Genomes Comparision via de Bruijn graphs

Student: Ilya Minkin
Advisor: Son Pham

St. Petersburg Academic University

September 18, 2012

Synteny Blocks: Algorithmic challenge

- Recent advancements in sequencing technologies and genome assembly algorithms
- Multiple strains of the same species:
  - Mycobacterium tuberculosis: Some strains are susceptible to Tuberculosis treatment, and others with multiple drugs resistance
  - Pseudomonas aeruginosa: can express a variety of virulance determinants.
  - 1001 strains of arabidopsis

Synteny Blocks: Algorithmic challenge

- Recent advancements in sequencing technologies and genome assembly algorithms
- Multiple strains of the same species:
  - Mycobacterium tuberculosis: Some strains are susceptible to Tuberculosis treatment, and others with multiple drugs resistance
  - Pseudomonas aeruginosa: can express a variety of virulance determinants.
  - 1001 strains of arabidopsis
  - **similarities and differences** of these genomes help to clarify their different/common phenotypes.

Synteny Blocks: Algorithmic challenge

- Recent advancements in sequencing technologies and genome assembly algorithms
- Multiple strains of the same species:
  - Mycobacterium tuberculosis: Some strains are susceptible to Tuberculosis treatment, and others with multiple drugs resistance
  - Pseudomonas aeruginosa: can express a variety of virulance determinants.
  - 1001 strains of arabidopsis
  - **similarities and differences** of these genomes help to clarify their different/common phenotypes.
- Genomes from multiple species (G10K)
  - How are they **evolutionary related**?
  - How to prove Whole Genomes Duplication?
  - How many rounds of WGDs have occurred?

SyntenyFinder

- These questions lead to a sheer demand for a synteny block generating program that:
  - Able to compare genomes from multiple strains within the same species
  - Able to allow evolutionary studies of genomes belong to different species

SyntenyFinder

- These questions lead to a sheer demand for a synteny block generating program that:
  - Able to compare genomes from multiple strains within the same species
  - Able to allow evolutionary studies of genomes belong to different species
- We introduce **SyntenyFinder** as the first step of addressing these problems

SyntenyFinder

- These questions lead to a sheer demand for a synteny block generating program that:
  - Able to compare genomes from multiple strains within the same species
  - Able to allow evolutionary studies of genomes belong to different species
- We introduce **SyntenyFinder** as the first step of addressing these problems
- **SyntenyFinder** constructs synteny blocks on genomes represented as **sequence of nucleotides**.

General Idea: de Bruijn Graph

- We are given an alphabet Σ and a string S over it, $|\Sigma| = m$
- A substring T, $|T| = k$ is called *k-mer*
- De Bruijn graph is a multigraph $G_k = (V, E)$, where
  $V = \Sigma^{k-1} = \{\text{all possible } (k-1)\text{-mers}\}$
- If k-mer T is presented in S, then we add an oriented edge $(T[1, k-1], T[2, k])$ to the graph
- Create de Bruijn graph from the nucleotide sequence
- Conserved regions will yield non-branching paths

Challenges

- Variations in synteny blocks generate cycles, so we need to simplify the graph
- Double strandness: conserved regions may occur on both strands. Example:
  5' AACCGGTT 3'
  3' TTGGCCAA 5'
  Such blocks are reverse complementary to each other $\Rightarrow$ no non-branching paths
- We need exact shared k-mers, so *directly* our approach can be applied to closely related species (different strains of a bacteria, etc)
- Efficiency

Colored graph

- We use colored de Bruijn graphs [Iqball et al., 2012] to handle double-strandness
- Suppose that S^+ and S^- are positive and negative strands of the chromosome
- Colored de Bruijn graph is a multigraph $G_k = (V, E)$ where $V = \Sigma^{k-1}$
- For each k-mer T^+ in S^+ add edge $(T^+[1, k-1], T^+[2, k])$ to G_k and mark it *blue*
- For each k-mer T^- in S^- add edge $(T^-[1, k-1], T^-[2, k])$ to G_k and mark it *red*

Edge labeling

- Note that our graph is built from a string, not set of reads
- Each walk in the graph represents a string
- We are interested only in walks that represent substrings of the source string
- Assign to each edge e label $L(e) =$ position of the corresponding k-mer on the positive strand
- Walk $W = (v_1\, e_1\, v_2\, e_2 \ldots)$ is considered valid iff:
  1. e_i and e_{i+1} are of the same color
  2. $|L(e_i) - L(e_{i+1})| = 1$

Example

5' ACCTGTCAGT 3'
3' TGGACAGTCA 5'

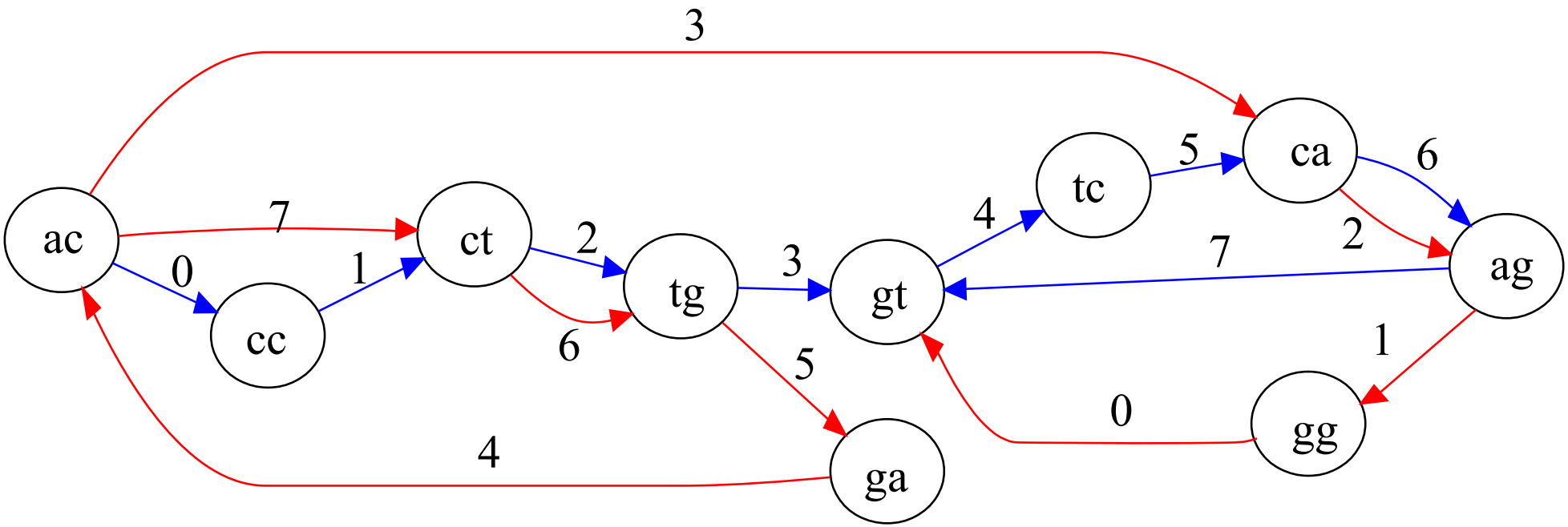


Figure 1: Colored de Bruijn graph built from two strands

Graph simplification

- Bulges spoil long non-branching paths and indicate indels/mismatches
- A pair of walks (W_1, W_2) is a bulge iff:
  1) Start and end vertices of W_1 and W_2 coincide
  2) W_1 and W_2 have exactly 2 common vertices
  3) There are no edges $u \in W_1$ and $v \in W_2$ such that $L(u) = L(v)$
  4) $|W_1| \leq \delta$ and $|W_2| \leq \delta$

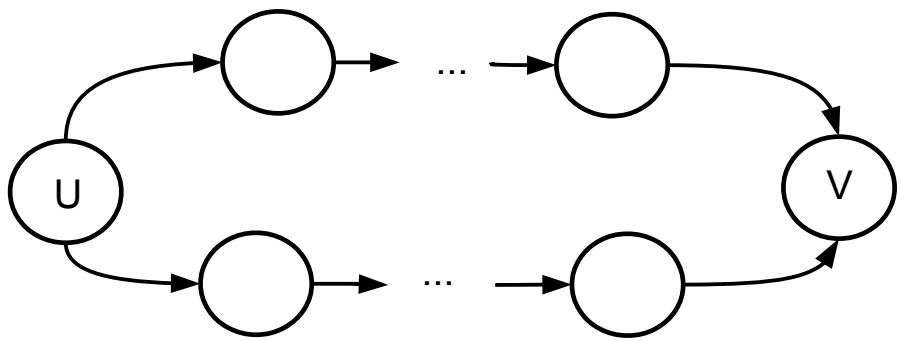


Figure 2: A bulge

General pipeline

- Build de Bruijn graph from the genome
- Remove bulges
- Bulges are removed by replacing one branch with another
- Output non-branching paths

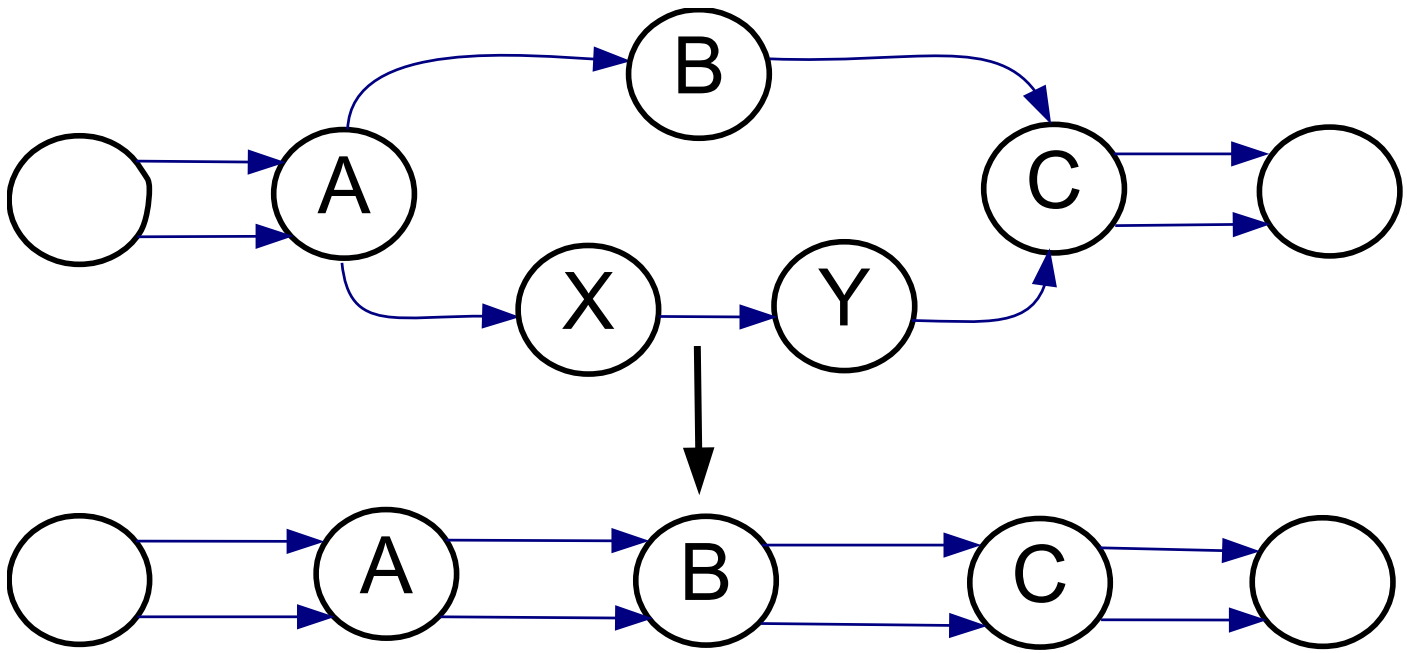
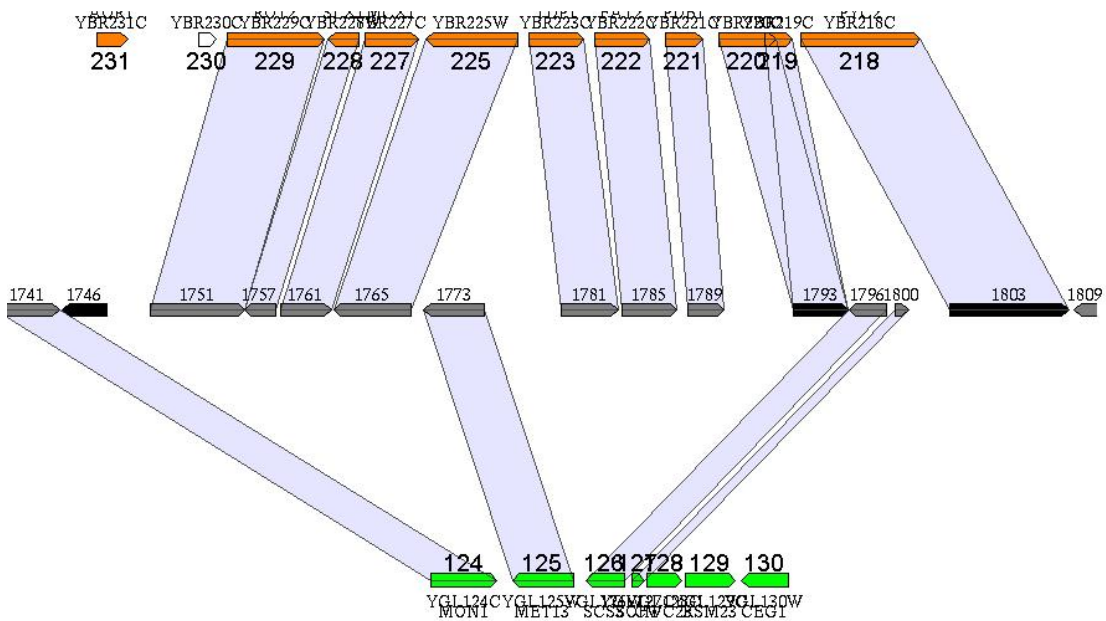


Figure 3: Bulge removal illustration

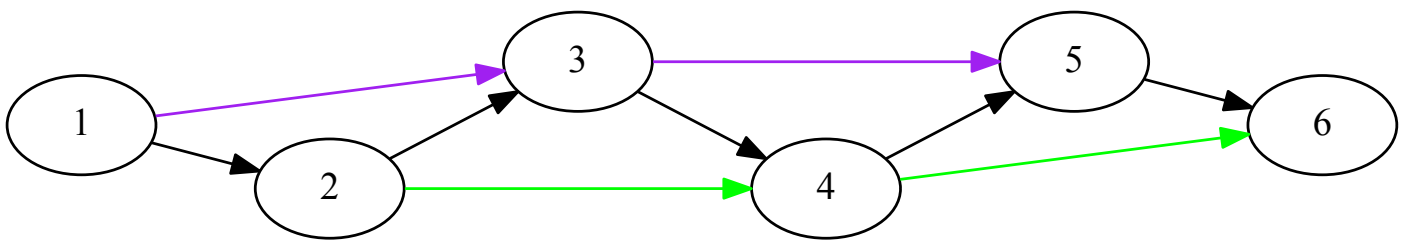
Bulge removal strategy

- It does matter which bulge's branch we replace
- Synteny blocks with muliplicity > 2 can have no k-mers shared across all instances of a same block
- Example: a synteny block from yeasts [Kellis et al., 2004]



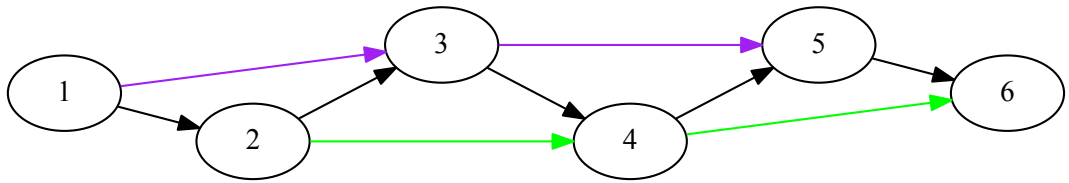
Bulge removal strategy

- Suppose that we have 3 synteny blocks, k-mers are denoted by integers:
  2 4 6
  1 2 3 4 5 6
  1 3 5
- Let's build de Bruijn graph for this situation:

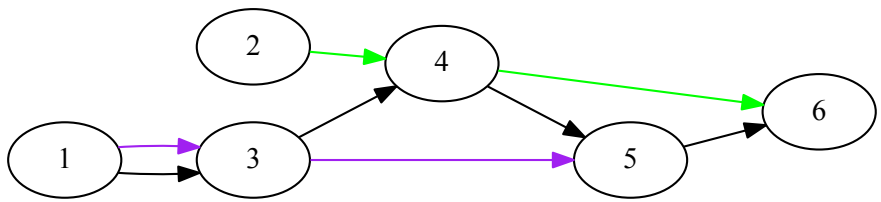


Wrong strategy

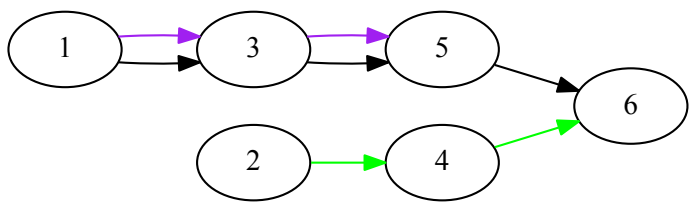
- Initial situation:



- Replace 1 2 3 by 1 3:

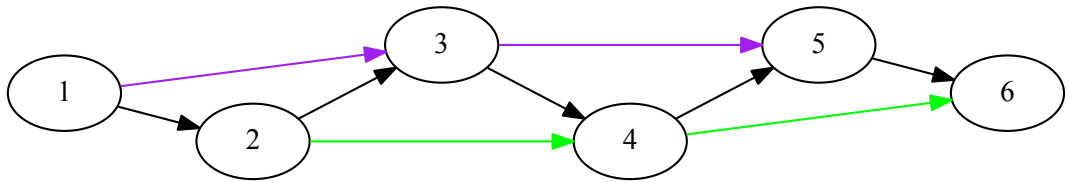


- Replace 3 4 5 by 3 5:

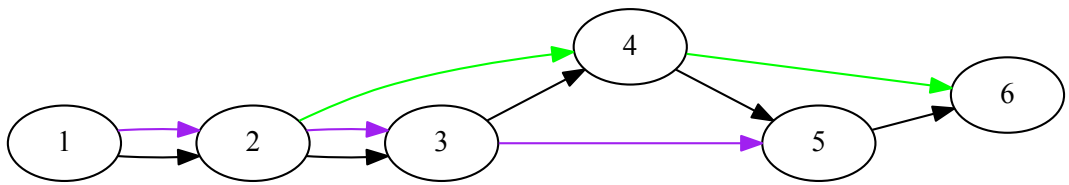


Proper strategy

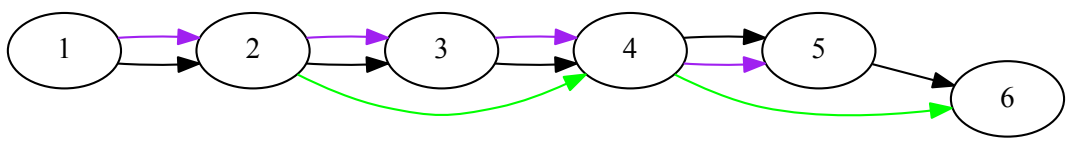
- Initial situation:



- Replace 1 3 by 1 2 3:

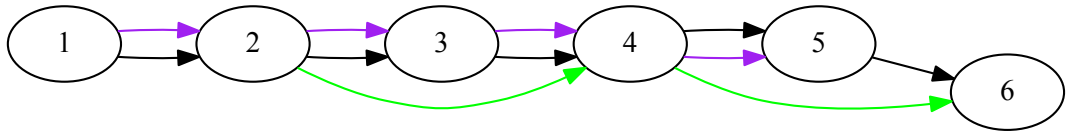


- Replace 3 5 by 3 4 5:

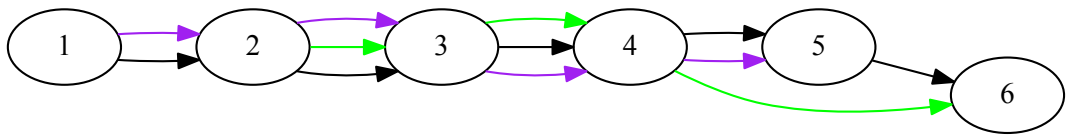


Proper strategy

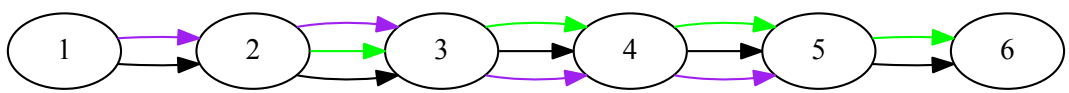
- From previous slide:



- Replace 2 4 by 2 3 4:



- Replace 4 5 by 4 5 6:



Proper strategy

- We incapsulate this intuition into a heuristic
- A vertex v in the graph is called *bifurcation* if there are at least two ingoing (outgoing) edges incident v that spell different k-mers.
- Let's denote by $MaxBif(p)$ maximum degree of bifurcation that lies on path p
- If two paths p_1 and p_2 form a bulge, then we replace p_1 by p_2 iff $MaxBif(p_1) > MaxBif(p_2)$, otherwise we replace p_2 by p_1
- And it seems to work

Results: three bacteria dataset

- Three strains of *Mycobacterium tuberculosis H37Rv*:
  Laboratory reference strain *H37Rv*
  CCDC5180
  CCDC5079
- One of this strains has multiple drugs resistance
- We used $K = 1000$ and $\delta = 5000$ and found 19 blocks shared between all three genomes
- Blocks with multiplicity 3 cover 96% of the genome
- We see an evidence of rearrangements.

When there are not many shared k-mers.

- Well known research about yeasts *S. cerevisiae* and *K. waltii* [Kellis et al., 2004]
- This paper shows evidence of so called *double conserved synteny* regions: one region (block) in k.waltii corresponds to two regions in cerevisiae
- We enrich number of shared k-mers by using **alignment tools**
- With $k = 1000$ and $\delta = 20000$ we cover 67% of the genome by blocks with multiplicity 3. Our blocks match 185 blocks from overall 252 in Kellis paper.
- Most uncovered blocks are small

Conclusions

- **SyntenyFinder** is applicable for reconstructing synteny blocks in closely related species.
- It can be extended to handle more complicated cases.
- **SyntenyFinder** was introduced at **WABI 2012.**

Ongoing work and Future Plan

- **In progress**: Paper writing
- Perform additional tests and evaluations on additional datasets (with interesting biological stories)
- Release software for finding synteny blocks in closely related species (end of September)
- Incorporate into MGRA website
- Incorporate a local alignment tool and extend the software for more complicated genomes.

Acknowledgements

- Pavel Pevzner
- Steve O'brien
- Nikolay Vyahhi
- Matt schultz (SCRIPPS)
- Dinh Diep (Bioengineering UCSD)

References

- 1. Pevzner P and Tesler G, (2003) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution.
- 2. Pham S and Pevzner P, (2010) DRIMM-Synteny: Decomposing Genomes into Evolutionary Conserved Segments
- 3. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G, (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs
- 4. Arabidopsis Genome Initiative, (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana
- 5. Kellis M, W. Birren B, S. Lander E, (2004)

Thank you!