Xenome—a tool for classifying reads from xenograft samples

Thomas Conway¹,*, Jeremy Waizn,¹ Andrew Bromage,¹ Martin Tymms,² Dhanya Sooraj,² Elizabeth D. Williams² and Bryan Beresford-Smith¹

¹NICTA Victoria Research Laboratory, Department of Computer Science and Software Engineering, The University of Melbourne, Parkville and ²Monash Institute of Medical Research, Monash University, Clayton, Australia

ABSTRACT

Motivation: Shotgun sequence read data derived from xenograft material contains a mixture of reads arising from the host and reads arising from the graft. Classifying the read mixture to separate the two allows for more precise analysis to be performed.

Results: We present a technique, with an associated tool Xenome, which performs fast, accurate and specific classification of xenograft-derived sequence read data. We have evaluated it on RNA-Seq data from human, mouse and human-in-mouse xenograft datasets.

Availability: Xenome is available for non-commercial use from http://www.nicta.com.au/bioinformatics

Contact: tom.conway@nicta.com.au

1 INTRODUCTION

Xenograft models are an important tool for many areas of biomedical research, including oncology, immunology and HIV pathology. A typical scenario, drawn from oncology research, is that of a human prostate cancer grown in an immunocompromised mouse model. Doing so allows researchers to investigate aspects of the cancer that are not necessarily preserved in cell lines, and it allows investigations into the interactions between the cancer and the surrounding stromal tissue. The mouse may be biopsied or harvested and samples of cancer and/or stroma collected at various time points during an experiment.

Difficulties arise, when sequencing the genome or transcriptome of the samples because host (mouse) material (i.e. DNA/RNA) will inevitably comeling with the graft (human) material. If a sufficiently careful section is taken, it has been generally assumed that the level of host contamination is low enough that it may be ignored. This may be a dangerous assumption, however, since the level of gene expression is non-uniform. If the overall level of host contamination in a graft sample is measured to be 10% overall, it may still be the case for a given gene that the host homologue accounts for most or all of the expression.

Contamination may be minimized by physical or biochemical techniques such as conservative sectioning, cell sorting or laser capture micro-dissection, but these techniques can be a significant source of technical bias, or in some cases may require infeasibly large amounts of starting material. Further, in the case of transcriptomic investigation, classifying host and graft in vitro may fail to adequately capture the interactions between them.

An alternative strategy is to sequence an acknowledged mixture of host and graft, then use in silico methods to classify the individual sequence reads. This is the approach discussed here. We demonstrate a simple technique, based on an analysis of sequence reads using TopHat, and a more precise technique based on a k-mer decomposition of the host and graft reference sequences, Xenome. In both cases, the primary goal of the analysis is to classify reads into four classes: reads attributable to the host, reads attributable to the graft, reads which could be attributed to both and reads which are attributable to neither. To the best of our knowledge, there are no results in the literature examining the classification of high-throughput sequencing short reads from xenograft models.

The studies we know of are concerned with microarray expression profiles or alternative methods for estimating the amount of host material or cell types in the samples. For example, [Lin et al. 2011] investigate the use of species-specific variation in gene length and a multiplex PCR to ascertain the relative amount of mouse and human DNA. [Wang et al. 2011] use microarray gene profiling data and in silico techniques to estimate the quantity of various tissue components. In [Samuels et al. 2010], there is an analysis of a mouse xenograft model using microarray data. They conclude that if there is more than 90% human DNA then the expression profiles are not unduly skewed. They also describe an experimental method for removing homologous genes based on cross-hybridization analysis of the probes. Ding et al. (2010) use short read sequencing to study a cancer genome and identify mutations/deletions. They estimate tumour cellularity using pathological assessment, and state that their xenograft is 90% tumour cells. They also map NOD/SCID (mouse) genomic data to human and mouse genomes, reporting 3.17% and 95.85% mapping rates, respectively, and so apply no correction for the murine cells. We note that in the context of non-uniform RNA-Seq data ignoring the contribution of the murine expression can lead to biases.

Tools such as TopHat serve a different purpose than that of Xenome. The former aligns reads to a reference, and we can use those alignments for a variety of purposes, including the classification task we present here. In contrast, Xenome only performs the classification task itself. This is an important distinction, since an alignment must assign the read to zero or more positions in the genome; the classification merely has to decide if the read was more likely to arise from the genome than not.

For the remainder of the article, we will assume, unless otherwise stated, that sequence reads arise from RNA-Seq. However, the techniques we present are applicable to genomic DNA sequences (including ChIP-Seq and MeDIP-Seq) and also to other mixtures of DNA species.

2 METHODS

Under the assumption that a graft sample has only a low level of host material contamination, the simplest analysis is to use a regular mapping-based RNA-Seq analysis tool, such as TopHat and assume that either the observed expression is dominated by the graft, which has the greatest number...
of input cells, or that the homology between the host species and graft species is such that reads arising from host material will tend to map poorly, and the resultant inferred level of gene expression will be negligible.

In some cases, these assumptions may be true, but in the case of human xenografts in mice, for example, the second assumption is false for many transcripts, and a more precise technique is desirable.

Therefore, we have developed two techniques—one based on the canonical RNA-Seq resequencing tool Tophat (Trapnell et al., 2009) and some simple scripts. As will be apparent in Section 4, we will present our specific choice which has important performance ramifications.

This definition can be extended to a set of $k$-mers $S$ in the obvious way:

$$S = \{x \in \sigma \}$$

**Definition 2.** (Marginal inclusion). Consider a set of canonical $k$-mers $S$. We say that a $k$-mer $x$ is in marginal membership of $S$ if $x$ does not exist in $S$, but has a Hamming distance 1 neighbour $y$ such that $y$ is a member of $S$.

To aid our computation of marginal inclusion, we define the function $M_x: \mathbb{H} \rightarrow \mathbb{H}$ where $H = \{y \in \sigma : \exists x \in S \}$. Note that $H$ is the set of Hamming distance 1 neighbours of $x$.

For each reference $k$-mers tend to be large: with $k = 25$, there are 2.4 billion in the human genome, 2.1 billion in the mouse genome, and 4.5 billion in the union—only 12 million are shared.

A naive representation (2 bits per base, packed) would use 50 bits per $k$-mer. To compete a complete set of canonical $k$-mers $H$ and $G$ respectively. From these we compute a compact set of canonical reference $k$-mers $\hat{S} \subseteq \hat{G}$. Note that $\forall x \in \mathbb{G} \cup \mathbb{H}, |x| = k$.

**2.2.1 Definitions** Since in most sequencing protocols the reads are a mixture of forward and reverse complements with respect to the reference sequence, we cannot assume the orientation of $k$-mers drawn from reads will match the orientation of $k$-mers drawn from the reference. We could always consider both orientations, but that would entail a lot of double handling of information, so instead we normalize or canonicalize $k$-mers.

**Definition 1.** ($k$-mer canonicalization). Consider a $k$-mer $x$. We denote its reverse complement by $\bar{x}$.

A canonical $k$-mer $x$ is defined by a choice function $C$ giving a deterministic choice between $x$ and $\bar{x}$:

$$\forall x : C(x) \in \{x, \bar{x}\}$$

In principle, we can choose any such function: nor or max being obvious candidates, and the results of our method are identical for all choices. In Section 4 we will present our specific choice which has important performance ramifications.

2.2.2 Reference construction. For both host and graft reference sequences, we construct the set of canonical $k$-mers ($H$ and $G$) respectively. From these we compute a compact set of canonical reference $k$-mers $\hat{S} \subseteq \hat{G}$. Note that $\forall x \in \mathbb{G} \cup \mathbb{H}, |x| = k$.

The sets of canonical $k$-mers tend to be large: with $k = 25$, there are 2.4 billion in the human genome, 2.1 billion in the mouse genome, and 4.5 billion in the union—only 12 million are shared.

A naive representation (2 bits per base, packed) would use 50 bits per 25-mer, or about 26 GB. As discussed in our previous work (Conway and Bromage, 2009), information theory gives a lower bound for the memory usage. For a domain of $d^k$ possible $k$-mers, the minimum number of bits required to represent a set of $k$-mers is

$$ \log_2 (d^k/8) $$

In the case of the union $\mathbb{S}$ above, this is about 10 GB, or less than half what is required for the naive representation. As also discussed in our previous work, succinct data structures have been developed to give concrete representations that approach this theoretical lower bound. The one we use, due to Okanohara and Sadakane (2006), works very well, requiring about 13 GB.

For each reference $k$-mer, we determine whether it occurs in the host, the graft, both, or if it occurs in one and has a marginal occurrence in the other. These classes are denoted $h$, $g$, $b$ or $m$, respectively (Fig. 4). More formally, we compute the function $K$, for each $x \in \mathbb{S}$:

$$ K(x) = \begin{cases} h & \text{if } \exists \bar{y} \in \mathbb{G} \cup \mathbb{H} \text{ such that } K(\bar{y}) = h \\ g & \text{if } \exists \bar{y} \in \mathbb{G} \cup \mathbb{H} \text{ such that } K(\bar{y}) = g \\ b & \text{if } \exists \bar{y} \in \mathbb{G} \cup \mathbb{H} \text{ such that } K(\bar{y}) = b \\ m & \text{if } \exists \bar{y} \in \mathbb{G} \cup \mathbb{H} \text{ such that } K(\bar{y}) = m \\ \emptyset & \text{otherwise} \end{cases} $$

$K$ may be extended to project not just $k$-mers to classes, but also a set of $k$-mers $Q$ to a set of classes in the obvious way:

$$ K(Q) = \{K(x) : x \in Q\} $$

These classes are pre-computed and stored as a sequence of 2-bit values corresponding to the $k$-mers in the succincty store reference set.
Consider an idealized hash table with load factor $\lambda$ (i.e. the number of values stored in the table equals the number of slots in the hash table), with no collisions. If the width of the hash table is $2^k$, then $J$ bits of the key are implied by the slot chosen. For keys of $N$ bits, therefore, the entries of the hash table need store only $N-J$ bits. The simplest way this may be realized would be to just use $J$ bits of the key as the slot number, and store the remainder in that slot, but of course, this is likely to have an unacceptable number of collisions.

Instead, we use an invertible hash function [based on a single-stage Feistel network, see Luby and Rackoff (1988)] to turn a key $x$ into a slot number $s$ and a stored component $r$ in such a way that given $s$ and $r$ we can recover $x$ (assuming a hash function $f$):

$$F_r(s)=\left\{\begin{array}{ll}
((x \mod 2^J) \oplus f(s)) \mod 2^k & \text{if } 2^J \leq 2^s \\
((x \mod 2^J) \oplus f(s)) \mod 2^N & \text{otherwise}
\end{array}\right.$$

The size of the key data stored in the hash table is, then, $2^s(N-J)$ bits. If $2^J \leq 2^s$, this is within $\lfloor n \log_2(1+1/j) \rfloor$ bits, and hence succinct.

Of course, idealized hash tables are not possible if the set of keys is not known in advance. Hence we use cuckoo hashing [Pagh and RRR, 2005], in which collisions are resolved (as far as possible) using multiple hash functions.

To make the sorting of keys more efficient, our hash function preserves the $N-J$ most significant bits of the keys, which are then stored in the hash table (along with a few bits to determine which hash function was used). This means that we can perform an initial bucketing without inverting the hash functions; the buckets can then be processed independently, using multiple threads if required.

Now consider a set of random $k$-mers. For hash-based canonicalization, given a good hash function, there is an equal probability of observing any base in the most significant position of $k$. For lexicographic canonicalization, there is a probability of $\frac{1}{2}$ that it is $A$, $\frac{1}{2}$ that it is $C$, $\frac{1}{2}$ that it is $G$, and only $\frac{1}{4}$ that it is $T$. For a set of $N$ random $k$-mers, the expected entropy at the most significant base is 2.0 bits and 1.85 bits, respectively.

This matters, because hash functions are frequently vulnerable to poor behaviour in the presence of highly correlated sets of keys. In real, biologically derived sets of $k$-mers, the set of $k$-mers is non-random, so there is likely to be less entropy to begin with, and therefore less entropy due to canonicalization is exacerbated. This is especially problematic in the case of the above family of invertible hash functions, since it is precisely the most-significant bits of the key which are passed to the underlying hash function $f$.

The upshot of this is that $C_{\text{sync}}$ leads to highly correlated $k$-mers, which in turn lead to a high probability of unresolvable collisions even when the hash table is nearly empty, resulting in a large number of short runs. Since the set of canonical $k$-mers that result from $C_{\text{sync}}$ are less correlated, the hash table is unlikely to encounter unresolvable collisions until it is almost full (80-85% in practice).
name the host and graft files. If paired data is being classified, the --pairs flag should be given—pairs are classified by computing $K(Q)$ over all the $k$-mers in the pair.

```
$ xenome classify -T 8 -P 1 dx --pairs \
  --graft-name human --host-name mouse \
  --output-filename-prefix XYZ \
  -1 XYZ_1.fastq -2 XYZ_2.fastq
```

This yields the following set of files:

- `XYZ_ambiguous_1.fastq`
- `XYZ_ambiguous_2.fastq`
- `XYZ_human_1.fastq`
- `XYZ_human_2.fastq`
- `XYZ_mouse_1.fastq`
- `XYZ_mouse_2.fastq`
- `XYZ_neither_1.fastq`
- `XYZ_neither_2.fastq`

If the total size of the index files is larger than RAM, xenome will perform poorly, but the flag --H can be supplied with the maximum desired working set size, and the classification will be done in multiple passes each using less memory.

On a single server with 8 AMD Opteron cores running at 2 GHz and with 32 GB of RAM xenome processes ~15 000 read pairs per second.

Having classified the reads, typical usage would be to then run Tophat and Cufflinks to perform intron-aware gapped alignments and compute gene expression. Instead of running Tophat on all of the reads, having separated the reads according to their origin, we can run it with the human genome just against the human fraction of the reads, and against the mouse genome on the mouse fraction of the reads. It may well be desirable to combine the both and ambiguous fractions with the human fraction to run against the human genome, and also with the mouse fraction to run against the mouse genome. If this is done, attention should be paid to homologous genes, since it is possible that the human and mouse homologues may be represented by the same reads. In some cases, the correct attribution of the gene expression may be apparent from the nature of the experiment. For example, a sample from a human prostate cancer grown in mouse may show ambiguous expression in MYH genes which would be reasonably attributed to the stromal mouse tissue. Instead of running Tophat and Cufflinks for each fraction with the human fraction to run against the human genome just on all of the reads, having separated "host" and "graft" files. If paired data is being classified, the --pairs flag should be given—pairs are classified by computing $K(Q)$ over all the $k$-mers in the pair.

```
$ xenome classify -T 8 -P 1 dx --pairs \
  --graft-name human --host-name mouse \
  --output-filename-prefix XYZ \
  -1 XYZ_1.fastq -2 XYZ_2.fastq
```

This yields the following set of files:

- `XYZ_ambiguous_1.fastq`
- `XYZ_ambiguous_2.fastq`
- `XYZ_both_1.fastq`
- `XYZ_both_2.fastq`
- `XYZ_human_1.fastq`
- `XYZ_human_2.fastq`
- `XYZ_mouse_1.fastq`
- `XYZ_mouse_2.fastq`
- `XYZ_neither_1.fastq`
- `XYZ_neither_2.fastq`

On a single server with 8 AMD Opteron cores running at 2 GHz and with 32 GB of RAM xenome processes ~15 000 read pairs per second.

Having classified the reads, typical usage would be to then run Tophat and Cufflinks to perform intron-aware gapped alignments and compute gene expression. Instead of running Tophat on all of the reads, having separated the reads according to their origin, we can run it with the human genome just against the human fraction of the reads, and against the mouse genome on the mouse fraction of the reads. It may well be desirable to combine the both and ambiguous fractions with the human fraction to run against the human genome, and also with the mouse fraction to run against the mouse genome. If this is done, attention should be paid to homologous genes, since it is possible that the human and mouse homologues may be represented by the same reads. In some cases, the correct attribution of the gene expression may be apparent from the nature of the experiment. For example, a sample from a human prostate cancer grown in mouse may show ambiguous expression in MYH genes which would be reasonably attributed to the stromal mouse tissue.

### 3 RESULTS

Our analysis examines two questions: whether or not it is technically feasible to separate host and graft reads in silico; and whether or not the fast technique we have proposed (Xenome) yields a worthwhile improvement over the mapping (Tophat) based technique.

In the first experiment, we take a sample of human cDNA sequence data (SRR342886), and a sample of mouse cDNA sequence data (SRR037689) and analyse them both with Tophat and Xenome, and compare the results. This allows us to evaluate the degree to which sequences are misclassified (assigned to human rather than mouse or vice versa), and the specificity of the classification—the proportion of sequences which are not classified as both. The use of pure human or mouse cDNA gives an experiment where the correct assignment of reads is known.

The second experiment runs the same analysis on sequence data from a human prostate cancer xenograft growing in a mouse host [BM18, see McCulloch et al. (2009)]. In this case, however, we not only classify the reads, but use the Tophat mappings to compute approximate levels of gene expression [measured in fragments per thousand bases of transcript per million mapped reads, or FPKM (Trapnell et al. 2010)] and use human species-specific quantitative RT-PCR (qRT-PCR) on selected genes to validate the results. In this instance, we have no gold standard by which we can judge the

---

**Fig. 2.** Summary of the results with Human cDNA. Each of the classes of reads is divided into those reads assigned to the class only by Xenome (Xenome), only by the Tophat analysis (Tophat) or by both Xenome and the Tophat analysis (Concordant) results, but the qRT-PCR will give some degree of validation, and known aspects of the biology of the cancer can give some qualitative corroboration.

Tophat uses a global analysis, combining the results of all the read mappings to locate exons, junctions and so on. In contrast, Xenome performs pre-computation on the two reference genomes, then classifies each read independently. Therefore, for each of the three sets of reads, we ran Tophat with the human reference genome and again with the mouse reference genome, then, as described in Section 2.3, the mappings were post-processed to determine which reads belonged to each of the four classes. Each of the four partitions of the sets of reads was then partitioned with Xenome to allow us to easily determine which reads were classified as the same by both procedures, and which were classified differently.

Figures 2A and 2B summarize the results. For the three samples, the proportion of reads receiving the same classification were 82, 87 and 84%, respectively. As can be seen from the human and mouse only figures, both techniques are accurate, in as much as they misclassify only a small proportion of the reads. In silico procedures, and which were classified differently.

Figures 2A and 2B summarize the results. For the three samples, the proportion of reads receiving the same classification were 82, 87 and 84%, respectively. As can be seen from the human and mouse only figures, both techniques are accurate, in as much as they misclassify only a small proportion of the reads. The worst case being that the latter yields better specificity—the fraction of reads classed as both is significantly smaller in the Xenome analysis. To check for false positives for the human cDNA dataset, we also used BLAT [29] to map the 1.1 million reads which Xenome classed as human but which were not mapped to either genome by Tophat. Most of them were successfully mapped with high quality to the human reference by BLAT (about 90%). BLAT also mapped about 18% of them, with very variable quality, to the mouse reference. This supports our confidence in the accuracy of the Xenome algorithm. From this we can conclude that the in silico classification of sequences is feasible and accurate.
The second experiment, using RNA-Seq data from a prostate cancer xenograft into a mouse demonstrates that the classification works on a real mixture. As described above, we performed the same process to partition the reads into classes, then we used the refGene genome coordinates as calculated by Tophat to assign reads to genes, from which we computed an expression level using the fragments per kilobase of transcript per million mapped reads formula (Trapnell et al., 2010):

$$\text{FPKM} = \frac{f \times 10^9}{zN}$$

where $f$ is the number of fragments (reads, for single ended data or pairs for paired data), $z$ is the combined length of the exons of the gene and $N$ is the total number of mapped fragments. Although this quantification is peripheral to the technique we are presenting, we have computed expression levels for the purposes of comparing with some qRT-PCR data for the same biological data. The qRT-PCR data were available for 18 genes: ABCG2, ALDH1A1, CD177, DLL1, DLL3, GLI1, GLI2, HES1, JAG1, JAG2, LGR5, NANOG, NOTCH1, NOTCH2, NOTCH3, PTCH1, PTCH2 and SMO. Figure 5 shows the log$_{10}$FPKM versus the difference of the $C_t$ for each target gene and the $C_t$ for actin (which was used as a housekeeping gene). With the exception of NANOG, the two methods correlate reasonably well (the Pearson’s correlation coefficient is 0.80). We have investigated the NANOG data, and cannot explain the low FPKM. Whether this is a sequencing issue or a biological variation in the mice is unknown, but the low level of expression does not appear to be related to the behaviour of Xenome or Tophat.

4 DISCUSSION

We have presented a simple read classification method based on Tophat, and our refined classification approach, Xenome. Xenome can be used to efficiently and effectively partition the read set for subsequent processing by tools such as Tophat.

What is not apparent from the results above is the relative behaviour at the level of a single gene. It should be expected that the distribution of ambiguously mapped reads (classed as both) should be non-uniform, since some genes in the two genomes are more highly conserved than others.
The first result we present in Figure 6 on this point is an in silico analysis showing the proportion of each human gene (ignoring introns) covered by k-mers that are not classed as human. It is clear that the vast majority of genes contain few or no k-mers that are not classed as human.

The fraction of k-mers which are ambiguous gives a worst-case view of how Xenome might be expected to perform. In order for a read to be classified as both, all of its k-mers must be of the both class, or there must be at least one k-mer from each of the two genomes (which happens less than 2% of the time in the samples we have tried). Conversely, a single host or graft k-mer is sufficient to classify the read into the respective class. Therefore for a read to be classified as both, the reference must contain a sufficiently long run of consecutive k-mers of class both and/or SNPs and sequencing errors must eliminate all the distinctively host or graft k-mers.

The second result we report on this point, presented in Figure 7, is the relative proportion of reads which are classed as both on a per-gene basis. What is evident in this figure is that although there are many genes for which the proportion of both reads is tightly correlated between Tophat and Xenome, there are a large number of genes for which the Tophat-based analysis has significantly more both reads. There are 15,591 genes for which there were at least 20 mapped reads in the BM18 xenograft sample. Of these, there were 65 for which Xenome assigned both or ambiguous to at least half the reads mapping to the gene; there were 498 for which the Tophat-based analysis assigned both to at least half the reads mapped to the gene. For the most highly conserved genes, there is not much that can be done with this data directly—further signal processing or other data would be required to determine the relative expression in the host and graft.

While we have developed Xenome with RNA-Seq on human mouse xenografts in mind, we anticipate it will be an effective tool for other similar mixtures. For example, capturing the differential methylation around genes between host and graft using MeDIP-Seq may shed light on the interaction between the two.

Fig. 6. An in silico analysis showing the degree of ambiguity in HG19 refGene, according to the k-mer based analysis used by Xenome. In this analysis, k = 25.

Fig. 7. A plot showing the distribution of human genes with respect to the proportion of xenograft reads which are classed as both by the Tophat-based analysis and the Xenome analysis. The reads considered are only those mapped by Tophat since Xenome does not yield mappings, so cannot be used to assign reads to genes. Only genes for which at least 20 reads mapped were considered. The horizontal axis corresponds to the number of reads classified as both or ambiguous by Xenome as a proportion of all the reads that might possibly be human (i.e. both, ambiguous or human). The vertical axis corresponds to the number of reads classified as both by the Tophat-based analysis, once again, as a proportion of all the reads that might possibly be human.

The need for our technique is substantially motivated by the fact that for a xenograft to be viable there must be very strong homology between the host and graft organisms. This leads to a situation where there is a high probability that a read may map to either genome, and it is this problem that Xenome specifically addresses. A side benefit is that the classification is done independently for each read, and results in groups of reads in each class; each group may then be processed independently with further tools (such as Tophat, Cufflinks [Roberts et al., 2011] or others). Given that many kinds of analysis require global processing of the input data, being able to process a coherent subset of the data can lead to a time/space gain. This benefit extends beyond the sphere of xenografts. For example, there are situations where a parasite or pathogen cannot be cultured independently (for example Chlamydia, and some fungi), so samples will generally contain a mixture of host and pathogen. In some examples, although the pathogen can be cultured independently, there are phenotypic differences between organisms growing in culture and those growing in a host. In both cases, there are benefits to being able to classify the two groups of reads, even though straight mapping based approaches will be less sensitive to cross-talk than xenograft data.

Precise alignment and alignment-free methods represent different points along a spectrum of possible classification techniques. Fundamentally, both rely on establishing homology between a read and the host and/or graft references. By substituting different algorithms for establishing homology (e.g. various alignment algorithms, k-mer spectrum methods, etc.), different sensitivity and specificity might be achieved.

Although our current technique is built on simple set-based classification, there is clearly scope to develop statistical models...
which allow for a more subtle classification procedure. These could, for example, be based on the frequency of k-mers of different classes, giving rise to the computation of the likelihood that a read originated from either the host or graft genomes. Indeed, if such likelihoods were used from read alignments (Tophat unfortunately does not produce such scores), a unified model allowing either our current k-mer based model or alignments could be used.

A further extension could take into consideration the established homology between genes in the host and graft organisms. Where reads are ambiguous, non-ambiguous reads associated with the same gene homologues could be used to help disambiguate the classification. EM-based methods such as those described in [Newkirk et al. 2011] would be a good basis for such an extension. We note however, that this would require significant conceptual changes to Xenome since it requires relatively more precise alignments.

It is instructive to consider a specific example of a gene where the Tophat based and Xenome analyses are very different. For the gene MYH3, in the human cDNA dataset, there are 29 reads classed as based and it requires relatively more precise alignments. Were the kind hinted at above may perform similarly to the Tophat analysis were classed as based by the latter. This is because there is a high level of conservation in this gene between the two species, and the reads therefore aligned to both. Were Tophat to yield meaningful alignment quality scores, a statistical approach of the kind hinted at above may perform similarly to the k-mer based approach.

ACKNOWLEDGEMENT

National ICT Australia (NICTA) is funded by the Australian Government’s Department of Communications; Information Technology and the Arts; Australian Research Council through Backing Australia’s Ability; ICT Centre of Excellence programs.

Funding: Prostate Cancer Foundation of Australia (EDW) and the Victorian Government’s Operational Infrastructure Support Program. EDW is supported by an Australian NHMRC Career Development Award [#519539] in part.

Conflict of Interest: none declared.

REFERENCES

Ashburner, Y. et al. (2010) Backyard cuckoo hashing: Constant worst-case operations with a succinct representation. In 2010 IEEE 51st Annual Symposium on Foundations of Computer Science, IEEE Computer Society, Los Alamitos California. pp. 787–796.

Chung, D. et al. (2011) Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. PLoS Comput. Biol., 7, e1002111.

Conway, T.C. and Bromage, A.J. (2011) Succinct data structures for assembling large genomes. Bioinformatics, 27, 479–486.

Ding, L. et al. (2010) Genome remodelling in a basal-like breast cancer metastasis and tumourigenesis. Nature, 464, 999–1005.

Hormozdiari, F. et al. (2010) Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics, 26, 1350–1357.

Kent, W. (2002) BLAT—the BLAST-like alignment tool. Genome Res., 12, 656–664.

Lin, Y. et al. (2011) Comparative studies of de novo assembly tools for next-generation sequencing technologies. Bioinformatics, 27, 2011–2017.

Li, R. et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res., 20, 265–272.

Luby, M. and Rackoff, C. (1988) How to construct pseudorandom permutations from pseudorandom functions. SIAM J. Comput., 17, 373–386.

McCullough, D. et al. (2005) BM18: a novel androgen-dependent human prostate cancer tumour model derived from a bone metastasis. Prostate, 65, 35–43.

Newkirk, D. et al. (2011) AREM: aligning short reads from ChIP-seqencing by expectation maximization. J. Comput. Biol., 18, 1499–1505.

Okanohara, D. and Sadakane, K. (2007) Practical entropy-compressed rank/select dictionary in. Proceedings of the Workshop on Algorithm Engineering and Experiments (ALENEX’07), http://www.siam.org/proceedings/alenex/2007/ alx07_007okanoharad2.pdf.

Pugh, R. and Rollett, F. (2004) Cuckoo hashing. J. Algorithms, 51, 122–144.

Roberts, A. et al. (2011) Identification of novel transcripts in annotated genomes using RNA-seq. Bioinformatics, 27, 2325–2329.

Samek, A. et al. (2010) Validation of a mouse tumour model system for gene expression analysis of human acute lymphoblastic leukemia. BMC Genomics, 11, 256.

Trapnell, C. et al. (2009) Tophat: discovering splice junctions with RNA-Seq. Bioinformatics, 25, 1105–1111.

Trapnell, C. et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol., 28, 511–515.

Wang, Y. et al. (2010) In-silico estimates of tissue components in surgical samples based on expression profiling data. Cancer Res.
