JLOH: Extracting Loss of Heterozygosity Blocks from Short-Read Sequencing Data

Matteo Schiavinato[#1], Valentina del Olmo[#2], Toni Gabaldón[#3]

[#]*Department of Life Sciences, Barcelona Supercomputing Center (BSC-CNS), Plaça Eusebi Güell, 1-3, Barcelona, Spain*

[1]matteo.schiavinato@bsc.es, [2]valentina.delolmo@bsc.es, [3]toni.gabaldon@bsc.es

***Keywords*— software, LOH, hybrids**

EXTENDED ABSTRACT

Introduction

Loss of heterozygosity (LOH) happens when a heterozygous genome loses one of the two alleles at a locus. This may have an evolutionary advantage in highly unstable genomes such as those of hybrids (Smukowski Heil et al., 2017). By extracting LOH from a hybrid we understand which alleles were beneficial in its evolution. This is important in wild, clinical, and industrial settings (Gabaldón, 2020). The genomic properties of hybrids are still, however, poorly understood. LOH are studied with reliable short-read sequencing data (Mixão et al., 2019), but the downstream analysis is often done with custom scripts that reduce reproducibility and do not to leverage the power of a computing cluster. Here we present a program called "JLOH" that streamlines LOH extraction from sequencing data maximizing parallel computing.

Materials and Methods

A series of divergent genomes (in short: *Sd*) were simulated from the *S. cerevisiae* genome (*Sc*) using JLOH's *sim* module, with increasing divergence (5%, 10%) and LOH (10%, 20%, 30%, 40%, 50%). Reads were simulated from *Sc* and *Sd* using wgsim (https://github.com/lh3/wgsim), joined together to simulate hybrid reads, and mapped against *Sc* and each *Sd* independently using hisat2 (Kim et al., 2019) with relaxed criteria (--score-min L,0.0,-1.0). The mapping results were filtered and used to call snps with bcftools (Danecek et al., 2021, --multiallelic-caller). The SNPs called were used as input for JLOH's *extract* module. Regions depleted of heterozygous SNPs are good candidates to be LOH blocks and are assessed by JLOH in terms of length, coverage in the up/downstream region, and homozygous SNP density. Called blocks were then compared against the LOH blocks originally introduced in the *Sd* genome to calculate precision and recall.

Results

True positives (TP), false positives (FP), false negatives (FN), precision and recall at each level of divergence and LOH are marked in Table 1.

TABLE I

Div.	LOH	TP	FP	FN	Precision	Recall
5%	10%	110	68	3	0.618	0.973
5%	20%	210	72	5	0.745	0.977
5%	30%	342	67	7	0.836	0.980
5%	40%	526	56	8	0.904	0.985
5%	50%	783	41	15	0.950	0.981
10%	10%	120	109	5	0.524	0.960
10%	20%	232	87	7	0.727	0.971
10%	30%	462	58	7	0.888	0.985
10%	40%	551	62	16	0.899	0.972
10%	50%	838	49	14	0.945	0.984

Conclusions

We conclude that JLOH successfully finds the LOH blocks that were artificially introduced in the simulated genomes using simple input files in common formats. We also conclude that the extent of LOH experienced from a genome may be a limit in JLOH's precision. While all introduced blocks are consistently found at any divergence or LOH rate, the number of false positives is anti-correlated with the level of LOH. Further development of the tool will hopefully reduce this effect.

References

[1] Smukowski Heil CS, DeSevo CG, Pai DA, Tucker CM, Hoang ML, Dunham MJ. Loss of heterozygosity drives adaptation in hybrid yeast. Molecular biology and evolution. 2017 Jul 1;34(7):1596-612.

[2] Gabaldón T. Hybridization and the origin of new yeast lineages. FEMS yeast research. 2020 Aug;20(5):foaa040.

[3] Mixão V, Hansen AP, Saus E, Boekhout T, Lass-Florl C, Gabaldón T. Whole-genome sequencing of the opportunistic yeast pathogen Candida inconspicua uncovers its hybrid origin. Frontiers in genetics. 2019 Apr 25;10:383.

[4] Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature biotechnology. 2019 Aug;37(8):907-15.

[5] Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Feb;10(2):giab008.

Author biography

Matteo Schiavinato was born in Italy in 1990. He received both the bachelor and the master degree in Molecular Biology at the University of Padova (Veneto, Italy), completing his studies in 2015. From January 2016 to February 2020 he has been a PhD student in bioinformatics at the University of Natural Resources and Life Sciences (BOKU) in Vienna, Austria. After the PhD, he briefly stayed as a postdoctoral researcher and then moved to the Core Facility Bioinformatics of the same university (BOKU), where he took the role of Operative Head.

Since October 2021 he has joined the research group of prof. Toni Gabaldón at the Barcelona Supercomputing Center (BSC-CNS) as a postdoctoral researcher. His research involves the investigation of the genomic changes underlying hybrid genome evolution.