BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 11 2011, pages 1481–1488
doi:10.1093/bioinformatics/btr184

Sequence analysis

Advance Access publication April 9, 2011

Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data

Andrew McPherson[1,2,*,†], Chunxiao Wu[3,†], Iman Hajirasouliha[1,†], Fereydoun Hormozdiari[1,†], Faraz Hach[1], Anna Lapuk[3], Stanislav Volik[3], Sohrab Shah[2], Colin Collins[3,*] and S. Cenk Sahinalp[1,*]

[1]School of Computing Science, Simon Fraser University, 8888 University Way, Burnaby, BC V5A 1S6, [2]BC Cancer Agency, 600th Avenue West, Vancouver, BC V5Z 4E6 and [3]Vancouver Prostate Centre, 899 12th Avenue West, Vancouver, BC V5Z 1M9, Canada

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Comrad is a novel algorithmic framework for the integrated analysis of RNA-Seq and whole genome shotgun sequencing (WGSS) data for the purposes of discovering genomic rearrangements and aberrant transcripts. The Comrad framework leverages the advantages of both RNA-Seq and WGSS data, providing accurate classification of rearrangements as expressed or not expressed and accurate classification of the genomic or non-genomic origin of aberrant transcripts. A major benefit of Comrad is its ability to accurately identify aberrant transcripts and associated rearrangements using low coverage genome data. As a result, a Comrad analysis can be performed at a cost comparable to that of two RNA-Seq experiments, significantly lower than an analysis requiring high coverage genome data.

Results: We have applied Comrad to the discovery of gene fusions and read-throughs in prostate cancer cell line C4-2, a derivative of the LNCaP cell line with androgen-independent characteristics. As a proof of concept, we have rediscovered in the C4-2 data 4 of the 6 fusions previously identified in LNCaP. We also identified six novel fusion transcripts and associated genomic breakpoints, and verified their existence in LNCaP, suggesting that Comrad may be more sensitive than previous methods that have been applied to fusion discovery in LNCaP. We show that many of the gene fusions discovered using Comrad would be difficult to identify using currently available techniques.

Availability: A C++ and Perl implementation of the method demonstrated in this article is available at http://compbio.cs.sfu.ca/.

Contact: andrew.mcpherson@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received and revised on March 2, 2011; accepted on March 29, 2011

1 INTRODUCTION

High-throughput sequencing of cDNA (RNA-Seq) is rapidly accelerating our understanding of the sequence content of the human transcriptome. RNA-Seq can be used for high-throughput quantification of transcript abundance as has been done previously using microarrays. However, microarray-based approaches require pre-existing knowledge of the transcriptome sequence. RNA-Seq, in contrast, can be used for *de novo* characterization of the transcriptome including unbiased discovery and nucleotide level characterization of novel transcripts. Compared to previous, Sanger sequencing-based approaches for discovering novel transcripts, RNA-Seq is higher throughput for lower cost (Wang *et al.*, 2009).

Applied to cancer genomics, RNA-Seq can be employed for the discovery of novel aberrant transcripts with implications for cancer biology. Maher *et al.* (2009a, b) used RNA-Seq to rediscover known gene fusions in chronic myelogenous leukaemia (CML) and prostate cell lines, and also discovered novel gene fusions in prostate tumours. Similarly, Berger *et al.* (2010) applied RNA-Seq to the discovery of gene fusions in melanoma. Pflueger *et al.* (2011) used a newly developed method called FusionSeq (Sboner *et al.*, 2010) to discover gene fusions in RNA-Seq data from prostate tumours, and Hu *et al.* (2010) developed PERAlign and used it to discover gene fusions in breast cancer cell lines. The general methodology used by all of these studies was the 'paired end' method: (i) RNA-Seq is used to sequence both ends of a set of cDNA fragments; (ii) the resulting sequence pairs are aligned to the reference genome or transcriptome; (iii) a chimeric transcript will produce chimeric fragments and those chimeric fragments will produce a pair of sequences (paired end reads) that align to different genes, thus, any paired end read for which one end aligns to one gene and the other end aligns to another gene is considered potential evidence of a gene fusion.

Despite the aforementioned methodological advances and associated discoveries, accurate prediction of gene fusions from RNA-Seq data remains a difficult problem. The large amount of sequence data produced by high-throughput sequencing and the complexity of the transcriptome make RNA-Seq data difficult to interpret. High expression levels combined with sequencing errors and novel splicing produce many sequence pairs that appear to have been produced from chimeric fragments (Sboner *et al.*, 2010). Furthermore, the reverse transcription step used to produce cDNA has been shown to produce chimeric fragments via the process of template switching (Houseley and Tollervey, 2010). Previous studies dealt with false positives produced by 'false' chimeric fragments with the application of heuristic filters (Berger *et al.*, 2010; Maher *et al.*, 2009a, b; Pflueger *et al.*, 2011; Sboner *et al.*, 2010). Another common practise was to discard paired end reads with multiple

[*] To whom correspondence should be addressed.

[†] The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors.

mappings to the genome (multi-map reads) (Berger *et al.*, 2010; Maher *et al.*, 2009a, b; Pflueger *et al.*, 2011; Sboner *et al.*, 2010). Although discarding multi-map reads and applying heuristic filters may reduce the false positive rate, the effects of these practises on the false negative rate have not been properly quantified.

Assuming successful prediction of fusion transcripts from RNA-Seq data, the significance of those transcripts may still depend on the discovery of an underlying genomic rearrangement. Fusions between adjacent genes not associated with a genomic rearrangements are a distinct class of fusion transcript known as transcription-induced chimeras (TICs) or read-throughs. Though a significant amount of recent work has identified tissue and tumour-specific read-throughs (Brooks *et al.*, 2009; Kato *et al.*, 2003; Rickman *et al.*, 2009; Wang *et al.*, 2007), the mechanisms for their heritability remain unclear, and their ubiquity in normal tissue (Akiva *et al.*, 2006; Parra *et al.*, 2006) impedes assessment of their functional significance in cancer. RNA-Seq alone cannot distinguish a read-through from a small deletion that brings together two genes. Instead of RNA-Seq, some investigators have sought to discover gene fusions using whole genome shotgun sequencing (WGSS) (Bashir *et al.*, 2008; Pleasance *et al.*, 2010). However, results produced from WGSS data suffer from the reverse problem, that is, the significance of any fusion discovered using WGSS data is difficult to determine without an understanding of its effects on expression and without knowing whether it produces a fusion transcript. Furthermore, the unfocused nature of whole genome sequencing make this method expensive at coverage levels required to accurately predict genomic rearrangements.

A natural progression, given the complementarity of genomic and transcriptomic data, would be an analysis that combine these two data types. For example, the study by Berger *et al.* (2010) combined RNA-Seq data with copy number data to identify gene fusions associated with deletions and unbalanced rearrangements. Unfortunately, copy number data is ineffective for the discovery of balanced rearrangements such as the reciprocal translocation that creates the BCR–ABL gene fusion associated with CML (Rowley, 1973). In contrast, WGSS data can effectively discover balanced and unbalanced rearrangements. However, there are currently no methods that combine whole genome sequence data with RNA-Seq data for the purposes of gene fusion detection.

The availability of methods for predicting gene fusions in RNA-Seq data and rearrangements in WGSS data make it conceivable that these existing methods could be combined for a joint analysis of RNA-Seq and WGSS data for accurate gene fusion prediction. However, as we demonstrate in this article, applying each tool independently and then combining the results would be inaccurate. All methods for analysis of either RNA-Seq or WGSS data use heuristic filters to discard low confidence predictions supported by marginal amounts of evidence in order to attain a reasonable true positive rate. Thus, any true fusion supported by only a marginal amount of evidence in either one or both datasets will be missed by independent analysis. We find that a joint analysis produces a limited number of results supported by both datasets, partially obviating the need for thresholding when searching for aberrant transcripts associated with genomic rearrangements. Similarly, the actual mapping location of multi-map reads may be difficult to resolve with an independent analysis of each dataset, even when using methods that effectively leverage multi-map reads, such as MoDIL (Lee *et al.*, 2009) or VariationHunter (Hajirasouliha *et al.*, 2010; Hormozdiari *et al.*, 2009, 2010) for WGSS analysis, and deFuse (McPherson *et al.*, 2011) or ShortFuse (Kinsella *et al.*, 2011) for RNA-Seq analysis. We show that an analysis that simultaneously considers all reads from both datasets is better able to resolve the alignment location of multi-map reads.

2 APPROACH

Comrad is a novel algorithmic framework for the integrated analysis of RNA-Seq and WGSS data for the purposes of discovering genomic rearrangements and aberrant transcripts. Comrad builds on the COMMON-LAW framework first proposed in related work by Hormozdiari *et al.* (2011) on structural variation discovery in multiple sequenced genomes. The Comrad method leverages the advantages of both types of data, providing accurate classification of rearrangements as expressed or not expressed and accurate classification of the genomic or non-genomic origin of aberrant transcripts. A major benefit of Comrad is its ability to accurately predict fusion transcripts and their associated genome rearrangements using low coverage WGSS data. As a result, a Comrad analysis can be performed at a cost comparable to that of two RNA-Seq experiments, significantly lower than an analysis requiring high coverage genome data. The algorithmic basis of Comrad, provided in detail in this article, is an integer programming formulation which can be solved exactly using branch and bound or approximately using the relaxation of the linear program. For larger datasets, Comrad provides the option of using a greedy algorithm that can yield efficient solutions with reasonable running times.

We have applied Comrad to the discovery of gene fusions and read-throughs in prostate cancer cell line C4-2, a derivative of the LNCaP cell line with androgen-independent characteristics. As proof of concept, we have used Comrad to rediscover 4 out of 5 fusions previously described in LNCaP and known to also exist in C4-2. We have also used Comrad to identify six novel fusion transcripts and associated genomic rearrangements. A simple extension to the Comrad framework has allowed us to discover reciprocal rearrangement breakpoints for the two translocations found in the C4-2 data, making Comrad the first method to allow for the systematic discovery of reciprocal rearrangements. Furthermore, since Comrad is not biased towards canonical fusion splice junctions or fusions between known exons, we are able to use Comrad to discover fusions exhibiting non-canonical splicing. Some of the fusions we identify are supported by multi-map reads, showing that Comrad can effectively leverage multi-map reads for fusion discovery. Finally, some of the rearrangement breakpoints discovered by Comrad have as few as one read of supporting evidence, showing that Comrad is effective at discovering fusion evidence in low coverage genome data.

3 METHODS

The Comrad method begins by enumerating all rearrangement breakpoints implied by the WGSS reads and all gene fusion splices implied by the RNA-Seq reads. Some of these breakpoints and fusion splices will be supported by multi-map reads, but a read can only originate from at most one genomic location. Thus, we require a robust method for determining the most likely origin for each read given the greater context of the alignments of all WGSS and RNA-Seq reads. Most rearrangements and gene fusions are specific to individual cell lineages, i.e. they occur at low levels of recurrence (Mitelman *et al.*, 2007). Furthermore, since the RNA-Seq and WGSS data originate from the same sample, fundamental differences between the two datasets (differences that are not the result of expression or splicing) are unlikely. Thus, when seeking to determine the most likely origin for each read, we seek as the most parsimonious solution, a global assignment (of reads) that minimizes three types of differences: differences between the WGSS dataset and the reference genome (rearrangement breakpoints), differences between the RNA-Seq dataset and the reference transcriptome (fusion splices) and differences between RNA-Seq dataset and WGSS dataset.

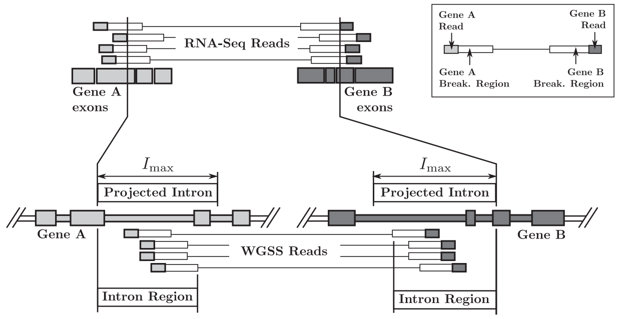

Fig. 1. Corroborating rearrangement breakpoints and fusion splices. Two conditions are required for RNA-Seq alignments and genome sequencing alignments to be considered evidence of the same rearrangement. The projected introns of the RNA-Seq alignments must overlap with the breakpoint regions of the genomic alignments, and the two intron regions must not overlap.

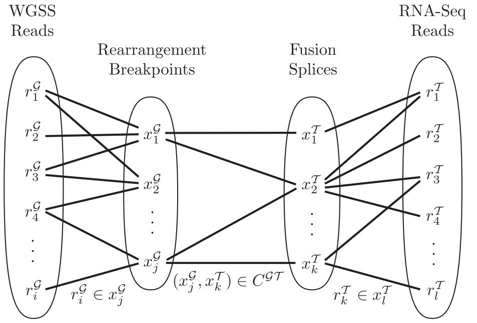

Fig. 2. Rearrangement support graph. The relationship between DNA reads, DNA clusters, RNA reads and RNA clusters can best be depicted using the rearrangement support graph.

3.1 Identifying potential rearrangement breakpoints and fusion splices

Analysis of the WGSS data begins by enumerating all rearrangement breakpoints implied by the WGSS reads, and forming clusters of WGSS reads that support each rearrangement breakpoint. WGSS reads are aligned to the genome (NCBI36). Concordantly, aligning reads are used to estimate the minimum and maximum DNA fragment length $L_{\min}$ and $L_{\max}$ (Berger *et al.*, 2010; Maher *et al.*, 2009a), and are subsequently discarded. All alignments of the remaining discordant reads are retained for further analysis. We then use an existing algorithm (Hormozdiari *et al.*, 2009; McPherson *et al.*, 2011) to identify all clusters of discordant alignments, where each cluster is a maximal set of reads that could be explained by a single pair of breakpoints.

Similar to the analysis of the WGSS data, analysis of the RNA-Seq data begins by enumerating all fusion splices implied by the RNA-Seq reads, and forming clusters of RNA-Seq reads that support each fusion splice. RNA-Seq reads are aligned to spliced transcript sequences and unspliced gene sequences as annotated by ensembl (version 54) (Bengtsson *et al.*, 2008). Aligning to all splice variants of each gene and also the unspliced gene sequence enables Comrad to handle alternative splicing in a natural way. Multiple alignments of RNA-Seq data will arise because of homology between genes and redundant inclusion of the same exon in multiple splice variants of the same gene. Selecting the most parsimonious set of unique alignments for RNA-Seq data, as described later, will select not only the most likely pair of genes involved in a fusion event, but also the most likely pair of splice variants for each gene. Maximal sets of discordant RNA-Seq alignments corroborating the same fusion splice are enumerated using analogous conditions and the same algorithm as described for WGSS alignments.

3.2 Corroborating rearrangement breakpoints and fusion splices

RNA-Seq and WGSS alignments corroborate the same rearrangement if there exists at least one plausible genomic breakpoint at each locus that could explain both the RNA-Seq and WGSS alignments. Splicing confounds this problem because it often results in RNA-Seq reads that align to the genome many kilobases from the corresponding rearrangement breakpoints. Thus, to accurately establish corroboration between RNA-Seq and WGSS data, the effects of splicing must be considered. We describe two conditions for corroboration that afford efficient computation and ensure the existence of rearrangement breakpoints that would explain the RNA-Seq and genome sequencing alignments.

From one end of a given set of RNA-Seq alignments, we define the *projected intron* as the I_{max} sized region starting at the most downstream genomic position of those alignments (Fig. 1). A set of RNA-Seq alignments is said to be corroborated by a set of WGSS alignments if the projected introns for the RNA-Seq alignments overlap with the breakpoint regions for the WGSS alignments. Thus, the *overlapping intron* condition is the condition that the pair of projected introns for RNA-Seq alignments must overlap with the pairs of breakpoint regions for the WGSS alignments.

We also define the *intron region* as the portion of the projected intron of a set of RNA-Seq alignments that is upstream from and including the breakpoint regions of a set of WGSS alignments (Fig. 1). The *non-conflicting intron* condition is the condition that the two intron regions for a potentially corroborating set of RNA-Seq and WGSS alignments cannot overlap. The *non-conflicting intron* condition disqualifies mutually exclusive sets of RNA-Seq and genome sequencing alignments that would otherwise satisfy the *overlapping intron* condition.

3.3 Selecting the most parsimonious set of alignments for ambiguously aligning reads

The relationship between WGSS reads, rearrangement breakpoints, RNA-Seq reads and fusion splices can be depicted in the *rearrangement support graph* as shown in Figure 2. The rearrangement support graph is formed by connecting WGSS reads to rearrangement breakpoints supported by those reads, and connecting RNA-Seq reads to fusion splices supported by those reads. Fusion splices and rearrangements breakpoints are connected based on the corroboration indicated by the overlapping intron condition and the non-conflicting intron condition.

The rearrangement support graph encodes the ambiguity of WGSS and RNA-Seq reads with multiple alignments. Since at most one alignment for each read is valid, we seek a transformation of the graph that removes edges such that the transformed graph has exactly one edge incident with each read. Given only RNA-Seq data, we previously attempted to produce a maximum parsimony solution with a minimum number of predicted fusions (McPherson *et al.*, 2011). More recently, we introduced combinatorial formulations to identify structural variation events in several donor *genomes* by means of minimizing a weighted sum of structural differences between the donor genomes as well as one reference genome (Hormozdiari *et al.*, 2011). Expanding on this principle, we now attempt to select a set of alignments

$\mathcal{M}$ so as to minimize the number of differences between the WGSS data, RNA-Seq data and the reference genome.

Let $X^{\mathcal{G}}$ be the set of all rearrangement breakpoints and let $X^{\mathcal{T}}$ be the set of all fusion splices. Let $C^{\mathcal{G}\mathcal{T}}$ be the set of all rearrangement breakpoint/fusion splice pairs $(x_j^{\mathcal{G}},x_k^{\mathcal{T}})\in X^{\mathcal{G}}\times X^{\mathcal{T}}$ that satisfy both conditions of corroboration. For each rearrangement breakpoint $x_j^{\mathcal{G}}\in X^{\mathcal{G}}$, let $\delta_j^{\mathcal{G}}\in\Delta^{\mathcal{G}}$ be a corresponding indicator variable, such that $\delta_j^{\mathcal{G}}=1$ if and only if at least one alignment has been selected that supports $x_j^{\mathcal{G}}$. In other words, $\delta_j^{\mathcal{G}}=1$ if and only if the corresponding rearrangement breakpoint vertex has at least one incident alignment edge in the transformed graph. Define fusion splice indicator variables, $\delta_k^{\mathcal{T}}\in\Delta^{\mathcal{T}}$, similarly.

We are now in a position to give precise definitions of the three types of differences being considered. For a given set of selected alignments, each $x_j^{\mathcal{G}}\in X^{\mathcal{G}}$ for which $\delta_j^{\mathcal{G}}=1$ implies a single difference between the WGSS data and the reference. Similarly, each $x_k^{\mathcal{T}}\in X^{\mathcal{T}}$ for which $\delta_k^{\mathcal{T}}=1$ implies a single difference between the RNA-Seq data and the reference. Fully enumerating the differences between the WGSS data and the RNA-Seq data would require assembly of at least one of these datasets, a method we do not consider here. Instead, we define a difference between the WGSS data and the RNA-Seq data as a difference found between one dataset and the reference that is not corroborated by a difference between the other dataset and the reference.

First define a corroboration indicator variable $\kappa_j^{\mathcal{G}}$ for each rearrangement breakpoint $x_j^{\mathcal{G}}$ that indicates whether a fusion splice has been selected that corroborates $x_j^{\mathcal{G}}$. More formally, $\kappa_j^{\mathcal{G}}=1$ if and only if there exists $(x_j^{\mathcal{G}},x_k^{\mathcal{T}})\in C^{\mathcal{G}\mathcal{T}}$ for which $\delta_k^{\mathcal{T}}=1$. Define a similar indicator variable $\kappa_k^{\mathcal{T}}$ for each fusion splice $x_k^{\mathcal{T}}$. A difference between the WGSS data and the reference not corroborated by a difference between the RNA-Seq data and the reference is given by $\delta_j^{\mathcal{G}}\cdot(1-\kappa_j^{\mathcal{G}})$. Conversely, an uncorroborated difference between the RNA-Seq data and the reference is given by $\delta_k^{\mathcal{T}}\cdot(1-\kappa_k^{\mathcal{T}})$.

Let $\mathcal{A}$ be the full set of alignments considered and let $\mathcal{M}\subseteq\mathcal{A}$ be any valid subset, that is, a set of alignments for which each read is aligned to exactly one mapping location. We seek to minimize the objective function $f(\mathcal{M})$ that calculates the total number of differences implied by $\mathcal{M}$ [Equation (1)].

$$\begin{aligned} f(M) &= \sum_j \delta_j^{\mathcal{G}}+\sum_k \delta_k^{\mathcal{T}}+\sum_j \delta_j^{\mathcal{G}}\cdot(1-\kappa_j^{\mathcal{G}})+\sum_k \delta_k^{\mathcal{T}}\cdot(1-\kappa_k^{\mathcal{T}}) \\ &= \sum_j \left[2\delta_j^{\mathcal{G}}-\delta_j^{\mathcal{G}}\cdot\kappa_j^{\mathcal{G}}\right]+\sum_k \left[2\delta_k^{\mathcal{T}}-\delta_k^{\mathcal{T}}\cdot\kappa_k^{\mathcal{T}}\right] \qquad (1)\end{aligned}$$

We seek an efficient algorithm that minimizes this objective function by reducing the rearrangement support graph. We propose two algorithms for solving this problem, an ILP formulation for use with an ILP solver and an approximation algorithm based on weighted set cover.

3.3.1 ILP formulation For each RNA-Seq read $r_i^{\mathcal{G}}\in x_j^{\mathcal{G}}$, a 0-1 integer variable $a_{ij}^{\mathcal{G}}$ indicates whether the edge between $r_i^{\mathcal{G}}$ and $x_j^{\mathcal{G}}$ is present in the transformed graph. Similarly, 0-1 integer variable $a_{lk}^{\mathcal{T}}$ represents whether the edge between $r_l^{\mathcal{T}}$ and $x_k^{\mathcal{T}}$ is present in the transformed graph. The ILP formulation for reducing the rearrangement support graph attempts to find a valid assignment for variables $a_{ij}^{\mathcal{G}}$ and $a_{lk}^{\mathcal{T}}$ that minimizes the objective function given in Equation (1). As described above, 0-1 integer variables $\delta_j^{\mathcal{G}}$ and $\delta_k^{\mathcal{T}}$ represent whether at least one read has been assigned to the respective rearrangement breakpoint $x_j^{\mathcal{G}}$ and fusion splice $x_k^{\mathcal{T}}$. Let $z_j^{\mathcal{G}}=\delta_j^{\mathcal{G}}\cdot\kappa_j^{\mathcal{G}}$ and $z_k^{\mathcal{T}}=\delta_k^{\mathcal{T}}\cdot\kappa_k^{\mathcal{T}}$ Also, let $c_{jk}=1$ if $(x_j^{\mathcal{G}},x_k^{\mathcal{T}})\in C^{\mathcal{G}\mathcal{T}}$, otherwise $c_{jk}=0$. The objective function for the ILP is given as in Equation (1), and the constraints are given below in Equations (2)–(9).

$$\forall r_i^{\mathcal{G}}\in R^{\mathcal{G}}\ :\ \sum_j a_{ij}^{\mathcal{G}}=1 \qquad (2)$$

$$\forall r_l^{\mathcal{T}}\in R^{\mathcal{T}}\ :\ \sum_k a_{lk}^{\mathcal{T}}=1 \qquad (3)$$

$$\forall r_i^{\mathcal{G}}\in R^{\mathcal{G}},\ x_j^{\mathcal{G}}\in X^{\mathcal{G}}\ :\ \delta_j^{\mathcal{G}}\geq a_{ij}^{\mathcal{G}} \qquad (4)$$

$$\forall r_l^{\mathcal{T}}\in R^{\mathcal{T}},\ x_k^{\mathcal{T}}\in X^{\mathcal{T}}\ :\ \delta_k^{\mathcal{T}}\geq a_{lk}^{\mathcal{T}} \qquad (5)$$

$$\forall j\ :\ z_j^{\mathcal{G}}\leq\delta_j^{\mathcal{G}} \qquad (6)$$

$$\forall k\ :\ z_k^{\mathcal{T}}\leq\delta_k^{\mathcal{T}} \qquad (7)$$

$$\forall j\ :\ z_j^{\mathcal{G}}\leq\sum_k c_{jk}\delta_k^{\mathcal{T}} \qquad (8)$$

$$\forall k\ :\ z_k^{\mathcal{T}}\leq\sum_j c_{jk}\delta_j^{\mathcal{G}} \qquad (9)$$

Constraints 2 and 3 ensure that each read is assigned to exactly one cluster in the transformed graph. Constraints 4 and 5 ensure that a cluster is considered as selected if at least one read has been assigned to that cluster. The constraint given by Equation (6) ensures that $z_j^{\mathcal{G}}=1$ only if $\delta_j^{\mathcal{G}}=1$, a consequence of defining $z_j^{\mathcal{G}}=\delta_j^{\mathcal{G}}\cdot\kappa_j^{\mathcal{G}}$. The constraint given by Equation (8) ensures that $z_j^{\mathcal{G}}=1$ only if there exists a k such that $\delta_k^{\mathcal{T}}=1$ and $c_{jk}=1$, that is, $z_j^{\mathcal{G}}=1$ for rearrangement breakpoint j only if a corroborating fusion splice has been selected. The constraints given by Equations (7) and (9) are analogous constraints for $z_k^{\mathcal{T}}$.

The above ILP formulation can be solved exactly using an ILP solver or can be solved approximately using randomized rounding applied to the LP relaxation of the problem. For the purposes of this study, we used the branch and bound-based exact ILP solver provided in the GLPK library (http://www.gnu.org/software/glpk/).

3.3.2 Greedy approximation algorithms The greedy approximation algorithm for reducing the rearrangement support graph can be used to provide an approximate solution to larger problems. The algorithm is based on a previous formulation for identifying structural variations in multiple genomes as proposed by Hormozdiari *et al.* (2011). However, we rework the cost function to allow a single rearrangement breakpoint to support multiple fusion splices and visa versa. For each corroborating rearrangement breakpoint/fusion splice pair $(x_j^{\mathcal{G}},x_k^{\mathcal{T}})\in C^{\mathcal{G}\mathcal{T}}$, form the read set $z=x_j^{\mathcal{G}}\cup x_k^{\mathcal{T}}$ and an associated indicator set $\Delta_z=\{\delta_j^{\mathcal{G}},\delta_k^{\mathcal{T}}\}$. For each uncorroborated rearrangement breakpoint $x_j^{\mathcal{G}}$ form the set $z=x_j^{\mathcal{G}}$ and an associated indicator set $\Delta_z=\{\delta_j^{\mathcal{G}}\}$. Form analogous sets for uncorroborated transcriptome clusters. Calculate the cost of each set z as given in Equation (10).

$$\text{cost}(z)\ =\ 2-\sum_{\delta_m\in\Delta_z}\delta_m \qquad (10)$$

Let U be the set of uncovered reads, initially empty. Also, all $\delta_j^{\mathcal{G}}$ and $\delta_k^{\mathcal{T}}$ are initially 0. At each step in the algorithm, select the set z_k that covers the largest number of reads for the lowest cost, that is, the set z_k that maximizes 11.

$$\frac{|z_k\setminus U|}{\text{cost}(z_k)} \qquad (11)$$

For each $\delta_m\in\Delta_{z_k}$, set $\delta_m=1$ if $|x_m\setminus U|>0$. That is, select cluster x_m by setting $\delta_m=1$ if x_m covers additional elements of U. Update all other Δ_z that contain δ_m and also update the cost of any set that may have changed due to changes in its associated Δ_z. For each read r in $z_k\setminus U$, remove all edges in the rearrangement support graph incident with r, retaining only the edge between r and the cluster used to create z_k. Add the reads in z_k to U, and repeat, selecting a new set z_{k+1} until U includes all WGSS and RNA-Seq reads. The greedy algorithm will provide a solution with cost at most $\log n\cdot\text{OPT}$, where OPT is the cost of the optimal solution and n is the total number of reads. For asymptotic analysis and proofs of complexity, please see Hormozdiari *et al.* (2011).

3.4 Modifying the breakpoint overlap function

One benefit of the given formulation of the problem is that it allows for the substitution of different rules for the corroborative relationship $(x_j^{\mathcal{G}},x_k^{\mathcal{T}})\in$

$C^{\mathcal{GT}}$. We explored one other possibility. Given a rearrangement breakpoint $x_j^{\mathcal{G}}$ and a fusion splice $x_k^{\mathcal{T}}$, we calculate the pair of genes potentially affected by the events represented by those clusters. We then define $C^{\mathcal{GT}}$ as $(x_j^{\mathcal{G}}, x_k^{\mathcal{T}}) \in C^{\mathcal{GT}} \iff \text{genepair}(x_j^{\mathcal{G}}) = \text{genepair}(x_k^{\mathcal{T}})$. We show in the Section 4 that this alternative corroborative relationship allows us to discover reciprocal translocations.

3.5 Assembling a prediction sequence

For each fusion splice and each rearrangement breakpoint, Comrad assembles a *prediction sequence*. Suppose a set of reads implies a fusion splice (or rearrangement breakpoint) between transcript *A* and *B* (or genomic loci *A* and *B*). Let $\{(s_i^A, e_i^A)\}$ and $\{(s_i^B, e_i^B)\}$ be the start and end positions for the alignments to *A* and *B*, respectively. Let S_A be the sequence in *A* in the range $[\min\{s_i^A\}, \max\{e_i^A\}]$, and let S_B be the sequence in *B* in the range $[\min\{s_i^B\}, \max\{e_i^B\}]$. If the alignments are to the + strand of gene *A* and the − strand of gene *B* (+− orientation), then the prediction sequence is $S_A \cdot S_B$. For −+, ++ and −− orientations, the predicted sequence is $\text{rc}(S_A) \cdot \text{rc}(S_B)$, $S_A \cdot \text{rc}(S_B)$ and $\text{rc}(S_A) \cdot S_B$, respectively, where rc() is reverse complementation.

3.6 Heuristic filtering

The heuristic filtering used by Comrad can be categorized as pre-filtering or post-filtering. Pre-filtering is applied before the application of the above algorithms, and is intended to remove reads that are unlikely to inform a gene fusion analysis. Post-filtering is applied to the results of the above algorithms in an attempt to remove predictions that are likely to be false positives or are unlikely to be novel.

3.6.1 Pre-filtering The pre-filtering used by Comrad involves aligning reads to a specific set of sequences using bowtie (Langmead *et al.*, 2009) and removing those reads from further consideration if their alignments satisfy a given criteria. RNA-Seq reads are aligned to the genome and UniGene clusters (Sayers *et al.*, 2011) and reads with concordant alignments are discarded. RNA-Seq data are often contaminated by a significant amount of ribosomal RNA (rRNA) (Sboner *et al.*, 2010). Thus, RNA-Seq reads are also aligned to ensembl annotated rRNA, and any read with one or both ends aligning to any rRNA is discarded. Comrad is not intended as a method for reconstructing immunoglobulin (IG) rearrangements. Thus, any RNA-Seq read that aligns with one end to one IG gene, and the other end to any other IG gene is discarded. Finally, Comrad discards RNA-Seq and WGSS reads for which each end aligns to a repeat region in the genome, and those repeat regions are of the same type.

3.6.2 False positive post-filtering False positive post-filtering attempts to remove predictions that are most likely to be false positives produced by spurious alignment artifacts. The *sequence concordance filter* aligns each prediction sequence to the appropriate reference sequences using blat (Kent, 2002). Fusion splices are aligned to spliced and unspliced gene sequences and rearrangement breakpoints are aligned to the genome. A prediction sequence is discarded if that sequence aligns to the reference with >80% identity. The *read concordance filter* uses blat to align all reads suggestive of an event to the appropriate reference (see above). A prediction is discarded if >10% of the supporting reads align concordantly to that reference. Genomic fusions require at least one supporting WGSS read and five supporting RNA-Seq reads to be considered in this study.

3.6.3 Novelty post-filtering Novelty post-filtering attempts to remove predictions that are unlikely to be novel. The sequence concordance filter is used to remove fusion splice predictions with significant alignments to ESTs [EST database retrieved from UCSC genome browser (Rhead *et al.*, 2010) November 26, 2010]. The EST island filter begins by aligning fusion splice prediction sequences to the genome using blat and allowing for a spliced alignment. Fusion splices are discarded if their prediction sequence aligns entirely within a region of the genome suggested as co-transcribed by clusters of overlapping spliced EST alignments (Rhead *et al.*, 2010).

4 RESULTS

We analyzed RNA-Seq and WGSS data produced from the C4-2 cell line, a derivative of the LNCaP prostate cell line. As a result of the close relationship between C4-2 and LNCaP, we hypothesized that fusions previously discovered in LNCaP would be useful as positive controls to be searched for in the C4-2 data. We also sought to use the C4-2 data as a proxy for discovering novel gene fusions in LNCaP. The WGSS and RNA-Seq data for C4-2 each consisted of 84 million 50 bp + 50 bp paired end reads. With an approximate fragment length of 500 bp, the WGSS data provide 7× physical coverage of a diploid human genome. Given that the LNCaP genome is tetraploid (Beheshti *et al.*, 2000), the physical coverage for C4-2 is more likely closer to 3.5×.

Previous analysis of LNCaP resulted in the discovery of six fusion transcripts (Maher *et al.*, 2009a, b), DLEU2-PSPC1, RERE-PIK3CD, MIPOL1-DGKB, MRPS10-HPR, C19orf25-APC2 and SLC45A3-ELK4. We used PCR to confirm five of these fusion transcripts as present in C4-2; DLEU2-PSPC1 could not be confirmed in C4-2. The five confirmed fusion transcripts serve as positive control fusion transcripts to be discovered by Comrad in the C4-2 data. Both C19orf25-APC2 and SLC45A3-ELK4 involve adjacent genes, and are thus potential read-through events, a possibility that was confirmed for SLC45A3-ELK4 in a more recent study (Rickman *et al.*, 2009). Comrad found only RNA-Seq evidence of SLC45A3-ELK4 in C4-2, providing further evidence that SLC45A3-ELK4 is not associated with chromosomal rearrangement. No RNA-Seq or WGSS evidence was found for C19orf25-APC2. A targeted search did not identify any RNA-Seq reads supporting a C19orf25-APC2 fusion transcript, suggesting that C19orf25-APC2 expression is lower than that required for detection at the sequencing depth provided by the C4-2 RNA-Seq data. The remaining three fusions, RERE-PIK3CD, MIPOL1-DGKB and MRPS10-HPR involve distant genes and are thus potentially caused by underlying genomic rearrangement. Comrad successfully identified the previously described (Tomlins *et al.*, 2007) rearrangement breakpoint for MIPOL1-DGKB and also identified rearrangement breakpoints for RERE-PIK3CD and MRPS10-HPR. The novel rearrangement breakpoints for RERE-PIK3CD and MRPS10-HPR were confirmed by PCR for both C4-2 and LNCaP.

Comrad predicted an additional 10 novel genomic fusions for C4-2 (Supplementary Table S1). We attempted to validate 9 of the 10 predictions, excluding AMACR-GUSBL1 as a likely read-through associated with a small 600 bp deletion. Of the nine novel Comrad predictions, six were validated in C4-2 by PCR and sanger sequencing. We also validated all six novel Comrad predictions in LNCaP, thereby showing that Comrad is more sensitive than previous methods that have been applied to fusion discovery in LNCaP. Evidence for the six previously identified fusion transcripts and the six novel Comrad predictions is shown in Table 1.

To identify potential false negatives, we also analyzed the C4-2 RNA-Seq data using deFuse, a method for identifying fusion transcripts in RNA-Seq data alone. The deFuse analysis produced 31 predictions, 8 of which are predicted to be interchromosomal or long-range intrachromosomal fusions (Supplementary Table S4).

Table 1. Known and novel fusions predicted by Comrad in C4-2 and validated in both LNCaP and C4-2

5′ gene	3′ gene	Event	Evidence	Reads	Multi-map
MIPOL1	DGKB	Transcript	RNA-Seq	30	0
		Translocation	WGSS	1	0
		reciprocal	WGSS	1	0
RERE	PIK3CD	Transcript	RNA-Seq	35	0
		Transcript	RNA-Seq	11	0
		Transcript	RNA-Seq	6	0
		Transcript	RNA-Seq	11	0
		inversion	WGSS	1	0
MRPS10	HPR	Transcript	RNA-Seq	67	67
		Translocation	WGSS	13	12
		Reciprocal	WGSS	5	4
DLEU2	PSPC1	Transcript	RNA-Seq	0	0
		Deletion	WGSS	0	0
SLC45A3	ELK4	Transcript	RNA-Seq	18	0
		Transcript	RNA-Seq	15	0
C19orf25	APC2	Transcript	RNA-Seq	0	0
TFDP1	GRK1	Transcript	RNA-Seq	10	10
		Deletion	WGSS	7	4
FAM117B	BMPR2	Transcript	RNA-Seq	12	0
		eversion	WGSS	2	0
ITPKC	PPFIA3	Transcript	RNA-Seq	6	0
		Deletion	WGSS	13	0
CCDC43	YBX2	Transcript	RNA-Seq	9	0
		Deletion	WGSS	8	1
GPS2	MPP2	Transcript	RNA-Seq	23	1
		eversion	WGSS	10	0
FAM190B	CYP2C19	Transcript	RNA-Seq	20	14
		Deletion	WGSS	2	0

The number of reads supporting each event is provided, in addition to how many of those reads multi-map to the genome. The first six fusions have been previously described.

Genomic evidence was identified by Comrad for 7 of these 8 fusions; genomic evidence for the singe remaining fusion transcript, ZDHHC20-TNFRSF19, could not be identified by Comrad. The predicted ZDHHC20-TNFRSF19 sequence exhibits canonical GT-AG splicing at the fusion boundary and is thus not likely to be the product of template switching during reverse transcriptase (Houseley and Tollervey, 2010). The lack of genomic evidence for ZDHHC20-TNFRSF19 makes this fusion a candidate trans-splicing event in C4-2 (Li *et al.*, 2008). Alternatively, ZDHHC20-TNFRSF19 could represent a false negative for Comrad.

4.1 Accurate discovery of gene fusions

Analysis of WGSS data for evidence of genomic rearrangement is made difficult by the repetitive nature of the genome, and the large amount of coverage required to reliably predict rearrangements (Chen *et al.*, 2009; Hormozdiari *et al.*, 2009). Conversely, RNA-Seq produces many spurious chimeric reads by at least two mechanisms: template switching during reverse transcriptase (Houseley and Tollervey, 2010) and the combined effect of read errors and high gene expression (Sboner *et al.*, 2010). Given RNA-Seq and WGSS data from the same sample, an integrated analysis using Comrad provides the ability to more accurately resolve multi-map reads, and more confidently identify real events even where those events have relatively little evidence.

Comrad accurately identifies WGSS evidence for gene fusions, even when the evidence consists of only a small number of reads. Five of the validated genomic breakpoints are supported by two or less WGSS reads (Table 1). The breakpoint for DGKB-MIPOL1, arguably the most biologically important fusion in the dataset, is supported by one read. An independent analysis of WGSS data using a threshold of one read would result in the prediction of 20 675 fusions between the genes considered in this study. Considering only uniquely aligned reads results in the prediction of 9949 fusions supported by at least one WGSS read.

The existence of 9949 fusions would indicate that C4-2 is very highly rearranged, especially since these 9949 fusions represent rearrangements involving genic regions only. However, aCGH data does not indicate this level of genomic rearrangement as only 41 copy number change points are predicted across the genome (Supplementary Table S5). Additionally, previous spectral karyotype results identified only nine structural aberrations (per diploid cell) for LNCaP (Beheshti *et al.*, 2000). Thus, it is likely that many of the 9949 fusions are either false positives or represent transposable elements as opposed to large-scale structural aberrations. Clearly, identification of true positives from this set of fusions would be difficult if not impossible, rendering a single discordant WGSS read unreliable in an independent analysis of WGSS data. However, that single discordant WGSS read could be used to identify an important rearrangement breakpoint when considered in conjunction with RNA-Seq data as is done by Comrad.

Comrad provides the ability to accurately identify fusions where evidence for those fusions does not map uniquely to the genome. For the 9 PCR confirmed genomic fusions, 38% of the RNA-Seq reads and 32% of the WGSS reads are multi-map reads. Removing these reads from the analysis would prevent the identification of the MRPS10-HPR and TFDP1-GRK1 fusion transcripts and would hinder our ability to properly identify the CYP2C19-FAM190B fusion transcript (14/20 multi-map reads) and both the forward and reciprocal rearrangement breakpoints for MRPS10-HPR (12/13 and 4/5 multi-map reads, respectively). The MRPS10-HPR and TFDP1-GRK1 fusion transcripts are supported entirely by multi-map RNA-Seq reads. Intron 2 of WDR32 contains a region of high sequence similarity to parts of MRPS10. As a result, the RNA-Seq evidence for MRPS10-HPR aligns to either a hypothetical WDR32-HPR fusion transcript or the MRPS10-HPR fusion transcript. Similarly, RNA-Seq evidence for TFDP1-GRK1 also supports a hypothetical BX842568-GRK1 fusion transcript. TFDP1 and BX842568 have 96% sequence similarity over a region of 1234 bp, making the two possibilities, TFDP1-GRK1 or BX842568-GRK1, equally likely without knowledge of the corroborating WGSS data identified by Comrad. Therefore, even an analysis of the RNA-Seq data using a method that considers multi-map reads could fail to correctly identify the MRPS10-HPR and TFDP1-GRK1 fusion transcripts. However, Comrad is able to resolve the correct alignment location of multi-map reads that support the MRPS10-HPR and TFDP1-GRK1 fusion transcripts by leveraging the relatively unambiguous WGSS evidence of associated rearrangement breakpoints.

4.2 MIPOL1-DGKB and MRPS10-HPR are caused by reciprocal exchanges in C4-2 and LNCaP

In order to search for multiple genomic breakpoints associated with gene fusions, we altered the breakpoint overlap function as described

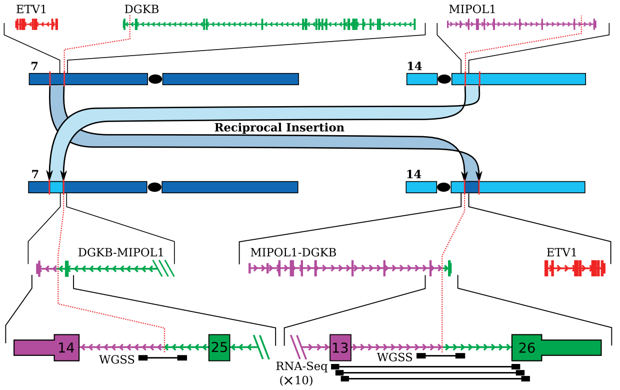

Fig. 3. Evidence for MIPOL1-DGKB as a reciprocal insertion. Insertion of the ETV1 locus into chromosome 14 is supported by 1 WGSS read and 30 RNA-Seq reads. The reciprocal insertion of the MIPOL1 locus into chromosome 7 is supported by 1 WGSS reads.

in Section 3.4. The MIPOL1-DGKB and MRPS10-HPR fusions were both found to have reciprocal rearrangement breakpoints in addition to the regular breakpoints that produce the MIPOL1-DGKB and MRPS10-HPR fusion transcripts. The reciprocal breakpoints were each positioned no further than 10 kb from the regular breakpoints, and were oriented so as to create reciprocal DGKB-MIPOL1 and HPR-MRPS10 fusion genes. Comrad did not detect a fusion transcript for either of the reciprocal DGKB-MIPOL1 and HPR-MRPS10 fusion genes. The presence of the reciprocal rearrangement breakpoints was confirmed by PCR for both C4-2 and LNCaP.

The regular and reciprocal breakpoints involving the MRPS10 gene at p21.1 on chromosome 6 and the HPR at q22.3 on chromosome 16 are almost certain to represent the t(6;16)(p21.1;q22) reciprocal translocation previously identified by spectral karyotyping of LNCaP (Beheshti *et al.*, 2000). The same spectral karyotype for LNCaP does not identify a reciprocal translocation between chromosomes 7 and 14 that would be necessary to explain the regular and reciprocal breakpoints identified for DGKB-MIPOL1. Extensive fluorescent *in situ* hybridization (FISH) experiments performed by Tomlins *et al.* (2007) also rule out the possibility of 7-14 reciprocal translocation, and instead Tomlins *et al.* hypothesize that the DGKB-MIPOL1 fusion is the result of an insertion. The new reciprocal breakpoint evidence identified by Comrad and validated by PCR strongly indicate that the DGKB-MIPOL1 fusion is not the result of a simple insertion. Given the previous spectral karyotype and FISH evidence, and the newly identified reciprocal breakpoint, a more likely hypothesis is that the DGKB-MIPOL1 fusion is caused by an underlying reciprocal insertion, by which genomic DNA is exchanged between chromosome 7 and chromosome 14 to produce DGKB-MIPOL1 and the reciprocal MIPOL1-DGKB (Fig. 3).

4.3 Genomic rearrangements create fusion transcripts with non-canonical splicing

The Comrad predictions include three validated fusion transcripts with non-canonical splicing in C4-2. The three transcripts predicted for the PIK3CD-RERE fusion include one for which an intronic region of RERE is not spliced out of the resulting transcript. The MRPS10-HPR fusion activates a cryptic splice site in the 3′ UTR of MRPS10 to produce a fusion transcript with non-canonical splicing. Finally, the CCDC43-YBX2 fusion transcript includes the full sequence of the genomic breakpoint, as no splicing occurs across the breakpoint. By analyzing the exon and intron expression obtained from RNA-Seq alignments, it is apparent that a significant proportion of CCDC43-YBX2 expression includes the first half of intron 3 of CCDC43 and the last half of intron 4 of YBX2 (Fig. 4). The intron retention found for CCDC43-YBX2 likely results because intron 3 of CCDC43 is an AT-AC intron (for both splice variants), whereas intron 4 of YBX2 is a GT-AG intron. The resulting fused intron, with an AT at the 5′ splice site and an AG at the 3′ splice site, is unlikely to be recognized by either the U2- or U12-dependent spliceosomes (Tarn and Steitz, 1996).

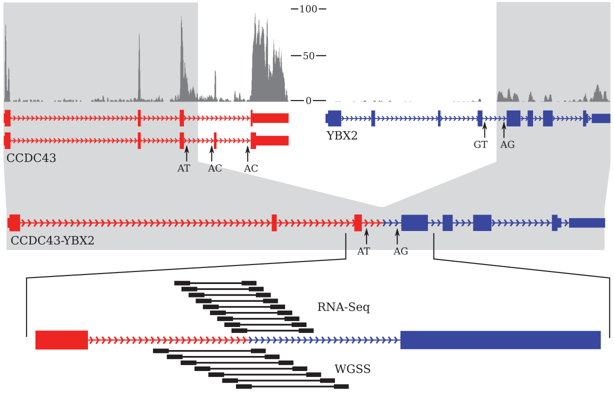

Fig. 4. Gene fusion CCDC43-YBX2 produces fusion transcripts with non-canonical splicing. Exon and intron expression estimated from RNA-Seq alignments is shown in dark grey above the CCDC43 and YBX2 gene models at the top. Sequences are shown for the 5′ and 3′ splice sites of introns involved in the fusion, in addition to the splice site sequences of the aberrant AT-AG intron of CCDC43-YBX2. RNA-Seq and WGSS reads supporting the fusion are shown at the bottom of the figure.

5 DISCUSSION

Comrad provides the first accurate computational method for simultaneous analysis of RNA-Seq and low coverage WGSS data for the purposes of identifying fused genes, and for differentiating fusions of genomic origin from those fusions that are non-genomic, i.e. co-transcription of adjacent genes or trans-splicing events. We have used the C4-2 data and a theoretical analysis to show that Comrad is able to discover fusions that other methods would not be capable of discovering given the same data. The advantages of Comrad are 2-fold. First, Comrad is able to leverage unambiguous WGSS data in order to correctly identify a fusion transcript supported by multi-map RNA-Seq data, and visa versa. Second, Comrad is able to accurately identify genomic rearrangements that result in gene fusions, even if that genomic rearrangement is supported by only one WGSS read. This second advantage means that genomic rearrangements producing gene fusions can be accurately identified in low coverage genome data, i.e. a Comrad analysis can be performed for roughly twice the cost of an RNA-Seq experiment.

As a proof of concept, we have shown that we are able to rediscover, in the C4-2 data, 4 of 6 fusions previously identified in the closely related cell line LNCaP. We then successfully validated the fusion transcripts and rearrangement breakpoints for 6 of the

10 novel genomic fusions nominated by Comrad. All six fusions were confirmed by PCR for both C4-2 and LNCaP, showing that Comrad is more sensitive to genomic fusions than previous methods that have been applied to LNCaP. Additionally, we used deFuse to identify fusion transcripts for which no genomic rearrangement is detected by Comrad, despite the fact that a genomic rearrangement is expected given the position and orientation of the genes involved. We identified and validated ZDHHC20-TNFRSF19, possibly a trans-splicing event, or possibly the product of a genomic inversion and thus a false negative for Comrad in the C4-2 data. The fact that only one high confidence fusion transcript could be identified as a potential false negative for Comrad implies that Comrad provides a sensitive method for the detection of fusion transcripts with associated rearrangement breakpoints.

Finally, we have used Comrad to gain new insight into the biology of rearrangement fusions. We have identified instances of non-canonical splicing in fusion transcripts produced by genomic rearrangements, including activation of cryptic splice sites, and intron retention due to incompatibility between the 5′ and 3′ splice sites of the rearrangement-induced intron. We have also used a simple modification of the Comrad framework to identify reciprocal evidence of rearrangement breakpoints. The modified framework led to the discovery of reciprocal evidence for both interchromosomal fusions identified by Comrad. For one of these fusions, MRPS10-HPR, the location of the reciprocal translocation coincides precisely with a translocation found in LNCaP by spectral karyotyping (Beheshti *et al.*, 2000). The other fusion, DGKB-MIPOL1, a reciprocal was previously thought to be a result of an insertion (Tomlins *et al.*, 2007); however, Comrad has provided evidence that this fusion could be the result of a reciprocal insertion.

RNA-Seq has already proven to be a powerful tool for the discovery of aberrant transcripts, and WGSS has already proven its utility when searching for rearrangements. We find that an integrated analysis of RNA-Seq and WGSS data minimizes some of the limitations of analyzing either RNA-Seq or WGSS data alone, and yields greater insight than either of these data types can provide independently. Given the possible existence of rearrangement fusions occurring at low levels of recurrence in cancer, the increased accuracy and decreased cost associated with a Comrad analysis may be useful when searching for these rearrangement fusions in a large number of tumours. The identification of such rearrangement fusions would then hopefully assist in the classification of molecular subtypes and the development of targeted therapies.

Funding: This project was funded by the Natural Sciences and Engineering Research Council (NSERC), the Michael Smith Foundation for Health Research (MSFHR), the British Columbia Cancer Research Centre (BCCRC), and the Canadian Institutes of Health Research (CIHR) Bioinformatics training program. Additional funding was provided by the Bioinformatics for Combating Infectious Diseases (BCID) project funded by the Simon Fraser University (SFU) Community Trust Endowment Fund (CTEF).

Conflict of Interest: none declared.

REFERENCES

Akiva,P. *et al.* (2006) Transcription-mediated gene fusion in the human genome. *Genome Res.*, **16**, 30–36.

Bashir,A. *et al.* (2008) Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput. Biol.*, **4**, e1000051.

Beheshti,B. *et al.* (2000) Identification of a high frequency of chromosomal rearrangements in the centromeric regions of prostate cancer cell lines by sequential giemsa banding and spectral karyotyping. *Mol. Diagn.*, **5**, 23–32.

Bengtsson,H. *et al.* (2008) Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, **24**, 759–767.

Berger,M.F. *et al.* (2010) Integrative analysis of the melanoma transcriptome. *Genome Res.*, **20**, 413–427.

Brooks,Y.S. *et al.* (2009) Functional pre- mRNA trans-splicing of coactivator coaa and corepressor RMB4 during stem/progenitor cell differentiation. *J. Biol. Chem.*, **284**, 18033–18046.

Chen,K. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.

Hajirasouliha,I. *et al.* (2010) Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics*, **26**, 1277–1283.

Hormozdiari,F. *et al.* (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.*, **19**, 1270–1278.

Hormozdiari,F. *et al.* (2010) Next-generation variationhunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, **26**, 350–357.

Hormozdiari,F. *et al.* (2011) Simultaneous structural variation discovery in multiple paired-end sequenced genomes. In Bafna,V. and Sahinalp,S. (eds), *Research in Computational Molecular Biology*, Vol. 6577 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, pp. 104–105.

Houseley,J. and Tollervey,D. (2010) Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PLoS One*, **5**, e12271.

Hu,Y. *et al.* (2010) A probabilistic framework for aligning paired-end rna-seq data. *Bioinformatics*, **26**, 1950–1957.

Kato,M. *et al.* (2003) Hodgkin's lymphoma cell lines express a fusion protein encoded by intergenically spliced mRNA for the multilectin receptor dec-205 (cd205) and a novel c-type lectin receptor dcl-1. *J. Biol. Chem.*, **278**, 34035–34041.

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

Kinsella,M. *et al.* (2011) Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. *Bioinformatics*, **27**, 1068–1075.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Lee,S. *et al.* (2009) Modil: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods*, **6**, 473–474.

Li,H. *et al.* (2008) A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science*, **321**, 1357–1361.

Maher,C.A. *et al.* (2009a) Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl Acad. Sci. USA*, **106**, 12353–12358.

Maher,C.A. *et al.* (2009b) Transcriptome sequencing to detect gene fusions in cancer. *Nature*, **458**, 97–101.

McPherson,A. *et al.* (2011) defuse: an algorithm for gene fusion discovery in tumor rna-seq data. *PLoS Comput. Biol.*, in press.

Mitelman,F. *et al.* (2007) The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer*, **7**, 233–245.

Parra,G. *et al.* (2006) Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res.*, **16**, 37–44.

Pflueger,D. *et al.* (2011) Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. *Genome Res.*, **21**, 56–67.

Pleasance,E.D. *et al.* (2010) A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, **463**, 184–190.

Rhead,B. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, 613–619.

Rickman,D.S. *et al.* (2009) Slc45a3-elk4 is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer. *Cancer Res.*, **69**, 2734–2738.

Rowley,J.D. (1973) Letter: a new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. *Nature*, **243**, 290–293.

Sayers,E.W. *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, 38–51.

Sboner,A. *et al.* (2010) FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biol.*, **11**, R104.

Tarn,W.Y. and Steitz,J.A. (1996) A novel spliceosome containing U11, U12, and U5 snrnps excises a minor class (AT-AC) intron in vitro. *Cell*, **84**, 801–811.

Tomlins,S.A. *et al.* (2007) Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature*, **448**, 595–599.

Wang,K. *et al.* (2007) RBM6-RBM5 transcription-induced chimeras are differentially expressed in tumours. *BMC Genomics*, **8**, 348–348.

Wang,Z. *et al.* (2009) RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.