NATURE | NEWS

Genome researchers raise alarm over big data

Storing and processing genome data will exceed the computing challenges of running YouTube and Twitter, biologists warn.

Erika Check Hayden

07 July 2015

The computing resources needed to handle genome data will soon exceed those of Twitter and YouTube, says a team of biologists and computer scientists who are worried that their discipline is not geared up to cope with the coming genomics flood.

Other computing experts say that such a comparison with other ‘big data’ areas is not convincing and a little glib. But they agree that the computing needs of genomics will be enormous as sequencing costs drop and ever more genomes are analysed.

By 2025, between 100 million and 2 billion human genomes could have been sequenced, according to the report[1], which is published in the journal *PLoS Biology*. The data-storage demands for this alone could run to as much as 2–40 exabytes (1 exabyte is 10^{18} bytes), because the number of data that must be stored for a single genome are 30 times larger than the size of the genome itself, to make up for errors incurred during sequencing and preliminary analysis.

The team says that this outstrips YouTube’s projected annual storage needs of 1–2 exabytes of video by 2025 and Twitter’s projected 1–17 petabytes per year (1 petabyte is 10^{15} bytes). It even exceeds the 1 exabyte per year projected for what will be the world’s largest astronomy project, the Square Kilometre Array, to be sited in South Africa and Australia. But storage is only a small part of the problem: the paper argues that computing requirements for acquiring, distributing and analysing genomics data may be even more demanding.

Major change

“This serves as a clarion call that genomics is going to pose some severe challenges,” says biologist Gene Robinson from the University of Illinois at Urbana-Champaign (UIUC), a co-author of the paper. “Some major change is going to need to happen to handle the volume of data and speed of analysis that will be required.”

Narayan Desai, a computer scientist at communications giant Ericsson in San Jose, California, is not impressed by the way the study compares the demands of other disciplines. “This isn’t a particularly credible analysis,” he says. Desai points out that the paper gives short shrift to the way in which other disciplines handle the data they collect — for instance, the paper underestimates the processing and analysis aspects of the video and text data collected and distributed by Twitter and YouTube, such as advertisement targeting and serving videos to diverse formats.

Nevertheless, Desai says, genomics will have to address the fundamental question of how much data it should generate. “The world has a limited capacity for data collection and analysis, and it should be used well. Because of the accessibility of sequencing, the explosive growth of the community has occurred in a largely decentralized fashion, which can't easily address questions like this," he says. Other resource-intensive disciplines, such as high-energy physics, are more centralized; they “require coordination and consensus for instrument design, data collection and sampling strategies”, he adds. But genomics data sets are more balkanized, despite the recent interest of cloud-computing companies in centrally storing large amounts of genomics data.

Coordinated approach

Astronomers and high-energy physicists process much of their raw data soon after collection and then discard them, which simplifies later steps such as distribution and analysis. But genomics does not yet have standards for converting raw sequence data into processed data.

The variety of analyses that biologists want to perform in genomics is also uniquely large, the authors write, and current methods for performing these analyses will not necessarily translate well as the volume of such data rises. For instance, comparing two genomes requires comparing two sets of genetic variants. “If you have a million genomes, you’re talking about a million-squared pairwise comparisons,” says Saurabh Sinha, a computer scientist at the UIUC and a co-author of the paper. “The algorithms for doing that are going to scale badly.”

Observational cosmologist Robert Brunner, also at the UIUC, says that, rather than comparing disciplines, he would have liked to have seen a call to arms for big-data problems that span disciplines and that could benefit from a coordinated approach — such as the relative dearth of career paths for computational specialists in science, and the need for specialized types of storage and analysis capacity that will not necessarily be met by industrial providers.

“Genomics poses some of the same challenges as astronomy, atmospheric science, crop science, particle physics and whatever big-data domain you want to think about,” Brunner says. “The real thing to do here is to say what are things in common that we can work together to solve.”

Nature | doi:10.1038/nature.2015.17912

References

1. Stephens, Z. D. *et al*. *PLoS Biol*. http://dx.doi.org/10.1371/journal.pbio.1002195 (2015).