



HAL
open science

A draft assembly of the almond genome

Tyler S. Alioto, Fernando Cruz, Jessica Gómez-Garrido, Leonor Frias, Paolo Ribeca, Marta Gut, Konstantinos Alexiou, Werner Howad, Jordi Morata, Josep Casacuberta, et al.

► **To cite this version:**

Tyler S. Alioto, Fernando Cruz, Jessica Gómez-Garrido, Leonor Frias, Paolo Ribeca, et al.. A draft assembly of the almond genome. 24. International Plant and Animal Genome Conference, Jan 2016, San Diego, United States. hal-01344926

HAL Id: hal-01344926

<https://hal.science/hal-01344926>

Submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Draft Assembly of the Almond Genome

Tyler Alloto (CNAG, Barcelona, Spain)


PAG XXIV
Saturday, January 9, 2016
Fruits/Nuts Workshop
Pacific Salon 3



cnag
Centre Nacional d'Anàlisi Genòmica
Centre National de Génétique
Centre National de Génétique

CRG
Centre de Recerca Genètica

The CNAG genomehenge





Sequencing capacity

- >1000 Gbases/day = 9-10 human genomes per day at 30x coverage

Equipment

- 11 Illumina HiSeq2500/2500/IT
- 1 Illumina MiSeq
- 4 Illumina cBios
- 2 Oxford Nanopore Minions
- Caliper/Eppendorf liquid handling robotics

- Bull 1250 core cluster super computer
- Maxeler Data Flow Engine
- 25 TiOps
- 2.7 petabyte disc space
- Barcelona Supercomputing Center (10 x 10 Gb/s)

Prunus persica




cnag
Centre Nacional d'Anàlisi Genòmica
Centre National de Génétique
Centre National de Génétique

1/19/2016

3

CRG
Centre de Recerca Genètica

Prunus dulcis



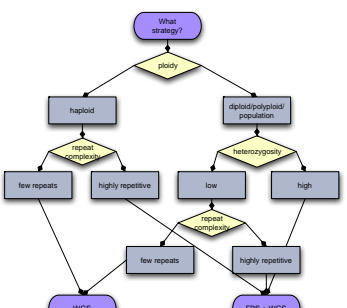
cnag
Centre Nacional d'Anàlisi Genòmica
Centre National de Génétique
Centre National de Génétique

1/19/2016

4

CRG
Centre de Recerca Genètica

Choosing an sequencing strategy



```

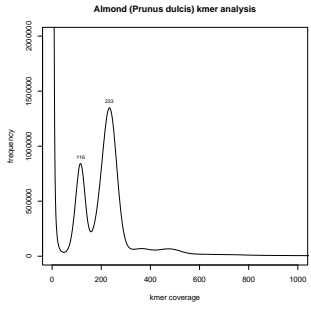
    graph TD
      A[What strategy?] --> B{ploidy}
      B --> C[haploid]
      B --> D[diploid/polyploid/population]
      C --> E{repeat complexity}
      E --> F[low repeats]
      E --> G[highly repetitive]
      D --> H{heterozygosity}
      H --> I[low]
      H --> J[high]
      F --> K[WGS]
      G --> L{repeat complexity}
      L --> M[low repeats]
      L --> N[highly repetitive]
      I --> O[PBS + WGS]
      J --> O
      M --> O
      N --> O
  
```

*kmer analysis used to determine extent of heterozygosity and repeat problem

cnag
Centre Nacional d'Anàlisi Genòmica
Centre National de Génétique
Centre National de Génétique

CRG
Centre de Recerca Genètica

Estimating heterozygosity and repeat structure using kmers



Almond (Prunus dulcis) kmer analysis

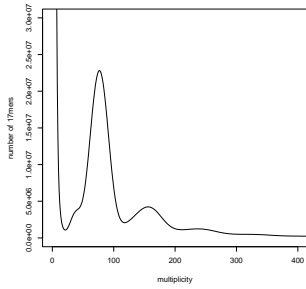
Frequency

kmer coverage

cnag
Centre Nacional d'Anàlisi Genòmica
Centre National de Génétique
Centre National de Génétique

CRG
Centre de Recerca Genètica

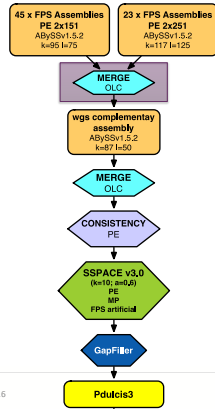
A mammal with low heterozygosity and more repeats



WGS Read Data

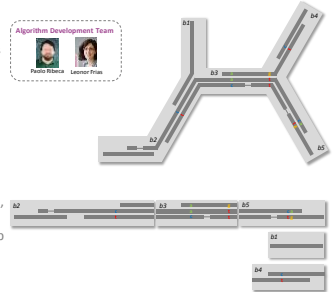
type	PE	PE	PE	MP	MP	FP (avg of 68 pools)
fragment size	300	360	263	5.2 kb	3.1 kb	310
read length	2x101	2x101	2x101	2x101	2x101	2x151/251
yield mBases	45,673	32,729	22,397	16,969	16,431	5,225
depth (x)	166	119	81	62	60	>200
avg percent unique	59	59	59	58	55	56
avg percent unmapped	4.9	5.1	5.1	3.6	3.1	8.0
avg difference* rate r1	2.8	2.8	2.8	3.5	3.3	3.6
avg difference* rate r2	2.8	3.2	2.9	3.4	3.4	3.6
avg percent duplicate	0.7	8.7	1.2	78.4	44.4	8.0
avg phix error r1	0.22	0.48	0.35	0.24	0.25	0.36
avg phix error r2	0.42	1.13	0.91	0.35	0.37	0.62

*Peach v1.0 used as reference for mapping



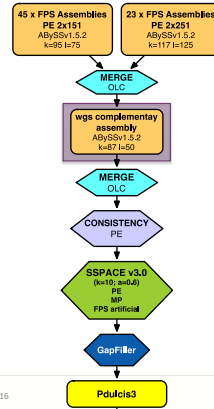
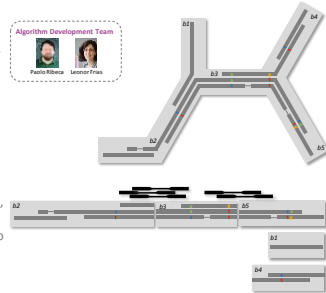
Fosmid Pool Assembly – Merging

- Merging strategy**
- FP contigs/scaffolds as long reads
 - * Length >= 1000 bp
 - OLC-like graph structure
 - Efficient implementation
 - High tolerance to mismatches and indels
 - NEW: uses read (scaffold) path to solve ambiguities
 - Time/memory requirements not dependant on overlap length, rather on the number of reads and the granularity of the overlap computation

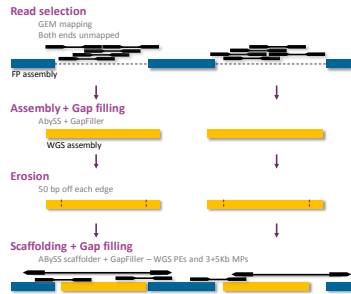


Fosmid Pool Assembly – Merging

- Merging strategy**
- FP contigs/scaffolds as long reads
 - * Length >= 1000 bp
 - OLC-like graph structure
 - Efficient implementation
 - High tolerance to mismatches and indels
 - NEW: uses read (scaffold) path to solve ambiguities
 - Time/memory requirements not dependant on overlap length, rather on the number of reads and the granularity of the overlap computation
 - Additional path ambiguities solved by scaffolding using FPS and WGS PEs (ABYSS or SSPACE)
 - Gap filling (GapFiller)

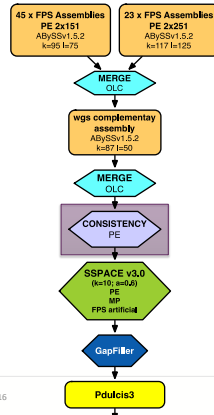


Complementary assembly



cnag

CRG



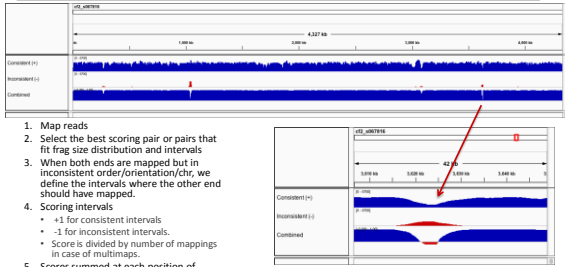
cnag

1/19/2016

14

CRG

Consistency-based misassembly detection



1. Map reads
2. Select the best scoring pair or pairs that fit frag size distribution and intervals
3. When both ends are mapped but in inconsistent order/orientation/chr, we define the intervals where the other end should have mapped.
4. Scoring intervals
 - +1 for consistent intervals
 - -1 for inconsistent intervals
 - Score is divided by number of mappings in case of multimaps
5. Scores summed at each position of genome
6. Determine intervals of positive and negative values.
7. Keep the positive, discard the negative and rescaffold.

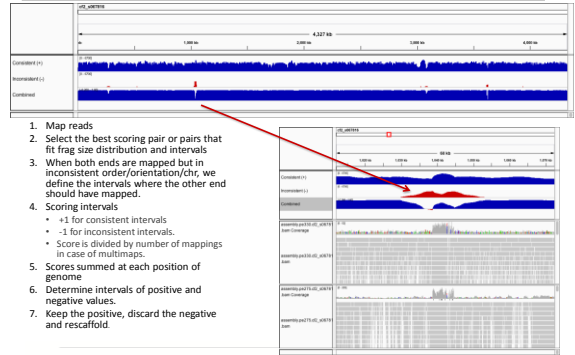
cnag

1/19/2016

15

CRG

Consistency-based misassembly detection



1. Map reads
2. Select the best scoring pair or pairs that fit frag size distribution and intervals
3. When both ends are mapped but in inconsistent order/orientation/chr, we define the intervals where the other end should have mapped.
4. Scoring intervals
 - +1 for consistent intervals
 - -1 for inconsistent intervals
 - Score is divided by number of mappings in case of multimaps
5. Scores summed at each position of genome
6. Determine intervals of positive and negative values.
7. Keep the positive, discard the negative and rescaffold.

cnag

1/19/2016

16

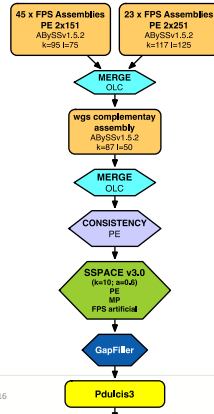
CRG

pdulcis3 assembly statistics

<i>pdulcis3</i> assembly	Contiguity (Nseries)				Gene Completeness (CEGMA)	
	N10	N50	N90	L_ass	Complete	Partial
contigs	147,071 (120)	47,112 (1402)	6627 (6248)	238,749,973 (19193)	-	-
scaffolds	582,101 (32)	235,959 (323)	51,188 (1140)	243,967,083 (8792)	99.19	100

cnag

CRG



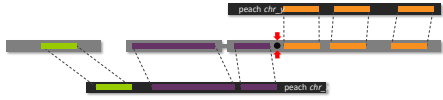
cnag

1/19/2016

18

CRG

Whole genome alignment misassembly detection



Strategy

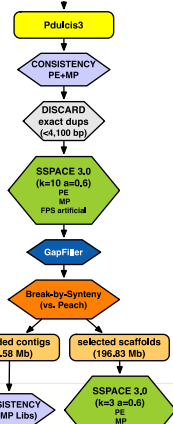
- Whole genome alignment to peach**
 Assembly fragmentation: ws = 5kb; ss = 2.5kb
 BLAT
 Alignment stitching
 Non-redundant alignment blocks

- Syntenic alignment blocks
- Dynamic programming algorithm
- Misassembly
- True structural rearrangement
- Alignment artifact

- Alignment chaining**
 Syntenic alignment blocks
 Dynamic programming algorithm

- Assembly fragmentation**
 Syntenic breakpoints

- Re-scaffold**
 SSPACE



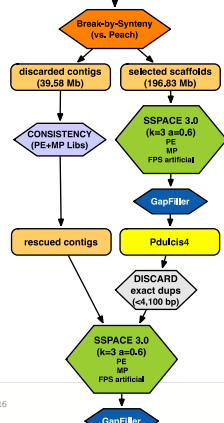
1/19/2016

20



Assemblies Statistics (Scaffolds)

Assembly	Contiguity (Nseries)				Gene Completeness		
	N10	N50	N90	L _{95s}	CEGMA % Complete	CEGMA % Partial	Transcriptome Mappings (>= 70%)
<i>Pdulcis3</i>	582,101 (32)	235,959 (323)	51,188 (1,140)	243,967,083 (8,792)	99.19	100	97.30 %
<i>Pdulcis4</i>	943,822 (17)	395,177 (170)	108,352 (561)	217,311,012 (3,430)	99.60	98.79	96.90 %
<i>Pdulcis5</i>	1,207,396 (16)	499,402 (147)	112,796 (510)	239,631,858 (4,840)	99.60	98.79	97.10 %



1/19/2016

22



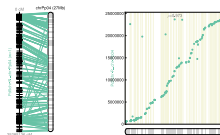
Assembly Statistics

		N10	N50	N90	total length	gaps
		pdulcis5	Contigs	129,054 (128)	32,093 (1,726)	3,645 (10,638)
Scaffolds	1,207,396 (16)		499,402 (147)	112,796 (510)	239,631,858 (4,840)	20,509
pdulcis6/7	Contigs	134,728 (125)	33,482 (1,664)	3,964 (9936)	228,136,029 (23,872)	
	Scaffolds	1,207,434 (16)	499,381 (147)	112,793 (510)	239,629,721 (4,840)	19,032

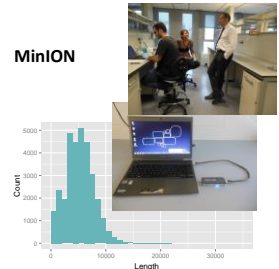


Validation – Anchoring and MinION sequencing

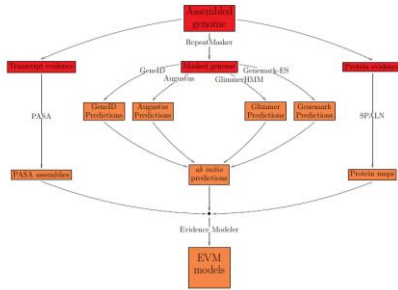
Anchoring to genetic map



MinION



Annotation Pipeline



cnag

CRG

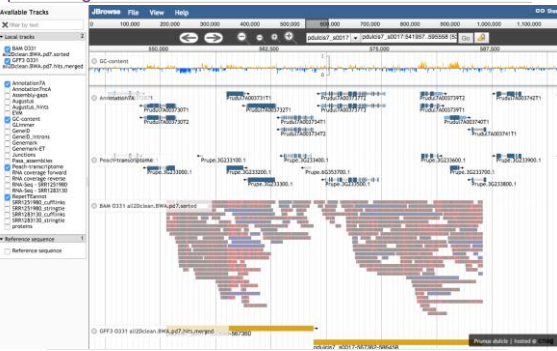
Pdulcis7 Annotation

	Almond	Peach
Genome size (bp)	239,629,721	227,411,381
Number of genes	24,290	26,873
Median gene length	2450	2626
Total genic length	76,978,498	87,059,967
Gene density (kb)	9.9	8.5
Number of exons	128,425	130,676
Median exon length	154	136
Exon GC content	43.5%	44.8%
Number of introns	104,135	103,803
Median intron length	170	162
Intron GC content	34.1%	33.9%
Number of transcripts	29,510	47,089
Exons per transcript	5.29	4.86
Introns per transcript	4.29	3.86
Transcripts per gene	1.22	1.75
Multi-exonic transcripts	79%	76%

cnag

CRG

pdulcis7 genome browser



cnag

CRG

Acknowledgements

- Pere Arús – leadership, coordination
- Marta Gut – sequencing
- Beatriz Galán, José L. Garcia – fosmid pools
- Fernando Cruz – assembly
- Leonor Frias, Paolo Ribeca – merger
- Jèssica Gómez – annotation
- Konstantinos Alexiou – T1/T2 phasing
- Jordi Morata, Josep M^a Casacuberta Suñer – repeat annotation
- Javier Gutierrez, Giulia Lunazzi, Ivo Gut – MiniON sequencing

Plus...

- Werner Howad, María José Rubio Cabetas, Amit Dhingra, Henry Duval, Ángel Fernández i Martí, Michelle Wirthensohn

cnag

1/19/2016

28

CRG

