HARP: an automated platform for targeted resequencing data analysis

Fernando Palluzzi✉

Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan, Italy

Motivation and Objectives

The fast-evolving scenario of Next Generation Sequencing (NGS) technologies caused an increasing demand of ready-to-use, costless and computationally powerful analysis systems, that could both represent a straightforward way to analyze huge amounts of data and offer a set of well assessed protocols to guide the user into an extensive landscape of different standards. In this session, I will present a simple tool called Hierarchical Assisted Resequencing Platform (HARP). HARP is an integrated NGS analysis platform, oriented especially towards resequencing experiments. HARP features allow the user to create personalized resequencing pipelines, using different tools and simplifying their usage and tuning; lead multiple projects at the same time; produce, manipulate, analyze and store data; a user-friendly interface, and finally create graphs, reports and benchmark protocols to assess the final results. Many general purpose platforms, such as Crossbow (Langmead et al. 2009), CloudBurst (Shatz 2009) or Galaxy (Goecks et al. 2010), have been successfully created, providing instruments for computationally intensive analyses directly on internet, without the need of huge hardware facilities. HARP has been prepared with the same purposes, but with the final goal of providing a risk evaluation parameter connected with the clinical and personal genetic profile of breast cancer affected patients.

Methods

HARP is almost completely implemented in Python (http://www.python.org/, last accessed on 22/09/2012) and Biopython http://biopython.org/wiki/Main_Page (last accessed on 22/09/2012), with a minor part of code written in _ Bash (http://www.gnu.org/software/bash/, last accessed on 22/09/2012) and R http://www.r-project.org/. A set of internal Python scripts has been used to create the HARP core. The core is composed by functions for environment management, format conversion, data pre-processing and wrappers for a set of third-party dependencies. Bash scripts has been used for sanity check, while R for statistics and graphics creation.

HARP interface has a modular structure, in which each module is presented to the user as a different menu, i.e. an independent task manager. In addition, there is a panel called experiment design. The experiment design manager allow the user to create personalized pipelines employing and interact with the whole HARP functionalities, by adjusting a relatively low number of basic parameters.

Third-party software include: Fastx toolkit http://hannonlab.cshl.edu/fastx_toolkit/, last accessed on 22/09/2012), for some data cleaning and manipulation procedure; SMALT http://www.sanger.ac.uk/resources/software/smalt/, last accessed on 22/09/2012), for reference-based alignments; Samtools http://samtools.sourceforge.net/, last accessed on 22/09/2012), to call variants; and finally the simulation tools set called ART http://www.niehs.nih.gov/research/resources/software/biostatistics/art/, last accessed on 22/09/2012), to perform results benchmarking. The different dependencies has been chosen regarding at different characteristics: the capability of working correctly with the main NGS standards, i.e.: fastq, SFF, SAM, BAM, BCF, VCF; the possibility of analyzing data from different experiments, e.g.: single-end or paired-end libraries; flexibility for different purposes, such as whole genome resequencing, amplicon resequencing or exome sequencing; and finally a straightforward usage.

Results and Discussion

In the table below are reported the results of a test analysis performed on a multiplexed sample, containing BRCA1/2 sequences from seven patients affected by breast cancer (BC). All these patients presented at least one BC variant. Specificity is expressed as the number of verified variants (i.e. the variants present in dbSNP) over the number of detected variants.

The risk assessment tool has been developed in R, but currently is not tested due to delays in obtaining real clinical data. However, the HARP risk assessment tool is an implementation of the Gail

Table 1: example of test analysis. This table shows the results of a test performed on an SFF file produced by a multiplexed sequencing experiment with a 454 GS Junior instrument (MID stands for Multiplex ID). The table reports all the detected variants, all those found in dbSNP, those variants that failed the quality check (QC) and the BC-related variants. The specificity is expressed as the ratio between the verified variants (i.e. present in dbSNP) over all the detected ones. Among these patients, only MID4 presents a novel variant (an indel), with unknown relation with BC.

MID	All variants	dbSNP	QC-failed	BC SNPs	Specificity
2	9	7	2	1	77.78%
3	8	8	0	2	100%
4	6	5	0	1	83.33%
5	15	15	0	5	100%
6	16	16	0	4	100%
7	9	9	0	2	100%
8	9	9	0	3	100%

model for absolute risk evaluation, that is a parametric model based on a series of clinical parameters, that can be enhanced using genome information (Gail et al. 1989; Gail 2009). The limitation in using such parametric approaches is the low discriminatory accuracy achieved, that is around 63% when genome information is included (Gail 2010).

Currently, HARP interface is still not available on the web, but a command-line version of HARP, called Breast Cancer risk Pipeline (BCP), is available for testing on Sourceforge, at https://sourceforge.net/projects/bcpipeline/.

Acknowledgements

The author thanks Professor Jordi Villa-Freixa, Professor Elena Maestrini and the Biocoputing Group of University of Bologna, for their constant support.

References

1. Ellsworth R E, Decewicz D J, et al. (2010). Breast cancer in the personal genomics era. *Curr. Genomics*, **11**: 146. doi:10.2174/138920210791110951
2. Gail M H, Brinton L A, et al. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.*, **81**: 1879.
3. Gail M H (2009). Value of adding Single-Nucleotide Polymorphism genotypes to a breast cancer risk model. *JNCI*, **101**: 959. doi:10.1093/jnci/djp130
4. Gail M H (2010). Personalized estimates of breast cancer risk in clinical practice and public health. *Stat. Med.*, **30**: 1090. doi:10.1002/sim.4187
5. Goecks J, Nekrutenko A, et al. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, **11**: R86. doi:10.1186/gb-2010-11-8-r86
6. Langmead B, Schatz M C, et al. (2009). Searching for SNPs with cloud computing. *Genome Biol.*, **10**: R134. doi:10.1186/gb-2009-10-11-r134
7. Shatz M C (2009). CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics*, **25**: 1363. doi:10.1093/bioinformatics/btp236