Abstract
The recent revolution in sequencing technologies has led to the exponential growth of sequence data. As a result, most of the current bioinformatics tools become obsolete as they fail to scale with data. To tackle this “data deluge”, here we introduce the BioPig sequence analysis toolkit as one of the solutions that scale to data and computation. BioPig is built upon the Apache’s Hadoop map-reduce system and the Pig data flow language. Its programmability greatly reduces development time for parallel bioinformatics applications. Testing linear algorithms implemented in BioPig with up to 500 Gb sequences demonstrates that it scales linearly with size of data. We also demonstrate that BioPig can be ported without modification on two independent pieces of Hadoop infrastructure, the Magellan system at NERSC, and the Amazon Elastic Compute Cloud. In summary, BioPig represents a novel program framework with the potential to greatly accelerate data-intensive bioinformatics analysis.

Motivation
Cost of sequencing is decreasing faster than the cost of storage. We need better algorithms and methods to handle TeraBases of data.

Easy to program
Listing 1: Just 8 lines for K-mer histogram calculation.

Conclusion
BioPig is designed for processing large datasets in bioinformatics applications. We found that BioPig is easy to use -- for the gene discovery pipeline, it requires 61 lines of BioPig script code. It is scalable -- it is able to process datasets as large as 500 Gb. It is portable -- it can be ported to Amazon EC2 easily. It may be somewhat slower than handcrafted MPI. We note that performance is not the only criterion by which to evaluate the various methods. Indeed, for some methods it is their feasibility that is of highest importance. Some methods simply break down when applied to large data. When dealing with huge data sets, there is no way around having enough resources. BioPig shifts the need from expensive resources such as large memory (> 1TB RAM) machines and/or parallel programming expertise, to commodity hardware and/or the expertise to run in the Cloud. Publication pending.

Acknowledgements: Karan Bhatia, Shane Cannon, Tatyana Smirnova, Aijazuddin Syed, Zhong Wang