Découvrir des SNPs (ou autre) dans les

1/ **données de séquençage**

2/ **sans génome de référence**

3/ **sans sortir de son fauteuil**

Pierre Peterlongo
GenScale team
Inria Rennes Bretagne-Atlantique

GENSCALE

Gen2Bio®
Jeudi 3 avril 2014
à Saint-Malo

Inria

NGS context

READ:

ACGACGTACGCATCAACACGTCACAAAGATACGACTGACGCATCAACACG
ACGCATCGGCGGGACTCATCTCTAACGCGAGCGACGAACGACTACGACAT

NGS context

Assembly

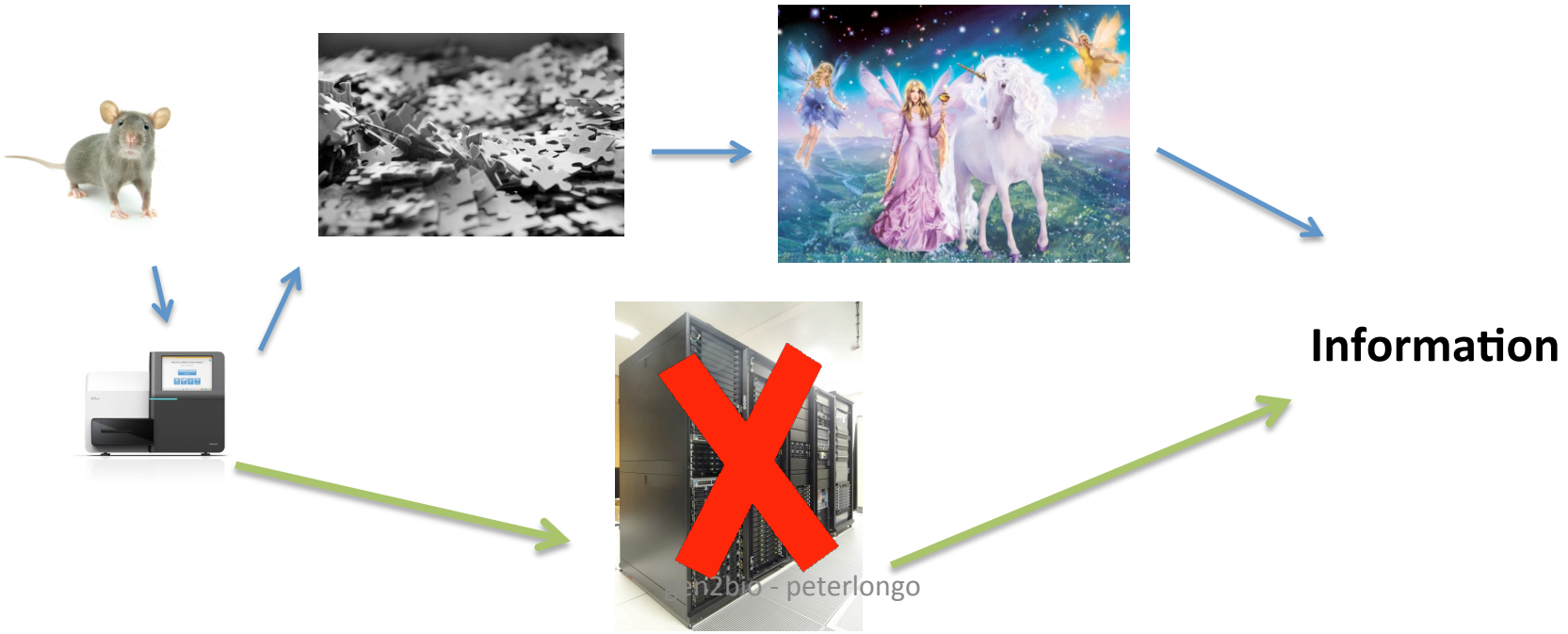
NGS context

Assembly

AACGTACGCATACGACGCATCAGACTACACGGAGACTAC
GACGACTACGACTACGACGACTACGCAGCATACACGCAC
AACTAGCTACGTACGTACGTGTGTCAGCTGCATATCGATC
GTCGATCGATCGATCGATCGTACGTACGTACGTACGTGT
AGGGGCAGACGACCGTACGTACGTGTGTCAGCTGCATAT
CGATCGTCGATCGATCGATCGATCGTACGTACGTACGTA
CGTGTAGGGGCAGACGACGCTACGACGCTACGACGACT
ACGACGACTACGCAGCATACACGCACAACTAGCTACGTA
CGACTACGCAGCATACACGCACAACTAGCTACGTACGTA
CGTGTGTCAGCTGCATATCGATCGTCGATCGATCGATCG
ATCGTACGTACGTACGTACGTGTAGGGGCAGACGACGCT
ACGACGACTACGACGACTACGCAGCATACACGCACAACT
AGCTACAACGTACGACTACGACGCTATCACATCAGCAGC
AGCATCATTCATACTCAACTCATCATACCAGCATCAT……..

What else?

- **Assembly, still**
  - Long
  - Huge memory footprint
  - Inexact
- **Avoid assembly?**
  - Bypass assembly to find biological meaning
  - Get bored with platforms ? ☺



What else?

- Alternative splicing (**kissplice**)
- SNPs (**discoSnp**)
- Inversion breakpoints (**takeABreak**)

Information

Raluca Uricaru
Guillaume Rizk
Vincent Lacroix
Elsa Quillery
Olivier Plantard
Rayan Chikhi
Claire Lemaitre
Pierre Peterlongo

DISCOSNP

De novo **disco**very of SNPs from raw NGS reads

DiscoSnp: SNPs

- "**Single Nucleotide** differs between members of a biological species or paired chromosomes"

- AACGGCATCAG**A**CGCGAGCATAC
  AACGGCATCAG**G**CGCGAGCATAC
- Importance of SNPs:
  - Markers, Personalized medicine, GWAS, forensic science, …

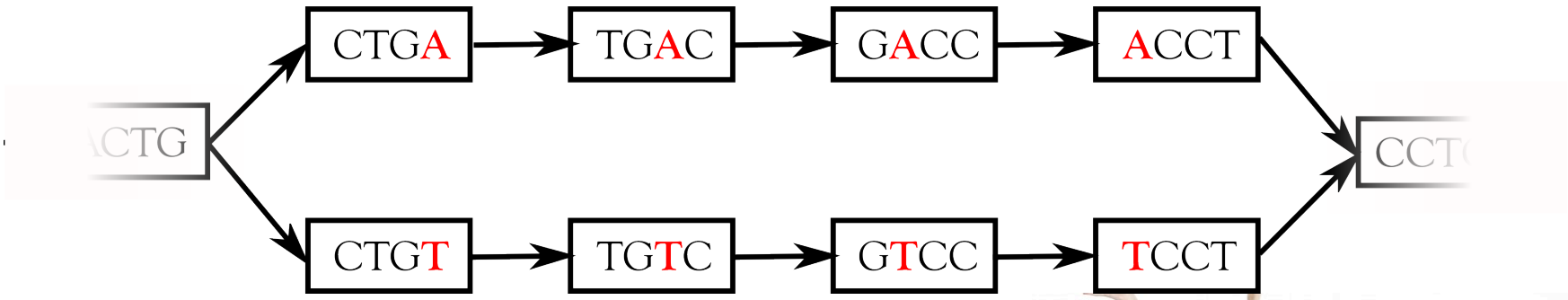
DiscoSnp: SNPs

DiscoSnp: Use cases

- **What you have**
  - sequenced reads
  - 1 to *n* sets (replicates, strains, individuals, …)
- **What you <u>don't</u> have**
  - reference genome (close or good)
- **What you want**
  - SNPs with their coverage/quality in each set
- **What you don't need**
  - genomic location

Methods at a glance

- Detect bubbles in *de-bruin* graph.



- All read sets mixed
- Uses minia impl. (**≤ 6GB for human**)

R. Chikhi, G. Rizk. Space-efficient and exact de Bruijn graph representation based on a Bloom filter, WABI 2012
K. Salikhov, G. Sacomoto and G. Kucherov. Using cascading Bloom filters to improve the memory usage for de Brujin graphs, WABI 2013

Overview

discoSnp

>SNP_higher_path_2|high|left_unitig_length_472|right_unitig_length_261|left_contig_length_472|right_contig_length_378|C1_8|C2_120|rank_0.88900
ttgcggataccgttgagacatcttataagtagacgcaatgcggaatcttatagaatcgcccgatagcgttgtgttggtggacacggctgattaccctctcacccgcgctattagcttccataccacctgcggccatccattaagatccgctgctcctcacgaaaaaagaattaataagaagtcccgtaacatgcggatttggtagtcgttatagacaactttactggggcgaactaaaacgcttgtggacagaattttggcagtggcaattaatctctaatgatgtgatattagggtctaaaatgtaagaattcggtgagttagattggacaaggggatccgaagatgttttggcgcagttagtcacaggggggagcccctgcctacaaaaagcgcttactgttgactgtctaggatacagcgaaagcggcagtcgttgaagcaaaagtgatatgtgcgacactgcatctagGCAGCGCAACAACGCAACAGCTCGAGGTGTACTTCGCAGAGAAACCGCACGTCCAGTTCTAacactctcatatgtgctcgtcgtttatgctttcggcgtgaaaactggtgcgccgggtgtctggagaccatccttcttgcgtatgactccaaggacagccatcacggtttgtgggttcactgggactgtcacgcttaaccggacggaactcgagaaggcatacgactggtcgtaagaccgctctgatccgacaccaccataacgcggcactcatgattatcatcacttttttagtccctattacagagctgccgggtggatgactctctaccgcgctctgtggaagtgcacttgatcgttttgctgtagaaaaaacttaataaacagaatgccgatgaaggcactactgtactaatagggccgggctacatgttaactac
>SNP_lower_path_2|high|left_unitig_length_472|right_unitig_length_261|left_contig_length_472|right_contig_length_378|C1_118|C2_6|rank_0.88900
ttgcggataccgttgagacatcttataagtagacgcaatgcggaatcttatagaatcgcccgatagcgttgtgttggtggacacggctgattaccctctcacccgcgctattagcttccataccacctgcggccatccattaagatccgctgctcctcacgaaaaaagaattaataagaagtcccgtaacatgcggatttggtagtcgttatagacaactttactggggcgaactaaaacgcttgtggacagaattttggcagtggcaattaatctctaatgatgtgatattagggtctaaaatgtaagaattcggtgagttagattggacaaggggatccgaagatgttttggcgcagttagtcacaggggggagcccctgcctacaaaaagcgcttactgttgactgtctaggatacagcgaaagcggcagtcgttgaagcaaaagtgatatgtgcgacactgcatctagGCAGCGCAACAACGCAACAGCTCGAGGTGTTCTTCGCAGAGAAACCGCACGTCCAGTTCTAacactctcatatgtgctcgtcgtttatgctttcggcgtgaaaactggtgcgccggtgtctggagaccatccttcttgcgtatgactccaaggacagccatcacggtttgtgggttcactgggactgtcacgcttaaccggacggaactcgagaaggcatacgactggtcgtaagaccgctctgatccgacaccaccataacgcggcactcatgattatcatcacttttttagtccctattacagagctgccgggtggatgactctctaccgcgctctgtggaagtgcacttgatcgttttgctgtagaaaaaacttaataaacagaatgccgatgaaggcactactgtactaatagggccgggctacatgttaactac

Main parameters

- *k* size of the used *k-mers*
- *c* minimal coverage

Output: fasta – sequence couples

>SNP_higher_path_2|high|left_unitig_length_472|right_unitig_length_261|
left_contig_length_472|right_contig_length_378|C1_8|C2_120|rank_0.88900

ttgcggataccgttgagacatcttataagtagacgcaatgcggaatcttatagaatcgcccgatagcgttgtgttggtggacacggctgatta
ccctctcacccgcgctattagcttccataccacctgcggccatccattaagatccgctgctcctcacgaaaaaagaattaataagaagtcccgt
aacatgcggatttggtagtcgttatagacaactttactggggcgaactaaaacgcttgtggacagaattttggcagtggcaattaatctctaa
tgatgtgatattagggtctaaaatgtaagaattcggtgagttagattggacaaggggatccgaagatgttttggcgcagttagtcacagggg
gagcccctgcctacaaaaagcgcttactgttgactgtctagggatacagcgaaagcggcagtcgttgaagcaaaagtgatatgtgcgacac
tgcatctagGCAGCGCAACAACGCAACAGCTCGAGGTGTACTTCGCAGAGAAACCGCACGTCCAGTTCTAacact
ctcatatgtgctcgtcgtttatgctttcggcgtgaaaactggtgcgccggtgtctggagaccatccttcttgcgtatgactccaaggacagccat
cacggtttgtgggttcactgggactgtcacgcttaaccggacggaactcgagaaggcatacgactggtcgtaagaccgctctgatccgacac
caccataacgcggcactcatgattatcatcacttttttagtccctattacagagctgccgggtggatgactctctaccgcgctctgtggaagtgc
acttgatcgttttgctgtagaaaaaacttaataaacagaatgccgatgaaggcactactgtactaatagggccgggctacatgttaactac

>SNP_lower_path_2|high|left_unitig_length_472|right_unitig_length_261|
left_contig_length_472|right_contig_length_378|C1_118|C2_6|rank_0.88900

ttgcggataccgttgagacatcttataagtagacgcaatgcggaatcttatagaatcgcccgatagcgttgtgttggtggacacggctgatta
ccctctcacccgcgctattagcttccataccacctgcggccatccattaagatccgctgctcctcacgaaaaaagaattaataagaagtcccgt
aacatgcggatttggtagtcgttatagacaactttactggggcgaactaaaacgcttgtggacagaattttggcagtggcaattaatctctaa
tgatgtgatattagggtctaaaatgtaagaattcggtgagttagattggacaaggggatccgaagatgttttggcgcagttagtcacagggg
gagcccctgcctacaaaaagcgcttactgttgactgtctagggatacagcgaaagcggcagtcgttgaagcaaaagtgatatgtgcgacac
tgcatctagGCAGCGCAACAACGCAACAGCTCGAGGTGTTCTTCGCAGAGAAACCGCACGTCCAGTTCTAacact
ctcatatgtgctcgtcgtttatgctttcggcgtgaaaactggtgcgccggtgtctggagaccatccttcttgcgtatgactccaaggacagccat
cacggtttgtgggttcactgggactgtcacgcttaaccggacggaactcgagaaggcatacgactggtcgtaagaccgctctgatccgacac
caccataacgcggcactcatgattatcatcacttttttagtccctattacagagctgccgggtggatgactctctaccgcgctctgtggaagtgc
acttgatcgttttgctgtagaaaaaacttaataaacagaatgccgatgaaggcactactgtactaatagggccgggcacatgttaactac

Output – bubble core

>SNP_higher_path_2|high|left_unitig_length_472|right_unitig_length_261|
left_contig_length_472|right_contig_length_378|C1_8|C2_120|rank_0.88900
ttgcggataccgttgagacatcttataagtagacgcaatgcggaatcttatagaatcgcccgatagcgttgtgttggtggacacggctgatta
ccctctcacccgcgctattagcttccataccacctgcggccatccattaagatccgctgctcctcacgaaaaaagaattaataagaagtcccgt
aacatgcggatttggtagtcgttatagacaactttactggggcgaactaaaacgcttgtggacagaattttggcagtggcaattaatctctaa
tgatgtgatattagggtctaaaatgtaagaattcggtgagttagattggacaaggggatccgaagatgttttggcgcagttagtcacagggg
gagcccctgcctacaaaaagcgcttactgttgactgtctagggatacagcgaaagcggcagtcgttgaagcaaaagtgatatgtgcgacac
tgcatctagGCAGCGCAACAACGCAACAGCTCGAGGTGTACTTCGCAGAGAAACCGCACGTCCAGTTCTAacact
ctcatatgtgctcgtcgtttatgctttcggcgtgaaaactggtgcgccggtgtctggagaccatccttcttgcgtatgactccaaggacagccat
cacggtttgtgggttcactgggactgtcacgcttaaccggacggaactcgagaaggcatacgactggtcgtaagaccgctctgatccgacac
caccataacgcggcactcatgattatcatcacttttttagtccctattacagagctgccgggtggatgactctctaccgcgctctgtggaagtgc
acttgatcgttttgctgtagaaaaaacttaataaacagaatgccgatgaaggcactactgtactaatagggccgggcacatgttaactac

>SNP_lower_path_2|high|left_unitig_length_472|right_unitig_length_261|
left_contig_length_472|right_contig_length_378|C1_8|C2_120|rank_0.88900
ttgcggataccgttgagacatcttataagtagacgcaatgcggaatcttatagaatcgcccgatagcgttgtgttggtggacacggctgatta
ccctctcacccgcgctattagcttccataccacctgcggccatccattaagatccgctgctcctcacgaaaaaagaattaataagaagtcccgt
aacatgcggatttggtagtcgttatagacaactttactggggcgaactaaaacgcttgtggacagaattttggcagtggcaattaatctctaa
tgatgtgatattagggtctaaaatgtaagaattcggtgagttagattggacaaggggatccgaagatgttttggcgcagttagtcacagggg
gagcccctgcctacaaaaagcgcttactgttgactgtctagggatacagcgaaagcggcagtcgttgaagcaaaagtgatatgtgcgacac
tgcatctagGCAGCGCAACAACGCAACAGCTCGAGGTGTTCTTCGCAGAGAAACCGCACGTCCAGTTCTAacact
ctcatatgtgctcgtcgtttatgctttcggcgtgaaaactggtgcgccggtgtctggagaccatccttcttgcgtatgactccaaggacagccat
cacggtttgtgggttcactgggactgtcacgcttaaccggacggaactcgagaaggcatacgactggtcgtaagaccgctctgatccgacac
caccataacgcggcactcatgattatcatcacttttttagtccctattacagagctgccgggtggatgactctctaccgcgctctgtggaagtgc
acttgatcgttttgctgtagaaaaaacttaataaacagaatgccgatgaaggcactactgtactaatagggccgggcacatgttaactac

- "core of length 2k-1"
- Polymorphic central position

Output – coverages

>SNP_higher_path_2|high|left_unitig_length_472|right_unitig_length_261|
left_contig_length_472|right_contig_length_378|**C1_8|C2_120**|rank_0.88900
ttgcggataccgttgagacatcttataagtagacgcaatgcggaatcttatagaatcgcccgatagcgttgtgttggtggacacggctgatta
ccctctcacccgcgctattagcttccataccacctgcggccatccattaagatccgctgctcctcacgaaaaaagaattaataagaagtcccgt
aacatgcggatttggtagtcgttatagacaactttactggggcgaactaaaacgcttgtggacagaattttggcagtggcaattaatctctaa
tgatgtgatattagggtctaaaatgtaagaattcggtgagttagattggacaaggggatccgaagatgttttggcgcagttagtcacagggg
gagcccctgcctacaaaaagcgcttactgttgactgtctagggatacagcgaaagcggcagtcgttgaagcaaaagtgatatgtgcgacac
tgcatctag**GCAGCGCAACAACGCAACAGCTCGAGGTGTTCTTCGCAGAGAAACCGCACGTCCAGTTCTA**acact
ctcatatgtgctcgtcgtttatgctttcggcgtgaaaactggtgcgccggtgtctggagaccatccttcttgcgtatgactccaaggacagccat
cacggtttgtgggttcactgggactgtcacgcttaaccggacggaactcgagaaggcatacgactggtcgtaagaccgctctgatccgacac
caccataacgcggcactcatgattatcatcacttttttagtccctattacagagctgccgggtggatgactctctaccgcgctctgtggaagtgc
acttgatcgttttgctgtagaaaaaacttaataaacagaatgccgatgaaggcactactgtactaatagggccgggcacatgttaactac

- Coverage per read set and per “allele”
- Quality not show on this example

>SNP_lowe[illegible]_unitig_length_472|right_unitig_length_261|
left_contig_[illegible]gth_472|right_contig_length_378|**C1_118|C2_6**|rank_0.88900
ttgcggataccgttgagacatcttataagtagacgcaatgcggaatcttatagaatcgcccgatagcgttgtgttggtggacacggctgatta
ccctctcacccgcgctattagcttccataccacctgcggccatccattaagatccgctgctcctcacgaaaaaagaattaataagaagtcccgt
aacatgcggatttggtagtcgttatagacaactttactggggcgaactaaaacgcttgtggacagaattttggcagtggcaattaatctctaa
tgatgtgatattagggtctaaaatgtaagaattcggtgagttagattggacaaggggatccgaagatgttttggcgcagttagtcacagggg
gagcccctgcctacaaaaagcgcttactgttgactgtctagggatacagcgaaagcggcagtcgttgaagcaaaagtgatatgtgcgacac
tgcatctag**GCAGCGCAACAACGCAACAGCTCGAGGTGTTCTTCGCAGAGAAACCGCACGTCCAGTTCTA**acact
ctcatatgtgctcgtcgtttatgctttcggcgtgaaaactggtgcgccggtgtctggagaccatccttcttgcgtatgactccaaggacagccat
cacggtttgtgggttcactgggactgtcacgcttaaccggacggaactcgagaaggcatacgactggtcgtaagaccgctctgatccgacac
caccataacgcggcactcatgattatcatcacttttttagtccctattacagagctgccgggtggatgactctctaccgcgctctgtggaagtgc
acttgatcgttttgctgtagaaaaaacttaataaacagaatgccgatgaaggcactactgtactaatagggccgggcacatgttaactac

Output – rank

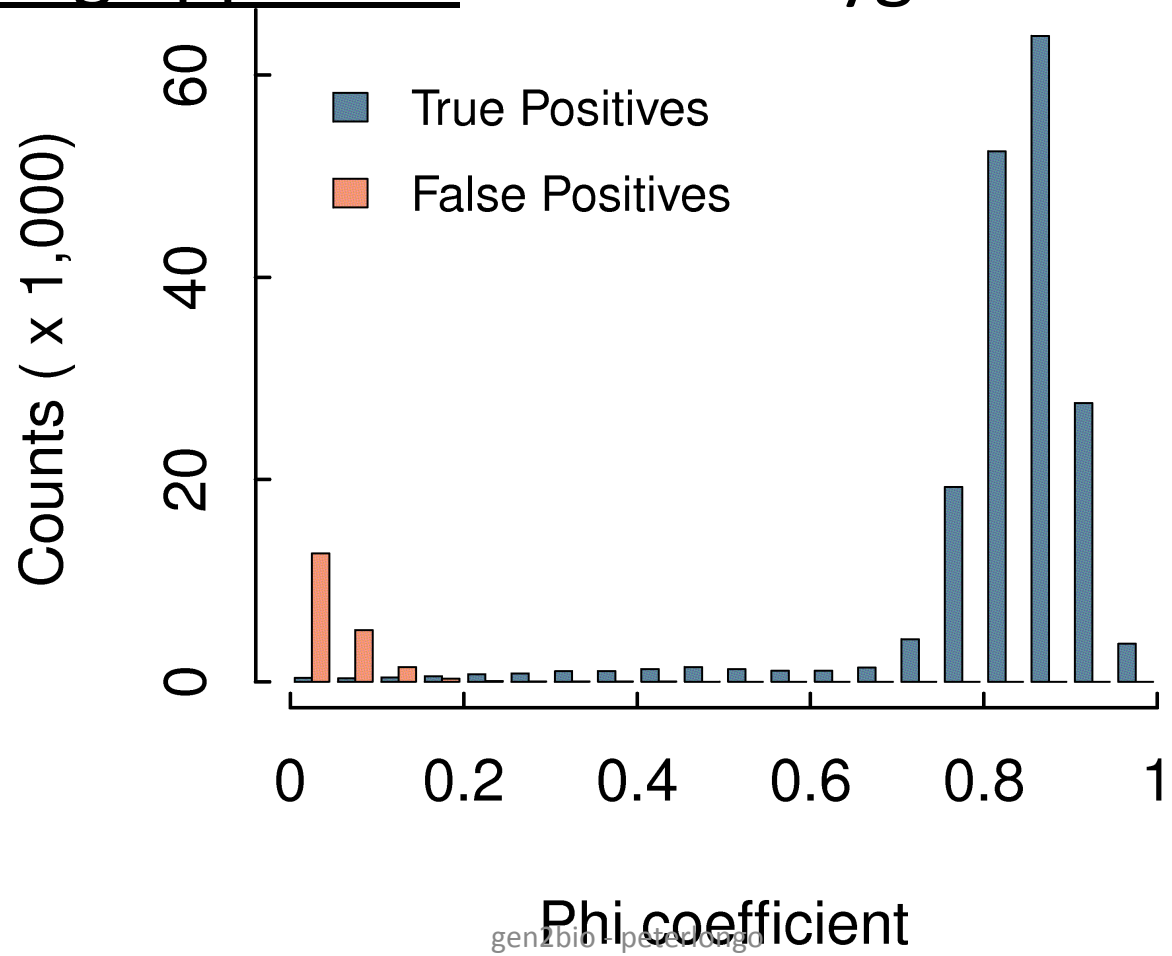
>SNP_higher_path_2|high|left_unitig_length_472|right_unitig_length_261|
left_contig_length_472|right_contig_length_378|C1_8|C2_120|rank_0.88900
ttgcggataccgttgagacatcttataagtagacgcaatgcggaatcttatagaatcgcccgatagcgttgtgttggtggacacggctgatta
ccctctcacccgcgctattagcttccataccacctgcggccatccattaagatccgctgctcctcacgaaaaaagaattaataagaagtcccgt
aacatgcggatttggtagtcgttatagacaactttactggggcgaactaaaacgcttgtggacagaattttggcagtggcaattaatctctaa
tgatgtgatattagggtctaaaatgtaagaattcggtgagttagattggacaaggggatccgaagatgttttggcgcagttagtcacagggg
gagcccctgcctacaaaaagcgcttactgttgactgtctagggatacagcgaaagcggcagtcgttgaagcaaaagtgatatgtgcgacac
tgcatctagGCAGCGCAACAACGCAACAGCTCGAGGTGTTCTTCGCAGAGAAACCGCACGTCCAGTTCTAacact
ctcatatgtgctcgtcgtttatgctttcggcgtgaaaactggtgcgccggtgtctggagaccatccttcttgcgtatgactccaaggacagccat
cacggtttgtgggttcactgggactgtcacgcttaaccggacggaactcgagaaggcatacgactggtcgtaagaccgctctgatccgacac
caccataacgcggcactcatgattatcatcacttttttagtccctattacagagctgccgggtggatgactctctaccgcgctctgtggaagtgc
acttgatcgttttgctgtagaaaaaacttaataaacagaatgccgatgaaggcactactgtactaatagggccgggcacatgttaactac

- 1 = SNP discriminative between read sets
- 0 = SNP non discriminative between read sets

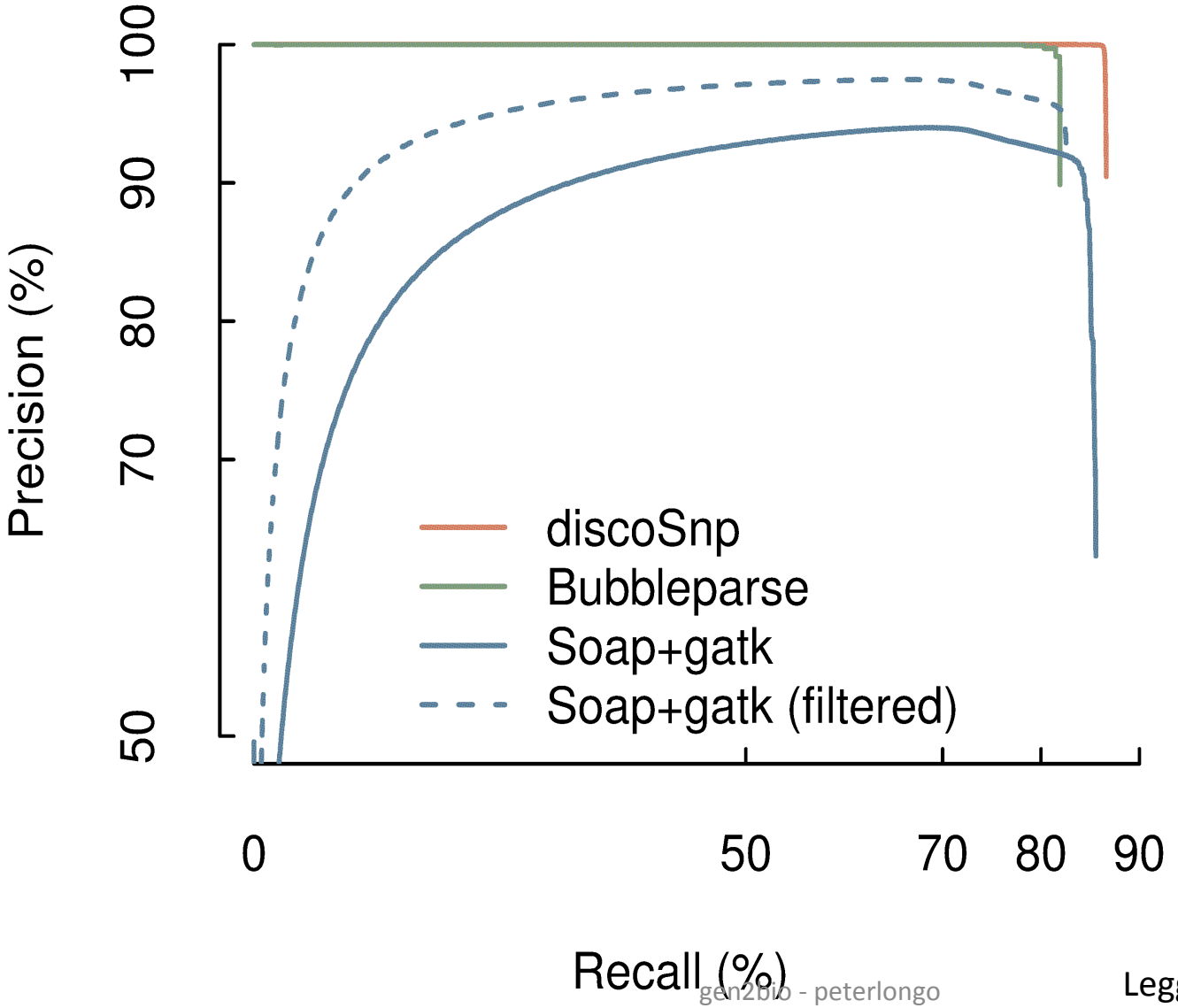
>SNP_lowe[illegible]_unitig_length_472|right_unitig_length_261|
left_contig_[illegible]gth_472|right_contig_length_378|C1_118|C2_6|rank_0.88900
ttgcggataccgttgagacatcttataagtagacgcaatgcggaatcttatagaatcgcccgatagcgttgtgttggtggacacggctgatta
ccctctcacccgcgctattagcttccataccacctgcggccatccattaagatccgctgctcctcacgaaaaaagaattaataagaagtcccgt
aacatgcggatttggtagtcgttatagacaactttactggggcgaactaaaacgcttgtggacagaattttggcagtggcaattaatctctaa
tgatgtgatattagggtctaaaatgtaagaattcggtgagttagattggacaaggggatccgaagatgttttggcgcagttagtcacagggg
gagcccctgcctacaaaaagcgcttactgttgactgtctagggatacagcgaaagcggcagtcgttgaagcaaaagtgatatgtgcgacac
tgcatctagGCAGCGCAACAACGCAACAGCTCGAGGTGTTCTTCGCAGAGAAACCGCACGTCCAGTTCTAacact
ctcatatgtgctcgtcgtttatgctttcggcgtgaaaactggtgcgccggtgtctggagaccatccttcttgcgtatgactccaaggacagccat
cacggtttgtgggttcactgggactgtcacgcttaaccggacggaactcgagaaggcatacgactggtcgtaagaccgctctgatccgacac
caccataacgcggcactcatgattatcatcacttttttagtccctattacagagctgccgggtggatgactctctaccgcgctctgtggaagtgc
acttgatcgttttgctgtagaaaaaacttaataaacagaatgccgatgaaggcactactgtactaatagggccgggcacatgttaactac

A few **results** – Ranking

- ranking approach for homozygous SNPs



A few (ranking) **results** - human

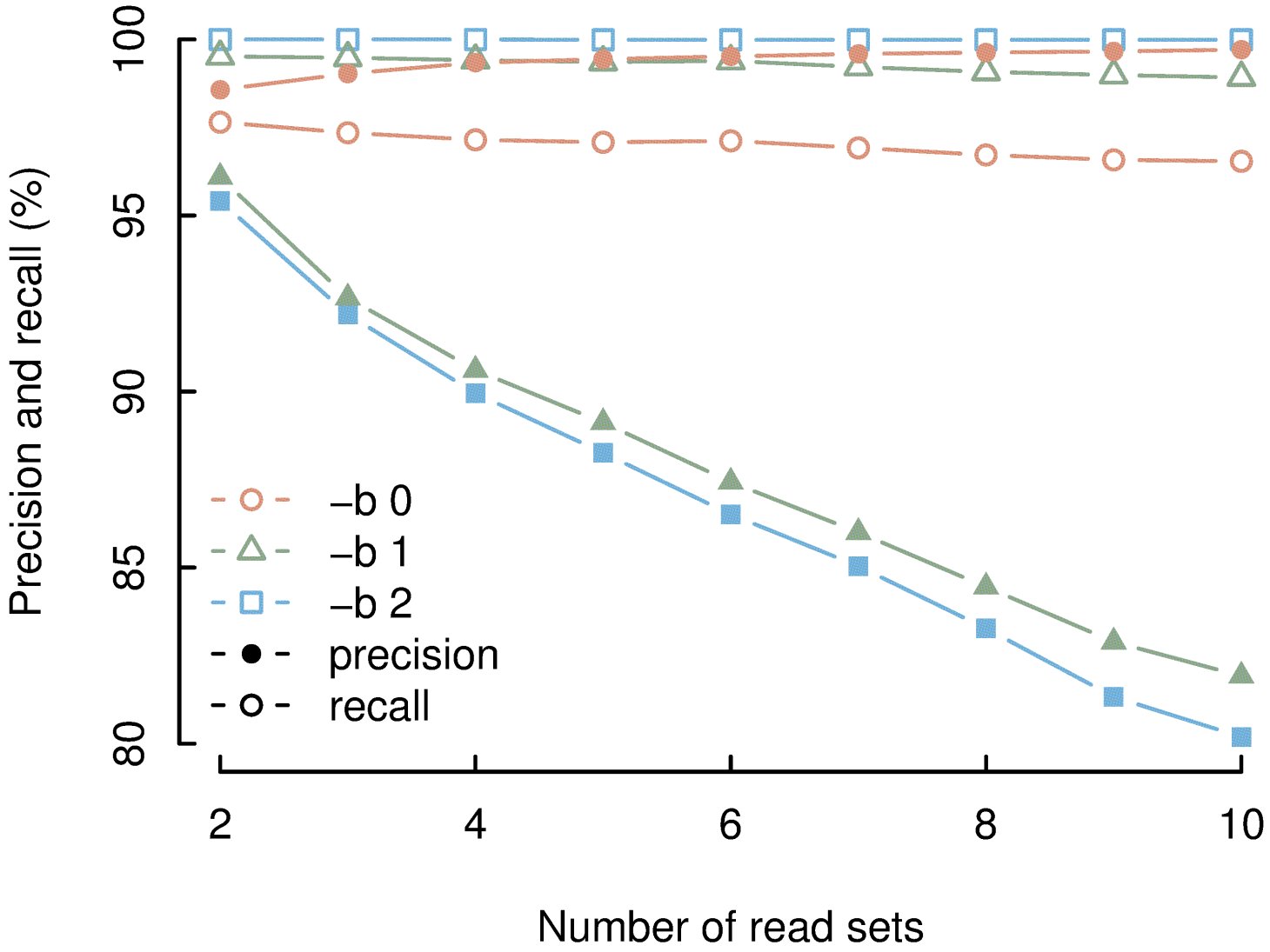


Alt. solutions:

- Bubbleparse
- Assembly +mapping

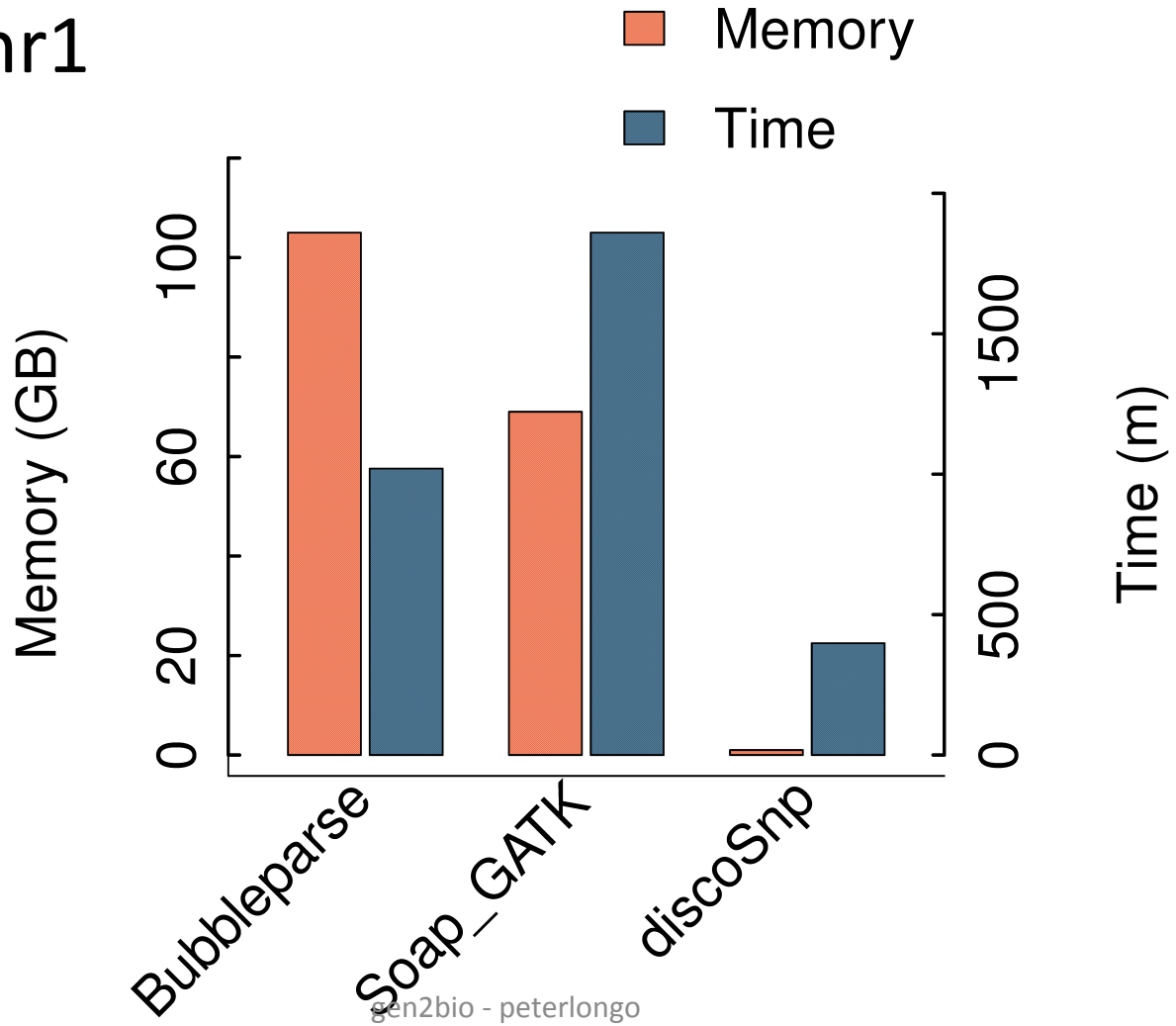
Leggett *et.al.* PLoS ONE, 2013 8(3)

A few results – *n* conditions



A few **results** – Time/memory

Human chr1



A few results – real data

- **Peapol**
  - 3.8 milions reads 454
  - Found 85% of biologically validated SNPs found with another method.

- **Mouse data set** (scaling test)
  - Near 3 billions Illumina reads
  - Less than 5 days
  - 5.7 GB of memory

A few results – real data

- **Tick data set** (454)
  - confirmed 96% of the predicted SNPs tested in vitro
  - Quillery *et al.* (2013)

    Molecular Ecology Resources

E. Quillery, O. Quenez, P. Peterlongo, O. Plantard. Development of genomic resources for the tick Ixodes ricinus: isolation and characterization of single nucleotide polymorphisms, Molecular Ecology Resources, 2013.

Availability

- **http://colibread.inria.fr/discosnp**
  - CeCILL License
- http://toolshed.genouest.org
- http://galaxy.genouest.org
- Packages debian and ubuntu (**GenOuest**)

Colib'read

Claire Lemaitre
Liviu Ciortuz
Pierre Peterlongo

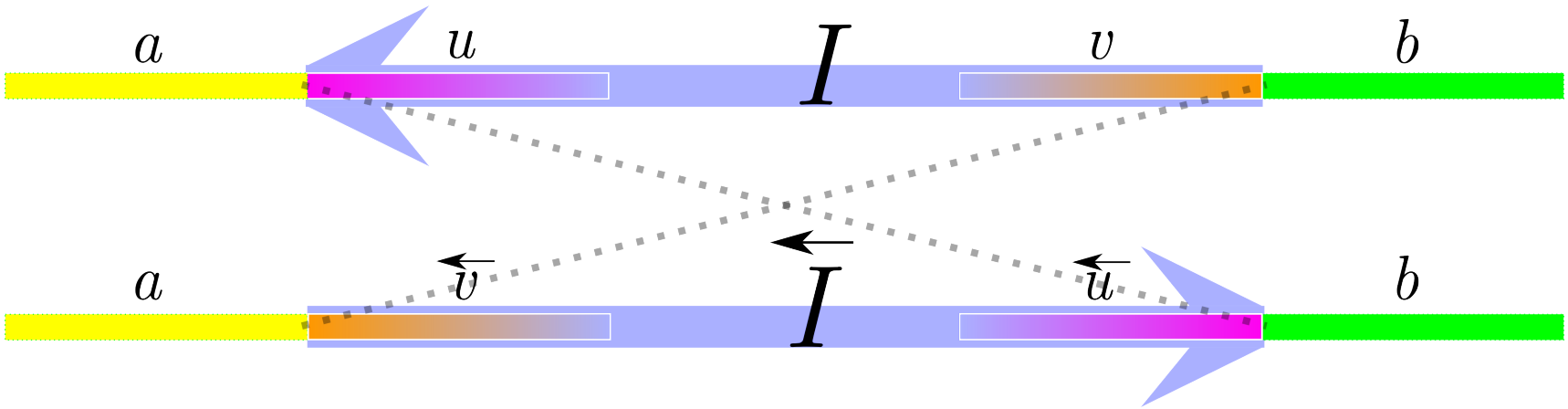
TakeABreak

- Mapping-free and assembly-free discovery of inversion breakpoints from raw NGS reads

Inria

GENSCALE

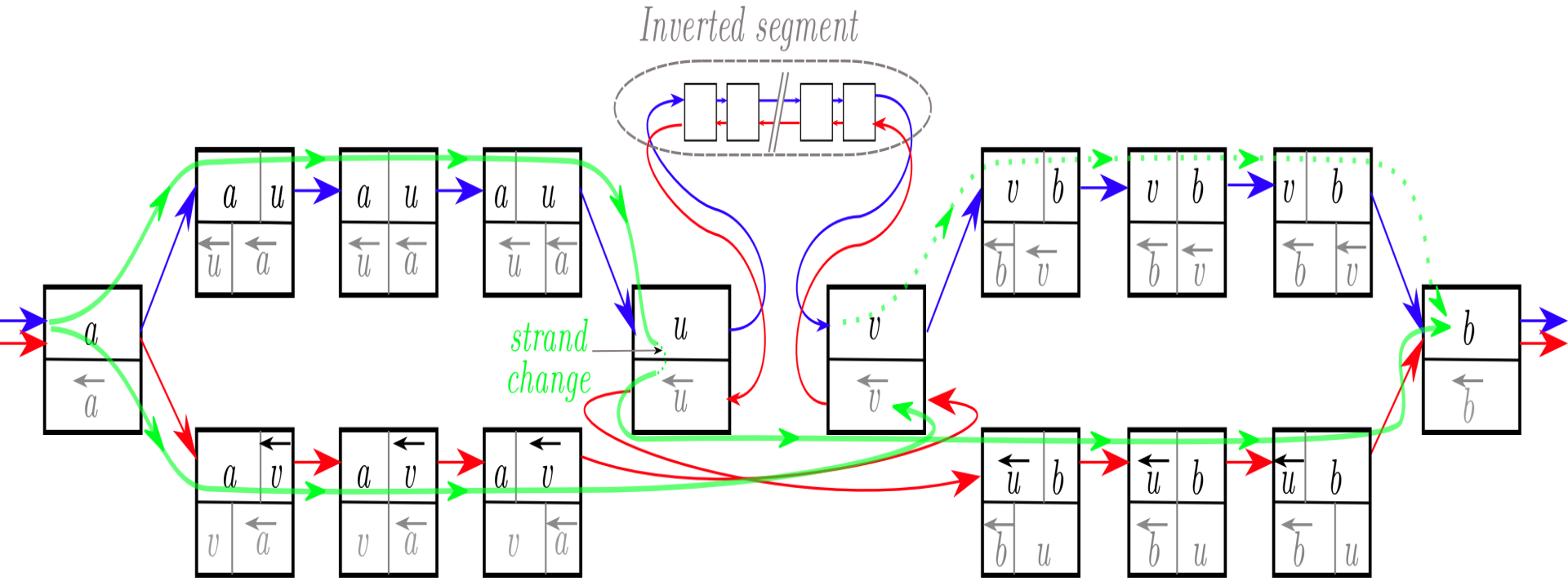
Inversion



Use cases

- **What you have**:
  - sequenced reads
  - 1 to *n* sets (replicates, strains, individuals, …)
- **What you don't have**:
  - reference genome (close or good)
- **What you want**:
  - Inversion break points
- **What you don't need**: genomic location

Inversion in the de Bruijn graph



TakeABreak

- Detects the inversion pattern
- Have several filters discarding approximate repeats from inversions
- Low memory footprint

TakeABreak - Results

E. Coli 100% precision 100% recall

C. Elegans 96% precision 99.07% recall

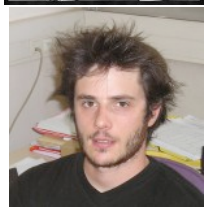
Human 87.6% precision 92.5% recall

Availability

- **http://colibread.inria.fr/takeabreak**
  - A-GPL License

Colib'read

Thanks



pierre.peterlongo@inria.fr

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR
ANR
Inria