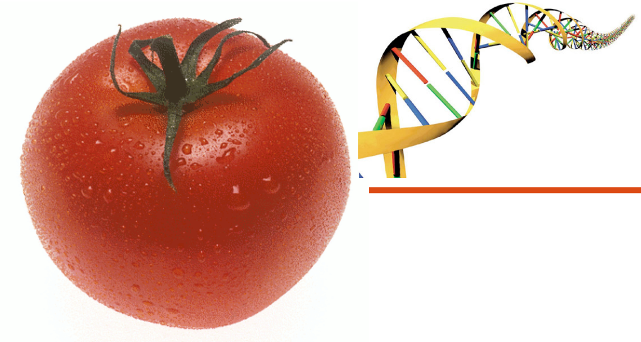


Assembly progress

PAG XVIII

Tomato Business Meeting

San Diego

Jan 12, 2010

Wageningen assembly team

Roeland van Ham, Erwin Datema, Jan van Haarst, Sandra Smit

CBSG 2012
Centre for BioSystems Genomics

eusol

Overview

- First annotated draft (= version 1.0, presented at PAG)
  - Assembly stats
  - Assembly validation
- From version 1.0 to version 2.0
  - Decisions taken at the second schiphol meeting (7 Dec 2009) + update
  - Base assembly

Newbler assembly v1.0
454 + SBM + bac/fosmid ends

Newbler: 2.3-PostRelease-11/19/2009

Input data for assembly 1.0

- 454 data
  - Non-redundant: 55 million reads, 20.5 Gb, 21.6X coverage
- SBM data
  - 3,797,957 reads, 3.1 Gb
- Clone ends
  - 459,789 reads,
    ~135,000 paired BAC ends, ~65,000 paired FOSMID ends

Newbler reports:
74,472,644 reads
22,565,532,344 bases

~ 23.7X genome coverage

Stats of assembly 1.0

	All contigs	Large contigs	Scaffolds
Number of seqs	118,692	62,716	7,409
Total seq. length	762,497,151	748,398,241	794,608,225
Average seq. length	6,424.17	11,933.13	107,249.05
Std. dev. Seq. length	19,868.05	26,128.42	801,095.9
Min. sequence length	100	500	1,998
Max. sequence length	575,502	575,502	20,687,090
Median seq. length	556	1,990	3,187
N50 sequence index	4,237	4,090	49
N50 sequence length	47,298	48,653	4,487,776
N95 sequence index	35,291	28,435	252
N95 sequence length	1,554	2,475	322,251
A content	32.82%	32.87%	29.94%
C content	17.20%	17.15%	15.47%
T content	32.77%	32.83%	29.92%
G content	17.21%	17.15%	15.48%
N content	0.00%	0.00%	9.19%

Validation assembly 1.0

- Validation based on various sources
  - External
    - SGN BAC contigs
    - SOLiD data
    - ESTs
    - Physical map
    - Per-base error rate (0.00035)
  - Internal
    - Clone ends (BAC & fosmid)
    - 454 matepairs
    - Coverage

SGN BAC contigs

- 364 iTAG contigs vs. 7409 assembled scaffolds
  - Blastn, e-value=0.0
- 7349 alignment pairs evaluated
- Cutoff 2 mismatches/Kbp:
  - from the 364 iTAG contigs, 364 are (partially) covered by scaffolds!
  - 789 iTAG_contig x Newbler_scaffold pairs
  - aln_block : 23,766,136
  - gap density : 0.240 (gap/Kbp)
  - mismatch density : 0.033 (mm/Kbp)
- Almost all gaps are due to homopolymer tracts

SOLiD data

Spain	2kb (25bp)	6kb (25bp)	10kb (50bp)
Reads mappable	75%	72%	33%
%coverage	93%	14%	83%

All libraries: 737,284,770 / 762,497,151 bases covered by SOLiD reads (97%)

UK	
Mapped reads	37,289,558 (13.1%)
Uniquely mapped reads	18,215,006 (6.4%)
Coverage	2.2 x (2.6 x)
Coverage based on uniquely mapped reads	1.1 x (1.3 x)
Perc. of bases covered	34.87% (38.40%)
Perc. of bases covered (uniquely mapped)	15.61% (17.18%)

ESTs

- Total number of *S.lycopersicum ESTs considered in the analysis: 265234*
- Data based on GenomeThreader cutoffs
  - coverage of EST = 80 %
  - identity = 90%
- *S. lycopersicum ESTs versus assembly version 1.0*
  - Number of ESTs mapped onto Newbler: 251022 (95%)
  - Number of ESTs with one match onto Newbler: 4642
  - 14212 S. lycopersicum ESTs have no match on the assembly

WGP physical map

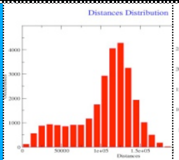
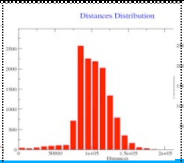
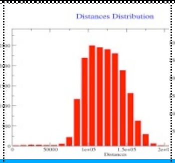
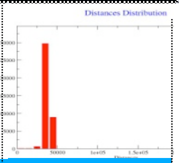
- WGP map:
  - 2,521 Contigs
  - 66,084 BACs
  - 26,1913 Tags
- Considering 52,617 BACs placed on WGP contigs
  - 236,670 tags (91.32% of tags in subset) map on unique location
  - For 71.1% of the BACs in the subset we find all mapped tags on a single scaffold within 200 Kb
  - These “correct” BACs cover 76% of all positions in the assembly

Clone ends

	all libraries
match on assembly	80,04%
discarded	19,96%

Parameters: identity >= 98%
coverage of query>= 90%
length of the ends >= 300 bp

Discarded: all queries below parameters' thresholds
Queries discarded for shortness: 6,60%

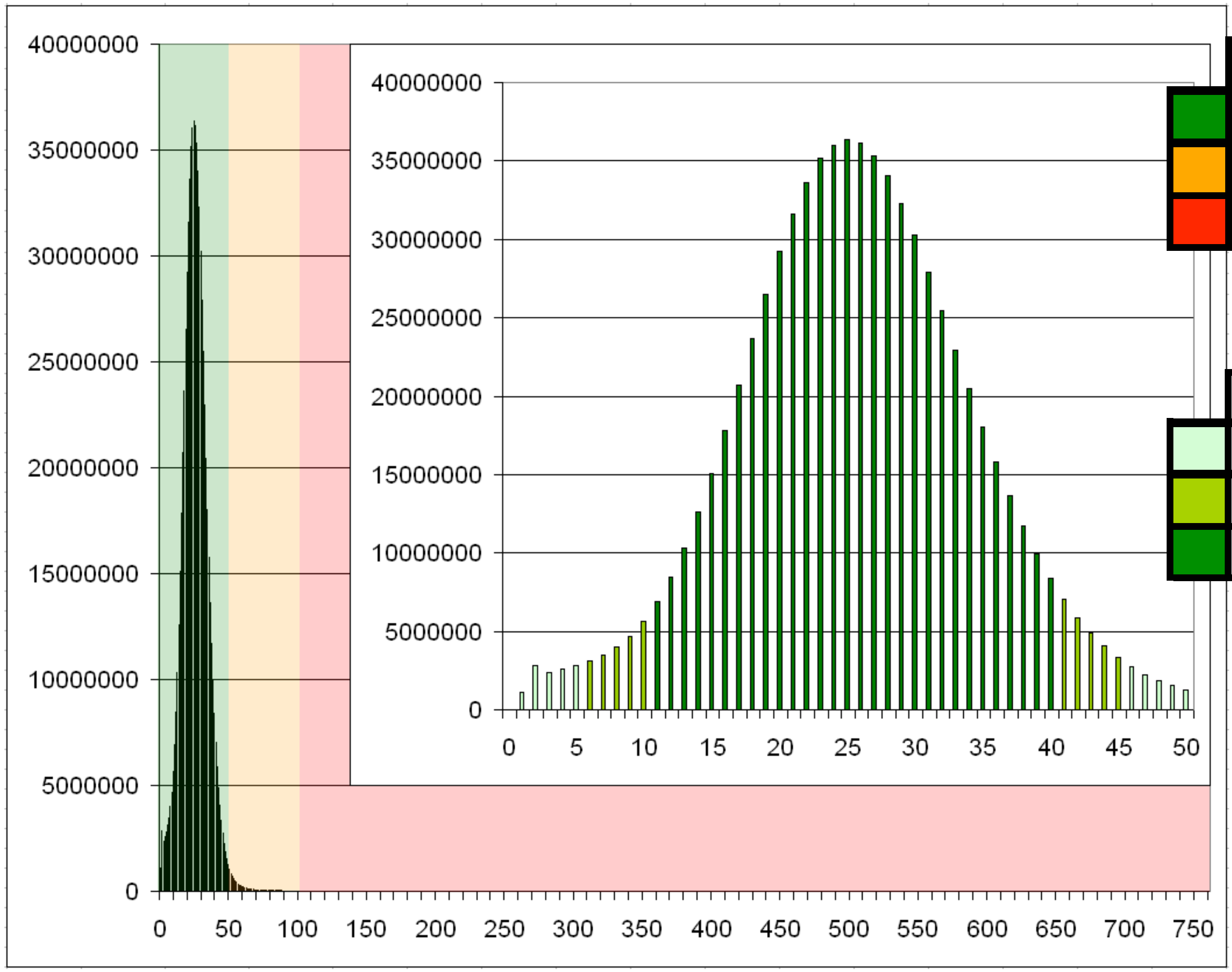
	LE_HBa	SL_EcoRI	SL_MboI	SL_FOS
correct orientation	45,10%	41,02%	52,20%	66,82%
incorrect orientation	0,02%	0,02%	0,03%	0,04%
discarded	54,88%	58,96%	47,77%	33,14%
different scaffolds	3,54%	3,03%	4,15%	3,55%
calculated distance				
median	114kb	102kb	120kb	37kb

Parameters: (both ends) identity >= 99%
coverage of query>= 95%
length of the ends >= 300 bp

incorrect orientation > 800bp
matepairs distance: < 250kb

	all libraries
correct orientation	99,95%
incorrect orientation	0,05%

Coverage



	from	to	%
🟩	1x	50x	98.89
🟧	51x	100x	0.98
🟥	101x	infinity	0.12

	from	to	%
🟩	6x	45x	96.10
🟨	11x	40x	90.04
🟩	16x	35x	75.23

From assembly version 1.0 to version 2.0

Decisions from Schiphol II (updated)

- Assembly v1.0 will function as the base assembly for v2.0
- Jan 31, 2010: version 1.1
  Single-base errors (substitutions, indels) fixed
- Feb 21, 2010: version 1.2
  Assembly consistent with SOLiD PE, 454 PE, clone ends
- Feb 28, 2010: base assembly and version 1.2 validated against SGN BACs
- March 7, 2010: version 1.3
  SGN BACs integrated
- March 21, 2010: version 1.4
  As many gaps as possible closed by SOLiD data
- March 31, 2010 (or 15): version 1.4 (or 1.3) anchored to genetic map and physical map

Base assembly

- Two issues with assembly version 1.0
  - *E. coli* contamination from SBM data
  - Latest 454 runs produced by Italy are not included
- Solution
  - Assembly version 1.01: same data as version 1.0, *E. coli* screening in addition
  - Assembly version 1.02: new filtered 454 data set and *E. coli* screening
  - Version 1.02 will replace the current public release
- Results
  - Assembly stats of version 1.02 are comparable to version 1.0
  - Version 1.01 and 1.02 contain basically no *E. coli* (1 single hit, not further investigated)

Base assembly stats

			Scaffolds		
	1.0		1.01		1.02
Number of sequences	7,409	-	6,783	-	7,237
Total sequence length	794,608,225	-	781,325,825	-	790,859,737
Average sequence length	107,249.05	+	115,188.83	+	109,280.05
Std. dev. sequence length	801,095.90	+	846,663.93	-	752,795.75
Min. sequence length	1,998	+	2,001	-	1,984
Max. sequence length	20,687,090	-	20,672,666	+	22,566,221
Median sequence length	3,187	-	3,169	+	3,206
N50 sequence index	49	-	45	+	56
N50 sequence length	4,487,776	+	4,662,615	-	4,298,623
N95 sequence index	252	-	225	-	242
N95 sequence length	322,251	+	434,162	+	438,110

Final note on the base assembly

- We just received the latest version of the newbler assembler
  - Improvements in the scaffolding algorithm
  - Improvements in the contigging phase
  - Reports additional stats
  - Some bug fixes
- Attempt to create assembly version 1.03
  - If finished before Jan 19 (Tues), this will be the base assembly