sv-channels: structural variant detection using deep learning

UMC Utrecht

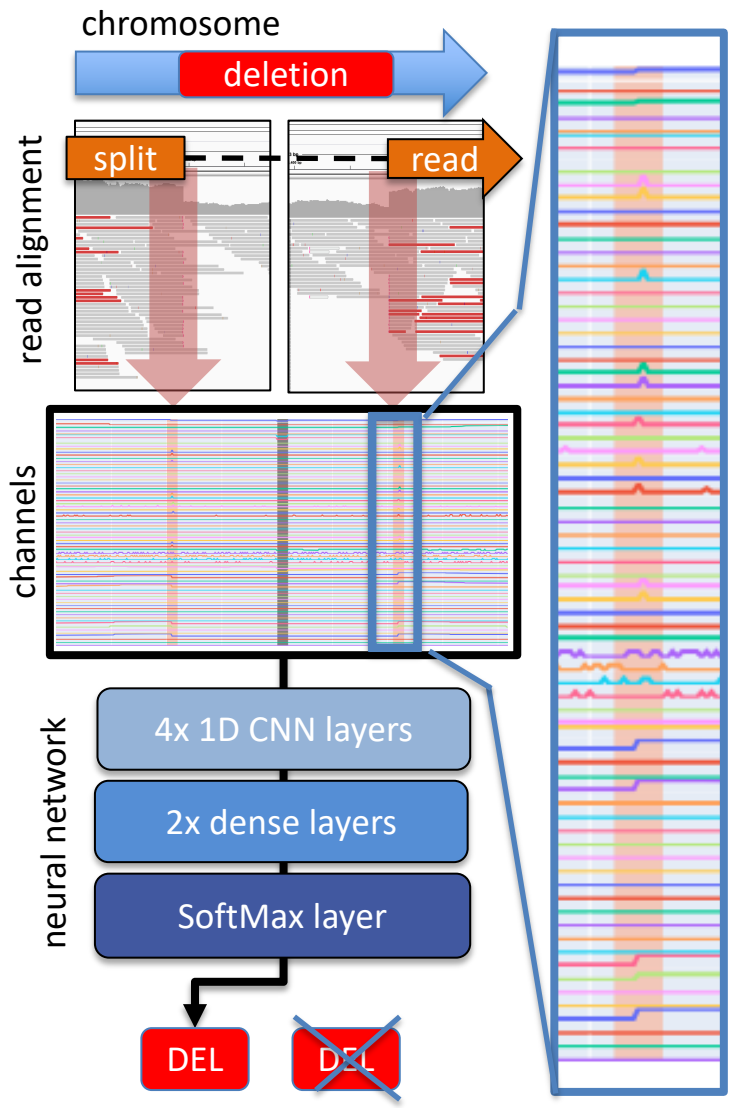
Luca Santuari[1], Sonja Georgievska[2], Arnold Kuzniar[2], Carl Shneider[1], Sarah Mehrem[1], Tilman Schaefers[1], Wigard Kloosterman[1] and Jeroen de Ridder[1]

1 Center for Molecular Medicine, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, Netherlands
2 Netherlands eScience Center, Science Park 140, 1098 XG Amsterdam, Netherlands

Introduction

Structural variants (SV) are large (> 50 bp) chromosomal rearrangements that have been implicated in many genetic diseases including cancer [1]. Current heuristics-based algorithms for SV detection (callers [2]) cannot capture the full set of SVs present in a human genome because signals at SV locations (breakpoints) closely resemble sequence and mapping artifacts. Deep Learning (DL) represents a promising methodology for SV detection, where relevant signals necessary to locate SVs can be learned automatically from the data instead of being hard coded. DL-based SV callers are currently limited to deletions and rely on complex architectures that are time-consuming to train. Here we present *sv-channels*, a DL approach that uses 1D Convolutional Neural Networks (CNN) to detect SVs of all the five major types.

1) Li *et al*., 2020, Nature
2) sv-callers, doi:10.5281/zenodo.1217111
3) Cai *et al.*, 2019, BMC Bioinformatics
4) Chowdury *et al.*, 2020, bioRxiv
5) sv-channels on GitHub
6) sv-gen, doi:10.5281/zenodo.3725663
7) Test data, doi:10.5281/zenodo.2663307

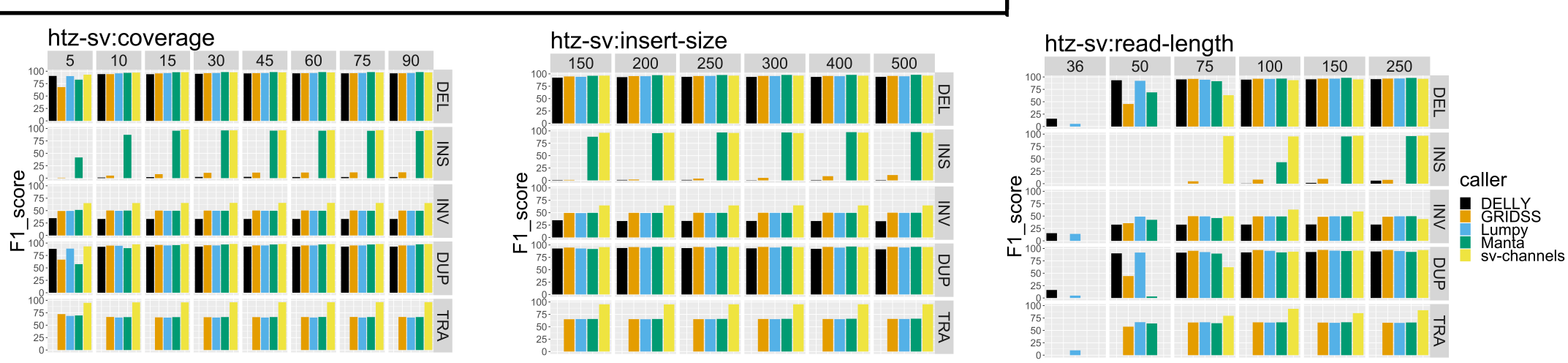


sv-channels workflow

SV signals are extracted from the reads aligned in 200 bp intervals (windows) centered at split read positions and converted into channels (one-dimensional vector representations). Channels (80 in total) include information on read and reference sequence properties, such as whether a read at a certain position are clipped or split on the left or on the right side, its orientation, its mapping quality, etc. Window pairs are labelled using ground truth SVs and used to train a CNN to classify pair of genomic positions as either SV breakpoints or not.

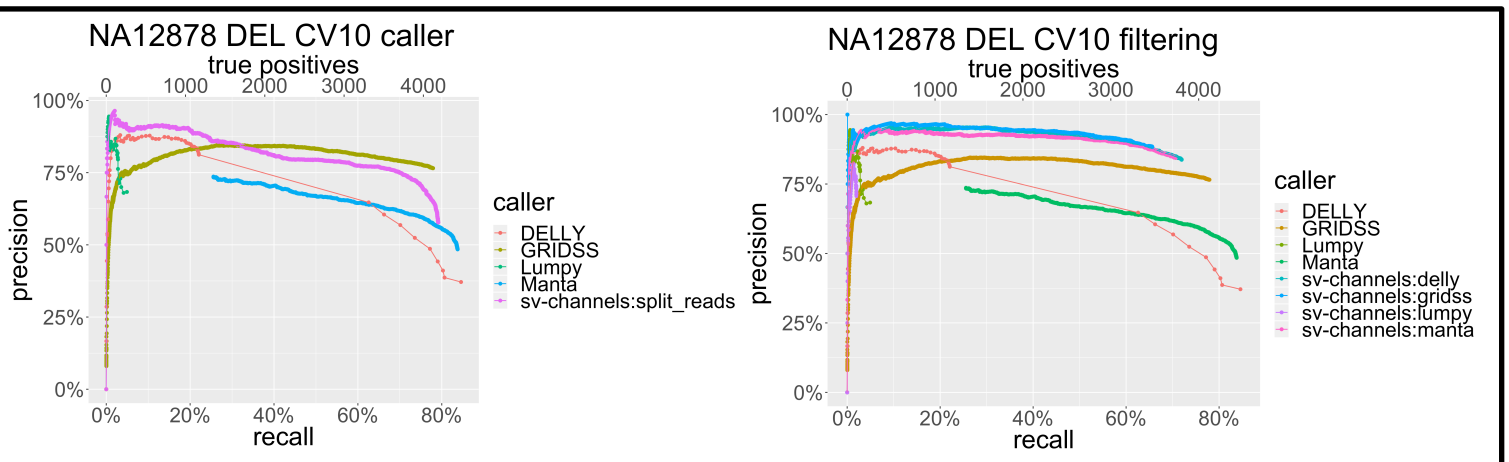
Simulated data

Using *sv-gen* workflow [6] we simulated read alignment data from chromosome 10 and 12 (hs37d5) where heterozygous SVs were inserted at known positions. Barcharts (below) show sv-channels performance (F1-score) at varying read coverage (left), insert size (center) and read length (right) compared to four state-of-the-art SV callers (GRIDSS, Manta, DELLY and Lumpy) that were run using the *sv-callers* workflow [2]. sv-channels is able to detect all SV types: deletions (DEL), insertions (INS), inversions (INV), tandem duplications (DUP) and inter-chromosomal translocations (TRA). sv-channels performance is either comparable to or higher than the other methods. An exception is at read lengths lower than 75 bp, when reads are too short to generate enough candidate split read positions for sv-channels.



Real data

We tested *sv-channels* using 10-fold cross validation on the germline deletions of the GiaB sample NA12878 [7]. As ground truth we used the sv-classify callset [7]. Left: *sv-channels* perfomance in **SV caller mode**, when split read positions are used as candidate positions. Right: *sv-channels* used in **filtering mode** to remove false positives from an SV callsets. In this case cancidate positions used in sv-channels are SV positions called by one of the other SV callers (DELLY, GRIDSS, Lumpy and Manta).



Precision and recall curves were obtained by varying the SV quality as threshold (as in Cameron et al., 2019, Nature Comm.). For *sv-channels*, the posterior probability of the DEL class was considered as SV quality score.

University Medical Center Utrecht
Oncode Institute
eScience center
L.Santuari@umcutrecht.nl