STATUS OF THE WATER BUFFALO GENOME

Giordano Mancini
Water Buffalo dataset overview

Number of reads

Illumina GAII paired end reads: 571,334,795 * 2

Illumina GAII jump libraries: 167,677,444 * 2 (insert size 4-6 kb)

Roche 454 Unmated reads: 10,228,343

Roche 454 Mate pairs: 2,416,466 * 2 (insert size 15-35 kb)

Total length of genomic DNA: ~300 Gbases

Clone Coverage: > 40 x
Outline of the assembly procedure

Preprocessed data set

- Removal of redundancy from 454 data
- Removal of redundancy and chimerism from Illumina Mate pairs

MSR-CA

- K-mer counting
  K=31 (Jellyfish)
- Error Correction → SuperReads creation (SR code)
- Genome assembly using SuperReads and OLC approach (CABOG)
Preprocessing: Roche 454 Draft Assembly

Data is cleaned removing reads with the same coordinates in a given unitig.

Results:
- 82% of the data is retained.
- 16% are discarded by sffToCA.
- 1% are redundant.

Computational resources:
- Computation time: 10 days
- Number of CPUs: 216 on 27 nodes
- RAM (each node): 16/32 Gb
- Disk space: 1.35 TB
Preprocessing: Illumina mate pairs alignment on *Bos taurus* genome

BT genome version: UMD 3.1
Starting mate pairs: 167,677,444
Mapping mate pairs (both mates): 63%
Chimeric reads: 10%
Redundant reads: 11%
Refined data set mate pairs: 32%
Alternative approach (assembly 2): alignment on *Bos taurus* after Error Correction

Illumina mate pairs are re-filtered and new genome assembly with CABOG is started.

**Results:**

- Insert size of properly paired reads:
  - 2% non mapping
  - 3% discarded
  - 17% removed after EC
  - 78% properly paired

The final jump library data set contained 58,679,256 reads → 29,339,628 mate pairs → 17.5% of initial (uncorrected) data.
Reads are renamed

K-mer counting (k=31) using Illumina paired ends

Paired ends and mate pairs are corrected using most frequent kmers

Paired ends after EC:

1.5B reads → 40M Super Reads

Computational resources:

Computation time: 8 days
Number of CPUs: 48 on 1 node with HyperThreading
RAM: ~400 GB (512 GB available)
Disk space: 1.5 TB
Starting and final average read length

- Paired ends
- Mate pairs
- Paired ends linking
- SJ cor clean
- Superreads
Genome Assembly with SuperReads and CABOG (assembly 1)

Computational resources:

- Computation time: 30* days
- Number of CPUs: 48 on 1 node with HyperThreading
- RAM: ~100 GB (512 GB available)
- Disk space: 7.0 TB

Currently we are in the scaffold merging phase of the OLC algorithm and we expect to obtain the final genome in a few days.
After the unitig-consensus step of CABOG a total of 11,214,882 unitigs have been obtained. Long unitigs (>200bp) have been aligned on Bos Taurus genome with NUCMER to estimate genome coverage and try a first reconstruction of scaffolds.

*Bos taurus* genome length (excluding gaps): ~2.63 Gb

Total length of BT genome covered joining unitigs longer than 200 bp: 2,379,273,793 ~ 90.5% of Bos Taurus genome
Conclusions

1) The first (draft) version of the water buffalo genome will be obtained in a few days, once the scaffolding still running terminates.

2) A second, better, version obtained including both significant improvements to the MSR-CA pipeline and better mate pair processing strategies (alignment after Error Correction) should be much quicker and we hope to obtain it in about a month.

3) Current data already allows to perform SNP discovery using Jellyfish and the Error Correction pipeline.
Acknowledgments

Aleksey Zimin, University of Maryland
Steven Schroeder, USDA
Giovanni Chillemi CASPUR
Tommaso Biagini, CASPUR
Fabrizio Ferrè, CASPUR
Susana Bueno, CASPUR
Francesco Strozzi, PTP
John Williams, PTP
Claudio Arlandini, CILEA
Alessio Valentini, University of Tuscia