NextGenScores:

Fast Alignment-based Estimation of Intragenomic Sequence Similarities

Philipp Rescheneder, Arndt von Haeseler, Niko Popitsch

Contact: philipp.rescheneder@univie.ac.at

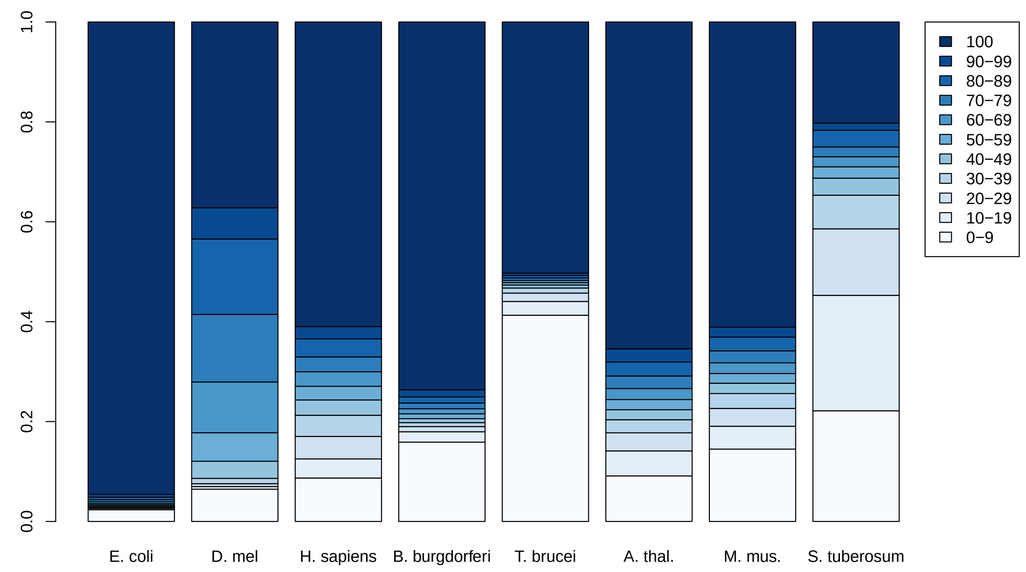
Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, Dr.-Bohr-Gasse 9, A-1030 Vienna, Austria

CIBIV MFPL
Center for Integrative Bioinformatics Vienna

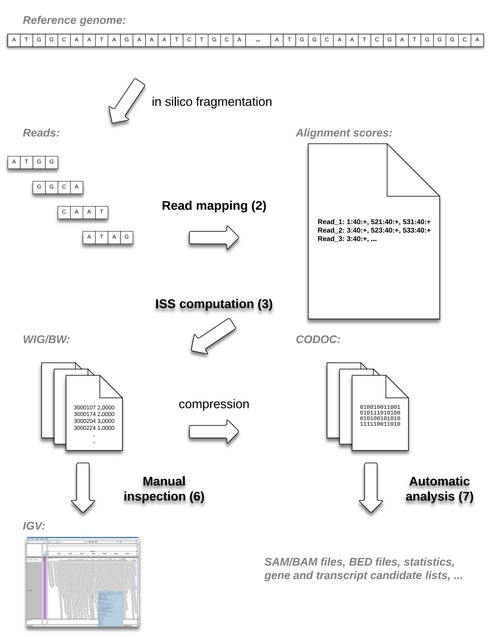
Abstract

A central step of high-throughput sequencing (HTS) analysis is mapping short reads to a reference genome. The accuracy of this mapping step, however, is heavily influenced by sequence similarities within a genome. To accurately identify regions of poor mapping accuracy, we have developed NextGenScore, a framework based on our highly-efficient alignment software NextGenMap. We calculate an unbiased score that represents the maximum similarity of a genomic region to any other region in the genome. Based on this score, we assessed the influence of intragenomic sequence similarities on the read mapping accuracy of several state-of-the-art read mapping programs. Furthermore, we show how our precomputed score can complement mapping quality in identifying ambiguously mapped reads. NextGenScore outputs a similarity-track that can be used to visually inspect regions of interest in state-of-the-art genome browsers. Additionally, it provides command-line tools to automatically post-process read alignments by filtering out spuriously mapped reads and determining mappable, yet uncovered genomic regions. By this, NextGenScore serves as valuable resource for quality assurance that can easily be integrated into existing HTS pipelines.

ISS distribution over the genome (5)



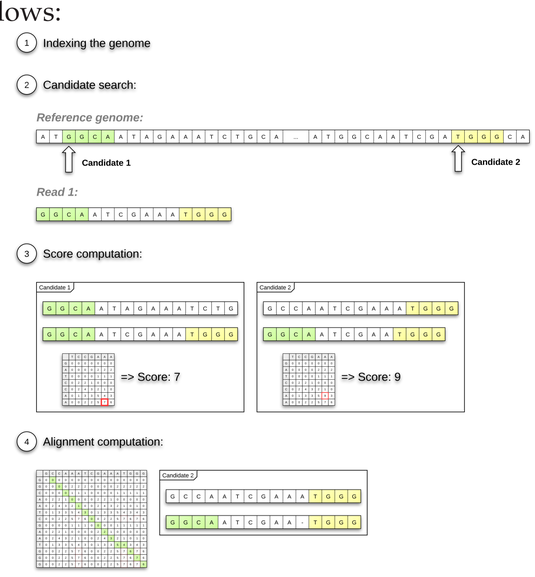
Pipeline (1)



NextGenScores only takes the reference sequence as input and can therefore be used for well-annotated as well as for newly assembled genomes. In addition, it is not biased by, potentially incomplete, prior knowledge or annotations.

Read mapping (2)

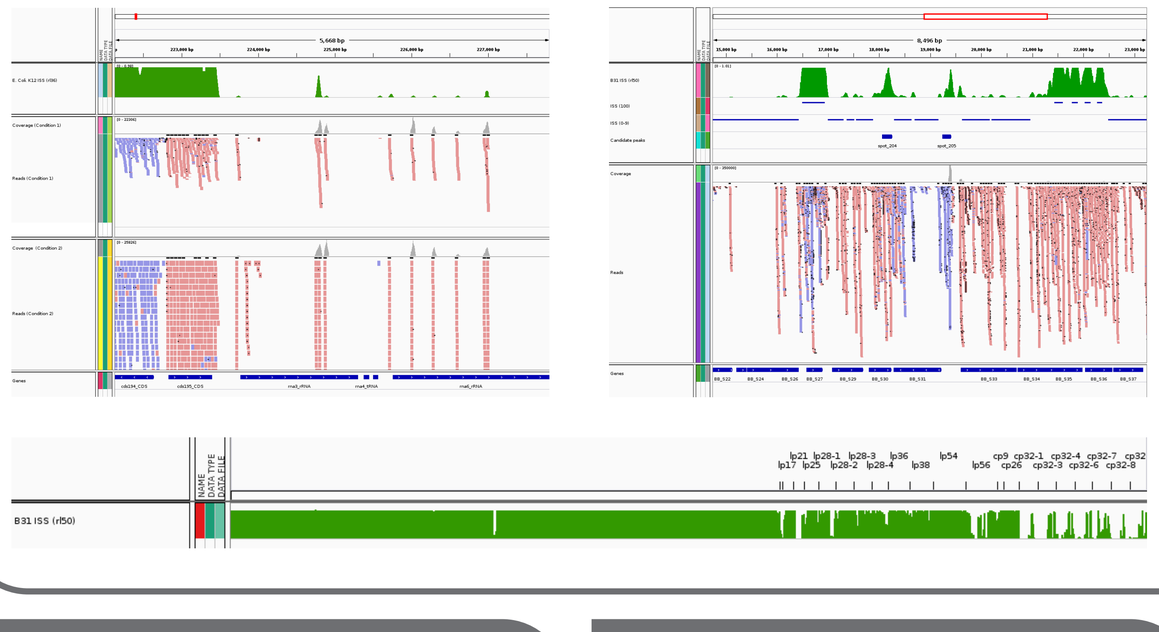
Read mapping is done using NextGenMap. The basic workflow of NextGenMap is as follows:



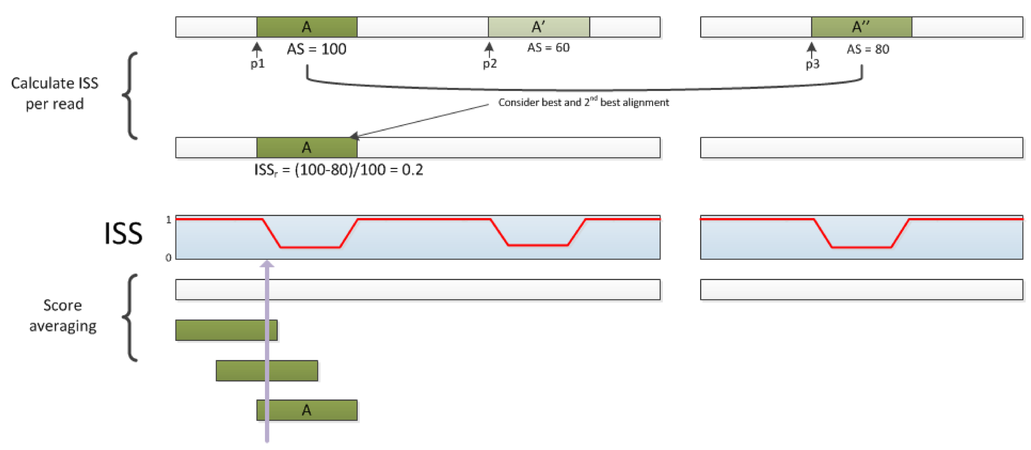
Step 4 is not strictly necessary for computing the IIS and is there fore skipped to reduce runtime. To save disk space, instead of SAM/BAM format NextGenMap outputs only the resulting alignment scores as a zipped file.

Manual inspection (6)

The ISS track helps to quickly identify problems with read mapping or biases caused by sequence similarities within a genome.



Intragenomic sequence similarity computation (3)



Automatic analysis (7)

NextGenScores can be used in existing HTS analysis pipeline for the following tasks:

- Automatically identify uncovered regions in the dataset.
- Filter spurious read mappings.
- Identify transcripts that are (partly) inaccessible to read mapping.
- Exclude SNPs in regions with low IIS from downstream analysis.
- Quickly identify unique regions in a set of genomes.
- Compute average score for a given set of annotations (e.g. genes, repeats, ...).

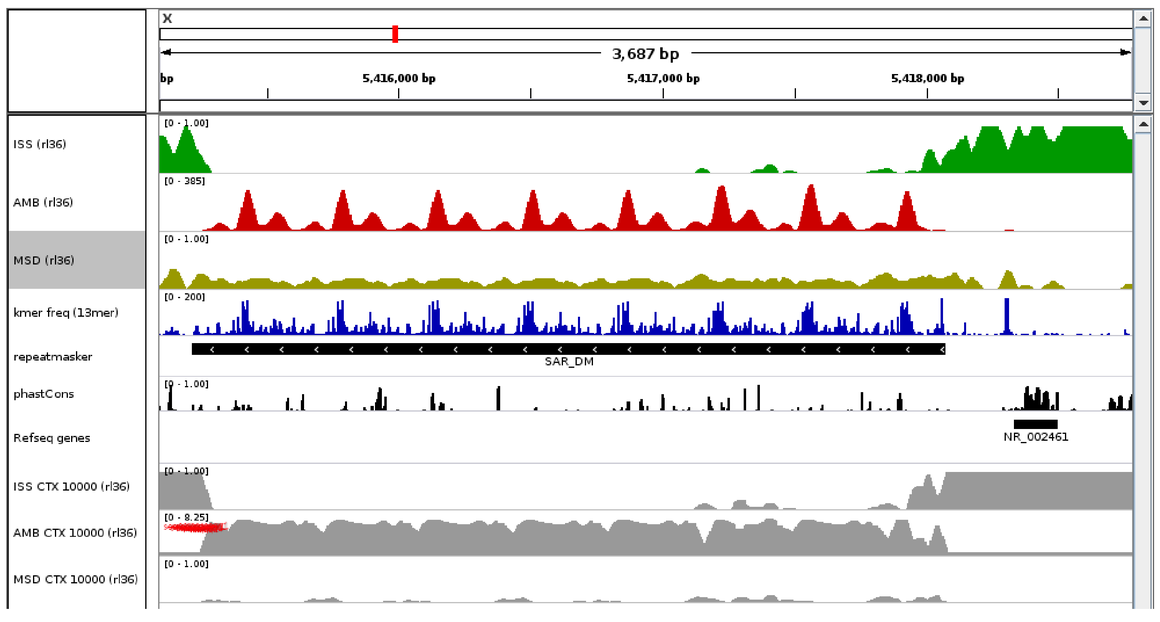
Mapping error (8)

ISS	Bowtie2	BWA	NGM
0	20.82 (77.46)	20.91 (78.83)	21.02 (78.72)
1	2.68 (65.60)	2.65 (66.71)	2.80 (66.62)
2	1.26 (64.20)	1.23 (65.27)	1.34 (65.18)
4	0.51 (61.28)	0.50 (62.22)	0.55 (62.18)
8	0.20 (58.52)	0.20 (59.37)	0.23 (59.37)
16	0.04 (55.84)	0.07 (57.16)	0.05 (56.65)
32	0.01 (53.08)	0.01 (54.09)	0.01 (53.85)
64	0.00 (50.98)	0.00 (51.80)	0.00 (51.72)

Percentage of incorrect (correct) mappings for 5 mio simulated T. brucei reads.

Outlook: ARGOS (9)

NextGenScores is the basis for ARGOS (Analysis of Repetitive Genome Sequences). In addition to the IIS, ARGOS incorporates three additional genome-wide signals to identify and characterize repeating regions without prior knowledge.



Comparison to Mapping Quality (4)