17 Sept 2015

Software article for Genome Biology

Dr. Louisa Flintoft

Executive Editor, Genome Biology

Dear Dr. Flintoft,

Following up from our presubmission enquiry (EOI) in June, we would like to propose the adjusted article for *Genome Biology*’s “Software” section titled “Population Scale Clustering of Genotype Information”, which discusses the use of the novel parallelization paradigm, Spark, on cloud compute infrastructure for large cohort genomic data analysis.

We addressed your comments to our EOI around

1. demonstrating that our approach is substantially different and faster than currently published methods by comparing against ADMIXTURE and the non-peer-reviewed method ADAM.
2. demonstrating our Spark implementation is substantially superior over traditional approaches by comparing against an R and Python implementation.
3. demonstrating utility for more than a single dataset by using data from the 1000 Genomes Project and the Personal Genome Project.

Furthermore, the here-introduced framework provides a general interface between genomic variant data (VCF files) and the machine learning library MLlib. It therefore enables a large area of population-scale genomic analysis that may be of interest to a wide range of *Genome Biology* readers.

Thank you in advance for your consideration.

Sincerely,

**Denis C. Bauer**

(on behalf of all authors)