Meta-alignment: Combining multiple sequence aligners to improve alignment quality

B. Wolf, P. Kuonen, *University of Applied Sciences Western Switzerland*
T. Dandekar, *University of Würzburg, Germany*
David Atlan, *Phenosystems, Belgium*

Introduction

Many different tools have been developed to perform sequence alignment, a central part of NGS data analysis. Choosing the right tool with the right parameters for a given dataset is a difficult task. Often different aligners perform better than others on parts of the same dataset, making it hard to choose the right tool and configuration.
We propose a new approach called meta-alignment, which combines the output of multiple sequence aligners to build the best possible alignment. Based on a score matrix, the best possible alignment for every aligned sequence is chosen among the different aligners.
Our approach has been tested on multiple simulated datasets, comparing the results of multiple sequence aligners with the results of meta-alignment which combines their individual results.

Methods

In our test setup, we used combined the output of 3 different aligners to create a combined alignment. We used BWA-MEM 0.7.12-r1039, Bowtie 2.2.6 and CUSHAW2 2.4.3.
The initial step of the meta-alignment, as shown in Figure 1, is the sorting of the alignment files by name. This allows for an efficient merging of the alignment files in the next step. During this step, the score for every alignment of a particular sequence is calculated based on a score matrix, and the best scoring alignment is saved. Depending on the user configuration, one or more aligners need to agree on the best alignment for it to be choosen. This allows us to either increase the coverage of the resulting alignment, or increase the amount of wrongly aligned reads.
The currently biggest limitation is the restriction to single end sequencing files. Paired end sequences will be supported at a later date.

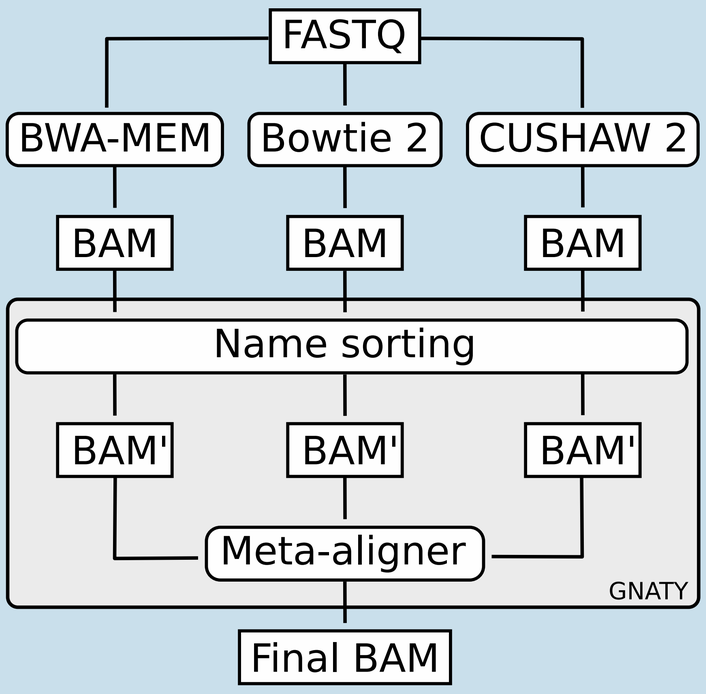

Figure 1: Meta aligner architecture, using 3 external aligners

Results

To test effect of meta-alignment on the alignment quality, we used two simulated datasets. The first one consisting of 11 datasets with increasing error rates (0-20%) with an increasing amount of indels (0-50%). We compare the alignment rate and the precision of all 3 aligners as well as 3 meta-alignment configurations (Meta 1 = best alignment, Meta 2 = 2 aligners agree, Meta 3 = all aligners agree).

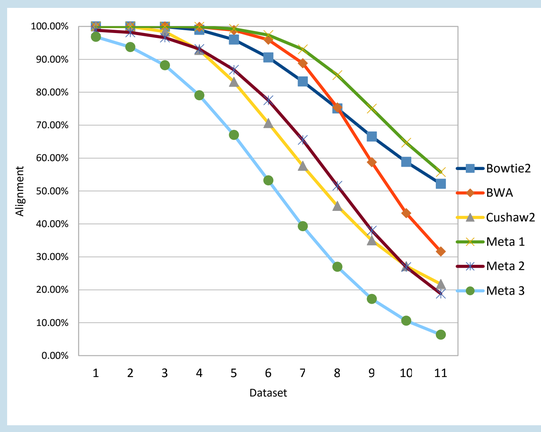

Figure 2: Aligned sequences (%)

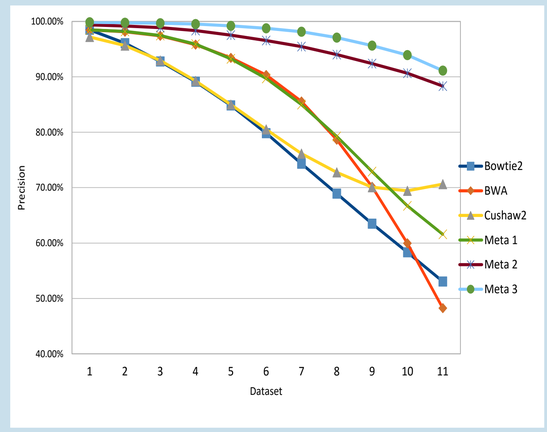

Figure 3: Correct sequences (%)

Especially at high error rates, we can observe the effect of meta-alignment on the resulting alignment file.
The second dataset comes from the genome comparison & analytic testing project (GCAT). This standardized dataset was tested the same way as the simulated datasets.

	Total	Correct	Wrong	Not aligned	Precision	Alignment rate
BWA-MEM	7'878'949	7'779'572	99'477	84'549	98.74%	97.69%
Bowtie 2	7'878'771	7'604'671	274'064	84'727	96.52%	95.49%
CUSHAW2	7'868'183	7'650'416	217'767	95'315	97.23%	96.07%
Meta 1	7'878'987	7'781'337	97'614	84'511	98.76%	97.71%
Meta 2	7'787'802	7'737'157	50'645	175'696	99.35%	97.16%
Meta 3	7'507'014	7'487'040	19'974	456'484	99.73%	94.02%

Figure 4: Dataset 2 tests

We can see the different tradeoffs made between the 3 meta-alignment configurations. The configuration requiring two aligners to agree has an interesting balance between alignment rate and precision.

Conclusion

We can observe promising results using our method, especially on datasets with high error rates. The ability for the user to choose between the quality of the alignment and the coverage is also interesting depending on the use-case.
Future works will concentrate on the following aspects of this new method:

- The ability to use paired end data
- Improved variant calling by using realignment

The meta alignment tool has been integrated into GNATY, a collection of NGS data analysis tools based on GensearchNGS. It is available for free for non-commercial usage at: http://gnaty.phenosystems.com

Contact: Beat wolf, beat.wolf@hefr.ch

Website GNATY:
http://gnaty.phenosystems.com

PhenoSystems
GensearchNGS
MOLECULAR GENETICS MADE EASY

Hes·so FRIBOURG FREIBURG
Haute Ecole Spécialisée de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences Western Switzerland

BIOZENTRUM UNIVERSITAT WÜRZBURG