SAHLGRENSKA ACADEMY
CORE FACILITIES

UNIVERSITY OF GOTHENBURG

PΨfinder: Identification of novel PΨ in DNA sequencing data

Sanna Abrahamsson[1], Anna Rohlin[2] and Marcela Dávila López[1]

[1]Bioinformatics Core Facility, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

[2]Department of Molecular and Clinical Genetics, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

Introduction

Pseudogenes are structures in the genome that have emerged from a parent gene and in most cases lost the ability to produce functional proteins. There are three types of pseudogenes[1]:

- Processed pseduogenes, (Derived from retrotransposition of mRNA), *see fig. 1*.
- Unprocessed pseudogenes (Derived from gene duplication)
- Unitary pseudogenes (Accumulated mutations in the parent gene)

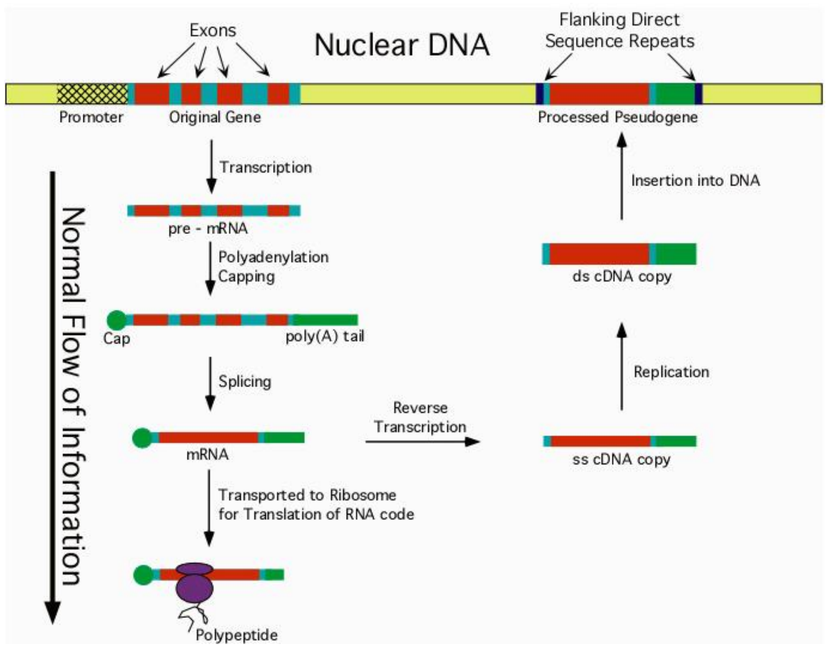


Figure 1: Formation of a processed pseudogenes: The mRNA is reverse transcribed and reinserted into the genome. *(Adapted from slide share)*

Formation of processed pseudogenes has been linked to a new class of mutations that occurs during cancer development. Additionally pseudogenes are important keys in evolutionary models.

Some approaches to detect pseudogenes involves primarily analysis of high throughput sequencing data and locus specific transcription evidence that involves manual curation[2].

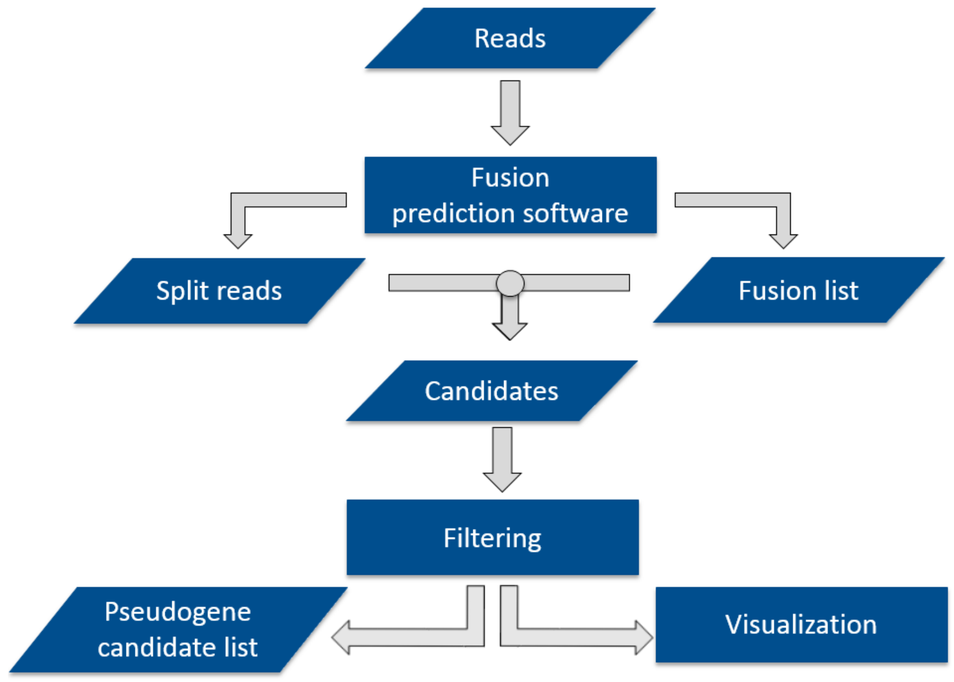
Here we aim to develop an automatic pipeline that use DNA paired-end sequencing data for identifying processed pseudogenes in the genome.

Software

The pipeline makes use of a fusion prediction software where the obtained fusions are linked to reads that are split across exons. The resulting candidates are filtered based on coverage and transcript stuctures. See workflow in *fig. 2.*

The output from the pipeline encompasses a list of pseudogene candidates together with visualization plots using circos[3], see *fig. 3.*

Figure 2: Workflow for the identification of processed pseudogenes using paired-end DNA sequencing data



Results

We have screened 120 NGS samples from hereditary colorectal cancer. We have identified several new processed pseudogenes which are under experimental validation. As an example we present the insertion of S*MAD4* processed pseudogene in the gene *SCAI*, see below.

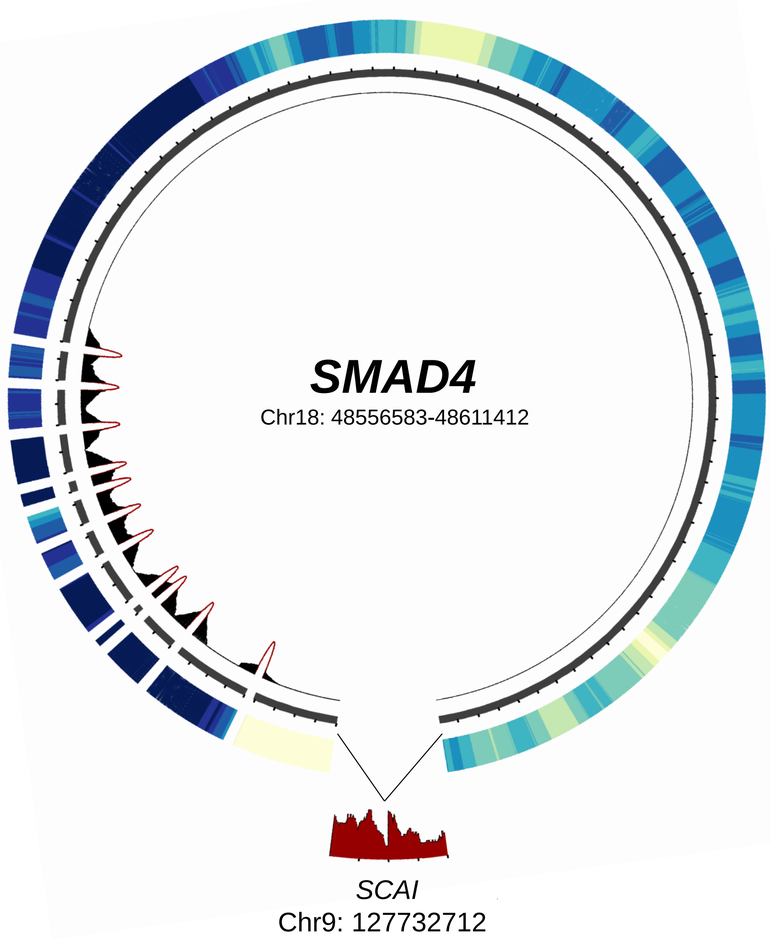


Figure 3: Insertion of the *SMAD4* processed pseudogene into *SCAI*. The outer heatmap displays the total genome coverage over the gene (blue gradient). The inner histogram (black) shows split reads supporting the presence of the pseudogene. The red outer histogram shows the fusion site in *SCAI*, where a drop in coverage is clearly shown.

Future Work

- Experimental and statistical validation
- Benchmarking

References:

1. Karro, J.E., et al., *Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation.* Nucleic Acids Res, 2007. **35**(Database issue): p. D55-60.
2. Cooke, S.L., et al., *Processed pseudogenes acquired somatically during cancer development.* Nat Commun, 2014. **5**: p. 3644.
3. Krzywinski, M., et al., *Circos: an information aesthetic for comparative genomics.* Genome Res, 2009. **19**(9): p. 1639-45

This project is financially supported by the Swedish Foundation for Strategeic Research

Contact: marcela.davila@gu.se