Comrad: a novel algorithmic framework for the integrated analysis of RNA-Seq and WGSS data

Andrew McPherson [1,2,4], **Chunxiao Wu** [3,4], **Iman Hajirasouliha** [1,4], **Fereydoun Hormozdiari** [1,4], **Faraz Hach** [1], **Anna Lapuk** [3], **Stanislav Volik** [3], **Sohrab Shah** [2], **Colin Collins** [3,4], **S. Cenk Sahinalp** [1,4]

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

Motivation:
Both paired end whole transcriptome shotgun sequencing (RNA-Seq), and paired end Whole Genome Shotgun Sequencing (WGSS), have been used to discover rearrangements in tumour genomes. However, to date no method exists that leverages both RNA-Seq and WGSS data for accurate discovery of rearrangements and their associated fusion transcripts.

Method:
We present Comrad, a novel algorithmic framework for the integrated analysis of RNA-Seq and 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 6 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. In addition, we identify novel evidence of the reciprocal nature of previously described translocations, and also identify 3 cases of non-canonical splicing induced by a rearrangement, including a novel mechanism for intron retention in a fusion transcript.

Availability:
A C++ and Perl implementation is available at http://compbio.cs.sfu.ca/

Contact:
andrew.mcpherson@gmail.com

Acknowledgement
This research was funded in part by NSERC and SFU CTEF supported BCID project as well as the MSFHR and CRC programs.