MASTER'S THESIS

UNIVERSITY OF RENNES 1

BIOINFORMATICS AND GENOMICS MASTER'S DEGREE

(2014 - 2015)

VISUAL REPRESENTATION AND COMPARATIVE EVALUATION OF SCAFFOLDING TOOLS FOR CIRCULAR GENOMES

INSTITUTE FOR RESEARCH IN IT AND RANDOM SYSTEMS, GENSCALE
263 AVENUE GENERAL LECLERC, 35000 RENNES, FRANCE

Author:
Alexandrina BODRUG

Supervisors:
Pr. R. ANDONOV
Pr. D. LAVENIER

May 28, 2015

Thanks

Abbreviations & github link

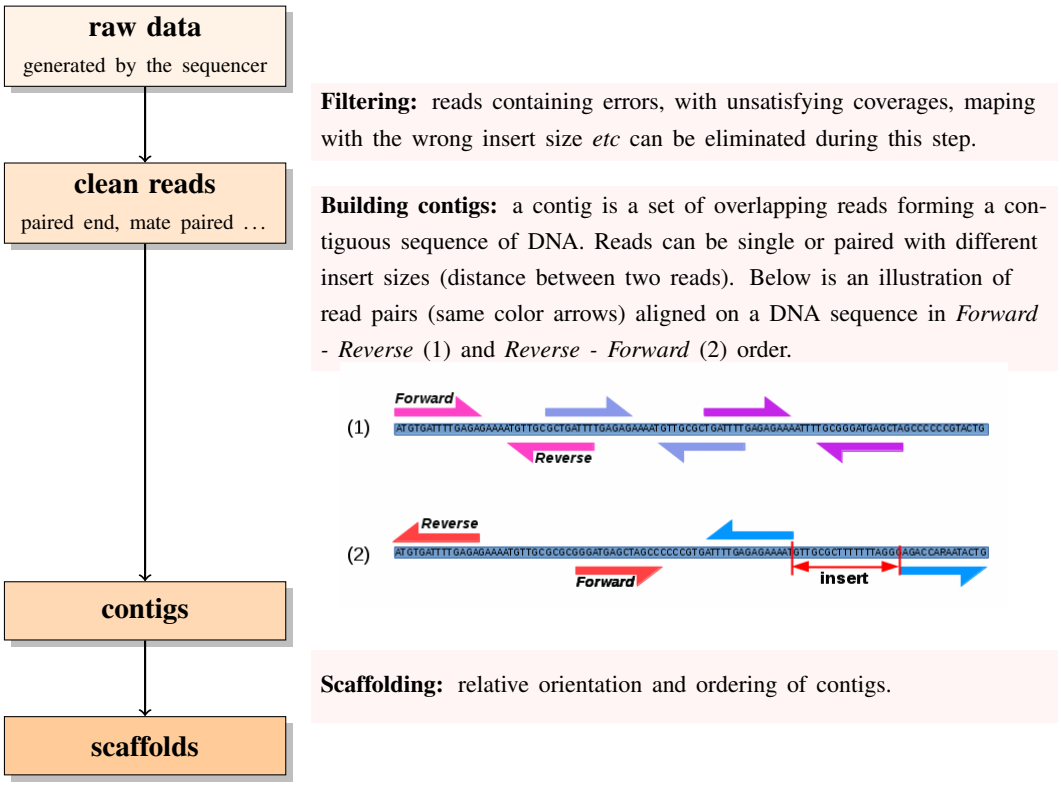
Contents

1 Introduction

1.1 Context

Whole genome shotgun assembly is the process which transforms the small fragmented DNA sequences called reads produced by Next Generation Sequencing methods into relatively unfragmented genomes and chromosomes. The uninterrupted genome sequence is a precious information in many research fields, which explains the plethora of NGS techniques and assembly tools. The assembly process is described in figure 1.1.

Figure 1.1: STEPS OF THE WHOLE GENOME SHOTGUT ASSEMBLY

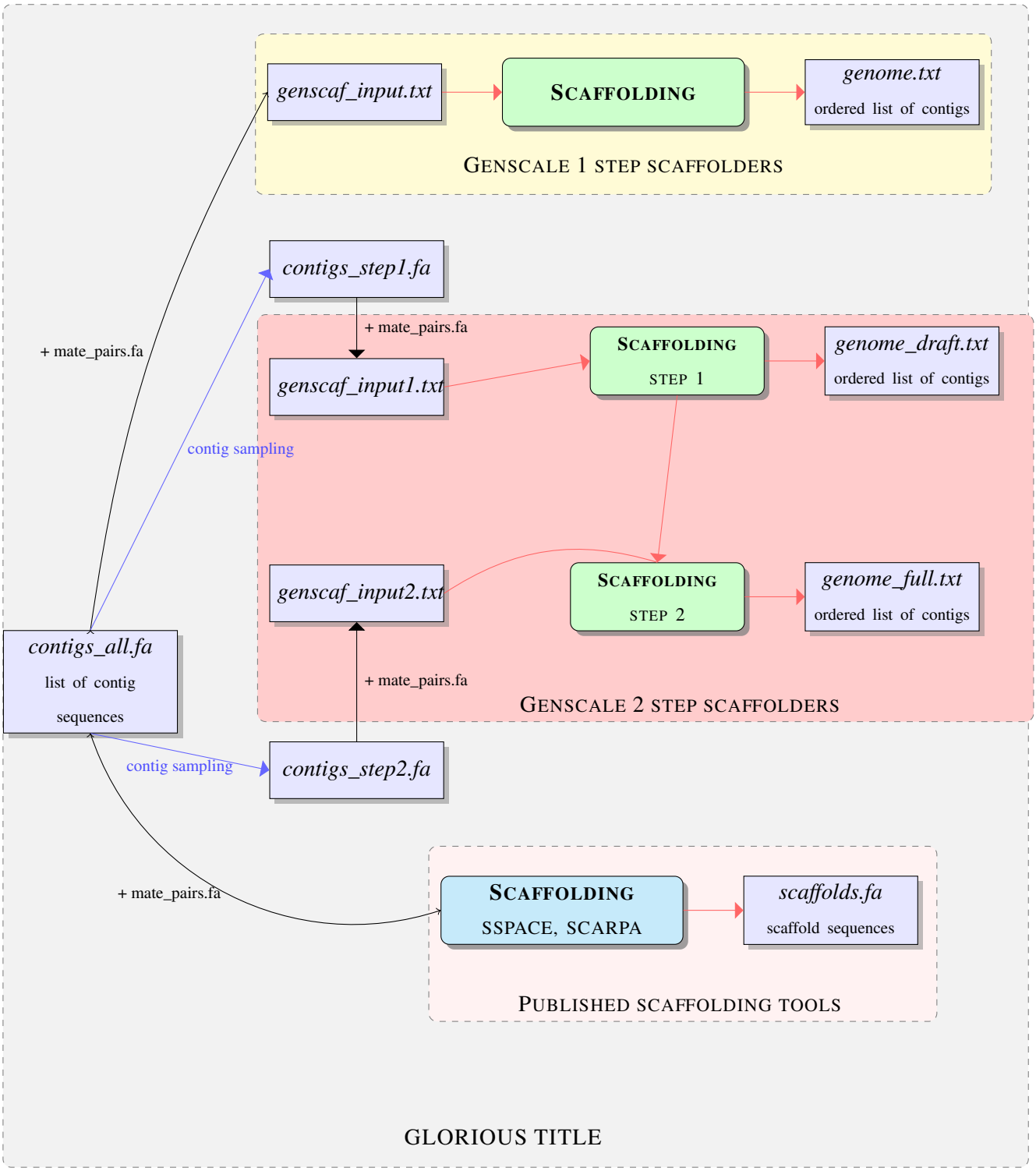


Building contigs and scaffolding are two very different problems that were often combined into a single black box tool, in the early years of genome assembly. Many assemblers are in fact just contig builders; the scaffolding methodology being hardly described or accessible. Since the first stand alone scaffolding tools, the importance of scaffolding was underlined. In 2014

researchers from The Wellcome Trust Sanger Institute (Cambridge) conducted a comprehensive evaluation of scaffolding tools then available. New tools are constantly being published. The aim is clear and appears as rather simple: to order and orient contigs. Most tools model the problem as a graph where vertices represent contigs and links represent bundles of pairs of reads linking two contigs. Erroneous data (fake links due to poorly assembled contigs, low quality libraries), missing data (low quality libraries, unfit insert size, low genome coverage) and inherent genome characteristics (repeated regions, heterozygosity) stand in the way of a perfect and easy scaffolding process. Thus, for an easier scaffolding process, the quality of contigs and the choice of paired end libraries are critical. However the difficulties arising from the sequence structure are to be dealt within the scaffolder.

1.2 Aims of the internship project

1.3 Workflow


genscaf_input.txt
SCAFFOLDING
genome.txt
ordered list of contigs
GENSCALE 1 STEP SCAFFOLDERS
contigs_step1.fa
+ mate_pairs.fa
+ mate_pairs.fa
genscaf_input1.txt
SCAFFOLDING
STEP 1
genome_draft.txt
ordered list of contigs
contig sampling
genscaf_input2.txt
SCAFFOLDING
STEP 2
genome_full.txt
ordered list of contigs
contigs_all.fa
list of contig
sequences
+ mate_pairs.fa
GENSCALE 2 STEP SCAFFOLDERS
contig sampling
contigs_step2.fa
+ mate_pairs.fa
SCAFFOLDING
SSPACE, SCARPA
scaffolds.fa
scaffold sequences
PUBLISHED SCAFFOLDING TOOLS
GLORIOUS TITLE

2 Challenges of the scaffolding problem

2.1 Definition

2.2 Proposed models

2.3 Workflow of the genscale scaffolding tools

2.4 Distinctive features of assembled genomes

3 Methods for benchmarking

3.1 Chosen scaffolding tools to benchmark against

3.2 Benchmarking workflow

3.3 Comparisons

3.3.1 QUAST

3.3.2 Comparison function

3.4 Visualization

3.4.1 MUMMER

3.4.2 Visualization tool

4 Results

4.1 Data sets compared

4.2 Comparisons

4.2.1 QUAST and comparison function

4.2.2 Visualization

5 Conclusion