COSMOS: cloud enabled NGS analysis

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.1186/1471-2105-16-S2-A2</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:14351280">http://nrs.harvard.edu/urn-3:HUL.InstRepos:14351280</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
COSMOS: cloud enabled NGS analysis

Yassine Souilmi1,2, Jae-Yoon Jung2, Alex Lancaster2, Erik Gafni3, Saaid Amzazi1, Hassan Ghazal4, Dennis Wall2,5, Peter Tonellato2*

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.

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

Published: 28 January 2015

References

Submit your next manuscript to BioMed Central and take full advantage of:

• Convenient online submission
• Thorough peer review
• No space constraints or color figure charges
• Immediate publication on acceptance
• Inclusion in PubMed, CAS, Scopus and Google Scholar
• Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit

© 2015 Souilmi et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.