Cirrus-NGS: Cloud-optimized next generation sequencing primary analysis pipeline

Guorong Xu, Mustafa Guler, Mengyi Liu, Roman Sasik, Amanda Birmingham, Kathleen M. Fisch*

Center for Computational Biology & Bioinformatics, Department of Medicine

University of California, San Diego, La Jolla, CA

Contact: g1xu@ucsd.edu

Motivation

Bioinformatics analysis of large-scale next-generation sequencing (NGS) data requires significant compute resources. While cloud computing makes such processing power available on-demand, the administration of dynamic compute clusters is a daunting task for most working biologists. To address this pain point, we have developed cirrus-ngs, a turn-key solution for common NGS analyses using Amazon Web Services (AWS).

Pipelines

Cirrus-ngs currently supports RNA-Seq, miRNA-Seq, ChIP-Seq and WGS/WES data with multiple version of genomes.

Features:
WGS/WES pipelines include bwa gatk and bwa mutect workflows for germline and somatic variants calling. The analysis steps include fastqc, trim, align, multiqc, sort, dedup, split, postalignment, haplotype, somatic_variant_calling , merge, combine_vcf.

RNA-seq pipelines include star_rsem, star_htseq, kallisto and star_gatk workflows. The analysis steps include fastqc, trim, align_count, multiqc, merge_counts and variant_calling.

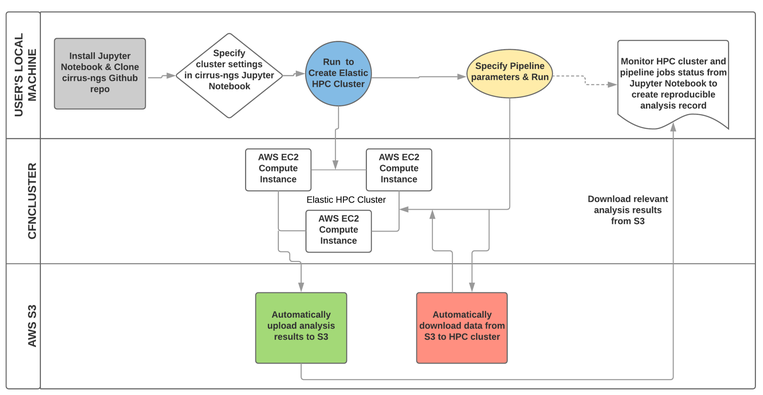
ChiP-seq pipeline includes Homer workflow. The analysis steps include fastqc, trim, align, multiqc, make_tag_directory, make_UCSC_file, find_peaks, annotate_peaks, pos2bed, find_motifs_genome.

miRNA-seq pipeline includes bowtie2 workflow. The analysis steps include fastqc, trim, cut_adapt, align_and_count, multiqc.

Reliability:
To avoid data loss, the intermediate results of each step are saved to S3 and users can rerun any step of the pipeline if one step is failed because of spot instances loss or unknown reasons.

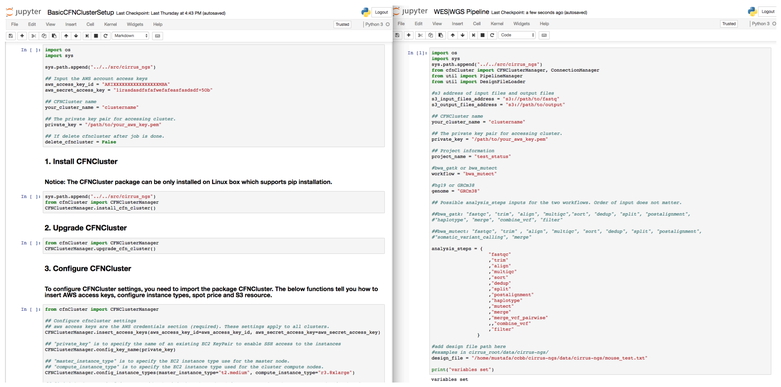
Extensibility:
Each workflow is designed independently but shares the common steps. The developers/users can easily modify or integrate their own codes to Cirrus-ngs by following the design and configuration of framework.

Schematic

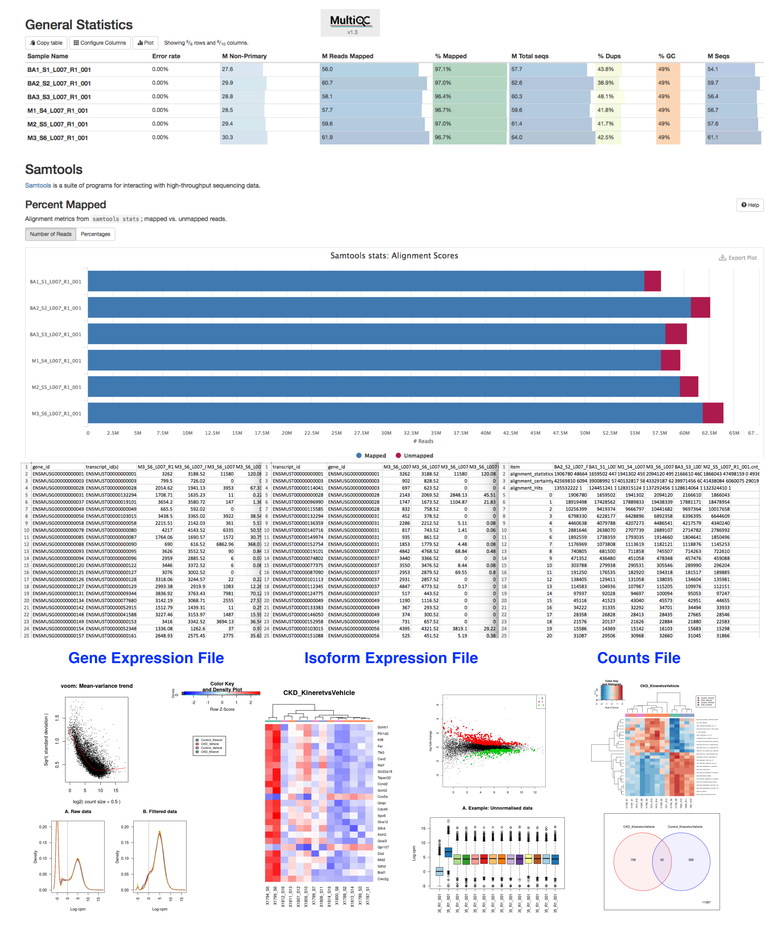


Configurations

Notebooks



Analysis



Performance

Data Type	Num of Reads (M)	Running Time (H)	Instance Type	Running Cost ($)
WGS	680	30	r3.8xlarge	15
WES	43	3.6	r3.8xlarge	1.7
RNA-seq	35	2	r3.8xlarge	1
ChiP-seq	45	2	r3.8xlarge	1
miRNA-seq	15	0.5	r3.4xlarge	0.25

Note: the numbers are per sample

Conclusion

Cirrus-ngs is a lightweight, reproducible tool to perform scalable NGS primary analyses on the cloud. Cirrus-ngs has been optimized for use on AWS HPC clusters, which dynamically scale depending on the compute resources required for the various steps in the pipelines to minimize the per sample compute cost. All computation is performed using AWS EC2 compute instances and all results are uploaded to AWS S3 storage.

Availability

Cirrus-ngs is developed in Python and is available for free use and extension under the MIT License. Source code and extensive documentation are on GitHub at https://github.com/ucsd-ccbb/cirrus-ngs

Funding/Disclosures

Supported by the UC San Diego Clinical and Translational Research Institute Grant UL1TR001442.

References

[1] 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. **Nature** 526:68-74. doi: 10.1038/nature15393

[2] cfncluster (https://github.com/awslabs/cfncluster)

[3] H. Li, R. Durbin Fast and accurate short read alignment with Burrows-Wheeler transform **Bioinformatics**, 25 (2009), pp. 1754-1760

[4] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G.Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, 25 (2009), pp. 2078-2079

[5] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K.Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristoThe Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data **Genome Res.**, 20 (2010), pp. 1297-1303

[6] D. Koboldt, D. Larson, R. Wilson. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. Curr Protoc **Bioinformatics**. 2013;44:15.14.11–17. doi: 10.1002/0471250953.bi1504s44.

[7] Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–219 (2013).

[8] Z. Lai, A. Markovets, M. Ahdesmaki, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research **Nucleic Acids Res**, 44 (2016), p. E108

[9] K. Wang, M. Li & H. Hakonarson ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. **Nucleic Acids Res**. 38, e164 (2010)

[10] S. Heinz et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. **Mol Cell** 38(4):576-589. PMID: 20513432 (2010)

[11] N. Bray et al. Near-optimal probabilistic RNA-seq quantification, **Nature Biotechnology 34**, 525–527 (2016), doi:10.1038/nbt.3519