

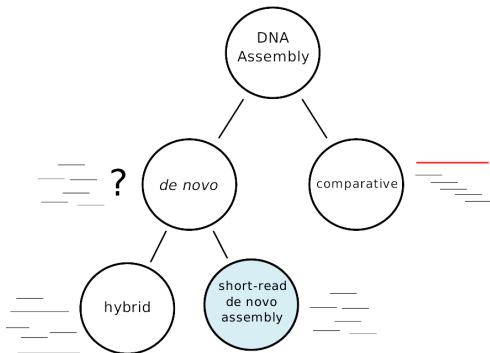
de novo paired-end short reads assembly

Rayan Chikhi
ENS Cachan Brittany
Symbiose, Irisa, France



THESIS FOCUS

- ▶ Graph theory for assembly models
- ▶ Indexing large sequencing datasets
- ▶ Practical implementation in the Monument assembler



OUTLINE

Assemblers

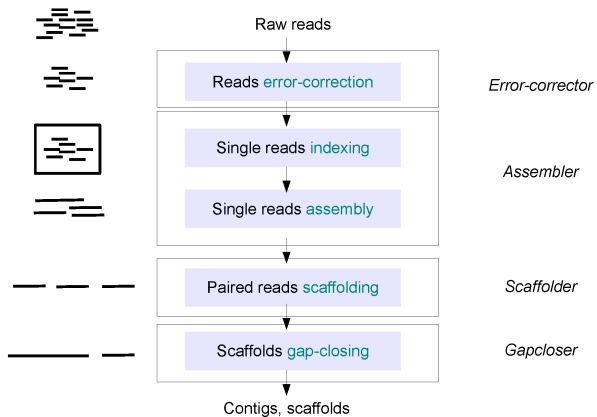
Measures and benchmarks

Monument

Mapsembler

Conclusion

A typical assembly pipeline

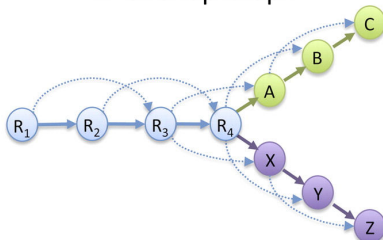


REMINDER : ASSEMBLY MODELS

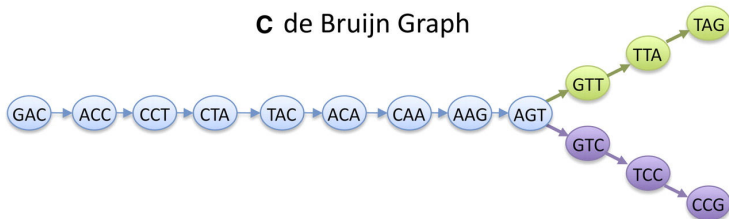
A Read Layout

R₁: GACCTACA
 R₂: ACCTACAA
 R₃: CCTACAAG
 R₄: CTACAAGT
 A: TACAAGTT
 B: ACAAGTTA
 C: CAAGTTAG
 X: TACAAGTC
 Y: ACAAGTCC
 Z: CAAGTCCG

B Overlap Graph



C de Bruijn Graph



THE ASSEMBLY PROBLEM²

- ▶ Let G be the *string graph*¹ of the reads
- ▶ **Assembly Problem** : find a *generalized Hamiltonian path* (visit each node at least once) of minimum length

1. overlap graph without transitive edges and contained reads
2. [Myers 2005]

WHY GREEDY/OLC/DBG INSTEAD OF AP ?

R number of reads

r read length

S genome length

k overlap length or k -mer length

Structure sizes :

- ▶ Overlap graph : $|V| = R, |E| \approx R$, label length r
- ▶ de Bruijn graph : $|V| \approx |E| \approx S$, label length k
- ▶ Greedy : array structure, $\approx S$ elements, k mer keys

Practically : $R \gg S$ and $r > k$.

overlap graph implementation (Newbler) : 4 bytes per read base³.

compressed de Bruijn graph : 12 bytes per graph edge⁴.

3. J. Knight (Roche). Assembly and Finishing of Large/Complex Genomes. 2009 talk

4. <http://arxiv.org/pdf/1008.2555v1>

LANDSCAPE OF ASSEMBLY

- ▶ Before the Illumina Hi-Seq :

Long reads (>200bp), **low coverage**

string graphs

Newbler, Cabog

Short reads (<100bp), **high coverage**

de Bruijn graphs

Velvet, SOAPdenovo, AbySS, ALLPATHS

- ▶ Now and future : **100 bp reads, high coverage**, mate pairs :
which data structure for assemblers ?
- ▶ New trend : greedy assemblers with ad-hoc structure (Ray, PE-Assembler, Meraculous, Monument)

SHORT-READ ASSEMBLERS

Assembler	Method	Error Corr.	Remarks
Euler	de Bruijn	pre-assembly	Pioneer
Velvet	de Bruijn	in-assembly	(still) Popular
ABYSS, CLC-bio⁵, SOAPdenovo	de Bruijn	in-assembly	Parallel, large genomes
Allpaths LG	de Bruijn	pre-assembly	Needs short/long inserts
IDBA⁶	de Bruijn	pre-assembly	Multiple- <i>k</i>
Cabog, Newbler	String	in-assembly	Long reads
Ray	de Bruijn	in-assembly	Parallel short/long reads
PE-Assembler⁷	ad-hoc	pre-assembly	Shorty-like, no graph
SGA⁸	String	pre-assembly	FM-index, promising

5. <http://www.clcdenovo.com/>

6. <http://i.cs.hku.hk/~alse/idba/>

7. <http://www.comp.nus.edu.sg/~bioinfo/peasm/>

8. <https://github.com/jts/sga>

PERSONAL EXPERIENCE (FOR ILLUMINA ASSEMBLY)

General purpose SOAPdenovo

Best quality Allpaths-LG

Genome too large ABySS

Easy to run (once installed) Ray

DE NOVO METAGENOMIC / RNA ASSEMBLERS

de novo metagenomic assemblers :

Genovo : Assembles up to 10^5 454 reads⁹.

Uses a probabilistic model + ICM method.

MetaVelvet : based on Velvet¹⁰

de novo RNA assemblers :

Oases : Actually a post-processing step for Velvet.

Trinity : new name for Ananas¹¹

9. Recomb 2010, <http://cs.stanford.edu/genovo/>

10. unpublished, <http://metavelvet.dna.bio.keio.ac.jp/>

11. <http://trinityrnaseq.sourceforge.net/>

OUTLINE

Assemblers

Measures and benchmarks

Monument

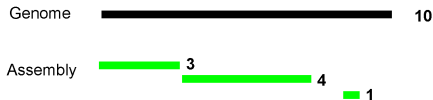
Mapsembler

Conclusion

N50

N50 = Scaffold/contig length at which you have covered 50% of total **assembly** length

NG50 = Scaffold/contig length at which you have covered 50% of total **genome** length



OTHER MEASURES

Reference-based :

- ▶ **Global accuracy** (% of 10 kbp blocks which align with $> 90\%$ identity)¹²
- ▶ **Coverage**¹³
- ▶ **Errors** : small substitutions, small indels, chimeric joins¹⁴

Without reference :

- ▶ **Internal consistency** : read coherence and happy pairs¹⁵

12. Allpaths/bin/AssemblyAccuracy

13. Allpaths/bin/AssemblyCoverage

14. Custom MUMmer-based script

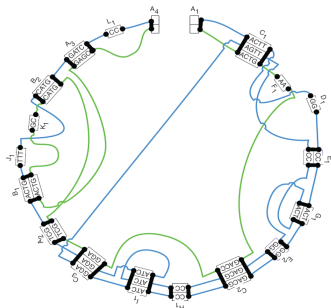
15. AMOS

DIPLOID ACCURACY MEASURE

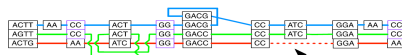
Scaffolds path N50¹⁶

Adjacency graph
(diploid genome)

[Paten et al., 2011,
submitted]



Assembly (red line):



Scaffold path: maximal paths with consistent edges

Edge not consistent

16. Code at <http://compbio.soe.ucsc.edu/assemblathon1/>

Assembly challenges

Goal: given a dataset of reads, produce the best possible assembly

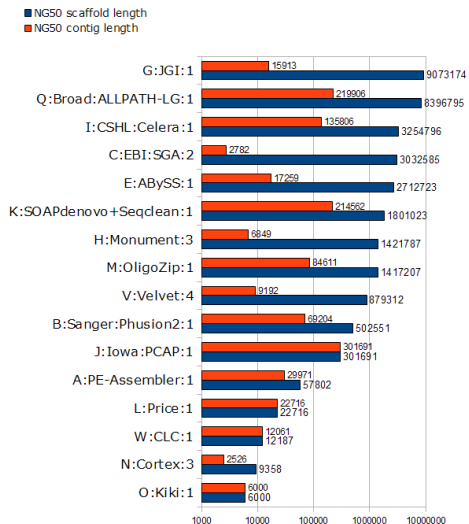
Assemblathon

- Feb 2011
- 100 Mbp diploid genome
- Hi-Seq reads, 80x + mate-pairs
- Organized by **UCSC/UCSD**
- **17** participants

dnGASP

- Mar 2011
- 2 Gbp diploid genome
- Hi-Seq reads, 44x + mate-pairs
- Organized by **CNAG**
- **9** participants

ASSEMBLATHON : BEST SCAFFOLD NG50

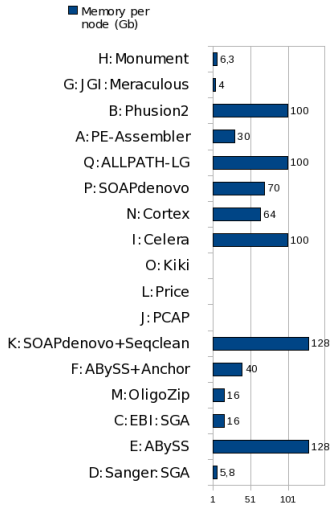
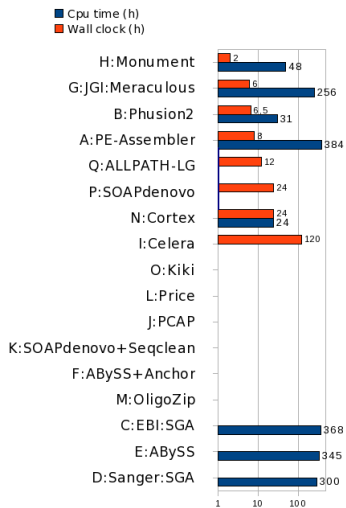


NA50: « alignment » N50

SPA50: scaffold path NA50

ID	# Contigs	N50	NA50	SPA50	HPA50	BNA50	Σ Errors
Q1	1,946	208,256	217,104	141,905	72,336	3,764	1,614
P1	4,566	343,889	343,889	106,510	81,390	3,858	2,061
J1	4,791	301,691	298,640	76,698	13,367	3,201	6,410
I2	37,571	139,666	149,087	63,119	41,968	3,656	6,091
I1	1,798	151,121	135,002	55,591	37,703	3,267	4,120
F5	28,683	56,660	64,245	53,957	46,668	3,684	2,244
M1	4,477	65,510	84,116	52,845	51,033	3,805	3,400
M3	45,200	58,916	82,867	52,321	50,529	3,209	5,337
F3	29,300	48,200	54,510	47,299	46,169	3,678	2,011
F4	32,134	47,737	53,151	47,112	35,944	3,621	2,641
M4	2,672	67,017	68,146	42,862	40,946	3,804	2,180
B1	4,502	66,967	69,214	37,273	37,273	3,719	4,253
F2	33,437	35,510	40,038	34,694	34,277	3,614	2,162
F1	45,487	34,247	38,946	34,249	34,197	3,603	2,030
M5	13,998	36,938	39,490	31,932	31,520	3,655	4,472
M2	5,780	36,443	37,753	31,828	31,706	3,767	2,295
K2	2,796	214,562	213,516	30,431	30,273	3,547	6,750
K3	926	216,393	213,516	30,408	30,250	3,551	6,656
K1	15,689	209,662	213,516	30,403	30,187	3,325	7,071
A1	9,741	25,383	29,791	24,930	24,930	3,299	969
D2	7,904	26,304	26,834	24,267	24,267	3,713	721
X2	141,144	16,191	21,386	19,773	19,765	3,243	1,844
X1	151,852	15,165	19,938	18,562	18,548	3,137	1,725
D3	11,310	18,021	18,871	17,403	17,403	3,618	1,246
D1	11,067	17,882	18,300	16,819	16,819	3,626	1,217
G1	14,817	15,585	15,814	16,787	11,396	3,081	953
D4	23,037	17,519	18,216	16,708	16,708	3,589	1,742
B2	3,040	153,374	158,964	15,796	15,766	3,394	15,125
E1	17,341	16,897	17,087	15,367	15,316	3,277	1,267
E2	11,093	17,129	17,097	15,354	15,354	3,310	1,144
H4	12,316	17,117	17,186	14,172	12,616	3,026	4,248
E3	1,937	2,575,286	2,575,286	14,105	14,105	3,229	13,245
X4	175,163	14,169	15,889	14,057	13,953	3,157	2,028
X5	174,679	14,168	15,785	14,028	13,931	3,163	2,028
L1	14,822	20,165	22,620	13,434	13,434	2,533	27,134
C1	20,229	10,819	10,561	9,929	9,221	3,252	2,940
H5	16,190	12,130	11,963	9,620	8,946	2,825	6,303
W11	17,979	11,777	11,930	9,432	9,396	3,147	9,297
W8	18,725	11,004	11,155	8,744	8,725	3,125	9,027
W1	17,759	11,765	11,941	8,718	8,699	3,061	10,400
H1	19,446	11,024	10,864	8,645	8,120	2,711	5,904
W9	19,862	10,472	10,639	8,394	8,394	3,096	6,349
W7	18,929	11,013	11,136	8,178	8,157	3,036	10,435
W5	20,561	10,181	10,388	8,109	8,087	3,075	9,088
V4	51,760	8,856	9,125	6,956	5,559	2,552	10,361
V5	53,949	8,141	8,419	6,924	5,715	2,581	7,797
W10	24,778	8,190	8,324	6,760	6,760	2,955	7,102
W6	26,531	7,715	7,833	6,478	6,463	2,892	8,915
W3	25,328	8,169	8,292	6,425	6,415	2,888	11,057
H3	25,709	7,057	6,849	5,861	5,540	2,479	7,028
H2	25,981	6,980	6,783	5,787	5,535	2,478	7,094
O1	14,994	7,847	5,928	5,518	5,518	2,034	11,387
V6	65,834	5,467	5,624	4,857	4,700	2,440	9,332
X3	328,797	4,611	5,217	4,808	4,808	2,330	1,857
X6	311,185	4,601	5,165	4,759	4,759	2,321	1,882
C2	53,273	3,003	2,782	2,744	2,357	1,882	4,763
N1	86,428	2,387	2,662	2,646	2,644	1,424	11,464
N3	103,555	2,196	2,512	2,496	2,494	1,322	8,644
N2	69,948	2,204	2,117	2,108	2,107	1,352	5,795

ASSEMBLATHON : RESOURCES



DNGASP : PRELIMINARY STATS

Edited, probably not public

WHAT'S NEXT ? ASSEMBLATHON 2

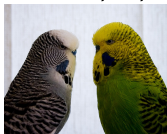
Maylandia zebra, Illumina short and long inserts, ≈ 100 bp reads, 192x coverage, 1 Gbp genome.



Red tailed boa constrictor, Illumina short and long inserts, ≈ 100 bp reads



common pet parakeet, Illumina short and long inserts, 454



Submissions deadline : Sept. 1st.

OUTLINE

Assemblers

Measures and benchmarks

Monument

Mapsembler

Conclusion

MOTIVATION FOR A NEW MODEL

Common problems with current assembly methods :

- ▶ Usually high memory footprint
- ▶ Computation-intensive for longer genomes
- ▶ Error correction dilemma
 - ▶ pre-assembly : no efficient algorithm, loss of information
 - ▶ in-assembly : creates too many vertices in graphs

In the future :

- ▶ Longer read lengths : de Bruijn-based assemblers will be inadequate
- ▶ Higher throughput : overlap graph-based assemblers will be inadequate

PAIRED ASSEMBLY PROBLEM¹⁷



Paired string graph : also represents paired links between reads.

Path-strings : string spelled by a path. Gaps allowed, e.g. $ab\Diamond defgh$.

Paired Assembly problem : find a minimum weight generalized H.P. s.t. path-string satisfies pairing constraints

17. to appear in WABI 2011 proceedings

Monument algorithm

Theoretical motivation: [can one build scaffolds directly?](#)
e.g. [localized assembly](#)

Starting read



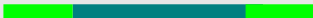
Monument algorithm

Theoretical motivation: **can one build scaffolds directly?**
e.g. **localized assembly**

Starting read



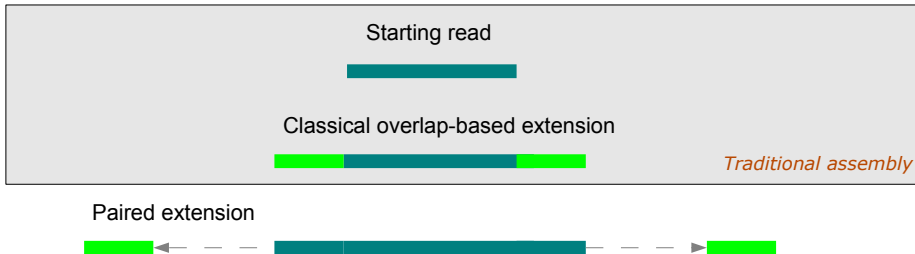
Classical overlap-based extension



Traditional assembly

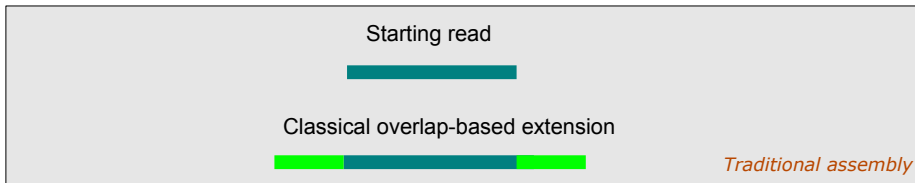
Monument algorithm

Theoretical motivation: **can one build scaffolds directly?**
e.g. **localized assembly**



Monument algorithm

Theoretical motivation: **can one build scaffolds directly?**
e.g. **localized assembly**



Paired extension



Repeat overlap extension, etc..



BACTERIAL RESULTS

Dataset	Software	Contig N50 (Kbp)	Scaffold N50 (Kbp)	Longest scaffold (Kbp)	Coverage (%)	Accuracy (%)
Experimental (1)	Monument	38.0	101.8	236.0	96.4	96.7
	Velvet	26.3	95.3	267.9	96.9	99.1
	Ray	69.5	87.3	174.4	97.4	98.4
Simulated with variants (2)	Monument	113.3	134.1	340.5	91.0	95.0
	Velvet	30.8	132.6	327.2	87.9	92.3
	Ray	10.2	10.2	41.2	89.2	100.0

Resources for bacterial assembly : 7 minutes, 0.5 GB RAM.

MONUMENT HIGHLIGHTS

- ▶ First software able to do targeted assembly of scaffolds
 - ▶ Collaboration for targeted assembly of chilean grape SVs
- ▶ Human assembly in < 80 GB RAM
- ▶ 1 Gbp/day on 6 nodes cluster

OUTLINE

Assemblers

Measures and benchmarks

Monument

Mapsembler

Conclusion

Mapsembler: focus on specific information



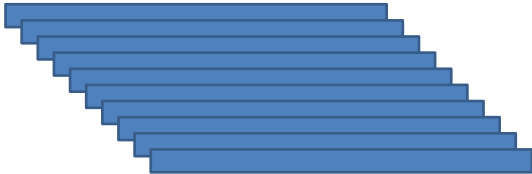
Mapsembler: focus on « known » information

In

- A fragment = starter.



- A set of NGS reads



Biological information is in the reads -

Peterlongo HTS2011

Mapsembler

Out

1. Is the starter coherent with a subset of reads ?

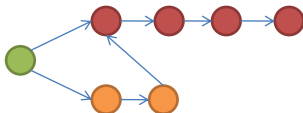
ATGCGGCATTGCA
CATGGACATGCGGC
TGGACATGCGGCAT
GCATTGCATAGC
GCATAGCTGACT

2. If yes:

1. give the context: **contig** containing the starter

...CAACGGACGCATATGGACATGCGGCATTGCATAGCTGACTACTGCATCATAAC...

2. give the *complex* context: **graph**



Mapsembler

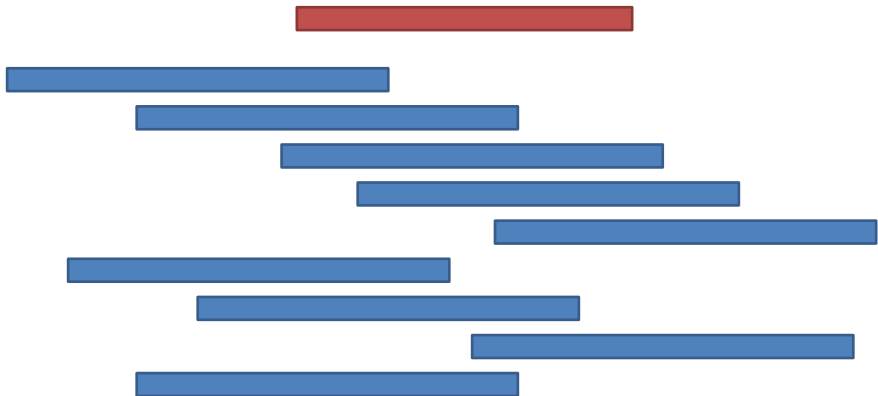
Usages

- Validate a k -mer assembly
- Is this gene has an homologue is this read set?
- RNA seq: is this splicing event present/absent?
- Metagenomic: is this enzyme present in this sample?
- Remove “annoying” reads (contamination, symbiont, ...)
- Enrich unmappable reads
- ...
- “Zero”memory. (small) desktop computer



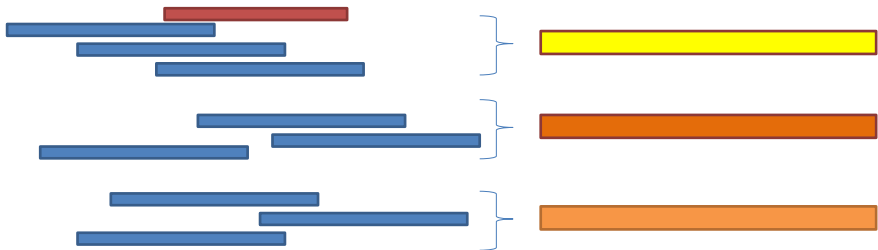
Mapsembler

Phase 1a: map reads on starter
(each read: at most d substitutions)



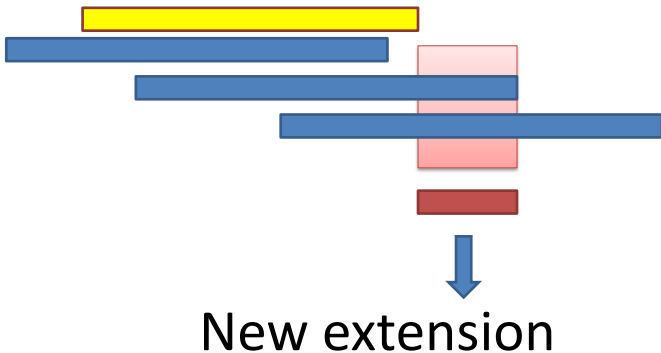
Mapsembler

Phase 1b: generates starters “read-coherent” at most d substitutions



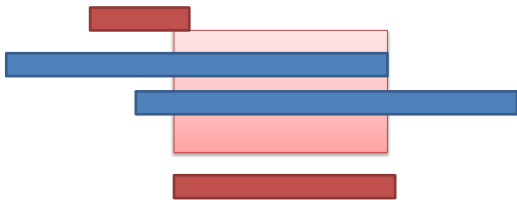
Mapsembler

Phase 2a: Extensions Extend each Starter'



Mapsembler

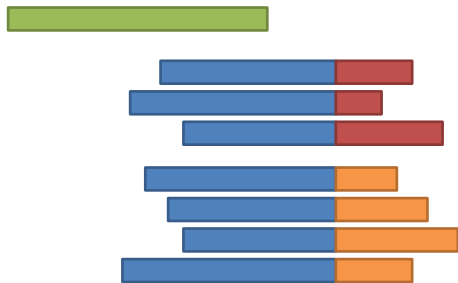
Phase 2b: Extensions
extend each extensions...



New extension...

Phase 2b: Extensions when do we stop ?

- V1: If case of branching: 2 possible extensions

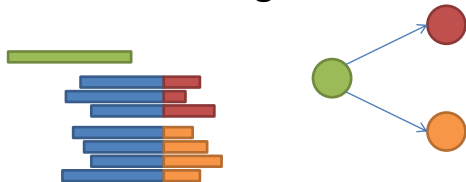


- Outputs fasta file

Mapsembler

Phase 2b: Extensions when do we stop ?

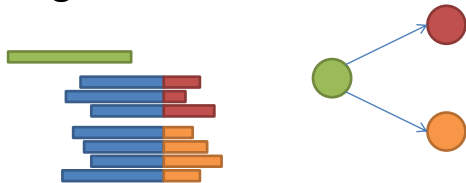
- V2: Never.: will generate a tree...



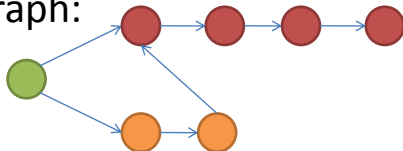
Mapsembler

Phase 2b: Extensions when do we stop ?

- V2: generate a tree...



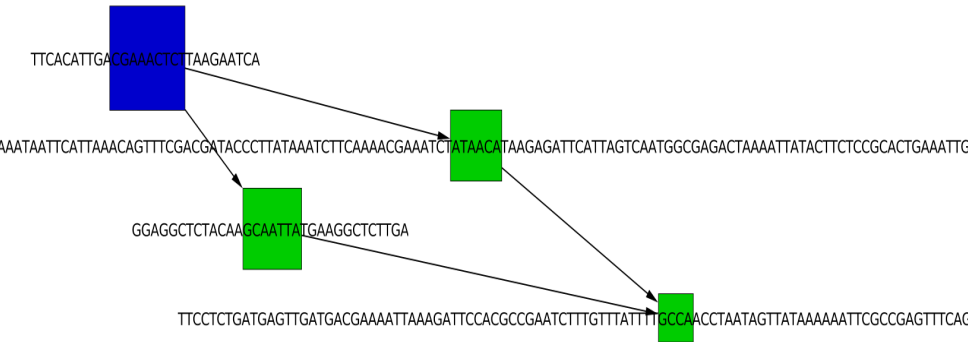
- ... and finally a graph:



- Outputs a XGMML graph (Cytoscape)

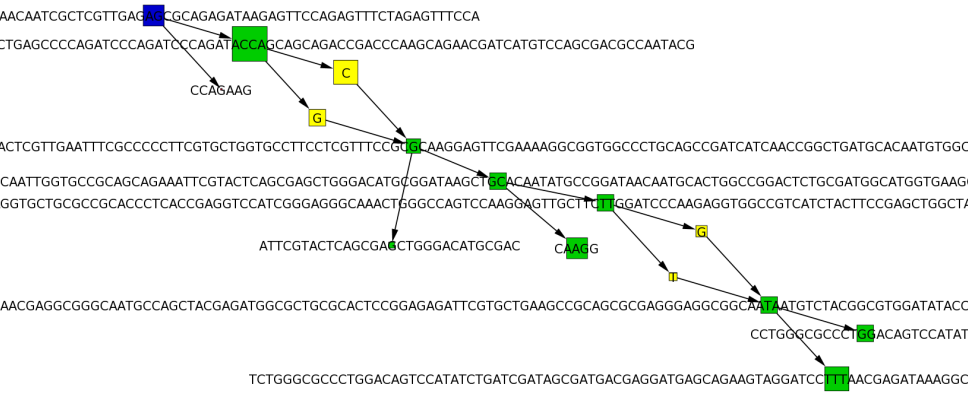
Mapsembler: some examples

Exon skipping (drosophila)



Mapsembler: some examples

SNPs (drosophila)



RECENT APPLICATION

Detection of 24 new candidate fusion genes implicated in human breast cancer.

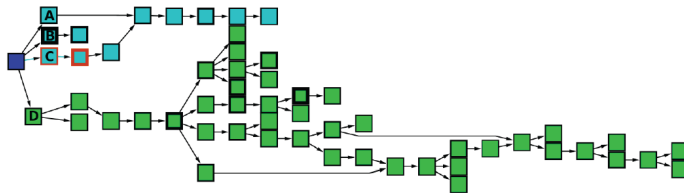


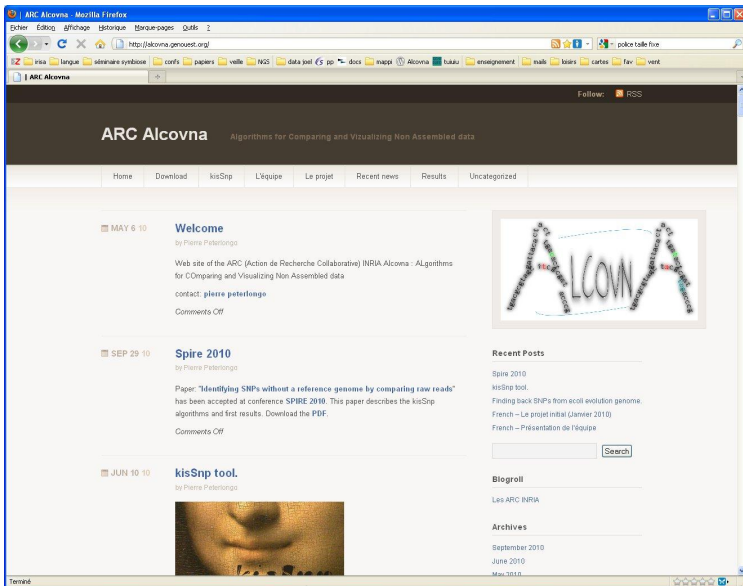
Fig. 3. Extension graph of exon VAPB (20:56,962,507-56,966,573). The dark blue node is the starter. The light blue nodes are exons on chromosome 17, gene IKZF3. Nodes with red borders correspond to fusion exon found in [2]. Green nodes correspond to exons with no fusion genes, also found in normal tissue.

MAPSEMBLER DEMO MAYBE ?

Let's conclude

- Non model species
- Avoid assembly:
 - Keep all information
 - Desktop computer
- Answer specific question
- Complementary to whole genome assembly approaches

Infos and downloads: <http://alcovna.genouest.org/>



The screenshot shows a Mozilla Firefox browser window displaying the website <http://alcovna.genouest.org/>. The browser's address bar shows the URL, and the page title is "ARC Alcovna". The website header features the text "ARC Alcovna" and the tagline "Algorithms for Comparing and Visualizing Non Assembled data". Below the header is a navigation menu with links for Home, Download, kisSnp, L'équipe, Le projet, Recent news, Results, and Uncategorized. The main content area displays three blog posts:

- MAY 6 10** **Welcome** by Pierre Peterlongo. The post text reads: "Web site of the ARC (Action de Recherche Collaborative) INRIA Alcovna : Algorithms for Comparing and Visualizing Non Assembled data. contact: [pierre peterlongo](#). Comments Off".
- SEP 29 10** **Spire 2010** by Pierre Peterlongo. The post text reads: "Paper: 'Identifying SNPs without a reference genome by comparing raw reads' has been accepted at conference SPIRE 2010. This paper describes the kisSnp algorithms and first results. Download the PDF. Comments Off".
- JUN 10 10** **kisSnp tool.** by Pierre Peterlongo. Below the title is a small image showing a close-up of a person's mouth.

On the right side of the page, there is a "Recent Posts" section with links to "Spire 2010", "kisSnp tool", and "Finding back SNPs from ecoli evolution genome". Below this is a search box with a "Search" button. Further down are sections for "Biogrol" (with a link to "Les ARC INRIA") and "Archives" (with links for "September 2010", "June 2010", and "May 2010").

Biological information is in the reads -
Peterlongo HTS2011

OUTLINE

Assemblers

Measures and benchmarks

Monument

Mapsembler

Conclusion

SOFTWARE FROM SYMBIOSE NGS TEAM

- ▶ **GASSST**
Fast, parallel reads alignment with arbitrary gap length
- ▶ **KissSNP**
Localize SNPs between two datasets SNPs a reference
- ▶ **Alcovna project**
Report RNA splicing events without a reference
- ▶ **MAPPI project**
Efficient intersection of two (metagenomic) reads sets

ACKNOWLEDGEMENTS

- ▶ Dominique Lavenier, PhD advisor
- ▶ Pierre Peterlongo, collaboration on Mapsembler
- ▶ Symbiose team
- ▶ Biogenouest platform (<http://www.biogenouest.org/>)

SYMBIOSE

IRISA, INRIA/CNRS/ENS Cachan, Rennes

► NGS algorithms

- sequence alignment
- assembly (since 2008)
- targeted assembly (since 2010)
- metagenomic analysis (since 2010)

- Biologicals networks and models
- Proteins : structures and grammars
- BioGenouest platform

► Workflow and parallelization



Stratégies d'assemblage

- Graphe des reads chevauchants (overlap graph)

