BMC Bioinformatics

MEETING ABSTRACT Open Access

COSMOS: cloud enabled NGS analysis

Yassine Souilmi[1,2], Jae-Yoon Jung[2], Alex Lancaster[2], Erik Gafni[3], Saaid Amzazi[1], Hassan Ghazal[4], Dennis Wall[2,5], Peter Tonellato[2*]

From Tenth International Society for Computational Biology (ISCB) Student Council Symposium 2014
Boston, MA, USA. 11 July 2014

Background

The dramatic fall of next generation sequencing (NGS) cost in recent years positions the price in range of typical medical testing, and thus whole genome analysis (WGA) may be a viable clinical diagnostic tool. Modern sequencing platforms routinely generate petabyte data. The current challenge lies in calling and analyzing this large-scale data, which has become the new time and cost rate-limiting step.

Methods

To address the computational limitations and optimize the cost, we have developed COSMOS (http://cosmos.hms.harvard.edu) , a scalable, parallelizable workflow management system running on clouds (e.g., Amazon Web Services or Google Clouds). Using COSMOS [1], we have constructed a NGS analysis pipeline implementing the Genome Analysis Toolkit - GATK v3.1 - best practice protocol [2,3], a widely accepted industry standard developed by the Broad Institute. COSMOS performs a thorough sequence analysis, including quality control, alignment, variant calling and an unprecedented level of annotation using a custom extension of ANNOVAR. COSMOS takes advantage of parallelization and the resources of a high-performance compute cluster, either local or in the cloud, to process datasets of up to the petabyte scale, which is becoming standard in NGS.

Conclusion

This approach enables the timely and cost-effective implementation of NGS analysis, allowing for it to be used in a clinical setting and translational medicine. With COSMOS we reduced the whole genome data analysis cost under the $100 barrier, placing it within a reimbursable cost point and in *clinical time*, providing a significant change to the landscape of genomic analysis and cement the utility of cloud environment as a resource for Petabyte-scale genomic research.

[2]Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
Full list of author information is available at the end of the article

Authors' details
[1]Department of Biology, Faculty of Sciences of Rabat, Morocco. [2]Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA. [3]INVITAE, San Francisco, CA 94107, USA. [4]Department of Biology, Mohamed First University, Oujda/Nador, Morocco. [5]Department of Pediatrics, Division of Systems Medicine, Stanford University, Stanford, CA 94305, USA.

Published: 28 January 2015

References

1. Gafni E, Luquette LJ, Lancaster AK, Hawkins JB, Jung J-Y, Souilmi Y, Wall DP, Tonellato PJ: **COSMOS: Python library for massively parallel workflows.** *Bioinformatics* 2014, btu385.
2. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet* 2011, **43**:491-498.
3. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA: **From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline.** 2013.

doi:10.1186/1471-2105-16-S2-A2
Cite this article as: Souilmi *et al.*: COSMOS: cloud enabled NGS analysis. *BMC Bioinformatics* 2015 **16**(Suppl 2):A2.