Bioinformatics Journal Club

Ulvi Talas

January 10, 2014

Bioinformatics (2014) 30 (1): 9-16.
doi: 10.1093/bioinformatics/btt255
First published online: May 17, 2013

Specificity control for read alignments using an artificial reference genome-guided false discovery rate.

Sven H. Giese, Franziska Zickmann and Bernhard Y. Renard*

Research Group Bioinformatics (NG4), Robert Koch-Institut,
Nordufer 20, 13353 Berlin, Germany

Proposed software: ARDEN
Specificity Control for Read Alignments Using an Artificial Reference

We introduce ARDEN (Artificial Reference Driven Estimation of false positives in NGS data), a novel benchmark that estimates error rates based on real experimental reads and an additionally generated artificial reference genome. It allows the computation of error rates specifically for a dataset and the construction of a ROC-curve.

Thereby, it can be used:

✧ to optimize parameters for read mappers,

✧ to select read mappers for a specific problem

✧ or also to filter alignments based on quality estimation.

Applications:

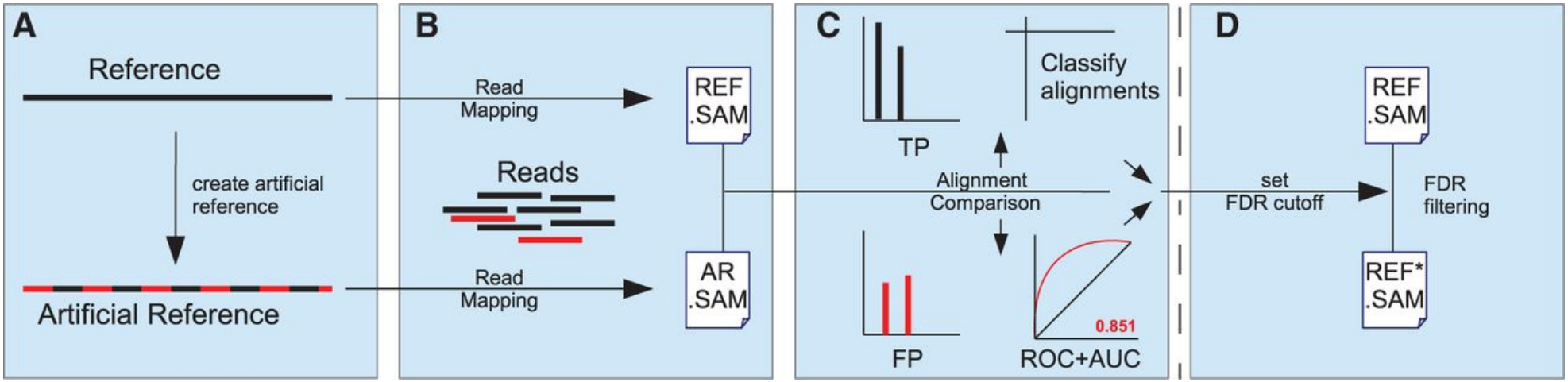
In this paper the use of ARDEN is demonstrated in:

- a general read mapper comparison,
- a parameter optimization for one read mapper,
- an application example in SNP discovery with (a significant?) reduction in the number of false positive identifications.

⇒***aims to provide a method of evaluation and quality control to find an optimal setting for a NGS read mapper.***

Availability: the source code can be downloaded at http://sourceforge.net/projects/arden/

ARDEN workflow.



Giese S H et al. Bioinformatics 2014;30:9-16

Bioinformatics

Creating an artificial reference genome

The artificial reference genome is close to the original sequence, but contains substitutions in a pre-defined distance. These mismatches are randomly chosen, but fulfill the (optional) conditions that a substitution does not change the following properties between A and R:

(i) the nucleotide distribution and thus the GC-content,
(ii) the amino acid distribution,
(iii) the amino acid neighborhood,
(iv) any putative start/stop codons.

The algorithm works as follows:

① Choose randomly a position n in the translated protein sequence (first frame translation of the complete genome) and its corresponding codon c_n.

② Store the amino acids at positions n − 1 and n + 1, as well as the corresponding codons c_{n+1} and c_{n+1} .

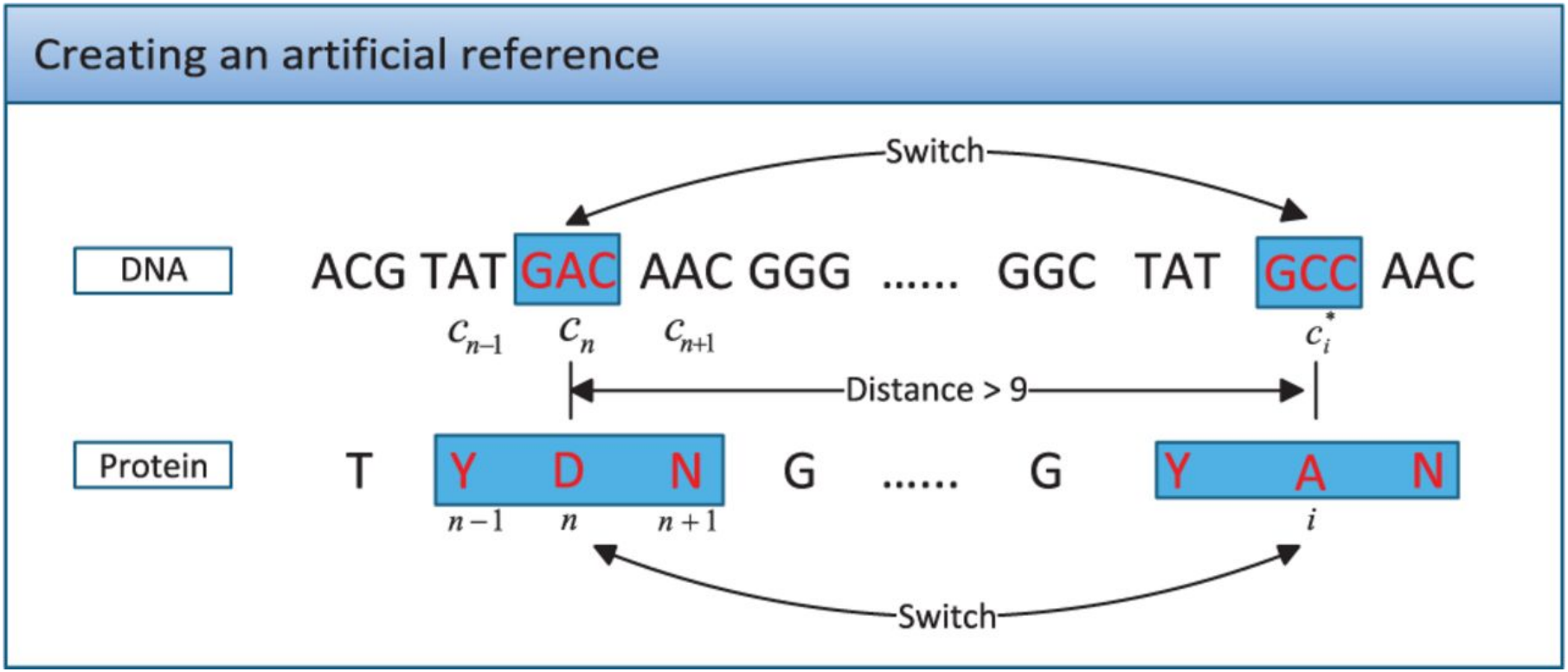
③ Generate a list of possible amino acids whose codon C_i^* has Hamming distance =1 to C_n

④ Search for every amino acid triplet corresponding to C_{n-1}, C_i^*, C_{n+1} from (3) in the protein sequence [respecting the constraints (i)–(iv)] and stop when one is found at a position pos_i.

⑤ Switch the codon at position n and pos_i.

⑥ Start again with (1) until no valid starting positions are left.

Example iteration for creating an artificial reference genome.



Giese S H et al. Bioinformatics 2014;30:9-16

Bioinformatics

A measure of sensitivity (Sn) is calculated as:

$$Sn = \frac{PTP}{JH} \cdot M,$$

where **M** denotes the fraction of mapped reads*.

PTP – probably true positive read mapping hits

JH – joint (true plus false) read mapping hits

**This fraction serves as a normalization constant to compensate the fact that some alignment strategies map more reads than others.*

For false positives (FP) read mappings, a measure of specificity (Sp) is defined as:

$$Sp = 1 - \left(\frac{FP}{JH} \cdot M\right).$$

***This follows the intuition that mappers cannot be specific if they tend to map more reads distinctly on the artificial reference.*

Creating ROC curves:

To compute positions on the ROC curve for different trade-offs between sensitivity and specificity, we **filter all reads according to various alignment features.** Thus, features are selected that **uniquely define points on the ROC curve.**

The set of features comprises:

- the number of gaps in the alignment,
- the number of mismatches
- and a read quality score (RQS), which is calculated as:

$$RQS = \frac{\bar{q}_r}{max(q_b) - min(q_b) + 1},$$

Where $\bar{q}_r$ denotes the average quality of read **r**, and $\mathbf{q_b}$ is the base quality for a base of the sequence of read r.

2 Envelope calculation

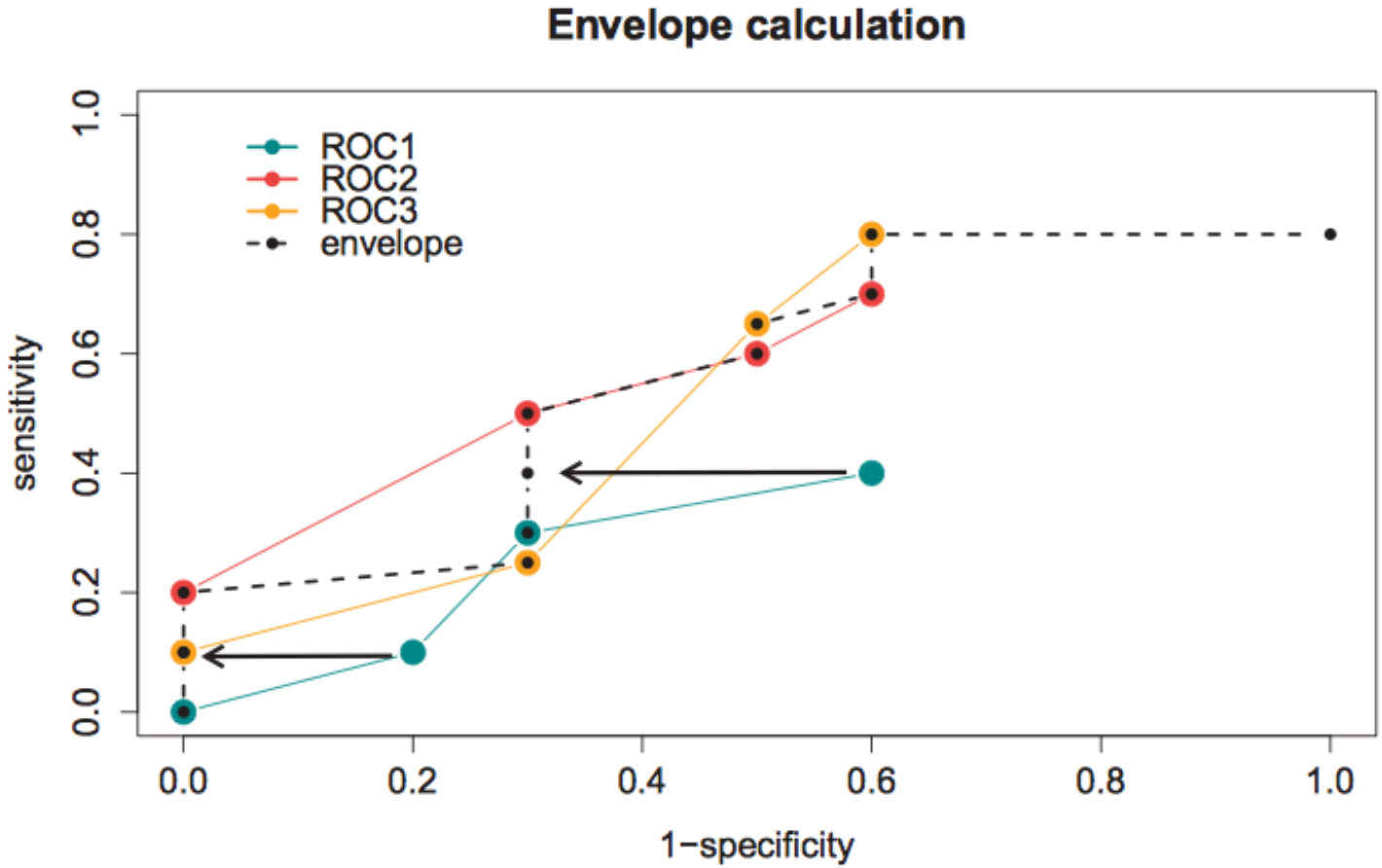
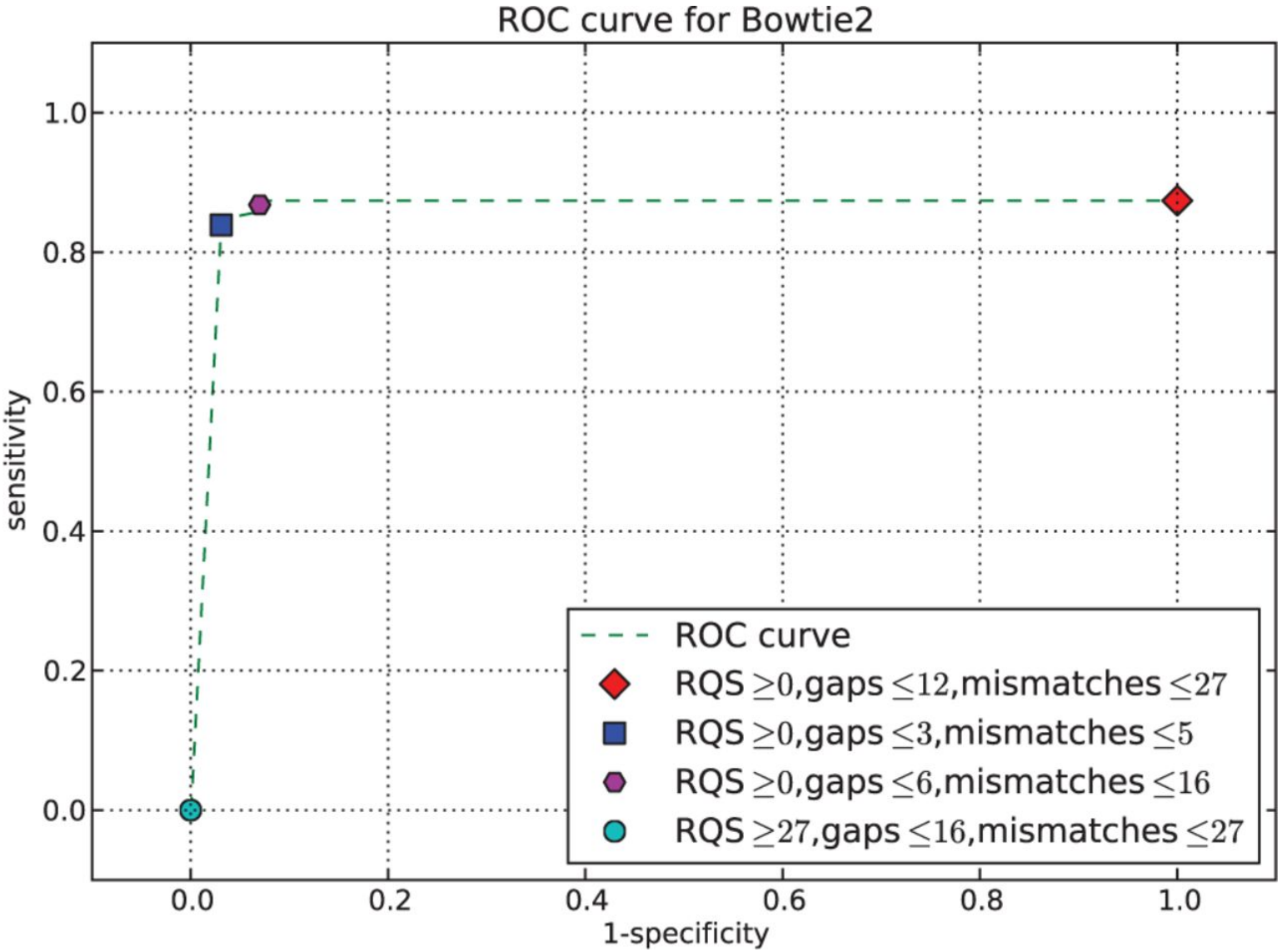


Figure 1: Illustration for combining ROC points with no linear relationship. Black dotted line indicates the final ROC curve. ROC1, ROC2 and ROC3 are subsets of all points within a subclass,where a linear relationship can be observed. Combining the points of the theoretical curves ROC1-ROC3 yields the final ROC curve. Lines and coloring are drawn for illustrative purpose only. Arrows indicate the point projection.

Example of a ROC curve for Bowtie2 generated by ARDEN. Here, we used a C.elegans dataset with 1 million single-end reads.



Giese S H et al. Bioinformatics 2014;30:9-16

Bioinformatics

Specificity control for read alignments using an artificial reference genome-guided false discovery rate

Table 1.
Comparison of the sensitivity (*Sn*) and specificity (*Sp*) of different read mappers using a *C.elegans* dataset with 1 million single-end reads

Mapper	PTP	FP	Sn	Sp	AUC	M
BWA	981 515	25 438	0.895	0.977	0.896	0.919
RazerS2	4 590 324	25 151	**0.921**	0.995	**0.92**	0.926
Bowtie2	945 238	77 084	0.874	0.929	0.859	**0.945**
mrsFAST	**6 528 165**	**328**	0.92	**1**	**0.92**	0.92

Note: M refers to the fraction of mapped input reads. The exact parameters for each mapper are available in the Supplementary Material. The analysis was performed using the analysis module of ARDEN. All values are rounded to three decimal digits. For each column best values are marked in bold.

Specificity control for read alignments using an artificial reference genome-guided false discovery rate

Table 2.
Excerpt from a resulting ROC table using Bowtie2 and ARDEN

RQS	GAPS	MM	PTP	FP	Sn	Sp	M
0	16	27	945 238	77 084	0.874	0.929	0.945
0	16	10	941 245	73 445	0.866	0.932	0.941
0	3	10	936 150	72 564	0.857	0.934	0.936
0	16	5	930 631	34 217	0.847	0.969	0.931
0	3	5	926 103	33 771	0.839	0.969	0.926
27.215	3	5	775	33	0.0	0.99	0.001

Note: The columns RQS, GAPS and MM indicate the cut-off parameters to divide the alignments in sub-classes. For instance, the first row includes a sub-class that includes all alignments that have an RQS ≥ 0, gaps ≤ 16 and mismatches (MM) ≤ 27.

Specificity control for read alignments using an artificial reference genome-guided false discovery rate

Table 3.
Comparison of different parameterizations for Bowtie2

Setting	PTP	FP	Sn	Sp	AUC	M
Very fast	944 473	**63 344**	**0.885**	**0.941**	**0.875**	0.945
Default	945 238	77 084	0.874	0.929	0.859	0.945
Custom	944 768	78 304	0.873	0.928	0.856	0.945
Very sensitive	**945 487**	84 281	0.868	0.923	0.851	**0.946**

Note: The settings reflect pre-defined configurations of Bowtie2 (*very fast*, *default* and *very sensitive*), as well as a *custom* configuration that adds the $-N = 1$ option to the default setting. For each column, best results are highlighted in bold.

Specificity control for read alignments using an artificial reference genome-guided false discovery rate

Table 4.
Comparison of SNP calling using *all* alignments and SNP calling with a set of *filtered (Filt.)* alignments defined by ARDEN on a modified *E.coli* genome

Mapper	True positives			False positives		
	All	Filt.	Δ in %	All	Filt.	Δ in %
BWA	127	127	0	198	188	−5.05
Bowtie2	126	126	0	225	224	−0.44
RazerS2	130	130	0	701	196	−72.04

Note: The ground truth contained 150 simulated SNPs. ARDEN decreases the number of FP while retaining all TPs. The effect of filtering depends on the particular mapper and the respective results of ARDEN. For Bowtie2, BWA and RazerS2, the percentage of all alignments that have been removed by the filter are $\approx 6.8\%$, $\approx 2.5\%$ and $\approx 3.4\%$, respectively. The relative difference between the All and Filt. category is denoted as Δ.

Specificity control for read alignments using an artificial reference genome-guided false discovery rate

Table 5.
Comparison of SNP calling using *all* alignments and SNP calling with a set of *filtered (Filt.)* alignments defined by ARDEN on a modified chromosome 21 of *H.sapiens*

Mapper	True positives			False positives		
	All	Filt.	Δ in %	All	Filt.	Δ in %
Bowtie2	45 342	45 805	+1.02	10 191	10 144	−0.46
RazerS3	46 592	44 069	−5.42	56 954	26 010	−54.33
BWA	48 715	45 058	−7.51	15 681	9612	−38.7

Note: TPs were compared with a simulated ground truth containing 1000 simulated SNPs and to public available SNP data (a more detailed distinction is available in the Supplementary Material). For RazerS3 and BWA, the filtering with ARDEN considerably reduced the numbers of FPs along with a comparably small loss of TPs. For Bowtie2, the number of FPs is decreased along with a gain in TPs. The relative difference between the All and Filt. category is denoted as Δ.

9 SNP discovery on human chr. 21

The following tables provide an distinction between public available SNP data (UCSC) and the simulated ground truth (GT).

Ground Truth / Mapper	true positives		false positives		false negatives	
	all.	filt.	all.	filt.	all.	filt.
Bowtie2	906	904	54,627	55,045	94	96
RazerS3	895	891	102,651	69,188	105	109
BWA	910	905	63,486	53,765	90	95

Table 6: Comparison of SNP calling using all alignments and SNP calling with a set of filtered (filt.) alignments defined by ARDEN on a modified chr. 21 of *H. sapiens*. TP were compared to a simulated ground truth containing 1000 SNPs.