

# Haplotype Assembly

Lesson of Bioinformatics

Milano-Bicocca, 6/5/2015

Simone Zaccaria

[simone.zaccaria@disco.unimib.it](mailto:simone.zaccaria@disco.unimib.it)

# Overview

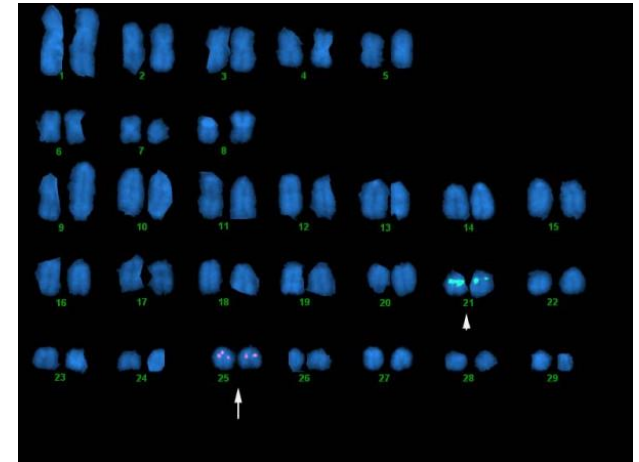
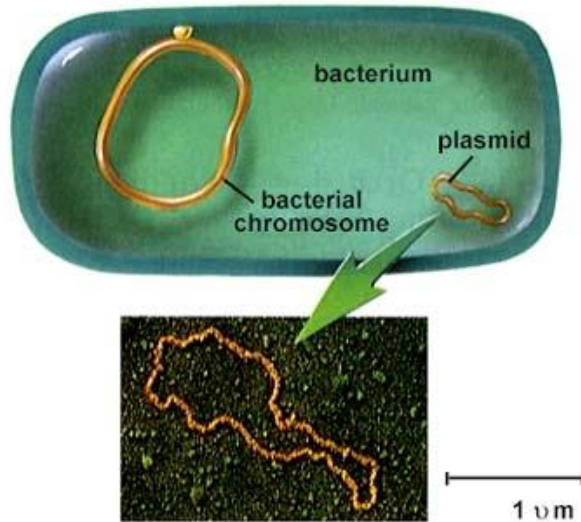
- Problem: introduction and motivation
- Formulations of the problem
- Approaches: FPT and approximation
- Proposal of Thesis

# Genomes

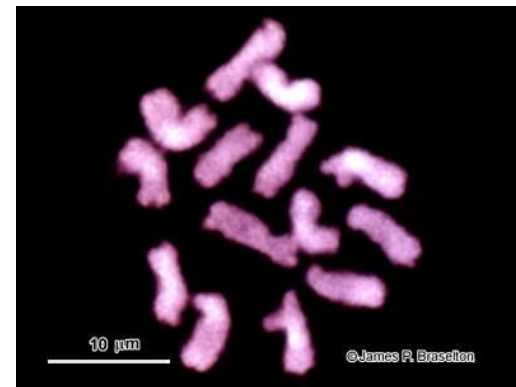
- Genome = collection of chromosomes



Human



Atlantic Salmon



Pea plant

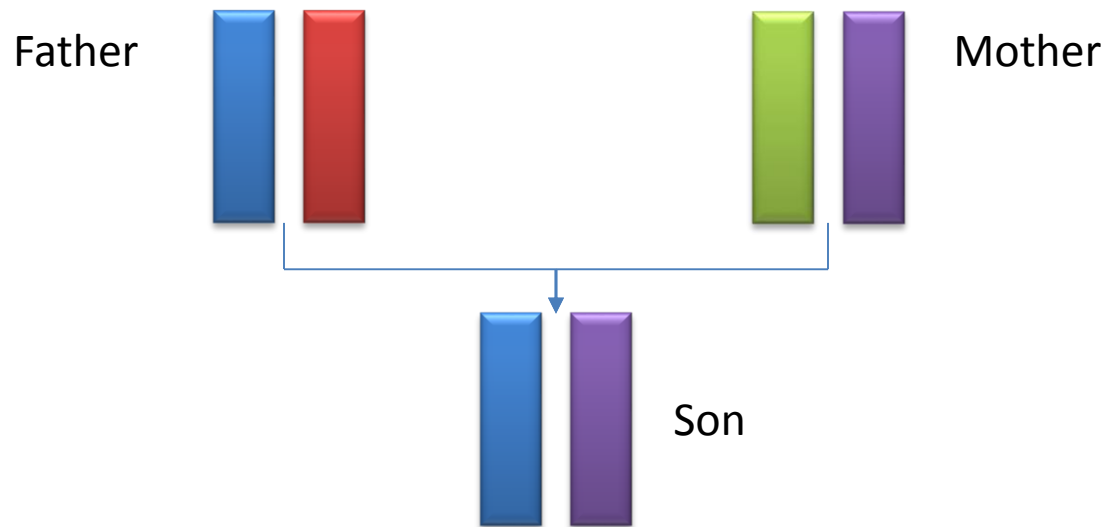
# Diploid Organisms

- Diploid = two sets of homologous chromosomes



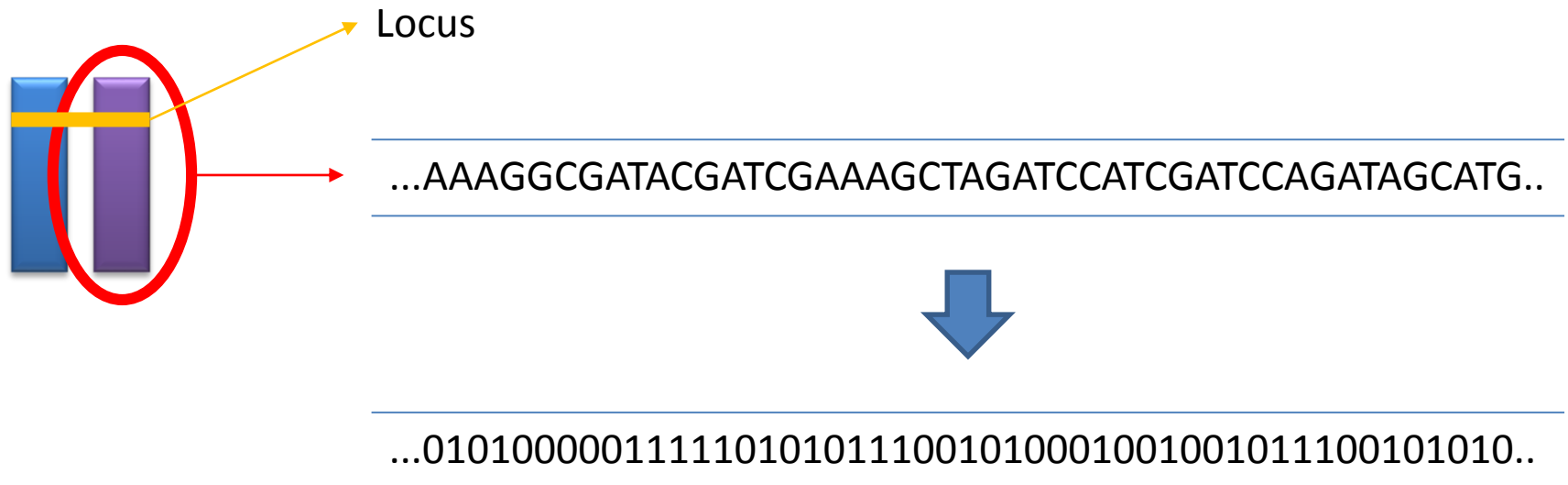
# Mendelian Laws

- For each chromosome there are 2 copies:
  - 1 inherited from the mother
  - 1 inherited from the father
- For now, ignore recombinations



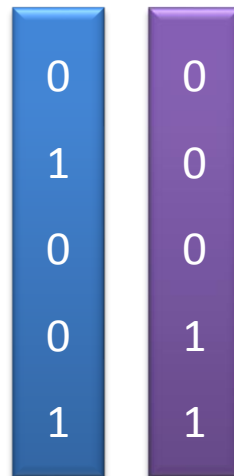
# Haplotype

- Haplotype = One copy of a chromosome
- Locus = genome position
- Haplotype is represented as a binary vector (0 major/1 minor allele)



# Single Nucleotide Polymorphisms (SNPs)

- Each pair of homologous haplotypes exhibits differences in terms of **Single Nucleotide Polymorphisms (SNPs)**
- SNPs = Heterozygous positions

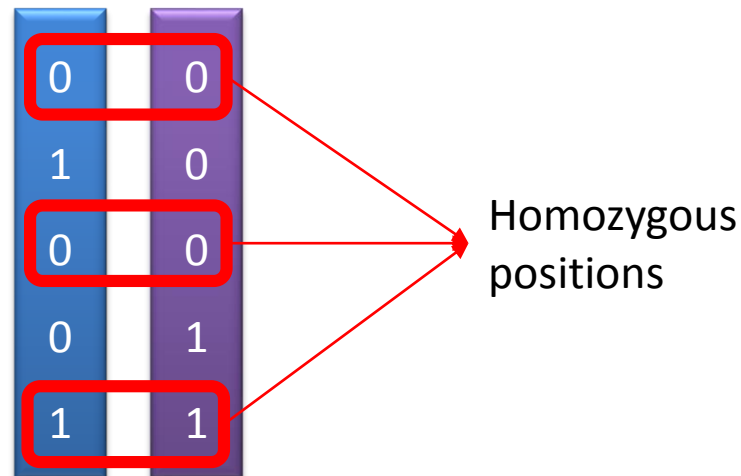


The diagram shows two vertical bars representing homologous haplotypes. The left bar is blue and contains the binary sequence 0, 1, 0, 0, 1 from top to bottom. The right bar is purple and contains the binary sequence 0, 0, 0, 1, 1 from top to bottom. The positions where the values differ (the second, third, and fourth positions) represent Single Nucleotide Polymorphisms (SNPs).

0	0
1	0
0	0
0	1
1	1

# Single Nucleotide Polymorphisms (SNPs)

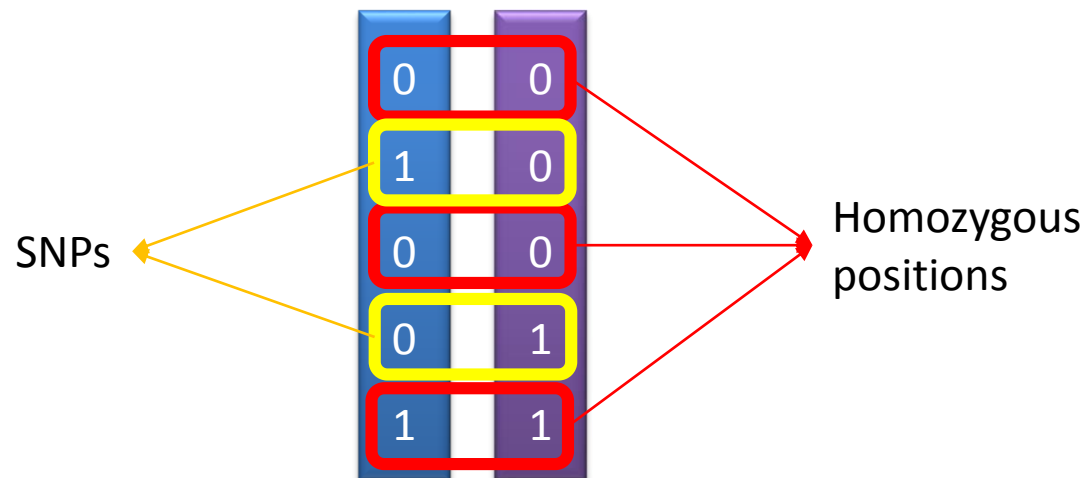
- Each pair of homologous haplotypes exhibits differences in terms of **Single Nucleotide Polymorphisms (SNPs)**
- SNPs = Heterozygous positions





# Single Nucleotide Polymorphisms (SNPs)

- Each pair of homologous haplotypes exhibits differences in terms of **Single Nucleotide Polymorphisms (SNPs)**
- SNPs = Heterozygous positions



# Motivations

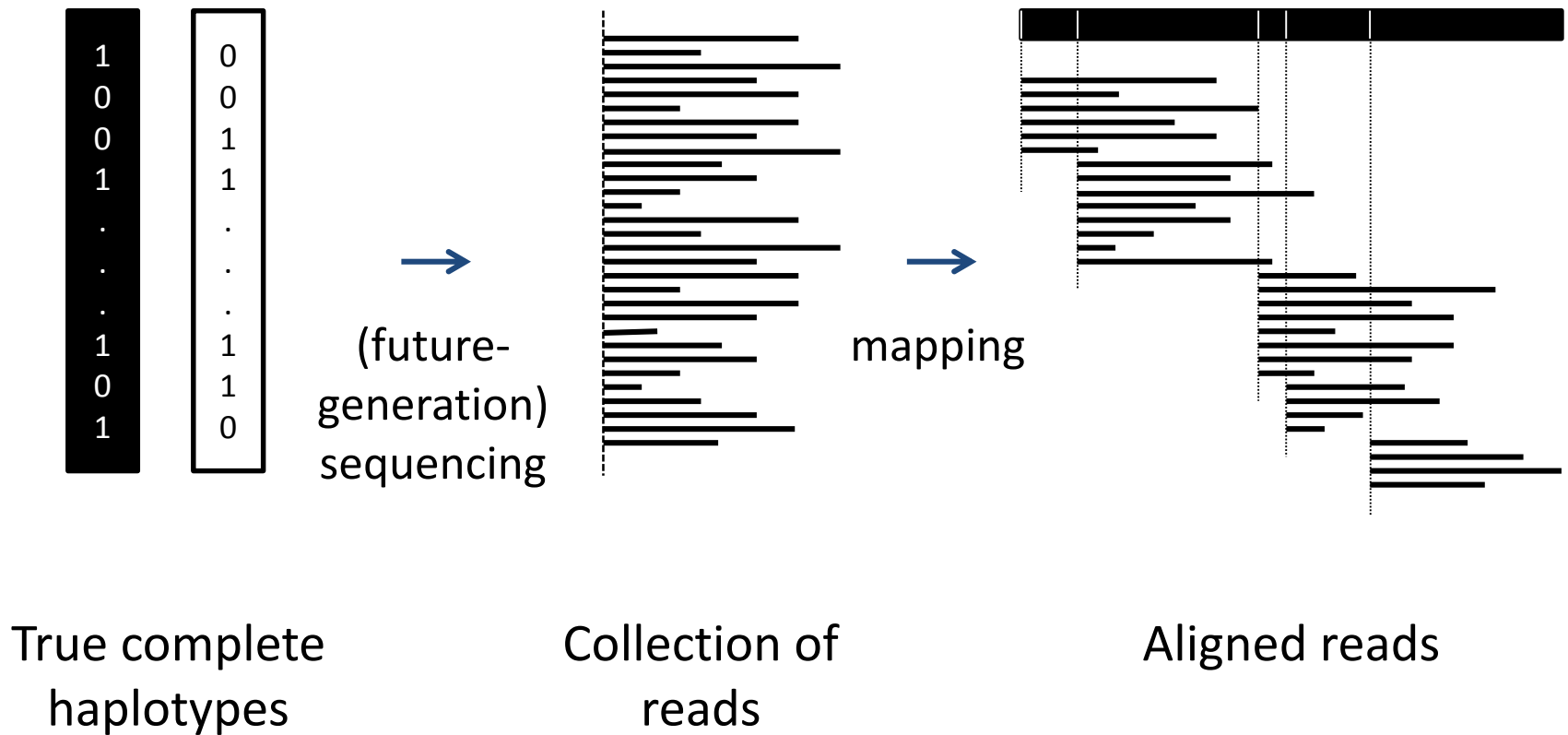
- The haplotypes are of fundamental importance for many applications (Bronwing and Browning, Nature Reviews, 2011):
  - genetic variations / gene function
  - genetic variations / disease susceptibility
  - Genetic variations / drug resistance
  - Etc...
- A whole-experimental reconstruction of the haplotypes is not cost effective => **Reconstruction of haplotypes** from easier collectable data is necessary:
  - Computational reconstruction (i.e., Statistical approach)
  - Experimental reconstruction through combinatorial approaches

# Sequencing

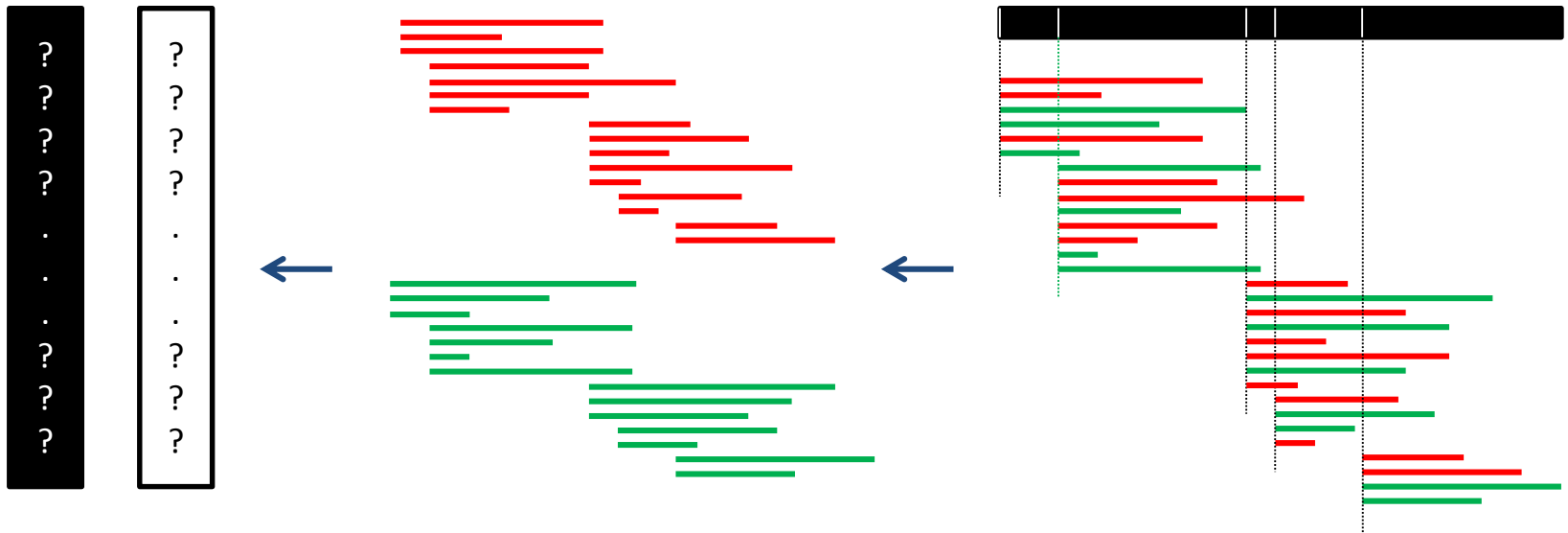
- Read = A fragment of a single strand of DNA (i.e., chromosomes)
- Sequencing = collect reads from a copy of a chromosome



# Pipeline



# Pipeline



- The reads are bipartite in order to reconstruct the two haplotypes

# Fragment Matrix

← SNP positions →

← reads →	1	0	-	1	1	-	-	0	1	-
	-	1	0	0	0	-	-	1	-	1
	-	-	1	1	1	-	1	0	1	-
	-	-	0	0	0	-	1	1	-	1
	-	-	-	0	0	0	1	1	0	1

- Each row corresponds to a read
- Each column correspond to a SNP position
- NOTICE: Homozygous positions may be ignored since they do not give information on the reconstruction/bipartition

# Fragment Matrix

Conflict

← SNP positions →

← reads →

1	0	-	1	1	-	-	0	1	-
-	1	0	0	0	-	-	1	-	1
-	-	1	1	1	-	1	0	1	-
-	-	0	0	0	-	1	1	-	1
-	-	-	0	0	0	1	1	0	1

- Conflict = different values on the same position
- The conflict may guide the reconstruction of the bipartition

# Fragment Matrix

← SNP positions →

← reads →	1	0	-	1	1	-	-	0	1	-
	-	1	0	0	0	-	-	1	-	1
	-	-	1	1	1	-	1	0	1	-
	-	-	0	0	0	-	1	1	-	1
	-	-	-	0	0	0	1	1	0	1

- Conflict = different values on the same position
- The conflict may guide the reconstruction of the bipartition



# Errors

← SNP positions →

← reads →	1	0	-	1	1	-	-	0	1	-
	-	1	0	0	0	-	-	1	-	1
	-	-	1	0	1	-	1	0	1	-
	-	-	0	1	0	-	1	1	-	0
	-	-	-	0	0	0	1	1	0	1

- Sequencing or mapping errors => the reads cannot be unambiguously bipartited
- The errors lead to an optimization problem

# MFR and MSR

## Minimum Fragment Removal (MFR)

1	0	-	1	1	-	-	0	1	-
-	1	0	0	0	-	-	1	-	1
-	-	1	0	1	-	1	0	1	-
-	-	0	1	0	-	1	1	-	0
-	-	-	0	0	0	1	1	0	1



1	0	-	1	1	-	-	0	1	-
-	1	0	0	0	-	-	1	-	1
-	-	-	0	0	0	1	1	0	1

## Minimum SNP Removal (MSR)

1	0	-	1	1	-	-	0	1	-
-	1	0	0	0	-	-	1	-	1
-	-	1	0	1	-	1	0	1	-
-	-	0	1	0	-	1	1	-	0
-	-	-	0	0	0	1	1	0	1



1	0	-	1	-	-	0	1
-	1	0	0	-	-	1	-
-	-	1	1	-	1	0	1
-	-	0	0	-	1	1	-
-	-	-	0	0	1	1	0

# Minimum Error Correction (MEC)

← SNP positions →

	1	0	-	1	1	-	-	0	1	-
← reads →	-	1	0	0	0	-	-	1	-	1
	-	-	1	0	1	-	1	0	1	-
	-	-	0	1	0	-	1	1	-	0
	-	-	-	0	0	0	1	1	0	1

- **Input:** Fragment matrix
- **Output:** Minimum number of corrections that allow to unambiguously bipartite the reads
- A weighted variant (wMEC) assigns a weight to each element and minimize the total correcting weight => improve accuracy

# Minimum Error Correction (MEC)

← SNP positions →

	1	0	-	1	1	-	-	0	1	-
	-	1	0	0	0	-	-	1	-	1
	-	-	1	0	1	-	1	0	1	-
	-	-	0	1	0	-	1	1	-	0
	-	-	-	0	0	0	1	1	0	1
← reads →										

- **Input:** Fragment matrix
- **Output:** Minimum number of corrections that allow to unambiguously bipartite the reads
- A weighted variant (wMEC) assigns a weight to each element and minimize the total correcting weight => improve accuracy

# MEC Variants

**Binary MEC**  
(no holes, no gaps)

0	0	1	0	1	1	0	1	0	1
0	0	0	1	0	1	0	1	0	1
0	0	1	1	1	0	1	0	1	0

**Gapless MEC**  
(holes, no gaps)

0	0	1	0	1	1	-	-	-	-
-	0	0	1	0	1	0	-	-	-
-	-	-	-	1	0	1	0	1	0

**Gap MEC**  
(holes, gaps)

0	0	-	-	1	1	-	-	-	-
-	0	-	-	0	-	0	-	-	-
-	1	1	-	-	1	0	-	-	-

# MEC Variants

**Binary MEC**  
(no holes, no gaps)

- **Computational Complexity?**
- **Scheme of Approximation**

**Gapless MEC**  
(holes, no gaps)

- **NP-Hard**
- **Approximation?**

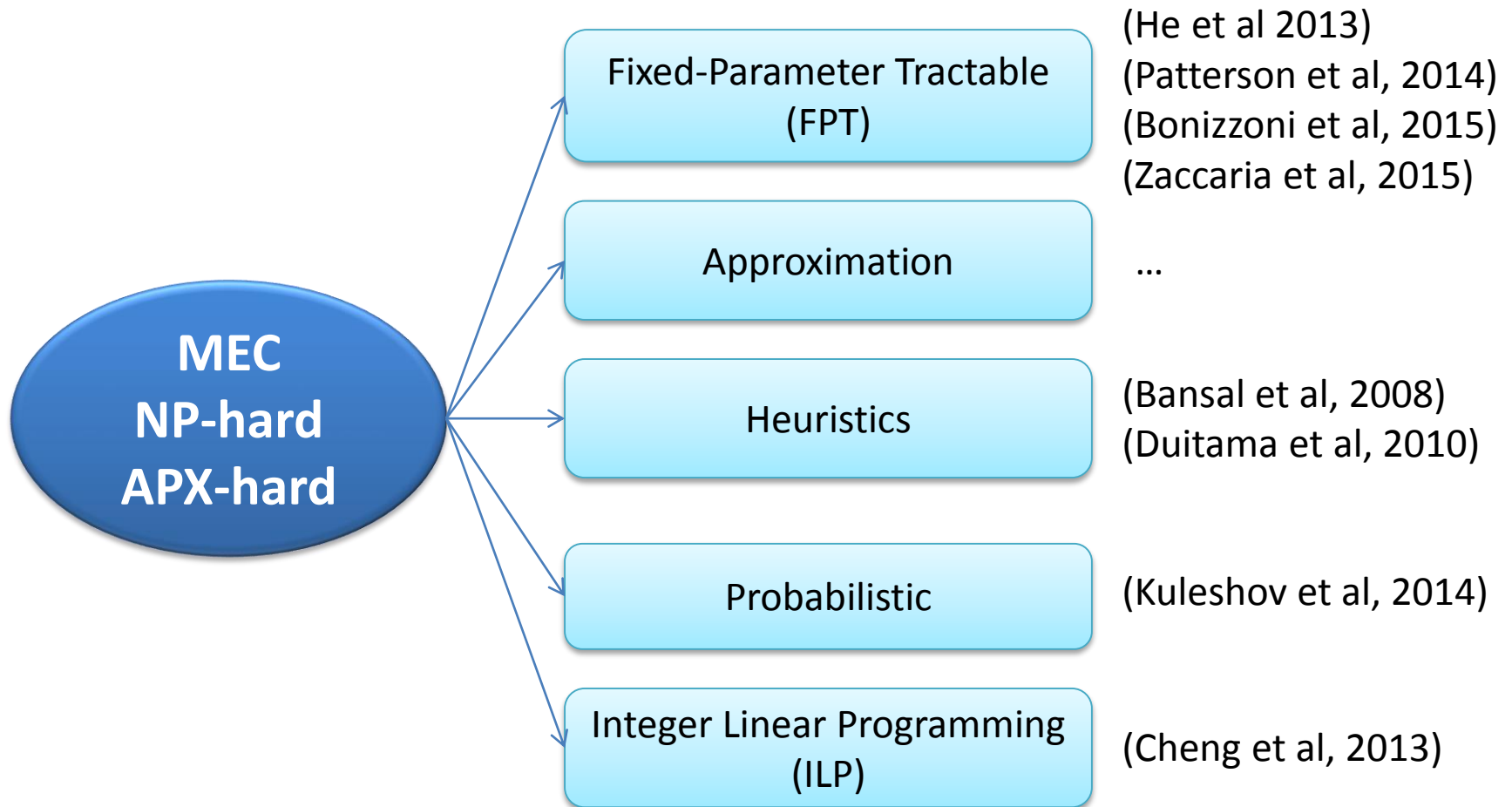
**Gap MEC**  
(holes, gaps)

- **NP-Hard**
- **APX-Hard (recently, not in APX)**

\*Cilibrasi *et al.*, *The Complexity of the Single Individual SNP Haplotyping Problem*, Algorithmica, 2007.

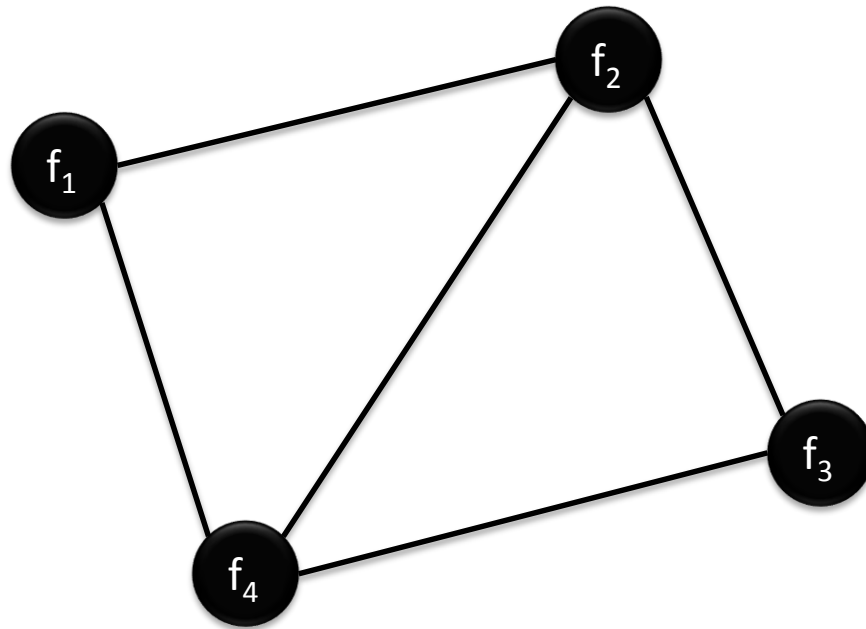
\*Bonizzoni *et al.*, 2015

# Approaches



# Graph Model: Fragment Conflict Graph

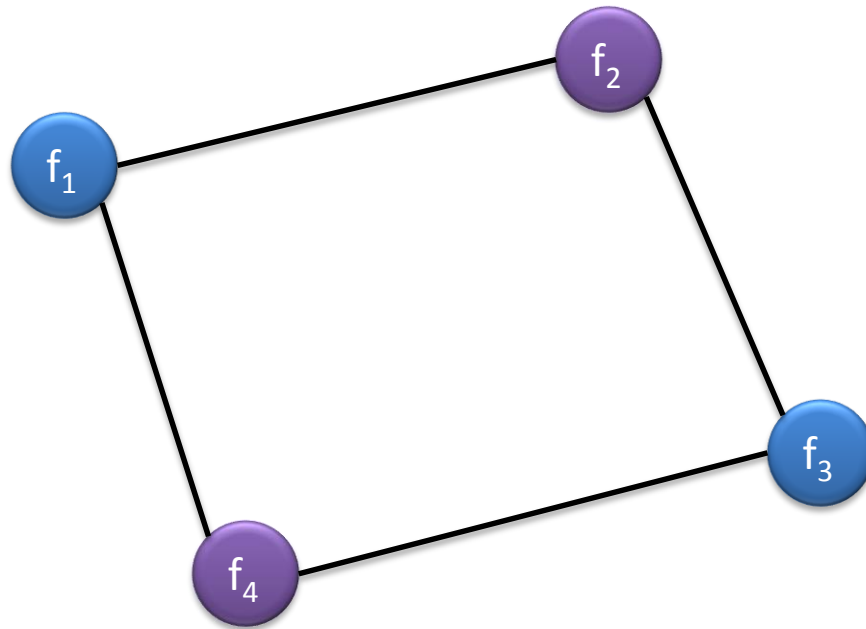
	$p_1$	$p_2$	$p_3$	$p_4$
$f_1$	0	1	-	-
$f_2$	-	0	1	1
$f_3$	-	-	0	1
$f_4$	1	-	-	0





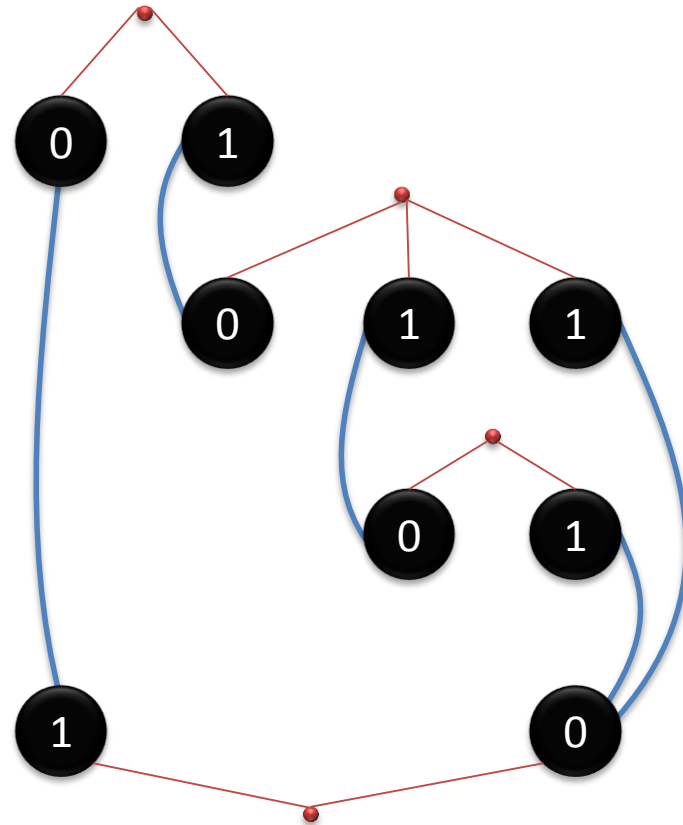
# Graph Model: Fragment Conflict Graph

	$p_1$	$p_2$	$p_3$	$p_4$
$f_1$	0	1	-	-
$f_2$	-	0	1	0
$f_3$	-	-	0	1
$f_4$	1	-	-	0



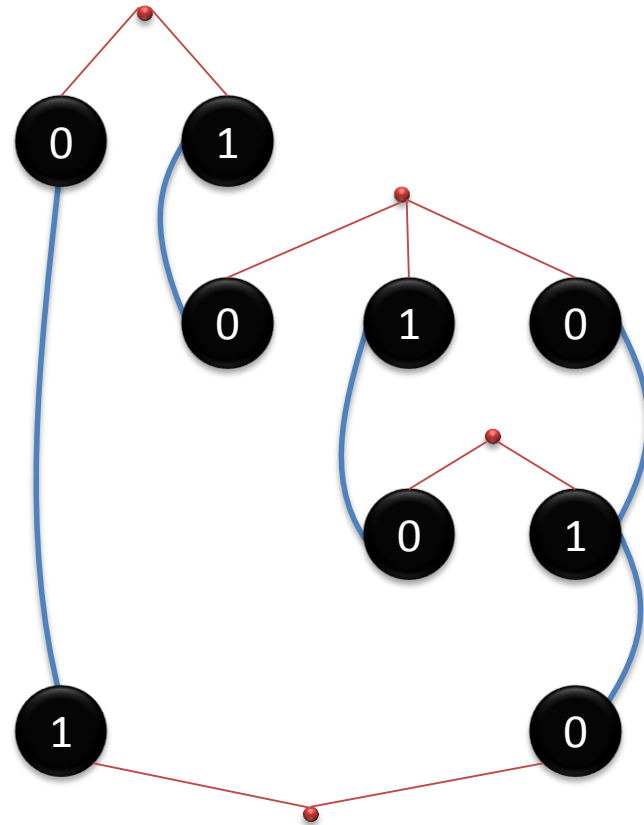
# Graph Model: SNP Conflict Graph

	$p_1$	$p_2$	$p_3$	$p_4$
$f_1$	0	1	-	-
$f_2$	-	0	1	1
$f_3$	-	-	0	1
$f_4$	1	-	-	0



# Graph Model: SNP Conflict Graph

	$p_1$	$p_2$	$p_3$	$p_4$
$f_1$	0	1	-	-
$f_2$	-	0	1	0
$f_3$	-	-	0	1
$f_4$	1	-	-	0



# Proposal of Thesis 1

- K-ploid Haplotype Assembly (fundamental for k-organisms or tumors, etc...)

	$p_1$	$p_2$	$p_3$	$p_4$
$f_1$	A	T	-	-
$f_2$	-	A	T	G
$f_3$	-	-	A	T
$f_4$	T	-	-	A

- **Theoretical Focus:**
  - Computational Complexity
  - Fixed-Parameter Tractability
  - Approximation Complexity
- **Algorithmic Focus:**
  - Fixed-Parameter Tractable algorithm
  - Approximation algorithm
  - Euristicstichs
- **Experimental Focus:**

# Proposal of Thesis 2

- multiallelic Minimum Error Correction (mMEC)

	$p_1$	$p_2$	$p_3$	$p_4$
$f_1$	A	T	-	-
$f_2$	-	A	T	G
$f_3$	-	-	A	T
$f_4$	T	-	-	A

- **Theoretical Focus:**

- Computational Complexity
- Fixed-Parameter Tractability
- Approximation Complexity

- **Algorithmic Focus:**

- Fixed-Parameter Tractable algorithm
- Approximation algorithm
- Heuristics

# Proposal of Thesis 3

- Design of an extension of the FPT algorithm presented in (Bonizzoni et al., 2015) exponential in the read length and following a dynamic programming approach:
  - Extension in order to manage weights and extend the algorithm in order to deal with gaps in the most general MEC
  - Introduce additional constraints in order to improve performance and bound the searching space
  - Implement the algorithm in a tool and experiment it on real and realistically simulated data, comparing with other current state-of-the-art approaches

# Proposal of Thesis 4

- Implement the 2-APX algorithm presented in (Bonizzoni et al., 2015), implement the other approaches proposed in literature and compare their performance and accuracy results.
- Analyze the possibility to extend the 2-APX algorithm presented in (Bonizzoni et al, 2015) to the gapless variant of the MEC problem

# Proposal of Thesis 5

- Carry on the study of the computational complexity, approximability, and tractability of the variants of MEC problem:
  - Computational complexity of Binary MEC
  - Approximability of Gapless MEC
  - FPT of the variants in the remaining parameters (MEC on the read length)
- Search for:
  - New models (es. graph models) for MEC problems
  - New heuristics for the MEC problem and compare it with the existing ones



Good Luck!

For information or question, correspondence to:

**Simone Zaccaria**

**[simone.zaccaria@disco.unimib.it](mailto:simone.zaccaria@disco.unimib.it)**

**Website:**

**<http://algotlab.eu/simone-zaccaria/>**