Models and algorithms for genome comparison and sequence alignment

Géraldine Jean

Towards a modern analysis of omics data of the Ocean - mission Microbiome: CEODOS and AtlantEco expeditions

May 15-18th, 2023, Valparaiso-Chile

Warning

- Computer scientist in computational biology

Warning

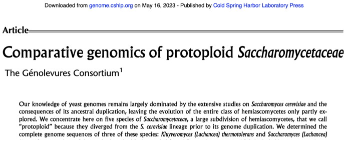
- Computer scientist in computational biology
- Mainly doing theoretical work

Warning

- Computer scientist in computational biology
- Mainly doing theoretical work
- But involved in applied projects

Warning

- Computer scientist in computational biology
- Mainly doing theoretical work
- But involved in applied projects



Downloaded from genome.cshlp.org on May 16, 2023 - Published by Cold Spring Harbor Laboratory Press

Article

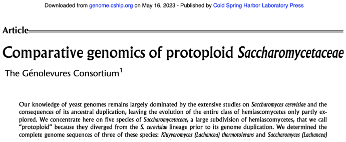
Comparative genomics of protoploid *Saccharomycetaceae*

The Génolevures Consortium[1]

Our knowledge of yeast genomes remains largely dominated by the extensive studies on *Saccharomyces cerevisiae* and the consequences of its ancestral duplication, leaving the evolution of the entire class of hemiascomycetes only partly explored. We concentrate here on five species of *Saccharomycetaceae*, a large subdivision of hemiascomycetes, that we call "protoploid" because they diverged from the *S. cerevisiae* lineage prior to its genome duplication. We determined the complete genome sequences of three of these species: *Kluyveromyces (Lachancea) thermotolerans* and *Saccharomyces (Lachancea)*

Warning

- Computer scientist in computational biology
- Mainly doing theoretical work
- But involved in applied projects



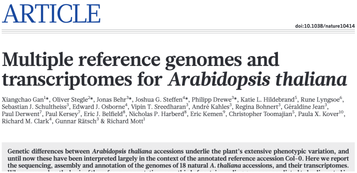
Downloaded from genome.cshlp.org on May 16, 2023 - Published by Cold Spring Harbor Laboratory Press

Article

Comparative genomics of protoploid *Saccharomycetaceae*

The Génolevures Consortium[1]

Our knowledge of yeast genomes remains largely dominated by the extensive studies on *Saccharomyces cerevisiae* and the consequences of its ancestral duplication, leaving the evolution of the entire class of hemiascomycetes only partly explored. We concentrate here on five species of *Saccharomycetaceae*, a large subdivision of hemiascomycetes, that we call "protoploid" because they diverged from the *S. cerevisiae* lineage prior to its genome duplication. We determined the complete genome sequences of three of these species: *Kluyveromyces (Lachancea) thermotolerans* and *Saccharomyces (Lachancea)*



ARTICLE

doi:10.1038/nature10414

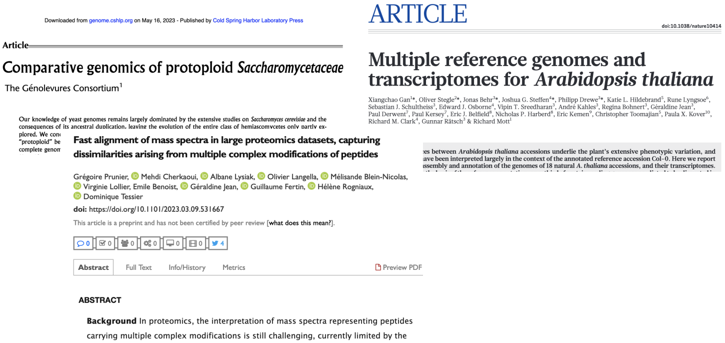
Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*

Xiangchao Gan[1]*, Oliver Stegle[2]*, Jonas Behr[3]*, Joshua G. Steffen[4]*, Philipp Drewe[3]*, Katie L. Hildebrand[5], Rune Lyngsoe[6], Sebastian J. Schultheiss[3], Edward J. Osborne[4], Vipin T. Sreedharan[3], André Kahles[3], Regina Bohnert[3], Géraldine Jean[3], Paul Derwent[7], Paul Kersey[7], Eric J. Belfield[8], Nicholas P. Harberd[8], Eric Kemen[9], Christopher Toomajian[5], Paula X. Kover[10], Richard M. Clark[4], Gunnar Rätsch[3] & Richard Mott[1]

Genetic differences between *Arabidopsis thaliana* accessions underlie the plant's extensive phenotypic variation, and until now these have been interpreted largely in the context of the annotated reference accession Col-0. Here we report the sequencing, assembly and annotation of the genomes of 18 natural *A. thaliana* accessions, and their transcriptomes.

Warning

- Computer scientist in computational biology
- Mainly doing theoretical work
- But involved in applied projects



Downloaded from genome.cshlp.org on May 16, 2023 - Published by Cold Spring Harbor Laboratory Press

Article

Comparative genomics of protoploid *Saccharomycetaceae*

The Génolevures Consortium[1]

Our knowledge of yeast genomes remains largely dominated by the extensive studies on *Saccharomyces cerevisiae* and the consequences of its ancestral duplication, leaving the evolution of the entire class of hemiascomycetes only partly explored. We con
"protoploid" be
complete genom

ARTICLE

doi:10.1038/nature10414

Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*

Xiangchao Gan[1]*, Oliver Stegle[2]*, Jonas Behr[3]*, Joshua G. Steffen[4]*, Philipp Drewe[3]*, Katie L. Hildebrand[5], Rune Lyngsoe[6], Sebastian J. Schultheiss[3], Edward J. Osborne[4], Vipin T. Sreedharan[3], André Kahles[3], Regina Bohnert[3], Géraldine Jean[3], Paul Derwent[7], Paul Kersey[7], Eric J. Belfield[8], Nicholas P. Harberd[8], Eric Kemen[9], Christopher Toomajian[5], Paula X. Kover[10], Richard M. Clark[4], Gunnar Rätsch[3] & Richard Mott[1]

es between *Arabidopsis thaliana* accessions underlie the plant's extensive phenotypic variation, and
ave been interpreted largely in the context of the annotated reference accession Col-0. Here we report
ssembly and annotation of the genomes of 18 natural *A. thaliana* accessions, and their transcriptomes.

Fast alignment of mass spectra in large proteomics datasets, capturing dissimilarities arising from multiple complex modifications of peptides

Grégoire Prunier, Mehdi Cherkaoui, Albane Lysiak, Olivier Langella, Mélisande Blein-Nicolas, Virginie Lollier, Emile Benoist, Géraldine Jean, Guillaume Fertin, Hélène Rogniaux, Dominique Tessier

doi: https://doi.org/10.1101/2023.03.09.531667

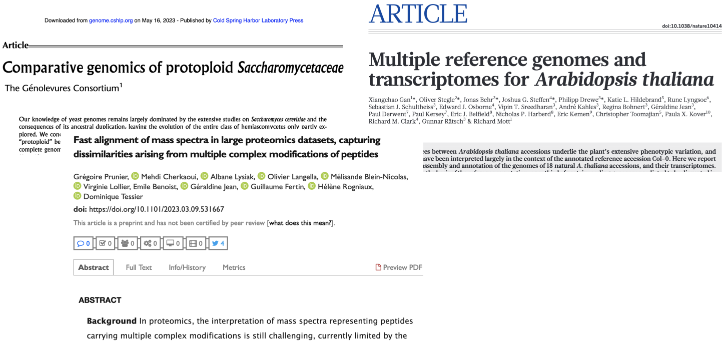
This article is a preprint and has not been certified by peer review [what does this mean?].

ABSTRACT

Background In proteomics, the interpretation of mass spectra representing peptides carrying multiple complex modifications is still challenging, currently limited by the

Warning

- Computer scientist in computational biology
- Mainly doing theoretical work
- But involved in applied projects



Downloaded from genome.cshlp.org on May 16, 2023 - Published by Cold Spring Harbor Laboratory Press

Article

Comparative genomics of protoploid *Saccharomycetaceae*

The Génolevures Consortium[1]

Our knowledge of yeast genomes remains largely dominated by the extensive studies on *Saccharomyces cerevisiae* and the consequences of its ancestral duplication, leaving the evolution of the entire class of hemiascomycetes only partly explored. We con
"protoploid" be
complete genom

ARTICLE

doi:10.1038/nature10414

Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*

Xiangchao Gan[1]*, Oliver Stegle[2]*, Jonas Behr[3]*, Joshua G. Steffen[4]*, Philipp Drewe[3]*, Katie L. Hildebrand[5], Rune Lyngsoe[6], Sebastian J. Schultheiss[3], Edward J. Osborne[4], Vipin T. Sreedharan[3], André Kahles[3], Regina Bohnert[3], Géraldine Jean[3], Paul Derwent[7], Paul Kersey[7], Eric J. Belfield[8], Nicholas P. Harberd[8], Eric Kemen[9], Christopher Toomajian[5], Paula X. Kover[10], Richard M. Clark[4], Gunnar Rätsch[3] & Richard Mott[1]

ces between *Arabidopsis thaliana* accessions underlie the plant's extensive phenotypic variation, and
ave been interpreted largely in the context of the annotated reference accession Col-0. Here we report
ssembly and annotation of the genomes of 18 natural *A. thaliana* accessions, and their transcriptomes.

Fast alignment of mass spectra in large proteomics datasets, capturing dissimilarities arising from multiple complex modifications of peptides

Grégoire Prunier, Mehdi Cherkaoui, Albane Lysiak, Olivier Langella, Mélisande Blein-Nicolas, Virginie Lollier, Emile Benoist, Géraldine Jean, Guillaume Fertin, Hélène Rogniaux, Dominique Tessier

doi: https://doi.org/10.1101/2023.03.09.531667

This article is a preprint and has not been certified by peer review [what does this mean?].

ABSTRACT

Background In proteomics, the interpretation of mass spectra representing peptides carrying multiple complex modifications is still challenging, currently limited by the

- Not working with environmental data (yet) ;)

General Interests

Using methods from **algorithms on strings** and **graph theory** to study...

Comparative Genomics

- Rearrangement scenario/distance
- Sequence alignment

Next-Generation Sequencing Problems

- Sequence alignment
- De novo assembly of repeats

Mass spectrometry

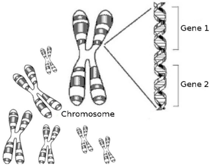
- De novo analysis of the spectrum for identification of unknown metabolite
- Peptide identification and protein inference

Scientific Context

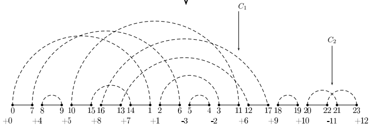
1. Biological objects
   genomes, genes, RNA sequences, spectrum...

2. Combinatorial objects
   string, tree, graph, permutation...

3. Algorithmics tools
   3.1 Computational complexity analysis
   3.2 algorithm development: approximation algorithms, FPT algorithms, heuristics...



genome A = +4 +5 + 8 +7 +1 -3 -2 +6 +9 +10 -11
genome B = +1 +2 +3 +4 +5 +6 +7 +8 +9 +10 +11

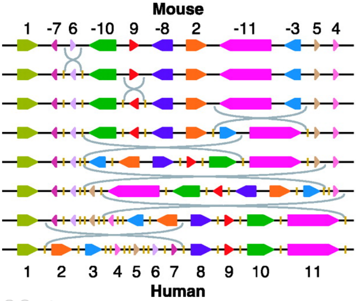


Focus

- Evolution through Genome Rearrangements
  - Is it interesting (and possible) to have such distances between MAGs?
  - Can we compare these distances to other types of distance?
- RNA-seq read alignments with PALMapper
  - Does the sequence I use exist in the database? is it certified?
  - How can we align considering SNPs?

Definition

- Genome = **ordered sequence** of genes
- GR = large-scale (=gene level) evolutionary events modifying the genome (thus the genes order)
- Studying evolution through GR:
  - take **2 species** (=2 genomes)
  - infer **minimum number** of GR between them (=distance)

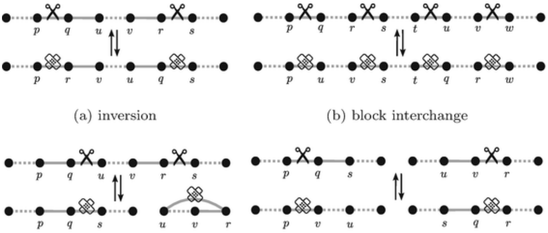


Source: https://www.pnas.org/content/100/13/7672

Definition

Heavily studied problems [FLR^{+}09] from 90's:

- different genomes: linear, circular, multichromosomal, with or without gene duplications, strand information or not...
- different genome models: (signed) permutations, (signed) strings, paths and cycles in graphs...
- different GR: inversion, transposition, double cut and join (DCJ)...



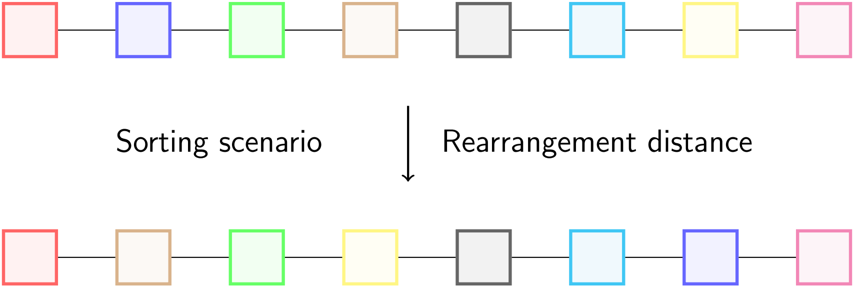
(a) inversion (b) block interchange

(c) fusion/fission (d) translocation

Source: [HMB18]

GR on both gene order and intergenic regions

- Standard models not realistic enough [BKBT16, BGKT16]
- Systematic underestimate of the distance

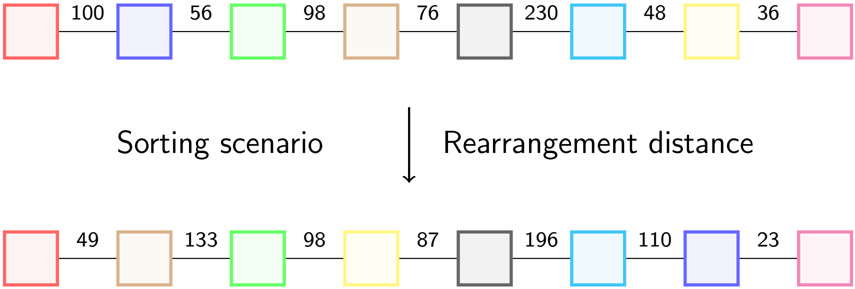


Sorting scenario

Rearrangement distance

GR on both gene order and intergenic regions

- Standard models not realistic enough [BKBT16, BGKT16]
- Systematic underestimate of the distance
- Genes separated by intergenic regions of different sizes
- Intergenic regions should be considered for computing rearrangement distances



Results

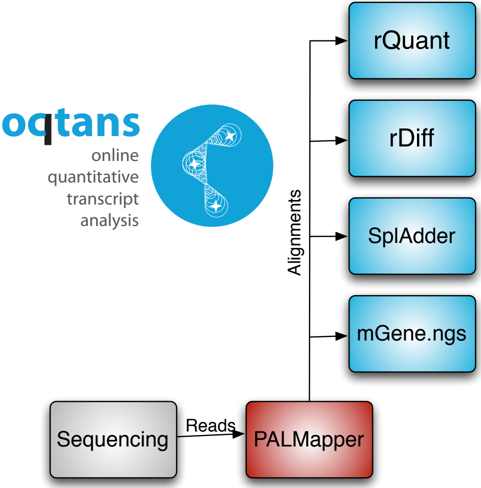
- Original paper [FJT17]
  - Same content of genes without duplication and unique operation (DCJ)
  - Already "difficult" even for this simplistic model

- Extended work (collaboration U. Campinas - Brazil) [OJF+21, BJF+20, OJF+20b, BJF+19, OJF+20a]
  - different gene contents
  - unbalanced intergenic sizes
  - different sets of operations

Focus

- Evolution through Genome Rearrangements
  - Is it interesting (and possible) to have such distances between MAGs?
  - Can we compare these distances to other types of distance?
- RNA-seq read alignments with PALMapper
  - Does the sequence I use exist in the database? is it certified?
  - How can we align considering SNPs?

PALMapper in oqtans



Isoform quantitation & bias modeling [BR10]

Tests for differential isoform expression [Stegle et al. 2010, 2012 i.p.]

Graph construction & sampling [Behr et al. 2012 i.p.]

Gene finding with RNA-seq evidence [Behr et al., 2010, 2012 i.p.]

Accurate spliced alignments [De Bona et al., 2008, Jean et al. 2010, 2012 i.p.]

Accuracy of downstream analysis drastically depends on accuracy of read mapping

Motivations

- Improve alignments by using more information:
  - Accurate splice site models
  - Intron length model
  - Quality scores model

  Idea: Use a machine learning method to infer an optimal scoring function

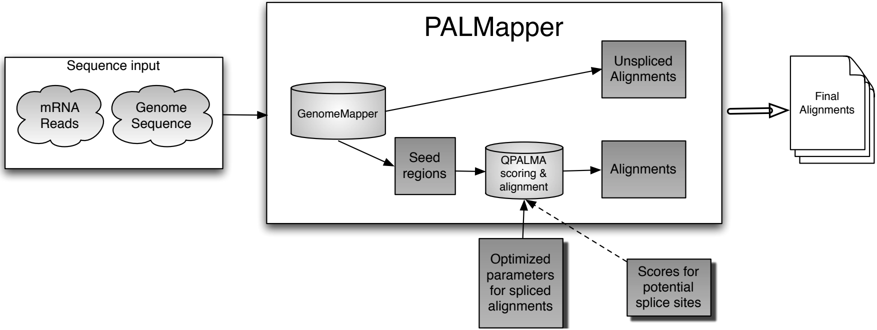
- Align reads efficiently:
  - Use a genomic mapper to find seed regions
  - Restrict the length of the genome to align against

  Idea: Adapt dynamic programming algorithm to RNA-seq specificities

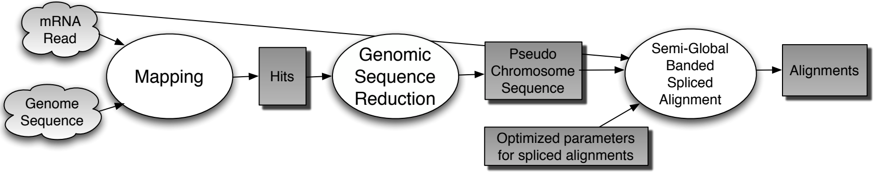
RNA-seq Read Alignments with PALMapper

PALMapper pipeline [JKS+10]

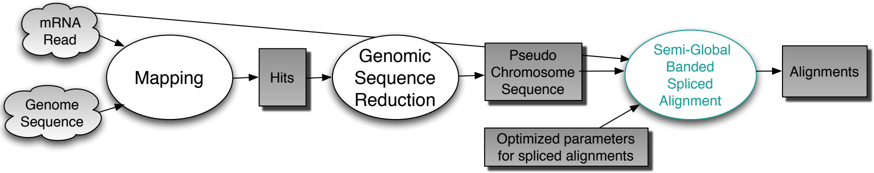
- **GenomeMapper** identifies unspliced alignments and *seed regions* for spliced reads [SHO+09]
- **QPALMA** infers spliced alignments from *seed regions* [DBOSR08]



Spliced alignments with PALMapper


mRNA Read
Genome Sequence
Mapping
Hits
Genomic Sequence Reduction
Pseudo Chromosome Sequence
Semi-Global Banded Spliced Alignment
Alignments
Optimized parameters for spliced alignments

Dynamic Programming Algorithm


mRNA Read
Genome Sequence
Mapping
Hits
Genomic Sequence Reduction
Pseudo Chromosome Sequence
Semi-Global Banded Spliced Alignment
Optimized parameters for spliced alignments
Alignments

The *seed position* inferred from the seed region guides a dynamic-programming-based alignment algorithm:

Semi-Global

Align the whole read with a portion of the pseudo chromosome sequence

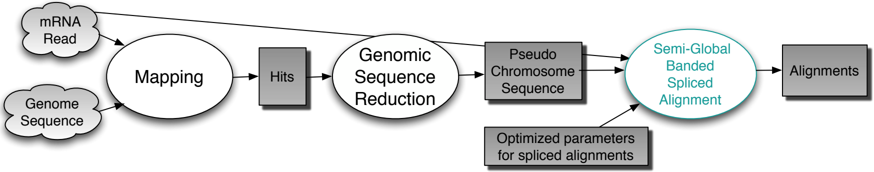
Banded

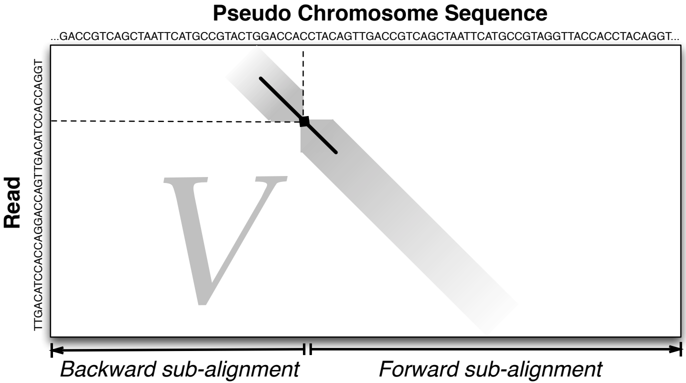
Limit the number of gaps from the perfect alignment

Spliced

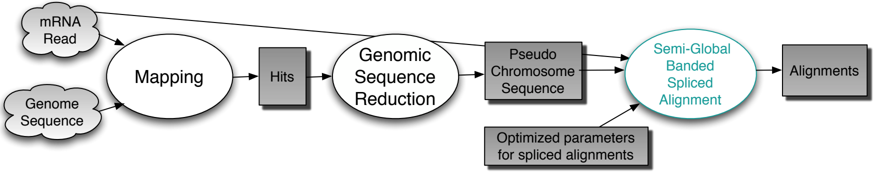
Allow long gaps corresponding to introns

Dynamic Programming Algorithm


mRNA Read
Genome Sequence
Mapping
Hits
Genomic Sequence Reduction
Pseudo Chromosome Sequence
Semi-Global Banded Spliced Alignment
Optimized parameters for spliced alignments
Alignments


Pseudo Chromosome Sequence
...GACCGTCAGCTAATTCATGCCGTACTGGACCACCTACAGTTGACCGTCAGCTAATTCATGCCGTAGGTTACCACCTACAGGT...
Read
TTGACATCCACCAGGACCAGTTGACATCCACCAGGT
V
Backward sub-alignment
Forward sub-alignment

Dynamic Programming Algorithm



mRNA Read

Genome Sequence

Mapping

Hits

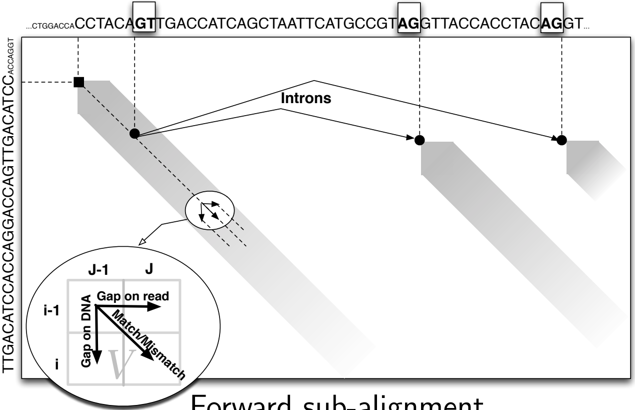
Genomic Sequence Reduction

Pseudo Chromosome Sequence

Semi-Global Banded Spliced Alignment

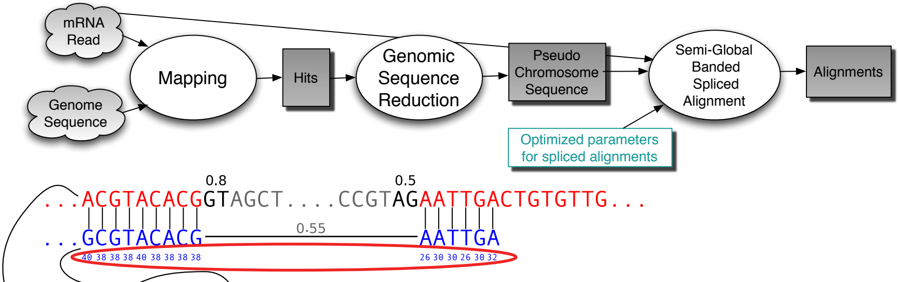
Optimized parameters for spliced alignments

Alignments


...CTGGACCACCTACAGTTGACCATCAGCTAATTCATGCCGTAGGTTACCACCTACAGGT...
TTGACATCCACCAGGACCAGTTGACATCCACCAGGT
Introns
J-1
J
i-1
i
Gap on read
Gap on DNA
Match/Mismatch
V

Forward sub-alignment

QPALMA extended scoring model


mRNA Read
Genome Sequence
Mapping
Hits
Genomic Sequence Reduction
Pseudo Chromosome Sequence
Semi-Global Banded Spliced Alignment
Alignments
Optimized parameters for spliced alignments
0.8
0.5
...ACGTACACGGTAGCT....CCGTAGAATTGACTGTGTTG...
...GCGTACACG
0.55
AATTGA
40 38 38 38 40 38 38 38 38
26 30 30 26 30 32

	gap	A	C	G	T	N
gap	0.33	$f_{-A}(\cdot)$	$f_{-C}(\cdot)$	$f_{-G}(\cdot)$	$f_{-T}(\cdot)$	$f_{-N}(\cdot)$
A	0.31	$f_{AA}(\cdot)$	$f_{AC}(\cdot)$	$f_{AG}(\cdot)$	$f_{AT}(\cdot)$	$f_{AN}(\cdot)$
C	0.44	$f_{CA}(\cdot)$	$f_{CC}(\cdot)$	$f_{CG}(\cdot)$	$f_{CT}(\cdot)$	$f_{CN}(\cdot)$
G	0.13	$f_{GA}(\cdot)$	$f_{GC}(\cdot)$	$f_{GG}(\cdot)$	$f_{GT}(\cdot)$	$f_{GN}(\cdot)$
T	0.55	$f_{TA}(\cdot)$	$f_{TC}(\cdot)$	$f_{TG}(\cdot)$	$f_{TT}(\cdot)$	$f_{TN}(\cdot)$
N	0.12	$f_{NA}(\cdot)$	$f_{NC}(\cdot)$	$f_{NG}(\cdot)$	$f_{NT}(\cdot)$	$f_{NN}(\cdot)$

Source of information

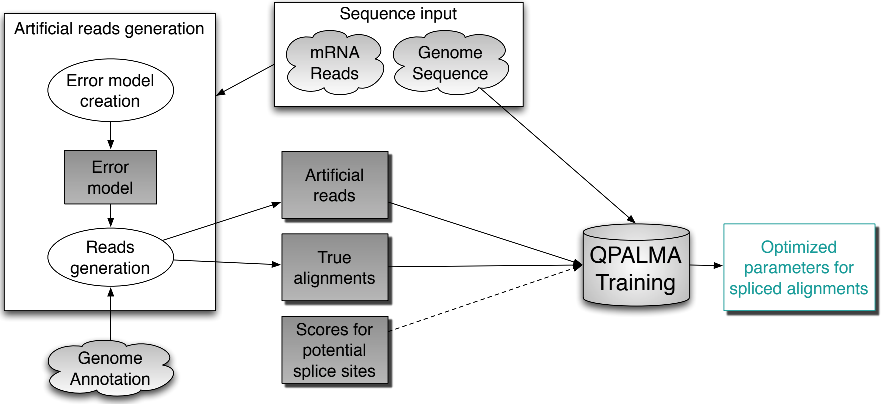
- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

Quality scoring $M : (\Sigma \times \mathbb{R}) \times \Sigma \to \mathbb{R}$

[DBOSR08]

QPALMA extended scoring model

Estimation of QPALMA scoring model via a large margin approach similar to SVMs



Results

Significant Speed Gain: QPalma *vs.* Palmapper

- Full sequencing run of *C. elegans* RNA-Seq data of 24 × 10^6 reads of 36-nucleotide length
- Evaluation of predicted introns

	TopHat	TopHat sen.	**QPALMA**	**PALMapper**
Recall	0.39	0.62	**0.65**	**0.65**
Precision	0.86	0.76	**0.88**	**0.88**
Running Time (min)	216	544	12000[1]	**186**

TopHat [TPS09]. [1] The running time for QPALMA was extrapolated.

Variant-Aware alignments

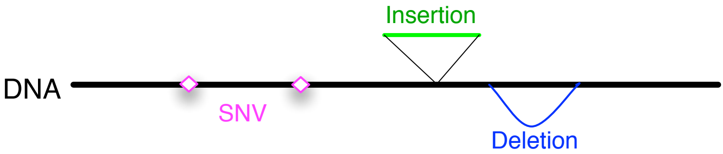
Motivation:

- Many reads may not have an alignment (errors, polymorphisms, RNA-editing)
  idea: Detect commonly occuring RNA/DNA differences and use during the alignment
- Genome of interest is unknown but a close relative is available
- Aligning against several close genomes is needed
  idea: Get variants between the genomes and use them during the alignment

Variant-Aware alignments

PALMapper strategy:

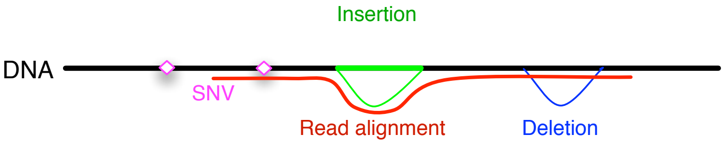
- Construct *super-sequence graph* containing all variants
- Use dynamic programming to align against all possible paths



Variant-Aware alignments

PALMapper strategy:

- Construct *super-sequence graph* containing all variants
- Use dynamic programming to align against all possible paths



- Strategy used in paper [DZM+14] about DNA methylation in *A. thaliana*

Conclusion

Thank you ! Gracias ! Merci !

El Morado January 2019

El Morado May 2023

References I

[BGKT16] Priscila Biller, Laurent Guéguen, Carole Knibbe, and Eric Tannier. Breaking Good: Accounting for Fragility of Genomic Regions in Rearrangement Distance Estimation. *Genome Biology and Evolution*, 8(5):1427–1439, 05 2016.

[BJF$^+$19] Klairton Lima Brito, Géraldine Jean, Guillaume Fertin, Andre Rodrigues Oliveira, Ulisses Dias, and Zanoni Dias. Sorting by Reversals, Transpositions, and Indels on Both Gene Order and Intergenic Sizes. In *International Symposium on Bioinformatics Research and Applications ISBRA 2019*, Bioinformatics Research and Applications, pages 28–39, Barcelona, Spain, June 2019.

[BJF$^+$20] Klairton Lima Brito, Géraldine Jean, Guillaume Fertin, Andre Rodrigues Oliveira, Ulisses Dias, and Zanoni Dias. Sorting by Genome Rearrangements on Both Gene Order and Intergenic Sizes. *Journal of Computational Biology*, 27(2):156–174, February 2020.

[BKBT16] Priscila Biller, Carole Knibbe, Guillaume Beslon, and Eric Tannier. Comparative genomics on artificial life. In Arnold Beckmann, Laurent Bienvenu, and Nataša Jonoska, editors, *Pursuit of the Universal*, pages 35–44, Cham, 2016. Springer International Publishing.

[BR10] R Bohnert and G Rätsch. rQuant.web: a tool for RNA-seq-based transcript quantitation. *Nucleic Acids Res*, 38(Web Server issue):348–351, Jul 2010.

[DBOSR08] Fabio De Bona, Stephan Ossowski, Korbinian Schneeberger, and Gunnar Rätsch. Optimal spliced alignments of short sequence reads. *Bioinformatics*, 24(16):i174–i180, August 2008.

[DZM$^+$14] Manu J Dubin, Pei Zhang, Dazhe Meng, Marie-Stanislas Remigereau, Edward J Osborne, Francesco Paolo Casale, Philipp Drewe, André Kahles, Géraldine Jean, Bjarni Vilhjálmsson, Joanna Jagoda, Selen Irez, Viktor Voronin, Qiang Song, Quan Long, Gunnar Rätsch, Oliver Stegle, Richard M Clark, and Magnus Nordborg. DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. *eLife*, 4:4:e05255, November 2014.

[FJT17] Guillaume Fertin, Géraldine Jean, and Eric Tannier. Algorithms for computing the double cut and join distance on both gene order and intergenic sizes. *Algorithms for Molecular Biology*, 12:16 (11 pages), 2017.

[FLR$^+$09] Guillaume Fertin, Anthony Labarre, Irena Rusu, Eric Tannier, and Stéphane Vialette. *Combinatorics of Genome Rearrangements*. Computational Molecular Biology. MIT Press, 2009.

References II

[HMB18] Tom Hartmann, Martin Middendorf, and Matthias Bernt. *Genome Rearrangement Analysis: Cut and Join Genome Rearrangements and Gene Cluster Preserving Approaches*, pages 261–289. Comparative Genomics: Methods and Protocols. Springer New York, New York, NY, 2018.

[JKS+10] G Jean, A Kahles, V T Sreedharan, F De Bona, and G Rätsch. RNA-seq read alignments with PALMapper. *Curr Protoc Bioinformatics*, 32(11):6.1–6.37, Dec 2010.

[OJF+20a] Andre Rodrigues Oliveira, Géraldine Jean, Guillaume Fertin, Klairton Lima Brito, Laurent Bulteau, Ulisses Dias, and Zanoni Dias. Sorting Signed Permutations by Intergenic Reversals. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.

[OJF+20b] Andre Rodrigues Oliveira, Géraldine Jean, Guillaume Fertin, Klairton Lima Brito, Ulisses Dias, and Zanoni Dias. A 3.5-Approximation Algorithm for Sorting by Intergenic Transpositions. In *7th International Conference, AlCoB 2020 (Algorithms for Computational Biology)*, pages 16–28, Missoula, United States, April 2020.

[OJF+21] Andre Rodrigues Oliveira, Geraldine Jean, Guillaume Fertin, Klairton Lima Brito, Ulisses Dias, and Zanoni Dias. Sorting Permutations by Intergenic Operations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–1, 2021.

[SHO+09] Korbinian Schneeberger, Jörg Hagmann, Stephan Ossowski, Norman Warthmann, Sandra Gesing, Oliver Kohlbacher, and Detlef Weigel. Simultaneous alignment of short reads against multiple genomes. *Genome biology*, 10(9):R98+, September 2009.

[TPS09] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.