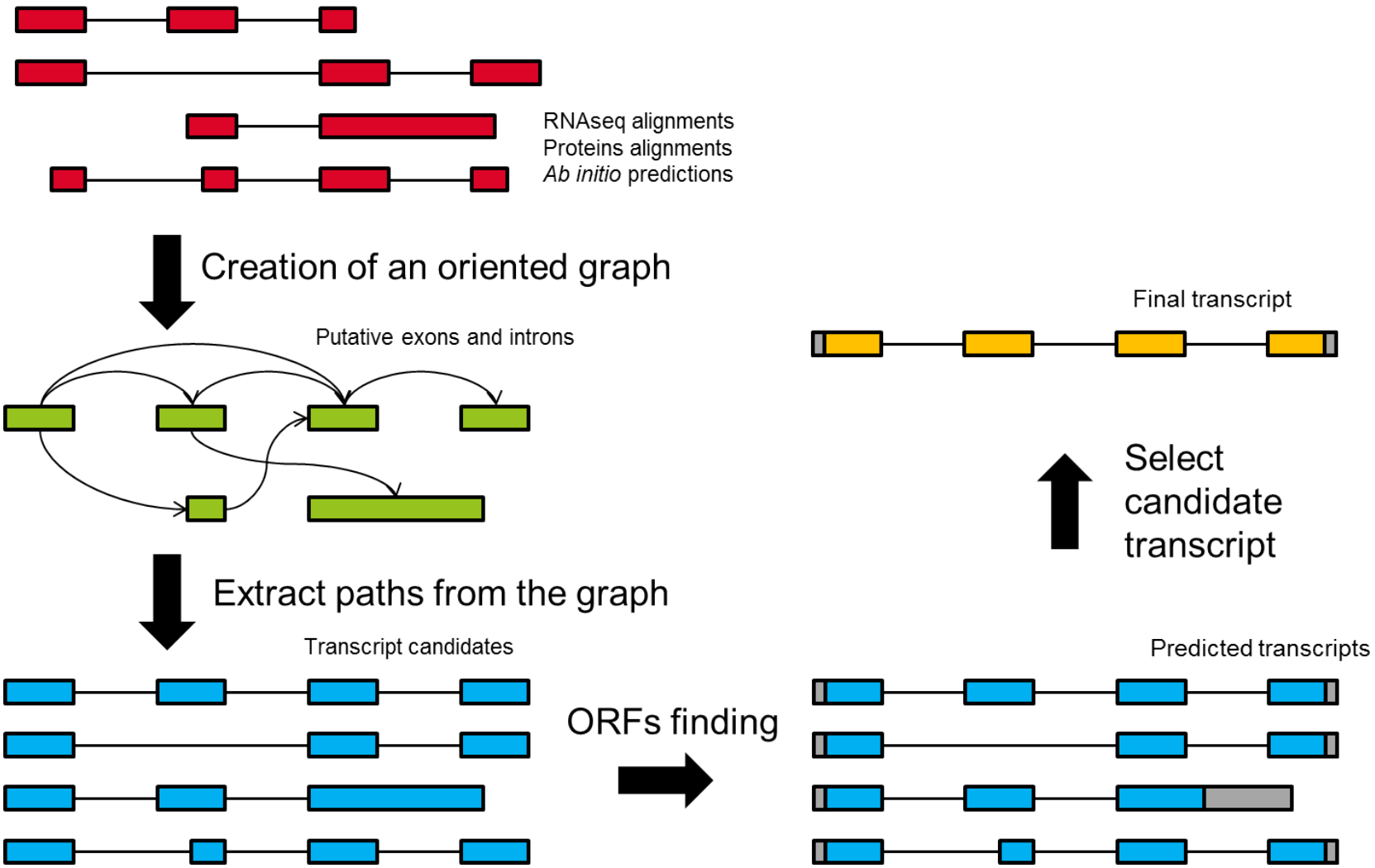
- reads assembly using Oases
- Filter 1
  - Best match (blat score) / contig
  - Identity percent $\geq$ 90%
  - Introns $\geq$ 100kb are splitted
- Filter 2
  - Identity percent $\geq$ 90%
  - Length ratio of aligned contig $\geq$ 85%

Protein mapping
annotation pipeline



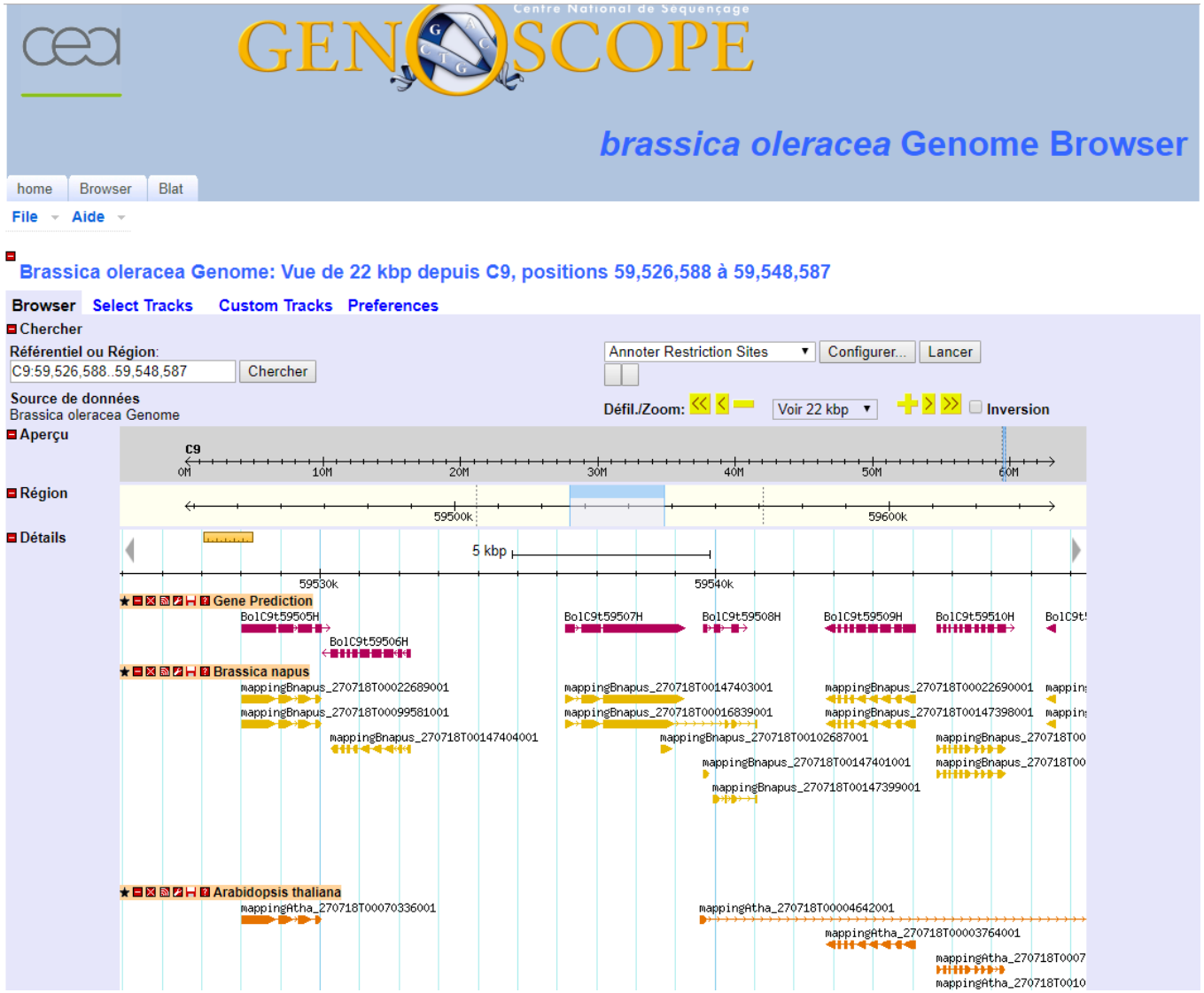
- homologous protein sequences (from UniProt)
- Protein sequences are masked (low complexity) using seg
- Filter 1
  - Best match (blat score) / contig
  - Matches with a score within 90% of the BM score
  - Introns $\geq$ 100kb are splitted
- Filter 2
  - Length ratio of aligned protein $\geq$ 50%

Gmove
annotation pipeline



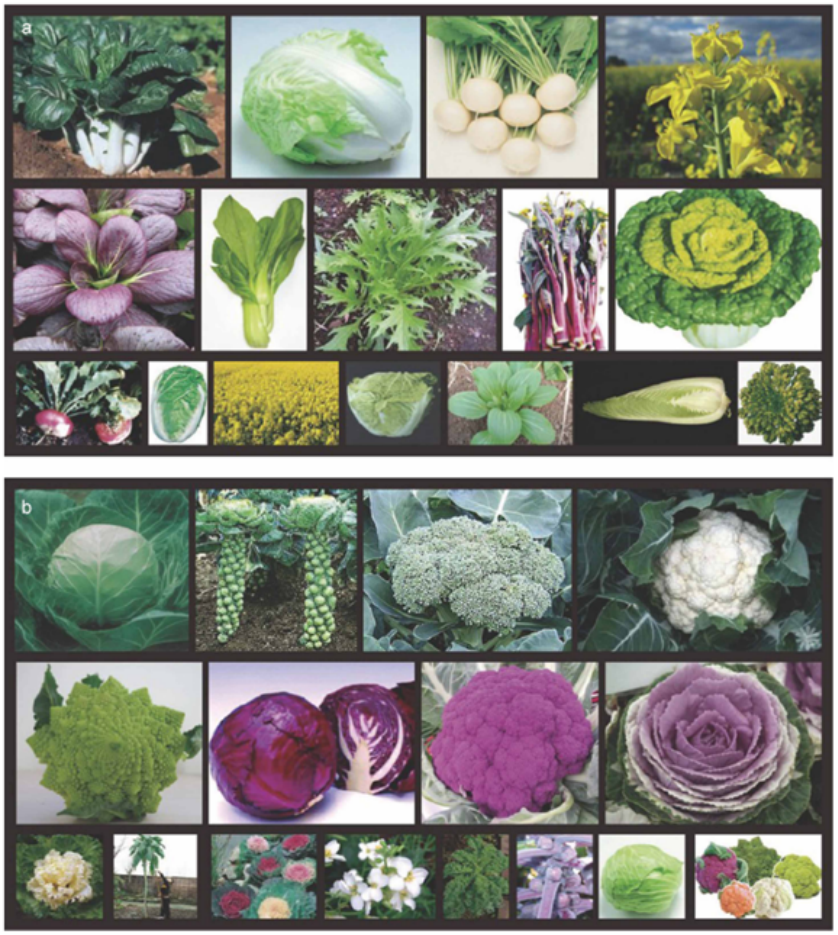
. github.com/institut-de-genomique/gmove

Annotation
Genome Browser



Brassicas and Musas

Example of a Project



Cheng et al. Nature 2014
Cirad

Brassicas and Musas
Annotation

several proteoms were used to annotate

- B.rapa
  - B.rapa Chiifu
  - B.napus
  - A.thaliana
  - panTranscriptome B.oleracea B.napus
- B.oleracea
  - B.oleracea TO1000
  - B.napus
  - A.thaliana
  - panTranscriptome B.oleracea B.napus
- M.schizocarpa
  - M.acuminata
  - O.sativa
  - P.dactyliphera

Brassicas and Musas

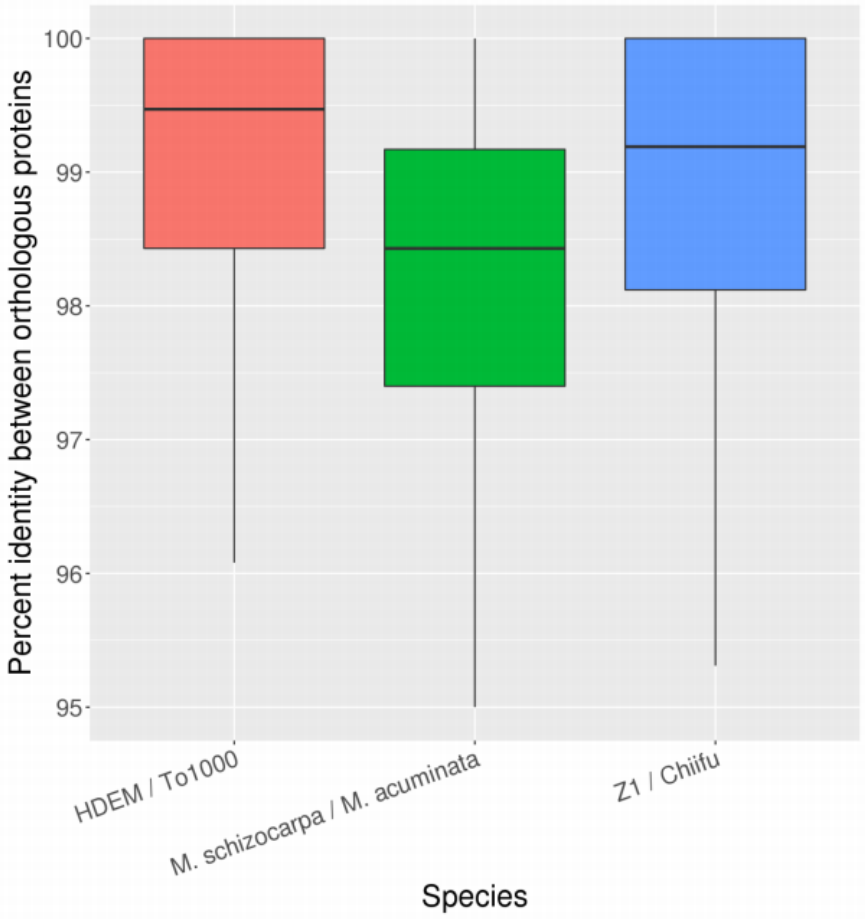
Annotation

	B.oleracea		*B.rapa*		*Musa spp.*	
	TO1000	HDEM	Chiifu	Z1	M.acuminata	M.schizocarpa
genes	59,225	61,279	41,019	46,721	36,542	32,809
exons/gene	5.54	5.47	6.15	5.94	6.05	6.19
CDS size	1,042.26 : 837	937.60 : 699	1,173.19 : 981	1,062.68 : 861	1,035.67 : 861	1,126.84 : 939
BUSCO%	95.1	95.8	96.3	96.6	86.8	92.3

Brassicas and Musas

BRH

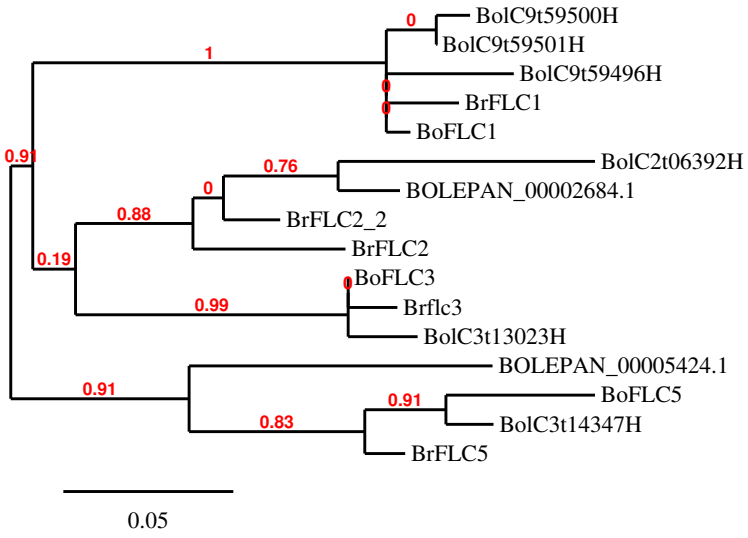
Supplementary Figure 16. Comparison of orthologous proteins. Distribution of the percent identity between proteins of each pair of genomes (N= 39,765 27,333 and 33,290 respectively for HDEM/To1000, M.schizocarpa/M.acuminata and Z1/Chiifu).

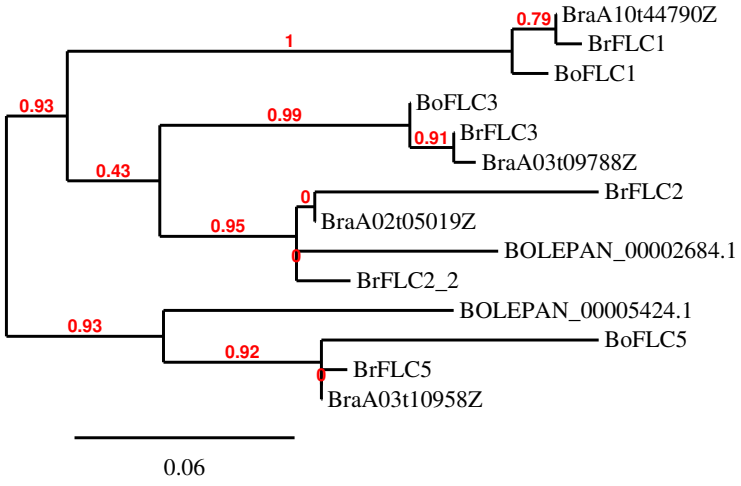


Brassicas

FLC annotation

Flowering locus C : regulator of vernilization and flowering time
FLC2 was found in B.oleracea : reported as specific to the Cauliflower morphotype And 3 tandemly repeated FLC1 in B.oleracea





Phylogenies of the FLC genes from B. oleracea (A) and B. rapa (B) annotations. The annotated genes from HDEM and Z1 are prefixed with Bol and Bra respectively.

Brassicas and Musas
Conclusion

- Combining sequencing technologies can improve genome sequencing
- HMW DNA could remain a challenge
- Gmove can annotate without RNAseq data

Conclusion

Gmove used in severals projects

Pocillopora meandrina

Oithona nana

Lymneae stagnalis

Brasssica rapa and B.oleracea

Elaies guineensis

Amoebophrya (algae parasite)

Acanthurus triostegus

Zanclus cornutus

Leptophaeria maculans (fungi)

Millepora platyphylla

Porites lobata

Tuber melanosporum

Musa schizocarpa (banana)

- Genoscope
- CIRAD Montpelier
- INRA Rennes
- CNRS Lille

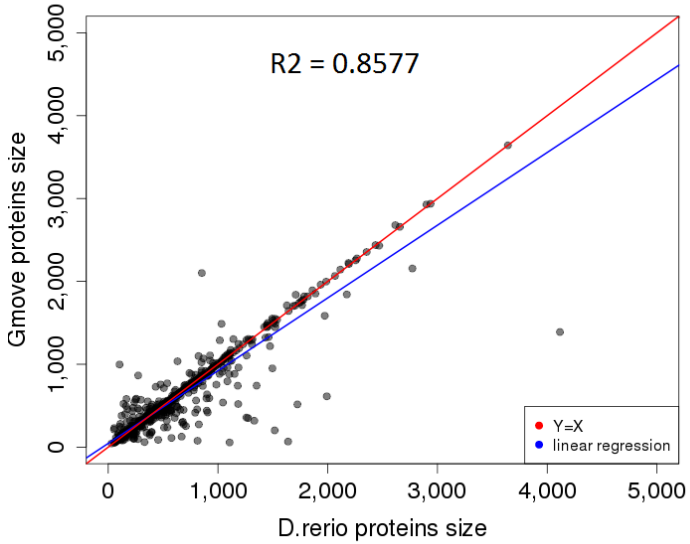
mdubarry@genoscope.cns.fr

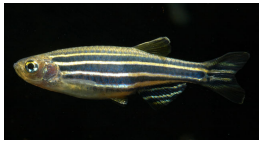
Gmove

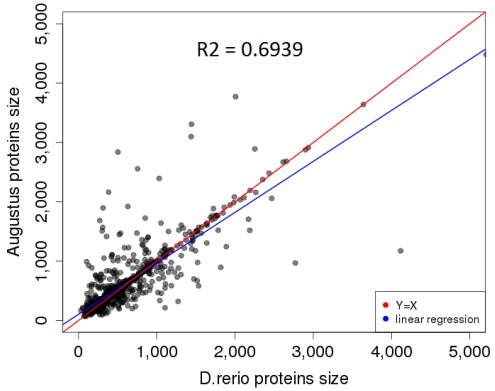
Benchmark with Augustus and Maker

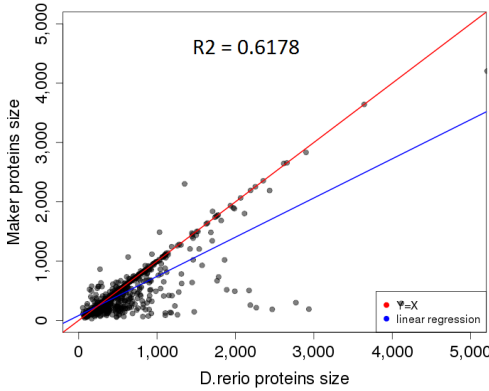
Annotation of *D.rerio* (Chr. 3) with Augustus, Maker2 and Gmove

	ref	Gmove	Augustus	Maker
Gene count	1293	1383	2059	1227
Gene without intron	104	394	495	322
gene length	20618.63 : 9376	17300.56 : 6339	17327.97 : 8043	7984.99 : 4301
exons / gene	8.27 :6	6.65 :4	5.76 :3	5.75 :3
CDS length	1469.01 : 1077	1189.75 : 873	1393.04 : 945	1117.13 : 798
coding base coverage	3.0%	2.6%	4.6%	2.2%
intron count	9397	7815	9798	5830
intron length	2634.93 : 898	2851.09 : 1011	3348.65 : 1331	1445.43 : 680
(SN+SP)/2		30	17	22









Annotation
gene order

Supplementary Figure 7. Gene order comparison. Synteny visualization between *Musa schizocarpa* and *Musa acuminata* (assembly version1) chromosomes.