SOFTWARE TOOL ARTICLE

Disambiguate: An open-source application for disambiguating two species in next generation sequencing data from grafted samples [version 1; peer review: 2 approved, 1 approved with reservations]

Miika J. Ahdesmäki¹, Simon R. Gray², Justin H. Johnson³, Zhongwu Lai³

¹AstraZeneca IMED Oncology, Cambridge, UK

²AstraZeneca R&D Information, Cambridge, UK

³AstraZeneca Oncology iMed, Waltham, USA

V1 First published: 22 Nov 2016, 5:2741
<https://doi.org/10.12688/f1000research.10082.1>
 Latest published: 24 Jan 2017, 5:2741
<https://doi.org/10.12688/f1000research.10082.2>

Abstract

Grafting of cell lines and primary tumours is a crucial step in the drug development process between cell line studies and clinical trials. Disambiguate is a program for computationally separating the sequencing reads of two species derived from grafted samples. Disambiguate operates on alignments to the two species and separates the components at very high sensitivity and specificity as illustrated in artificially mixed human-mouse samples. This allows for maximum recovery of data from target tumours for more accurate variant calling and gene expression quantification. Given that no general use open source algorithm accessible to the bioinformatics community exists for the purposes of separating the two species data, the proposed Disambiguate tool presents a novel approach and improvement to performing sequence analysis of grafted samples. Both Python and C++ implementations are available and they are integrated into several open and closed source pipelines. Disambiguate is open source and is freely available at <https://github.com/AstraZeneca-NGS/disambiguate>.

Keywords

NGS, patient derived xenograft, explant, disambiguation, sequencing

This article is included in the Python collection.

Open Peer Review

Approval Status

	1	2	3
version 2 (revision) 24 Jan 2017			 view
version 1 22 Nov 2016	 view	 view	 view

1. **Daniel Nicorici**, Orion Corporation Orion Pharma, Espoo, Finland
2. **Matthew D. Eldridge**, University of Cambridge, Cambridge, UK
3. **Gavin R. Oliver** , Mayo Clinic, Rochester, USA
Asha A. Nair, Mayo Clinic, Rochester, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Miika J. Ahdesmäki (miika.ahdesmaki@astrazeneca.com)

Competing interests: All authors are employees of AstraZeneca.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2016 Ahdesmäki MJ *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Ahdesmäki MJ, Gray SR, Johnson JH and Lai Z. **Disambiguate: An open-source application for disambiguating two species in next generation sequencing data from grafted samples [version 1; peer review: 2 approved, 1 approved with reservations]** F1000Research 2016, 5:2741 <https://doi.org/10.12688/f1000research.10082.1>

First published: 22 Nov 2016, 5:2741 <https://doi.org/10.12688/f1000research.10082.1>

Introduction

Xenografts, both cell line and primary tumour, are routinely profiled in preclinical and translational research. Xenografts are used to study everything from new target identification to responses to targeted therapeutics and mechanisms of resistance¹ in an environment that is more realistic than just 2D cell lines. However, due to mouse stromal contamination of the human tumour, not all the data resulting from studying the extracted samples are guaranteed to be of human origin.

Direct high throughput sequencing of grafted samples with a mixture of two species is routine practice. However with the high volume of data and computational challenges of alignment and kmer identification, new computational strategies are required to computationally separate the two species' components for more accurate downstream analysis¹, especially for the reduction of variant calling artefacts. However, the two-species alignment approach proposed in Bradford *et al.*¹ excludes reads that align to both organisms, clearly dismissing a large portion of the data as evidenced in [Table 1](#) and [Table 2](#) when observing cross species alignment rates.

Algorithms designed for disambiguating the host and tumour sequences include e.g. the Xenome tool², which is based on machine learning applied to k-mers from both species. However, the implementation is not readily available and is not free for non-academic users. In [3](#) the authors also aligned the reads to both species, but no attempt was taken to disambiguate the data and no implementation is readily available.

Table 1. Read pairs assigned human (hg19) and mouse (mm10) post disambiguation in BWA aligned DNA-seq data. The 'Ambiguous' column includes reads that aligned to neither or had equal quality scores for the alignments and could not be disambiguated.

Sample	mm10	hg19	Ambiguous	Read pairs total
SRR1176814	47197650	26157 [†]	88542	47312349
SRR1528269	11502 ^{††}	77102895	153767	77268164

[†]Down from 25638785 read pairs with alignment to hg19

^{††}Down from 39686392 read pairs with alignment to mm10

Table 2. Read pairs assigned human (hg19) and mouse (mm10) post disambiguation in STAR aligned RNA-seq data. The 'Ambiguous' column includes reads that aligned to neither or had equal quality scores for the alignments and could not be disambiguated.

Sample	mm10	hg19	Ambiguous	Read pairs total
SRR1930152	23126086	80694 [†]	849364	24056144
SRR387400	94289 ^{††}	49677937	9880844	59653070

[†]Down from 3005372 read pairs with alignment to hg19

^{††}Down from 6001230 read pairs with alignment to mm10

Here, an alternative approach using read alignment quality is proposed to further disambiguate reads that can be mapped to both species. Alignment is first performed to both species independently and the reads are disambiguated as a post-processing step. There is no requirement to maintain pseudo reference indices based on combinations of reference sequences. This approach shows a very high sensitivity and specificity on artificially generated samples obtained by mixing reads from the individual species. The *Disambiguate* tool is community supported and widely used in several open and closed source pipelines.

Methods

Implementation

The *Disambiguate* algorithm works by operating on natural name sorted BAM files from alignments to two species. Name sorting is a critical part in not having to read all the data from both species' alignments into memory simultaneously; the same read aligned to both species is disambiguated on the fly by going through both alignment files synchronously. For reads that have alignments to both species and therefore require disambiguation, the specific details of the disambiguation process are slightly different for the different aligners. Thus far the algorithm has been tested for BWA-MEM⁴ and Bowtie2⁵ for DNA-seq, and TopHat2⁶, STAR⁷ and Hisat2⁸ for RNA-seq. Illumina's paired end sequencing is preferred as the mate can often break a tie. [Figure 1](#) illustrates the disambiguation process.

Disambiguate assigns the reads on a per-pair basis, based on the highest quality alignment of the read pair. For BWA and STAR the alignment score (AS, higher better) is used as the primary disambiguation metric followed by edit distance (NM, lower better) to the reference. For TopHat2 and Hisat2 based alignments the sum (lower better) of edit distance, number of reported alignments (NH) and the number of gap opens (XO) is used.

Operation

The algorithm is implemented in Python (with dependency on the *Pysam* package) and C++ (with dependency on *BamTools*), with the C++ version being approximately four times faster than the Python code. 64 bit unix/linux systems are supported.

Given name sorted alignment (BAM) files aligned to the two species of interest (e.g. human and mouse), the algorithm infers

Figure 1. The disambiguation process illustrated. Alignment is first performed against both species. The disambiguation application then operates on the raw, natural name sorted BAM files to assign the read pairs into one of the two species or as ambiguous for unresolved cases.

for each read the most likely origin. The output contains BAM files for both species, BAM files for ambiguous reads and a text file describing how many read pairs were assigned to each BAM file. The simplest way to perform all of the alignment and disambiguation is by running *bcbio*, in which *Disambiguate* is integrated, on the raw sequencing data.

Results

To illustrate the utility of *Disambiguate*, raw publicly available human and mouse exome sequencing reads (100bp paired end Illumina data) were downloaded from the European Nucleotide Archive (ENA) with Run Accessions SRR1176814 and SRR1528269.

The reads were concatenated, aligned against hg19 and mm10 using BWA MEM, and processed using *Disambiguate*. Pre-disambiguation, for the human sample (SRR1528269), there were 39686392 read pairs (out of total 77268164), for which at least one read aligned to mouse. Similarly, for the mouse sample (SRR1176814), there were 25638785 read pairs (out of total 47312349) for which at least one read aligned to human. Table 1 summarises the post disambiguation results. As can be seen, the disambiguation algorithm correctly pulls apart virtually all of the read pairs. In other internal studies, *Disambiguate* has time and again highlighted samples with low human assigned component, correlating with poor extraction or lack of growth of the tumour cells in the host.

STAR aligned human (SRR387400) and mouse (SRR1930152) RNA-seq data was also analysed with very similar results, see Table 2.

Conclusions

In summary, *Disambiguate* provides an important tool for computationally separating sequence reads originating from two species. In human-mouse studies it also allows the study of the mouse stromal component for gene expression and DNA variation.

In addition to RNA-seq and whole genome sequencing, it is worth highlighting that for targeted hybridisation capture sequencing of xenograft samples, where baits from a single species are used, disambiguation is still highly recommended. This is best seen in Table 1 where a large number of human exome reads aligned to mouse and would potentially affect downstream interpretation without disambiguation.

Disambiguate has been well adopted in the open source community; it is integrated in the open source *bcbio* pipeline, and has been successfully used in both RNA and DNA sequencing of xenografts both at AstraZeneca and other research institutes. This is evidenced by the number of support tickets from a variety of organisations on the *bcbio-nextgen* Github page.

Data availability

The data used here is available from the European Nucleotide Archive with Run Accession numbers SRR1176814 and SRR1528269.

Software availability

Software integrating *Disambiguate* available from: <https://github.com/chapmanb/bcbio-nextgen>

Latest source code: <https://github.com/AstraZeneca-NGS/disambiguate>

Archived source code as at time of publication: DOI: [10.5281/zenodo.1660179](https://doi.org/10.5281/zenodo.1660179)

License: MIT.

Author contributions

MA authored the bwa and rna-star disambiguation algorithms, co-authored the manuscript and implemented the algorithms in Python. SG wrote the C++ implementation of the algorithms. JJ co-authored the manuscript. ZL designed and implemented the original Tophat (and Hisat2) disambiguation algorithm and co-authored the manuscript.

Competing interests

All authors are employees of AstraZeneca.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgments

The authors wish to thank Brad Chapman, Rory Kirchner and Eric Schellhorn for feedback and fixes on *Disambiguate*.

References

- Bradford JR, Farren M, Powell SJ, et al.: RNA-Seq Differentiates Tumour and Host mRNA Expression Changes Induced by Treatment of Human Tumour Xenografts with the VEGFR Tyrosine Kinase Inhibitor Cediranib. *PLoS One*. 2013; 8(6): e66003. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Conway T, Wazny J, Bromage A, et al.: Xenome—a tool for classifying reads from xenograft samples. *Bioinformatics*. 2012; 28(12): i172–i178. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rossello FJ, Tothill RW, Britt K, et al.: Next-generation sequence analysis of cancer xenograft models. *PLoS One*. 2013; 8(9): e74432. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Li H: Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *bioRxiv*, arXiv:1303.3997 q–bio.GN. 2013. [Reference Source](#)
- Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012; 9(4): 357–359. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kim D, Pertea G, Trapnell C, et al.: TopHat2: accurate alignment of

- transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 2013; **14**(4): R36.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics*. 2013; **29**(1): 15–21.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements**. *Nat Methods*. 2015; **12**(4): 357–360.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Ahdesmäki MJ: **AstraZeneca-NGS/disambiguate: Release for publication [Data set]**. *Zenodo*. 2016.
[Data Source](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 05 December 2016

<https://doi.org/10.5256/f1000research.10863.r17881>

© 2016 Oliver G et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Gavin R. Oliver

Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

Asha A. Nair

Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

We believe that overall the software tool article by Ahdesmäki *et al.* seems sound and provides a solution to a problem that appears to be inadequately addressed in the field currently.

Nonetheless, we believe the manuscript would benefit from some minor amendments in order to increase its utility and accessibility to readers.

In brief:

Intro/Background

Needs expanded slightly to better set the scene and describe the general approach of read disambiguation.

Methodology

The methodology should be expanded slightly and made more explicit.

Tables 1&2:

- Combine 1 & 2 into a single table and label the samples by data type, i.e DNA and RNA
- Show %s as well as numbers
- Clearly label the species in the tables
- Clearly label correctly mapped/incorrectly mapped reads in table
- Clearly label human and mouse genomes as such

- Tables should clearly show all numbers pre- and post- disambiguation, rather than having superscripted references in the table legend
- Essentially, a novice should be able to read the paper and extract relevant info more easily.

Figure 1

- Should be more granular, informative and descriptive of the process. Include read alignment etc. Describe the Disambiguate process
- Use same font size for all text in the Figure

Comparison with a competitor product

This is something that is clearly missing. If it is literally impossible to compare to a competitor because the software is not accessible, this should be stated clearly as a reason for the lack of comparison in the paper.

Tumor samples

It would be interesting to know how performance is affected by use of highly mutated tumor xenografts. This is arguably beyond the scope of the paper, but warrants at least some mention.

Competing Interests: No competing interests were disclosed.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 11 Jan 2017

Miika Ahdesmäki,

Dear Gavin and Asha,

Many thanks for the very detailed review and comments. We have addressed your points in v2 of the manuscript.

Into/background:

We have added the text in braces: "Direct high throughput sequencing of grafted samples with a mixture of two species is routine practice. {However, the origin species of each read or read pair is unknown and needs to be determined informatically.}" to better set the scene. Further, the operation of xenome is now updated and xenome is now included in a comparison study. We have more explicitly stated that "Alignment is first performed to both species independently and the reads are disambiguated as a post-processing step, {assigning reads to the species with higher quality alignments}"

Methodology:

We have clarified the methodology section by spelling out the disambiguation algorithm and giving the reasoning why two schemes are used.

Table 1&2:

We have combined Tables 1&2 and revised the contents to address these points.

Figure 1:

We have redrawn the figure to be more descriptive.

Comparison to competitor product:

We have now compared our approach to Xenome, which was recently open sourced, and included the results of the comparison in the updated table with discussion.

Tumor samples:

We agree that evaluating the performance of the disambiguation algorithm in a messy cancer genome like the highly rearranged MCF7 would be extremely interesting. If we get our hands on appropriate data we will consider publishing the results on the program Github page.

Competing Interests: NA

Reviewer Report 25 November 2016

<https://doi.org/10.5256/f1000research.10863.r17879>

© 2016 Eldridge M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Matthew D. Eldridge

Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK

This paper describes a computational tool for separating sequencing reads from a sample that contains DNA or RNA from two species. This is a necessary pre-processing step for genomic or transcriptomic analysis of patient-derived xenograft cancer models.

The approach is based on alignments of sequence reads to the reference genome sequences for the two species in question. The authors have tested their approach on DNA-seq data from publicly available human and mouse exome datasets concatenated to simulate a xenograft sample. The results presented in Table 1 show very good separation of reads from the two species datasets with only a small percentage of reads being assigned to the wrong species (0.06% and 0.01%) and a higher but still very low percentage of reads flagged as ambiguous, i.e. align equally well to both genomes. Similar results were presented for RNA-seq data, although here the percentages of incorrectly assigned and ambiguous reads are unsurprisingly higher than for DNA-seq.

Use of the alignment scores, and in the event of a tie the edit distance, is a reasonable approach to disambiguate reads and is the method used for BWA and STAR alignments. For TopHat2 and

HISAT2 a different scoring function is required, although the reasons for this are not given. Further, the choice of function (sum of edit distance, number of reported alignments and number of gap opens) is not completely obvious and raises the question of whether the authors have attempted to tune the function, e.g. by adjusting the weighting of each component.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 11 Jan 2017

Miika Ahdesmäki,

Dear Mathew,

Many thanks for reviewing our manuscript and the comments. We have modified v2 of the manuscript to address the points you raise, namely:

1. The aligner tags are very similar between BWA and STAR; and between TopHat2 and HISAT2. However, fairly different between BWA/STAR vs TopHat2/Hisat2 and therefore we couldn't use the same scheme originally developed for TopHat2 with BWA/STAR. With the appearance of HISAT2 especially for hg38 we decided to utilise the TopHat2 scheme for HISAT2 given their outputs are almost interchangeable. We have mentioned this in the updated text.
2. The sum of edit distance, number of reported alignments and number of hap opens has always worked for us well out of the box (as illustrated in the tables) and while tuning their weights may yield some minor benefits, it would risk overfitting to existing data. Any benefits of the weight tuning would have to be measured over a very long time, running multiple versions of weighted and the unweighted algorithms side by side. We have given this reasoning (complexity) in the text as our excuse of not tuning the weights further.

Thank you again for the comments and helping us improve the manuscript.

Competing Interests: NA

Reviewer Report 23 November 2016

<https://doi.org/10.5256/f1000research.10863.r17877>

© 2016 Nicorici D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Daniel Nicorici

Orion Corporation Orion Pharma, Espoo, Finland

This paper introduces a tool, named Disambiguate, for computationally separating the DNA/RNA sequencing reads of two species, like for example in case of xenograft samples. The tool takes as input BAM files from wide range of NGS aligners.

I have made the following minor observations:

1. The tool Disambiguate works on RNA-seq and DNA-seq data and this is mentioned for the first time in Methods section. Probably it would help to have this mentioned much earlier, like for example in the abstract too.
2. In order to improve the clarity, to the Tables 1 and 2 could be added also the percentages where is relevant, like for example, "26157" would become "26157 (0.0553%)" and so on.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 11 Jan 2017

Miika Ahdesmäki,

Dear Daniel,

Thank you for the review, your comments are much appreciated. We have addressed your points in v2 of the manuscript.

1. We have explicitly mentioned in the abstract and the introduction that the tool can be used for both DNA and RNA-seq data
2. We have added percentages into the tables as you suggested

Thank you for the review and helping us improve the manuscript.

Competing Interests: NA

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research