Al collegio docenti del Dottorato in Medicina Molecolare

Dott.: Davide Bolognini

Ciclo: *XXXIII*

Tutor: Prof. Alberto Magi

Scientific report (2nd year of PhD Fellowship)

Introduction

Tandem repeats (TRs) are described as stretches of DNA containing the same (or highly similar) sequence motif repeated multiple times in order. Short tandem repeats (STRs) are TRs containing repeated units of 1-6 bp. Due to their repetitive structure, STRs *loci* are subject to DNA polymerase slippage events that add or delete repeat units, resulting in mutation rates higher than those for other variant types. Expansions of STRs are implicated in more than 30 human hereditary disorders, including Huntington's disease and several ataxias. Robust tools have been created for detecting STRs in short-read sequencing data. However, given the length of the reads being analyzed, these methods are limited in the multiplicity and period size of TRs that can be resolved. In the past few years, third-generation sequencing technologies have arisen that offer substantially improved read lengths. Although they are obvious candidates for broadening the scope of detectable TRs, their high error profiles make it necessary to develop new approaches for accurate resolution of TR multiplicities.

Results and methods

We developed TRiCoLOR, a tool written in python 3.6 with supporting C++ and bash scripts, capable to detect TRs in error-prone long-read alignment files (BAM). BAM are quickly scanned to identify stretches with low entropy (that is, putative repetitive sequences). Identified regions are trimmed and, for each, a consensus sequence using an affine-gap penalties, multiple pairwise alignment is built in order to reduce their high error profiles. A regex search algorithm, modified to allow the identification of approximate TRs, is then run over each consensus sequence and the reference genome to accurately call motif and period of each repetition. Main output of the tool is a standard variant calling format file (VCF). The tool further includes methods to interactively visualize identified tandem repetitions in consensus sequences and to genotype repetitive *loci* in related individuals. For the time being, encouraging precision and recall both on simulated and real data has been demonstrated.

Abstracts

• *Title*: Large Genomic Alterations Occurring in the Transition from Chronic to Blast Phase of Chronic Myeloproliferative Neoplasms.
Authors: Niccolò Bartalucci, Alberto Magi, Elisa Contini, **Davide Bolognini**, Simone Romagnoli, Paola Guglielmelli and Alessandro Maria Vannucchi.
Conference site/date: 60th ASH (American Society of Hematology) Annual Meeting & Exposition. San Diego, CA. December 1-4,2018.

Conferences and courses (In addition to the ones attended during the 1st year of PhD Fellowship)

• *Title*: Visualising Biological Data (VIZBI 2019)

Conference site/date: EMBL Advanced Training Center. Heidelberg, Germany. March, 13–15, 2019.
• *Title:* Nuovi strumenti per l'analisi della risposta immunitaria alla vaccinazione e all'infezione tramite un approccio di 'systems biology'
Conference site/date: Presidio San Miniato. Siena, Italy. September, 5, 2019.

Scientific publications

• Magi A, **Bolognini D**, Bartalucci N, Mingrino A, Semeraro R, Giovannini L, Bonifacio S, Parrini D, Pelo E, Mannelli F, Guglielmelli P, Vannucchi AM. Nano-GLADIATOR: real-time detection of copy number alterations from nanopore sequencing data. Bioinformatics. 2019 Apr 5. pii: btz241.
• **Bolognini D**, Bartalucci N, Mingrino A, Vannucchi AM, Magi A. NanoR: A user-friendly R package to analyze and compare nanopore sequencing data.
PLoS One. 2019 May 9;14(5):e0216471.
• **Bolognini D**, Semeraro R, Magi A. Versatile Quality Control Methods for Nanopore Sequencing. Evol Bioinform Online. 2019; 15: 1176934319863068.
• **Bolognini D**, Sanders AD, Korbel JO, Magi A, Benes V, Rausch T. VISOR: a versatile haplotype-aware structural variant simulator for short and long read sequencing. Bioinformatics. *Accepted*.
• Sanders AD, Meiers S, Ghareghani M, Porubsky D, Jeong H, van Vliet MACC, Rausch T, Richter-Pechanska P, Kunz JB, Jenni S, **Bolognini D**, Longo GMC, Raeder B, Kinanen V, Zimmermann J, Benes V, Schrappe M, Mardin BR, Kulozik A, Bornhauser B, Bourquin JP, Marschall T, Korbel JO. Single cell tri-channel-processing reveals structural variation landscapes and complex rearrangement processes. Nat. Biotechnol. *Accepted*.

One-year stay abroad/scientific visit

From 01/10/2018 till 30/09/2019, visitor Bioinformatician at the European Molecular Biology Laboratory (EMBL) – European Bioinformatic Institute (EBI), Heidelberg, Germany. The visit is hosted by the Genomics Core Facility (head: Vladimir Benes) and scientifically supervised by Tobias Rausch, Senior Bioinformatician.