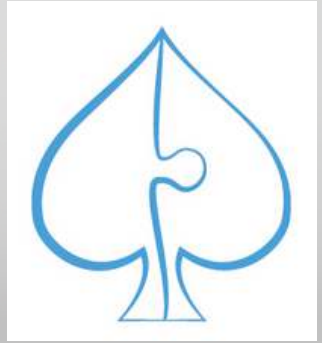
Assembly scaling bottlenecks

Advisor: S.Nurk

Student: T.Malygina

SPAdes

- designed for small genomes
- utilizes assembling algorithm based on deBruijn graph construction
- takes generally longer time and memory than other assemblers
- does an excellent job (probably the best to date) at assembling bacteria, either multi-cell or single-cell data, and small metagenomes



Problems && project goals

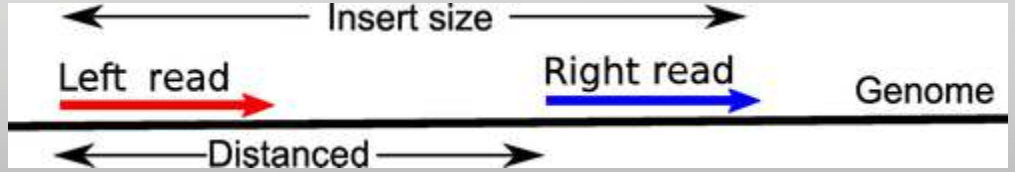
What if:

- we want to process larger genomes?
- we have insufficient memory and/or a large dataset to process?

Tasks: paired info optimizations I

Problem: high redundancy in selected pair information storage strategy leads to inefficient memory usage; we distinguish direct and conjugate edges in graph, corresponding pair information differs only in distance between paired edges

Idea: redefine structure of given class PairedInfoStorage, thus changing storage strategy



Tasks: paired info optimizations I

Initial: pair information is bounded to edges e1, e2, it is stored 2 times independently for each edge in PairInfoStorage class.

Pros: simple, no hacks needed

Cons: wastes memory

Algorithm:

1. select ‘minimal’ edge (in terms of edge ids) - e.g., e1
2. store pair information only for that edge
3. for corresponding pair edge, e2, we can obtain e1 by call

e1 = graph.conjugate(e2)

and restore pair info structure when we need it

Tasks: paired info optimizations III

Problem: Still too much space!

Idea 1: Use distance bins instead of precise distances
Idea 2: Don’t store distance 0 information from long edge to itself
Idea 3: Use counting bloom filters to cut off not significant information (filter edge pairs with little pair info)

Idea 4: having counting bloom filters for pair info, use it for k-mer filtering

Results

Paired info optimizations

- Changing of underlying structure for paired index
- Don’t store distance 0 information from long edge to itself
- Distance bins instead of precise values (even small bin sizes should lead to drastic improvements)
- Counting bloom filters to filter edge pairs with little pair info (should count stats first)