Finding fixed variations between the transcriptomes of two closely-related Rhinoceros species

Ryan C. Thompson

January 23, 2012

Motivation

There are several extant species of Rhinoceros, and all but the White Rhino are endangered or critically endangered in the wild. In particular, there are only a few thousand Black Rhino individuals remaining in the wild due largely to illegal poaching. Some rhino species, such as White Rhinos, are grazers that subsist primarily on grasses, while others, such as Black Rhino, are browsers, adapted to roaming in search of a great variety of succulent and broad-leaf plants. However, in captivity they are typically given the same feed as their grazing counterparts. Reasons for this includ the impracticality of importing from Africa the necessary quantities of native-habitat browse, as well as the use of domestic horses, which are grazers, as the dietary model for *all* rhinos. As a result, attempts to raise browsing rhinos, including Black Rhinos, in captivity are frustrated by their tendency to develop iron overload when fed a zoo diet. (In contrast, grazers do not exhibit this problem in captivity.) Black Rhinos in captivity must be periodically bled to reduce the amount of iron in their bloodstream.

There are several hypotheses to explain the occurrence of iron overload. The current best-guess hypothesis is that the plants that make up the Black Rhino diet in the wild produce large quantities of iron chelators such as tannins in order to make themselves less attractive to herbivores. In response, it is thought that Black Rhinos have evolved to make efficient use of the available iron, or to somehow get at the chelated iron. Since their diet in captivity contains significantly more iron and fewer tannins than their natural diet, this leads to excessive iron uptake and iron overload.

This hypothesis, while attractive, cannot hint at the biological pathway or pathways that have been modified in Black Rhinos to cope with the dearth of available dietary iron. There are many possibilities: perhaps Black Rhinos have a countermeasure to iron chelators; perhaps they simply have super-efficient uptake of un-chelated iron; perhaps they never excrete their iron and simply accumulate sufficient quantities gradually over time. Any number of biological

mechanisms could underlie the prevalence of iron overload among captive Black Rhinos.

Given that the browsing Black Rhino is closely related to the grazing White Rhino, it is likely that their genomes are mostly identical, especially in protein-coding regions. It should be possible identify and align orthologous genes from the two species, and from the alignments it should be possible to find protein-coding variations that are fixed within each species but variant between the two species. Such variations are good candidates for being a causitive mutation for iron overload in Black Rhinos, especially if the variant that is fixed in White Rhino also matches an outgroup such as domestic horse or human. A transcriptome-wide search for such variations could yield many promising targets that suggest potential mechanisms for iron overload, which could in turn aid the search for an effective treatment for the condition.

Approach

In order to look at protein-coding variation between White and Black Rhinos, RNA-sequencing was performed on samples from the livers and spleens of individuals of both species. The liver and spleen are good tissue choices because their known central roles in iron metabolism makes it likely that genes containing causative variations are expressed there. The general goal is to assemble transcripts from the RNA-seq data and then to establish correspondence between White Rhino transcripts and their orthologous Black Rhino transcripts. Note that there is no reference genome for anyrhino species. The closest species with even a draft genome is the domestic horse. Hence, the best option is to use a *de novo* assembly tool to assemble transcripts without a reference genome.

From there, the problem is analgous to intra-species variant calling, treating the White Rhino sequence as the "reference" and the Black Rhino sequence as the "affected". Candidate variations should be entirely invariant among all individuals of a single species but different between the two species, indicating that the variant is fixed within each species' population. Additionally, if the White Rhino variant matches the orthologous position in an outgroup, then the variation likely occurred in the Black Rhino lineage after Black and hite Rhinos diverged, making it an even better candidate. Naturally, a candidate variation that occurs in a gene with a known iron-related function would be ideal, but not every "iron gene" is annotated as such, and even a mutation in a gene not directly related to iron could still affect iron metabolism indirectly. For example, Black Rhinos could harbor a variant that allows a protease to degrade tannins, freeing the iron chelated inside. Thus, the search for a causitive variation must be performed across the whole transcriptome, rather than limiting the search to a subset of genes known to have a role in iron metabolism.

Table 1: Sequencing samples

Sample	Description
wr0001	Half-lane each of white rhino spleen & liver
br0492	Full lane of black rhino liver
br0656	Half-lane each of black rhino spleen & liver
br1177	Full lane of black rhino spleen

Methods

Samples

RNA was isolated from samples of liver and spleen tissue of one white rhino and three black rhinos. The RNA samples were sequenced on an Illumina HiSeq instrument. Sequencing was paired-end 100 bp. Each sequencing lane contained either a single sample or two multiplexed samples from the same individual. The sequenced samples are detailed in Table 1. Multiplexed lanes were demultiplexed into separate files for each sample before further analysis.

Data pre-processing

Prior to assembly with Trinity, each read was trimmed to just the first consecutive run of bases quality 20 and higher. That is, any initial run of bases below quality 20 was trimmed off the start of the read, and then the remainder was truncated right before the first base of quality less than 20. Reads shorter than 25 bp after trimming were discarded, since Trinity uses a k-mer size of 25. Any read whose mate was discarded was used as a single-end read. Trimming was performed using Trimmomatic[1].

The goal of this stringent quality filtering is to remove as many sequencing errors as possible in order to reduce the memory usage and run-time of the Trinity *de novo* assembler[2]. Since Trinity is a de-Bruijn graph assembler, its runtime and memory requirements depend primarily on the number of unique k-mers, even if most of those k-mers represent sequencing errors that are filtered out during the assembly process. With genome sequencing, errors could be filtered out or corrected before assembly by analyzing the k-mer frequency spectrum. However, this technique assumes uniform coverage and is therefore inadvisable for RNA-seq data, which is highly non-uniform. Hence, the purely quality-based filtering approach was used instead since it would not bias the assembly against low-abundance transcripts.

Trinity Assembly

To perform the assembly the recently-published Trinity was run on the quality-trimmed data set with the minimum contig length set to 100, the read length.

```
>IPI:IPI00000013.1 Gene_Symbol=CTSL2 Cathepsin L2

121 RKKGYVTPVKNQKQCGSCWAFSATGALEGQ M FRKTGKLVSLSEQNLVDCSRPQGNQGCNG 180 Ref

 79 .E.........G.................. T .....R.............A........ 138 BR
 79 .E.........G.................. T .....R.............A........ 138 BR
 79 .E.........G.................. T .....R.............A........ 138 BR
 79 .E.........G.................. T .....R.............A........ 138 BR

 79 .E.........G.................. . .....R.............A........ 138 WR
121 .E.........G.................. . .....R.............A........ 180 WR
 79 .E.........G.................. . .....R.............A........ 138 WR
121 .E.........G.................. . .....R.............A........ 180 WR
```

Figure 1: Fragment of multiple alignment of *H. Sapiens* Cathepsin L2 and all matching Trinity contigs. The position that is variant between the White and Black Rhino sequences is highlighted.

The quality trimming described above reduced memory requirements of Trinity by about 50%, despite removing only about 10% of the total nucleotides in the data sets.

Tblastn searches

After assembly, the resulting contigs from each sample were formatted as a BLAST database and the full set of protein sequences from the September 2011 release of IPI was queried against each sample database using tblastn. The same procedure was also performed with a curated set of known iron-related protein sequences from all species.

Clustal omega multiple alignments

The BLAST results were found to be inadequate for mutation finding, either by eye or by computational means. To derive a better alignment in a more standard format, a script was produced to extract the high-scoring hits (above 85% identity) from the BLAST XML file and use align them along with their corresponding query using Clustal Omega[3]. The result is a multiple sequence alignment of each query with its high-scoring hits from all samples, stored in the standard stockholm format. With the alignments in a standard format, a second script was produced to print each alignment in human-readable fashion.

Results

From a manual inspection of the multiple alignments, it is clear that the White and Black Rhino sequences are nearly identical to each other in all protein-coding regions. (Note that aligning at the protein level with tblastn automatically limits the search to nonsynonymous variations.) However, there are rare differences between them that are consistent across all contigs (and therefore

potentially represent fixed mutations). One such example, found by manual inspection, is shown in Figure 1. The Black Rhino sequences all have a threonine in the highlighted column, while the White Rhino and human sequences have a methionine. Note that this 60-residue segment of the alignment also contains 4 other variant sites where all rhino sequences are identical and different from the human sequence. The majority of variant sites are of this kind, reflecting the very close evolutionary relationship between these two rhino species.

Artifacts at contig ends

Manual inspection of the multiple alignments also revealed that the contigs produced by Trinity frequently had highly suspect sequences at the ends. This may be due to the omission of an adapter trimming step prior to assembly. In the future, a tool such as SeqPrep[?] or the adapter-trimming functionality of Trimmomatic[1] should be used to ensure that the input to the assembler is free of adapter contamination.

Assembly performance

As mentioned above, a the quality trimming step greatly reduced the memory requirements of Trinity (about 50%) with only a modest (10%) reduction in sequence quantity. This is highly useful, since it allows Trinity to run on significantly larger sequencing data sets and produce better assemblies without requiring additional memory. Some sort of error filtering step before assembly is highly recommended.

Future Work

The next step is to conduct an automated search through every column of each generated multiple alignment for columns that match the expected signature. Such a column would have the same amino acid for all White Rhino contigs, and a different amino acid for all Black Rhino contigs. Additionally, the outgroup reference's amino acid in that column should not match the Black Rhino one, and ideally it should match the White Rhino amino acid, as in Figure 1. After variations matching the expected signature are identified, they can be run through a functional prediction program such as SIFT or Polyphen to identify predicted damaging mutations.

References

[1] Anthony Bolger. Trimmomatic: A flexible read trimming tool for illumina ngs data, 2010.

[2] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen,

E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29:644–652, Jul 2011.

[3] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, 7:539, 2011.