


RESEARCH ARTICLE

Open Access



# What's in your next-generation sequence data? An exploration of unmapped DNA and RNA sequence reads from the bovine reference individual

Lynsey K. Whitacre<sup>1,2</sup>, Polyana C. Tizioto<sup>2,3</sup>, JaeWoo Kim<sup>2</sup>, Tad S. Sonstegard<sup>4,5</sup>, Steven G. Schroeder<sup>4</sup>, Leeson J. Alexander<sup>6</sup>, Juan F. Medrano<sup>7</sup>, Robert D. Schnabel<sup>1,2</sup>, Jeremy F. Taylor<sup>2\*</sup> and Jared E. Decker<sup>1,2\*</sup> 

## Abstract

**Background:** Next-generation sequencing projects commonly commence by aligning reads to a reference genome assembly. While improvements in alignment algorithms and computational hardware have greatly enhanced the efficiency and accuracy of alignments, a significant percentage of reads often remain unmapped.

**Results:** We generated *de novo* assemblies of unmapped reads from the DNA and RNA sequencing of the *Bos taurus* reference individual and identified the closest matching sequence to each contig by alignment to the NCBI non-redundant nucleotide database using BLAST. As expected, many of these contigs represent vertebrate sequence that is absent, incomplete, or misassembled in the UMD3.1 reference assembly. However, numerous additional contigs represent invertebrate species. Most prominent were several species of Spirurid nematodes and a blood-borne parasite, *Babesia bigemina*. These species are either not present in the US or are not known to infect taurine cattle and the reference animal appears to have been host to unsequenced sister species.

**Conclusions:** We demonstrate the importance of exploring unmapped reads to ascertain sequences that are either absent or misassembled in the reference assembly and for detecting sequences indicative of parasitic or commensal organisms.

**Keywords:** DNA sequencing, RNA sequencing, Unmapped reads

## Background

Next-generation sequencing technology has vastly increased the dimensionality of sequencing projects and routinely allows the generation of hundreds of millions or even billions of short reads. Analysis of these data requires that the short reads be assembled into contiguous sequences either using *de novo* or reference-guided assembly. For organisms with a reference genome, reads generated in the sequencing process are usually matched to the reference sequence with a variety of alignment algorithms. This is currently the most efficient way of transforming the raw sequence reads into a consensus

sequence. However, there are several limitations inherent to the alignment process, including alignment to repetitive regions, absent or misassembled sequence in the reference genome, and individual genetic divergence between the subject organism's genome and the reference genome [1]. Despite these challenges, the majority of reads produced from a sequencing experiment will adequately align to a reference assembly. Nevertheless, a small but significant fraction of reads frequently remain unmapped.

Unmapped reads have generally been disregarded and these data are often discarded. However, recent work has begun to focus on the development of bioinformatic tools for detecting pathogens in human sequence data by the computational subtraction of known human sequences [2–4]. Application of these pipelines in other

\* Correspondence: taylorjerr@missouri.edu; DeckerJE@missouri.edu

<sup>2</sup>Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA

<sup>1</sup>Informatics Institute, University of Missouri, Columbia, MO 65211, USA

Full list of author information is available at the end of the article



recent studies has suggested that potentially biologically relevant information can be extracted from the unmapped reads [5, 6]. Using an original alignment, assembly, and identification pipeline that can be applied to data from any species, we took advantage of a unique opportunity to explore the unmapped reads from the DNA and RNA sequencing of L1 Dominette 01449, the *Bos taurus* reference individual [7]. These data had not previously been used in the creation or annotation of the reference assembly.

Using sequence data produced from the reference individual, we minimized alignment challenges that are due to genetic variation among individuals. Thus, we expected to encounter meaningful biological information pertaining to sequences poorly represented in the bovine reference assembly and sequences indicative of parasitic or commensal non-vertebrate organisms. We identified DNA and RNA contigs that were assembled *de novo* from unmapped reads that could generally be classified into one of three categories: 1) sequence from bovine; 2) sequence from other vertebrate species that was homologous to bovine; and 3) sequence from non-vertebrate species. This analysis unequivocally demonstrates that the unmapped reads contain important data pertaining to sequences from the organism that are missing from the reference assembly, represented by categories 1 and 2, and sequences that can be used to identify microbiota members, putatively represented by category 3.

**Results**

**De novo assembly of unmapped reads**

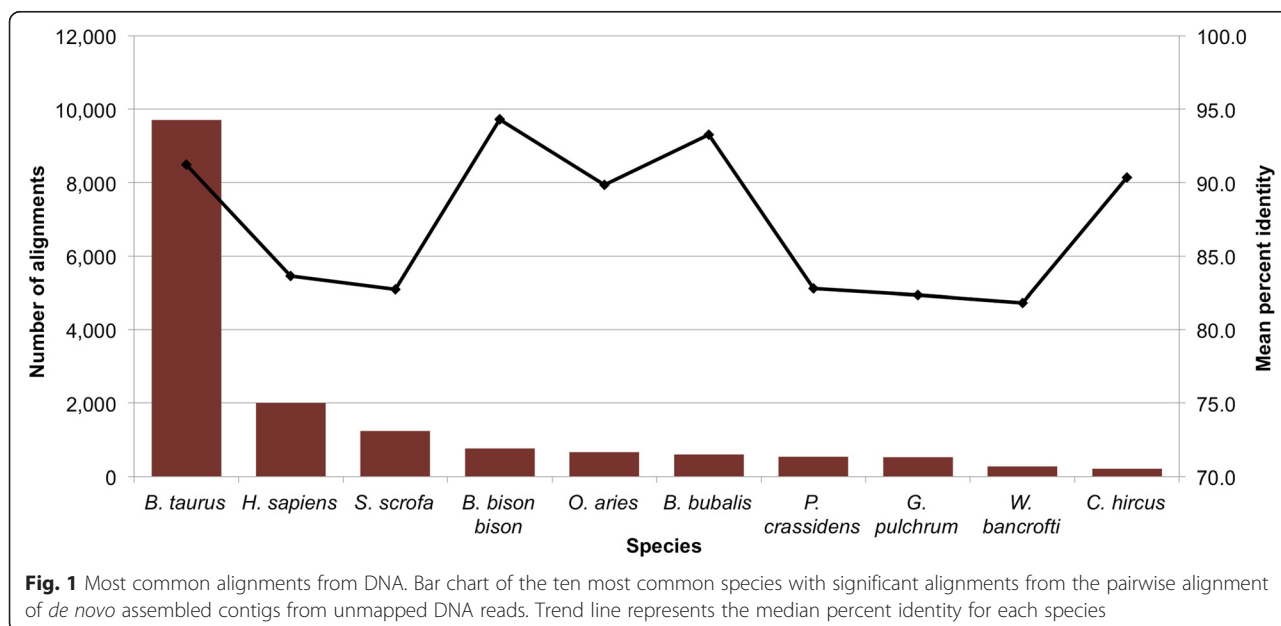
Approximately 111.7 million DNA sequence reads, 7.2 % of the total, remained unmapped after alignment to the

reference genome. A fraction of those reads could be used for assembly, due to a large number of sequences with low quality (Additional file 1: Table S1). However, approximately 1.4 million reads were incorporated into 69,230 contigs with an N50 of 737 bp. Overall, the contigs comprised approximately 46.6 Mb. Additional assembly statistics are provided in Additional file 1: Table S1.

A median of approximately 6.7 % of RNA-seq reads remained unmapped across each of the 17 tissue samples. *De novo* assembly of these reads yielded a total of 43,961 contigs, with a median of 1792 contigs per tissue and an N50 of 324.5 bp. Overall, the contigs spanned 14.8 Mb with a median of 603 Kb per tissue. Assembly statistics for each tissue are in Additional file 1: Table S2.

**Pairwise alignment of contigs assembled from unmapped DNA reads to the non-redundant nucleotide database**

Approximately 51 % of the contigs generated from the unmapped DNA reads produced a significant alignment when queried against the non-redundant nucleotide (*nt*) database. The most common alignment was to other *Bos taurus* sequences (Fig. 1). This result was expected given the draft quality of the bovine reference assembly and considering that we assembled paired reads if either one or both of the reads were unmapped to the reference assembly. However, the second most common alignment for these DNA contigs was to *Onchocerca ochengi*, a nematode known to infect indicine cattle that has been heavily researched due to its similarity to the parasite that causes African River Blindness in humans. We simulated paired-end sequence read data from the *O.*



*ochengi* genome assembly by randomly shearing the genome and then aligned the produced paired-end reads to the bovine reference assembly and concluded that the *O. ochengi* assembly is contaminated with cattle sequences (Additional file 2: Note 1). Consequently, we excluded the *O. ochengi* assembly from any further analyses.

With subsequent analyses preventing alignment to *O. ochengi*, approximately 44 % of the contigs produce a significant alignment against the *nt* database. A fraction of the contigs originally identified as *O. ochengi* were unambiguously matched to bovine sequences. However, the number of alignments to other filarial nematode sequences also increased. These included hundreds of contigs aligned to *Gongylonema pulchrum* and *Wuchereria bancrofti*, and a few to *Parascaris equorum*. *G. pulchrum* and *W. bancrofti* belong to the order Spirurida, as does *O. ochengi*, but that are known to only infect humans. The alignments to each of these species had a percent identity of approximately 82 % (Table 1), which is consistent with cattle not being a host for these nematodes and indicating that these alignments represent sequences from unsequenced sister species of *G. pulchrum* and *W. bancrofti*.

Also detected were sequences with high percent identities to *Babesia bigemina*, a blood-borne parasite known to cause bovine babesiosis, or Texas fever in cattle. While only 190 contigs aligned to *B. bigemina*, significantly less than the combined number of alignments to nematode species, ten were larger than 1000 bp and the median identity was 91.10 % (Table 1). A complete summary of significant alignments, both vertebrate and non-vertebrate, is presented in Additional file 1: Table S3.

Alignments to other vertebrates represent cattle sequences that are not currently well represented in the *Bos taurus* database. Thus, the number of alignments to these organisms is a function of the completeness of the available data for each species and the phylogenetic relationship between the species and cattle. For example, human (*Homo sapiens*), being the most complete, has the largest number of alignments of the other vertebrate species, followed by pig (*Sus scrofa*), and while bison (*Bison bison bison*) and water buffalo (*Bubalus bubalis*) are more closely related to cattle than human or pig, these bovinds have less sequence data available and thus do not produce as many alignments.

#### Pairwise alignment of contigs assembled from unmapped RNA-seq reads to the non-redundant nucleotide database

The pairwise alignment of the *de novo* assembled contigs generated from the unmapped RNA-seq reads to the *nt* database produced similar results to the alignment of the DNA contigs. Overall, 81 % of the RNA-seq contigs had significant alignments to sequences in the *nt* database. Across all tissues, *Bos taurus* produced the largest number of alignments. Also prominent were alignments to *Bison bison bison*, *Bubalus bubalis*, and *Bos mutus*, all species that are closely related to cattle (Fig. 2). Significant BLAST alignments of the RNA-seq unmapped read contigs to cattle or these other closely related species indicates the existence of coding regions that are missing or misassembled in the reference assembly. By mapping the GI number of the most significant BLAST alignment to a gene symbol, we detected alignments to 4412 *B. taurus* and 4029 *B. bison bison*, *B. bubalis*, or *B. mutus* genes. As the total number of *Bos taurus* genes reported by Ensembl is 19,994 [8], this suggests that as many as 42.2 % of the bovine protein coding genes are misassembled (although these misassemblies likely represent a small fraction of total transcriptome base pairs). Additionally, approximately 5 % of RNA alignments failed to map to a gene with an assigned symbol, likely corresponding to unannotated structural or regulatory RNAs. Further results and discussion of these analyses are included in Additional file 2: Note 2.

As was the case for the pairwise alignment of unmapped DNA contigs, there were also numerous alignments to other vertebrate and non-vertebrate species. The most common alignments to non-vertebrate species included uncultured bacterium, bovine herpesvirus 6, *Onchocerca flexuosa* and *B. bigemina* (Table 2). Bovine herpesvirus 6 was previously discovered as a contaminant in the UMD3.1 build by Merchant *et al.* [9], who concluded that Dominette must have been host to the virus. Alignments to *O. flexuosa* and *B. bigemina* support the hypothesis generated from the analysis of the unmapped DNA read contigs that Dominette was also host to a nematode of the Spirurida order and an unsequenced relative of *B. bigemina*. Several additional fungal and bacterial species were also detected in the unmapped read RNA-seq contigs at low levels. Nearly all of the

**Table 1** Top four non-vertebrate alignments to *de novo* assembled contigs from unmapped DNA sequence reads

Species	Number of Alignments	Median Identity (%)	Maximum Identity (%)	Median Length (bp)	Maximum Length (bp)	Median E-Value
<i>Gongylonema pulchrum</i>	516	82.33	100	641.0	1,008	2.00E-134
<i>Wuchereria bancrofti</i>	273	81.35	98.82	640.0	1,607	1.53E-143
<i>Babesia bigemina</i>	206	91.10	100.00	505.0	2,078	3.50E-179
<i>Parascaris equorum</i>	11	81.17	100	206.0	1,008	4.00E-41

**Table 2** Top four non-vertebrate alignments to *de novo* assembled contigs from unmapped RNA-seq reads

Species	Number of Alignments	Median Identity (%)	Maximum Identity (%)	Median Length (bp)	Maximum Length (bp)	Median E-Value
<i>Uncultured bacterium</i>	34	97.11	100.00	274.0	1,381	2.01E-116
<i>Bovine herpesvirus 6</i>	22	99.18	100.00	294.5	933	1.00E-143
<i>Onchocerca flexuosa</i>	13	87.74	89.12	224.0	510	3.00E-57
<i>Babesia bigemina</i>	12	94.35	99.51	379.5	926	1.00E-151

detected non-vertebrate organisms had alignments from multiple tissues, which would be expected for blood-borne parasites. The number of alignments for each tissue was a function of the total number of sequencing reads from that tissue. A complete summary of alignments from all 17 tissues is presented in Additional file 1: Tables S4 and S5.

#### No evidence of horizontal gene transfer

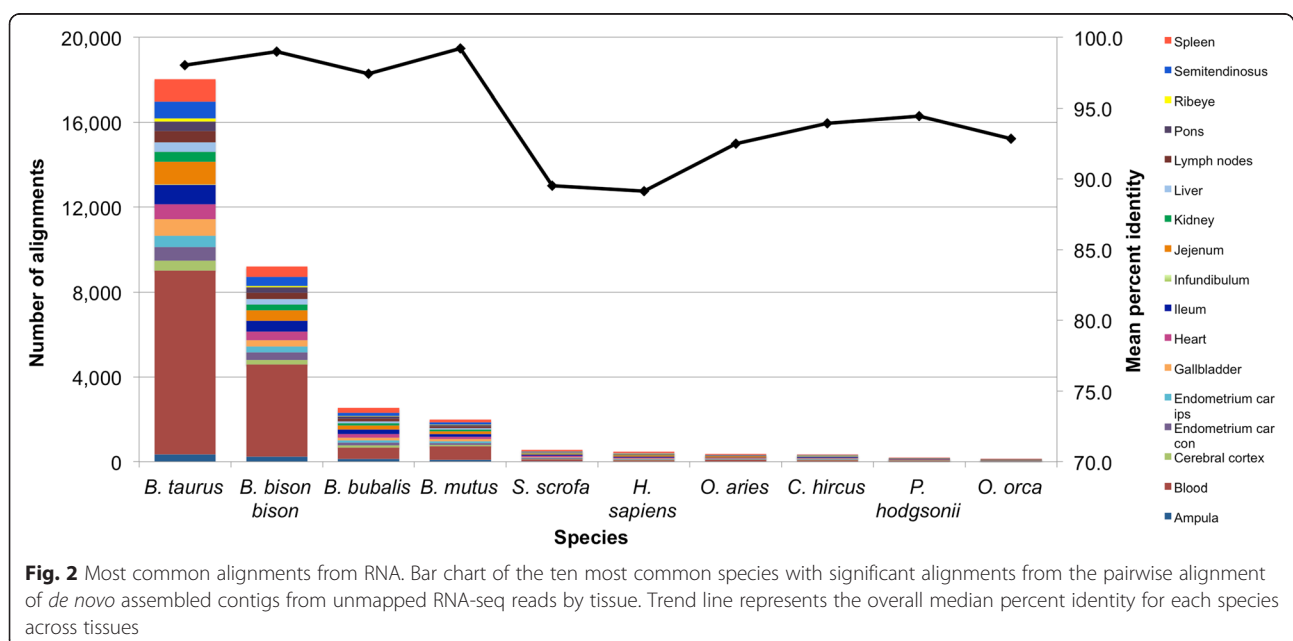
With deep sequencing, it is possible to expose rare horizontal gene transfer events. To address this possibility, we searched for mate-pair reads from the large insert DNA libraries where one mate was uniquely mapped to the cattle reference genome and the other mate mapped uniquely to a non-vertebrate sequence. No such mate-pairs were identified that met these criteria. Additionally, in our BLAST results we also searched for chimeric contigs (contigs that partially mapped uniquely to cattle and partially mapped uniquely to a non-vertebrate species). Again, no such contigs were identified that met these criteria.

#### Discussion

To our knowledge, this is the first formal investigation into the nature and identity of unmapped reads from the

resequencing of an individual used for the generation of a reference genome assembly. These data allowed us to directly compare reads to the reference assembly without alignment challenges due to genetic variation between the reference and the resequenced genome. Second, the opportunity to compare independently generated datasets from the same individual provided unequivocal support for our discovery of concordant non-vertebrate sequences within the whole genome and transcriptome sequences of the bovine host. In addition to our sequencing of cDNA generated from RNA isolated from 17 tissues, we also sequenced genomic DNA that had been isolated from both liver and white blood cells at three separate facilities. Endogenous contaminants were detected in the reads that were generated from all three sequencing runs. Nearly all of the contigs assembled *de novo* from unmapped reads that were identified as representing a non-vertebrate species were comprised of reads that originated from multiple libraries sequenced at separate facilities. These attributes facilitated both the discovery and validation of the parasitic and commensal species sequences found in this study.

Despite the continuing exponential increases in sequences submitted to NCBI's databases, the number of



represented species still comprises only a small proportion of existing species. While we detected several sequence alignments to spirurid nematodes in both the DNA and RNA sequence data, none of these species are known to be present in the US or to infect taurine cattle. Therefore, we postulate that the actual species present within the tissues of Dominette either represent undiscovered species or a previously recognized, but unsequenced, organism such as *Onchocerca gutturosa* or *Onchocerca lienalis* from the Spirurida order. Both *O. gutturosa* and *O. lienalis* are known to infect taurine cattle in various parts of the United States [10]. However, these species have not been sequenced other than for a few selected genes used to generate data for phylogenetic analyses [11–19]. In this study, we assembled nearly 1000 contigs that we believe represent novel sequence from a Spirurid nematode that infects taurine cattle in North America.

The precise identity of the species generating the sequence matching *B. bigemina* in both the RNA-seq and genomic DNA data is also ambiguous. As no fever like symptoms were reported in this cow who spent her life at a USDA research facility near Miles City, Montana and babesiosis has been reported to have been eradicated in the United States with vaccination no longer being required [20], we suspect that Dominette was asymptotically infected with a non-pathogenic strain of *Babesia* spp., as has previously been reported in Turkey [21], Syria [22] and Thailand [23]. Although it is currently not possible to determine the exact species of parasite, we can estimate the animal's parasite burden via deep sequence data by evaluating the number of species to which the contigs of unmapped reads align and the number of contigs that align to each species. Parasite burden negatively impacts animal health and profitability [24, 25] and can serve as a reservoir for later infections [22]. Although symptoms were not visible and the animal appeared healthy, the detection of subclinical parasite burden, even from non-pathogenic parasites, is important because a physiological response to the infection from the host must still occur. This response reduces fitness, causes a decrease in production traits such as feed intake and feed efficiency [25, 26] and can also influence the interpretation of RNA-seq experiments.

An alternate explanation for the identification of non-vertebrate sequences in a vertebrate animal is the actual integration of these DNA sequences into the animal's genome. Recently, horizontal gene transfer has been reported to occur at a low level in many animal species [27–29]. It has also been reported that there can be integration of foreign DNA released by dead cells into healthy host cells [30]. However, we were unable to find evidence for the integration of non-vertebrate DNA into this animal's genome and must exclude horizontal gene transfer based on our data.

## Conclusions

In conclusion, we alert researchers that many sequences of interest may be found in the reads that fail to align to a reference assembly. We demonstrate that the unmapped reads contain biologically significant information relative to genes that are either partially or completely missing from the reference assembly, as well as information regarding the identity and magnitude of commensal or parasitic organisms. The large number of missing or misassembled bovine protein coding genes must significantly impact the interpretation of RNA-seq studies, warrants further research, and is likely more severe in the less complete reference genomes of other livestock species. Continuation of unmapped read mining will also expand our knowledge of the extent of internal parasitic infections and may lead to the discovery of previously unknown symbiotic relationships. These metagenomic inferences are an additional source of information from whole-genome sequencing data that can be used as phenotypes or covariates in downstream analyses. As the quality of reference assemblies improves and the scope of sequenced microorganisms broadens, the detection of parasitic infections and other symbiotic relationships will become more explicit.

## Methods

### Ethics statement

Tissues from L1 Dominette 01449 were sampled according to IACUC No. 081711–1, which was approved by the USDA-ARS Fort Keogh Livestock and Range Animal Care and Use Committee.

### DNA and RNA sequencing

DNA was extracted from liver and whole blood samples from L1 Dominette 01449 (referred to here as “Dominette”), a Hereford cow used to generate the *Bos taurus* Sanger reference assembly [7], and was sent to three separate facilities for sequencing. Paired-end and mate-pair libraries were constructed and DNA was 2 x 100 bp sequenced to an average coverage of approximately 55X. Further details regarding the sequencing of each library is provided in Additional file 1: Table S6.

RNA was extracted using Trizol Reagent (Invitrogen, Carlsbad, CA) as described elsewhere [31] from 17 tissue samples including ampulla, blood, cerebral cortex, endometrium sampled from caruncular regions contralateral (car con) and ipsilateral (car ips) to the corpus luteum, gallbladder, heart, ileum, infundibulum, jejunum, kidney, liver, mesenteric lymph nodes, pons, ribeye muscle, semitendinosus muscle, and spleen. Preparation of the mRNA samples for sequencing was performed by Global Biologics (Columbia, MO) using the TruSeq Stranded mRNA Library Prep Kit (Illumina®, San Diego, CA) and sequenced 2 x 100 bp using Illumina technology, with

the exception of blood which used the TruSeq RNA Sample Preparation Kit and was sequenced 1 x 100 bp.

#### Pre-processing and alignment of reads

Error correction was performed on DNA sequence reads using the QuorUM error correction algorithm [32]. After filtering duplicate and low quality reads, 1,622,097,087 unique reads remained. Paired reads were aligned to the UMD3.1 cattle reference assembly using NextGENe 2.4.1 (SoftGenetics, LLC, State College, PA) requiring at least 35 contiguous bases with  $\geq 95.0$  % overall match, up to 2 allowable mismatched bases, and up to 100 allowable alignments of equal probability genome-wide.

RNA sequence reads were filtered for quality and adapter sequences and were then trimmed using a custom Perl script already described [31]. Computations were performed on the HPC resources at the University of Missouri Bioinformatics Consortium (UMBC). TopHat v2.0.6 [33] was used to map the reads to the *Bos taurus* UMD3.1 reference genome. A total of 2 mismatches and up to 3 bp indels were allowed in alignment.

#### De novo assembly of unmapped reads

Reads from DNA sequencing that remained unmapped following alignment to the reference genome were assembled using MaSuRCA 2.3.2 [34]. Reads from RNA sequencing that remained unmapped following alignment were assembled using Trinity version r20140717 [35]. For both DNA and RNA assemblies, the default parameters were used. To maintain a paired read file structure, reads where both the forward and reverse read were unmapped or where one of the reads was unmapped but the other was mapped were collectively used for assembly.

#### Pairwise alignment of unmapped contigs to the nt database

Prior to pairwise alignment, contigs assembled from the unmapped DNA reads were sorted by size and only contigs greater than 500 bases were aligned ( $n = 42,086$ ). Due to the smaller size of the RNA contigs, they were not filtered by size prior to pairwise alignment. Using the blastn algorithm of BLAST+ 2.2.30 [36, 37] and the command line options `-db nt, -max_target_seqs 1, -outfmt "6 qseqid sseqid staxids sscinames pident length mismatch gapopen value"`, each DNA and RNA contig was aligned to the NCBI non-redundant nucleotide database and the most significant alignment was returned. The `-negative_glist` option was used with a text file of all *O. ochengi* gi numbers in subsequent blast searches excluding *O. ochengi* sequences. The BLAST output was then parsed to determine the subject species, percent identity, length of match, number of mismatches, number of gaps, E-value, and overall score. Significant alignments were declared only if the length of the alignment was  $\geq 150$  bp for DNA or  $\geq 50$  bp for

RNA. Only the best match for each aligned contig was reported. This output was summarized according to the total number of alignments per species, mean, median, and maximum percent identity, mean, median, and maximum length of match, and mean and median E-value (Additional file 1: Tables S3 and S5).

#### Quantification and identification of coding regions within unmapped reads

Contigs from unmapped RNA-seq reads were aligned to contigs from unmapped DNA reads using NextGENe 2.4.1 requiring  $\geq 98$  % overall match to declare a match. Additionally, for the significant RNA alignments, the gene symbol corresponding to the GI accession number for the alignment was captured where possible and recorded using the db2db tool in bioDBnet [38]. A unique list of gene symbols was constructed and the number of significant alignments to each gene was tallied (Additional file 1: Tables S7 and S8).

#### Availability of data and materials

The data set supporting the results of this article is available in the SRA repository, SRA accessions SRX1177177 through SRX1177278.

#### Additional files

**Additional file 1: Table S1.** Statistics from the *de novo* assembly of unmapped reads from DNA sequencing. **Table S2.** Statistics from the *de novo* assembly of unmapped reads from RNA sequencing. **Table S3.** Summary of all significant alignments from pairwise alignment of *de novo* assembled contigs from DNA unmapped reads to the *nt* database. **Table S4.** Number of significant alignments per tissue from pairwise alignment of *de novo* assembled contigs from unmapped RNA-seq reads to the *nt* database. **Table S5.** Summary of all significant alignments from pairwise alignment of *de novo* assembled contigs from unmapped RNA-seq reads to the *nt* database. **Table S6.** DNA sequencing metadata. **Table S7.** Genes represented in alignments of *de novo* assembled contigs from unmapped RNA-seq reads to *Bos taurus*. **Table S8.** Genes represented in alignments of *de novo* assembled contigs from unmapped RNA-seq reads to *Bison bison bison*, *Bubalus bubalis*, or *Bos mutus*. (XLSX 731 kb)

**Additional file 2: Note 1:** The *Onchocerca ochengi* reference assembly is contaminated with bovine genomic sequence. **Note 2:** Estimation of the number of protein coding genes missing or misassembled in the UMD3.1 bovine reference assembly. (DOCX 41 kb)

#### Competing interests

J.F.T. is on the scientific advisory boards (SABs) of Recombinetics, Inc and Neogen Corporation.

#### Authors' contributions

Author contributions are as follows: JFT, JED, RDS, and LKW designed the experiments and interpreted the results of all analyses. L.K.W. built the analysis pipeline and analyzed the sequence data. PCT did alignments and *de novo* assemblies for RNA sequence data. LJA collected tissues and extracted nucleic acids. JWK extracted and quantitated RNA. TSS, SGS and JFM sequenced genomic DNA. JFT, JED and RDS sequenced DNA and RNA. LKW wrote the manuscript and JFT, JED and RDS edited the manuscript. All authors read the final manuscript.

### Acknowledgments

Funding for this study was provided in part from the bovine species coordinators of the USDA National Institute of Food and Agriculture supported NRSP-8 National Animal Genome Research Support Program and National Research Initiative Competitive Grants numbers 2011-68004-30214, 2011-68004-30367, 2012-67012-19743, 2013-68004-20364, 2015-67015-23183, MO-HAAS0027, and MO-MSAS0014 from the USDA National Institute of Food and Agriculture. The authors appreciate the contributions of the Beijing Genomics Institute in generating whole genome and tissue transcriptome sequence from the bovine reference animal.

### Author details

<sup>1</sup>Informatics Institute, University of Missouri, Columbia, MO 65211, USA. <sup>2</sup>Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA. <sup>3</sup>Embrapa Southeast Livestock, São Carlos, São Paulo 13560-970, Brazil. <sup>4</sup>Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD 20705, USA. <sup>5</sup>Recombinetics Inc., 1246 University Ave W #301, St Paul, MN 55104, USA. <sup>6</sup>USDA-ARS (retired), LARRL, Fort Keogh Miles City, Montana 59301, USA. <sup>7</sup>Department of Animal Science, University of California-Davis, Davis, CA 95616, USA.

Received: 26 August 2015 Accepted: 15 December 2015

Published online: 29 December 2015

### References

- Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet*. 2010;11:31-46.
- Bhaduri A, Qu K, Lee CS, Ungewickell A, Khavari PA. Rapid identification of non-human sequences in high-throughput sequencing datasets. *Bioinformatics*. 2012;28:1174-5.
- Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RGW, Getz G, et al. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat Biotechnol*. 2011;29:393-6.
- Isakov O, Modai S, Shomron N. Pathogen detection using short-RNA deep sequencing subtraction and assembly. *Bioinformatics*. 2011;27:2027-30.
- Tae H, Karunasena E, Bavara JH, McIver LJ, Garner HR. Large scale comparison of non-human sequences in human sequencing data. *Genomics*. 2014;104(6 Pt B):453-8.
- Gouin A, Legeai F, Nouhaud P, Whibley A, Simon J-C, Lemaître C. Whole-genome re-sequencing of non-model organisms: lessons from unmapped reads. *Heredity (Edinb)*. 2015;114(5):494-501.
- Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, et al. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol*. 2009;10:R42.
- Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. *Nucleic Acids Res*. 2014;43:D662-669.
- Merchant S, Wood DE, Salzberg SL. Unexpected cross-species contamination in genome sequencing projects. *PeerJ*. 2014;2:e675.
- Eberhard ML, Stilesi OFO, Eberhardt ML. Studies on the Onchocerca (Nematoda: Filarioidea) Found in Cattle in the United States. I. Systematics of *O. gutturosa* O. lionalis with a Description of *O. stilesi* sp. n. *J Parasitol*. 1979;65:379-88.
- Gill LL, Hardman N, Chappell L, Hu Qu L, Nicoloso M, Bachelier J-P. Phylogeny of *Onchocerca volvulus* and related species deduced from rRNA sequence comparisons. *Mol Biochem Parasitol*. 1988;28:69-76.
- Casiraghi M, Anderson TJC, Bandi C, Bazzocchi C, Genchi C. A phylogenetic analysis of filarial nematodes: comparison with the phylogeny of *Wolbachia* endosymbionts. *Parasitology*. 2001;122:93-103.
- Casiraghi M, Bain O, Guerrero R, Martin C, Pocacqua V, Gardner SL, et al. Mapping the presence of *Wolbachia pipientis* on the phylogeny of filarial nematodes: evidence for symbiont loss during evolution. *Int J Parasitol*. 2004;34:191-203.
- Xie H, Bain O, Williams SA. Molecular phylogenetic studies on filarial parasites based on 5S ribosomal spacer sequences. *Parasite*. 2014;1:141-51.
- Krueger A, Fischer P, Morales-Hojas R. Molecular phylogeny of the filaria genus *Onchocerca* with special emphasis on Afrotropical human and bovine parasites. *Acta Trop*. 2007;101:1-14.
- Garofalo A, Kläger SL, Rowlinson M-C, Nirmalan N, Klion A, Allen JE, et al. The FAR proteins of filarial nematodes: secretion, glycosylation and lipid binding characteristics. *Mol Biochem Parasitol*. 2002;122:161-70.
- Morales-Hojas R, Cheke RA, Post RJ. Molecular systematics of five *Onchocerca* species (Nematoda: Filarioidea) including the human parasite, *O. volvulus*, suggest sympatric speciation. *J Helminthol*. 2006;80:281-90.
- Morales-Hojas R, Cheke RA, Post RJ. A preliminary analysis of the population genetics and molecular phylogenetics of *Onchocerca volvulus* (Nematoda: Filarioidea) using nuclear ribosomal second internal transcribed spacer sequences. *Mem Inst Oswaldo Cruz*. 2007;102:879-82.
- Kulke D, von Samson-Himmelstjerna G, Miltsch SM, Wolstenholme AJ, Jex AR, Gasser RB, et al. Characterization of the Ca<sup>2+</sup>-gated and voltage-dependent K<sup>+</sup>-channel Slo-1 of nematodes and its interaction with emodepside. *PLoS Negl Trop Dis*. 2014;8:e3401.
- Bock R, Jackson L, de Vos A, Jorgensen W. Babesiosis of cattle. *Parasitology*. 2004;129(Suppl):S247-69.
- Altay K, Aydin MF, Dumanli N, Aktas M. Molecular detection of *Theileria* and *Babesia* infections in cattle. *Vet Parasitol*. 2008;158:295-301.
- Terkawi MA, Alhasan H, Huyen NX, Sabagh A, Awier K, Cao S, et al. Molecular and serological prevalence of *Babesia bovis* and *Babesia bigemina* in cattle from central region of Syria. *Vet Parasitol*. 2012;187:307-11.
- Simking P, Saengow S, Bangphoomi K, Sarataphan N, Wongnarkpet S, Inpankaew T, et al. The molecular prevalence and MSA-2b gene-based genetic diversity of *Babesia bovis* in dairy cattle in Thailand. *Vet Parasitol*. 2013;197:642-8.
- Corwin RM. Economics of gastrointestinal parasitism of cattle. *Vet Parasitol*. 1997;72:451-7. discussion 457-60.
- Hawkins JA. Economic benefits of parasite control in cattle. *Vet Parasitol*. 1993;46:159-73.
- Gunn A, Irvine RJ. Subclinical parasitism and ruminant foraging strategies-a review. *Wildl Soc Bull*. 2003;31:117-26.
- Dunning Hotopp JC. Horizontal gene transfer between bacteria and animals. *Trends Genet*. 2011;27:157-63.
- Syvanen M. Evolutionary implications of horizontal gene transfer. *Annu Rev Genet*. 2012;46:341-58.
- Crisp A, Boschetti C, Perry M, Tunnacliffe A, Micklem G. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biol*. 2015;16:50.
- Mitra I, Khare NK, Raghuram GV, Chaubal R, Khambatti F, Gupta D, et al. Circulating nucleic acids damage DNA of healthy cells by integrating into their genomes. *J Biosci*. 2015;40:91-111.
- Chapple RH, Tizioto PC, Wells KD, Givan SA, Kim J, McKay SD, et al. Characterization of the rat developmental liver transcriptome. *Physiol Genomics*. 2013;45:301-11.
- Marçais G, Yorke JA, Zimin A. Quorum: an error corrector for Illumina reads. *arXiv.org* 2013.
- Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25:1105-11.
- Zimin A, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome Assembler. *Bioinformatics*. 2013;29:2669-77. btt476-.
- Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 2013;8:1494-512.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403-10.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421.
- Mudunuri U, Che A, Yi M, Stephens RM. bioDBnet: the biological database network. *Bioinformatics*. 2009;25:555-6.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
www.biomedcentral.com/submit

