Reference Assisted Assembly and Annotation of the *Octopus vulgaris* Genome

GENOMICS CORELEUVEN
www.genomicscore.be

Koen Herten[1,2], Gregory Maes[1,2], Eve Seuntjes[3], Fiorito Graziano[4] and Joris R Vermeesch[1,2]

1) Genomics Core, UZ Leuven, KU Leuven
2) Laboratory for Cytogenetics and Genome Research, Center for Human Genetics, KU Leuven
3) Animal Physiology and Neurobiology, KU Leuven
4) Stazione Zoologica Anton Dohrn

KU LEUVEN

Introduction

Compared to any other species, the *Octopus vulgaris* or common octopus appears to be utterly different, with its eight prehensile arms and its complex nervous system, enabling clever problem-solving and observational learning abilities and amazing physical behaviours such as millisecond colour and shape change (camouflage). Strikingly, the genome encoding these alien features turned out to be equally alien.

Materials & Methods

Octopus vulgaris
Estimated genome size: between 2.4-4.6GB
Estimated number of genes: 33,000

Sequencing:
213Gb (Giga Bases (A,C,G,T))
+90x coverage
HiSeq1500 PE 95bp
Insert sizes: 170bp, 250bp, 500bp, 800bp

Computing:
Thinking: Ivybridge, 2x10 cores per node, 124GB RAM per node
Cerebro: Ivybridge, 10 cores per node, 250GB RAM shared

?

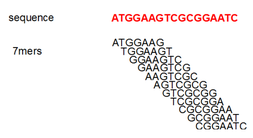
Can the genome be assembled by using the reference of the *Octopus bimaculoides*, and computationally annotated using the genes of the *Octopus bimaculoides*?

Workflow / Requirements* / Explanation / Results*

Pre-computation

Requirements:
Tool: SGA 0.10.14
cput: 55h + 10h
mem: 105GB + 64GB
Thinking, 1 node
Input: raw (txt, gz) 165GB
Output: txt 20GB

Explanation:
- Estimation of genome size
- k-mer analysis
- Estimation of GC content
- Heterozygousity estimation

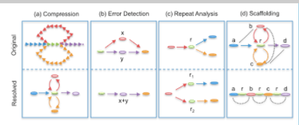


Results:
- Estimated genome size: 2.3GB
- Estimated GC content: 36%
- Estimated Heterozygousity Level: 1/30
- Best k-mer: 81

De novo Assembly

Requirements:
Tool: AbySS 1.9.0
cput: 2,638h
mem: 351GB
Cerebro, 5 nodes
Input: raw (txt, gz) 165GB
Output: txt 3.4GB, total: 53GB

Explanation:
- De Bruijn graph construction of raw data
- Creation of Contigs
- Creation of Scaffolds



Results:

Estimated Genome Size	3.39GB
# Scaffolds	7,452,219
N50	3072bp
L50	214,551
N75	1658bp
L75	479,686
#N's/100kbp	2551.2

Reduction & Scaffolding

Requirements:
Tool: Redundans 0.12
cput: 470h
mem: 293GB
Cerebro, 5 nodes
Input: raw (txt, gz) 165GB + txt 3.4GB
Output: txt 2.3GB, total: 449GB

Explanation:
- Removal of high similar Contigs/Scaffolds
- Re-Scaffolding using reduced Contig set and raw data

Results:

Estimated Genome Size	2.36GB
# Scaffolds	1,290,144
N50	3524bp
L50	179,163
N75	1931bp
L75	394,315
#N's/100kbp	1126.2

Reference Assisted Scaffolding

Requirements:
Tool: Blast 2.5.0
+ Chromosomer 0.1.3
cput: 51m + 30m
mem: 16GB + 2GB
Thinking, 1 node
Input: txt 3.4GB + txt 2.3GB
Output: txt 1.7GB

Explanation:
- Find location of scaffolds on the genome of the *Octopus bimaculoides*
- Connect scaffold-set to chromosome level scaffolds

Results:

Estimated Genome Size	1.85GB
# Scaffolds	84,018
N50	276,442bp
L50	1,602
N75	61,698bp
L75	4,939
#N's/100kbp	18,410.79

Repeat Masking

Requirements:
Tool: WindowMasker 2.2.22
cput: 50m
mem: 2GB
Thinking, 1 node
Input: txt 1.7GB
Output: txt 1.7GB

Explanation:
- Find an mask repeat regions (repeats of small fragments of nucleotides)

Results:
- No stats of the repeat regions are currently available

Annotation

Requirements:
Tool: splign 2.0.8
cput: 50m
Mem: 2.2GB
Thinking, 1 node
Input: txt 1.7 GB + txt 65MB
Output: 27MB

Explanation:
- Mapping of genes of the *Octopus bimaculoides* to the constructed *Octopus vulgaris* genome
- Finding the locations of the genes in the *Octopus vulgaris* genome

Results:
- Found 15,036 possible gene locations
- 15,036 of 38,556 *Octopus bimaculoides* genes have a location with high similarity on the new constructed *Octopus vulgaris* genome.

* All measurements and results are obtained by using the standard parameters for all tools.

Discussion

- High Performance Computing is needed in order to solve the Big Data problem of *de novo* genome assembly. The Big Data in this projects measures around 1TB including raw data, intermediate files and the end results. Computing time exceeded over 3,000h, the needed memory over 350GB.
- The genome of the *Octopus vulgaris* appears to be alien, due to its high heterozygousity level. This biological problem introduces a big computational challenge, therefor this non-standard assembly workflow was used.
- The use of an evolutionary close species (*Octopus bimaculoides*) for a Reference Assisted Scaffolding, resulted in a genome with less scaffolds. Due to the evolutionary distance and the quality of the *Octopus bimaculoides* genome, this step could only partly solve this complex problem.
- Computational Annotation of the genome, by using the genes of an evolutionary close species (*Octopus bimaculoides*), resulted in a big set of high similar genes between the species. Expected is that a high number of genes are not found due to evolutionary distance and possible novel genes.
- The quality of the assembled genome can be improved by further parameter optimization, or by the generation of new data with the use of novel techniques like PacBio.

Acknowledgement

The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation Flanders (FWO) and the Flemish Government – department EWI