

Genome analysis: the global bottleneck

The successful completion of the bovine genome sequencing project and associated hapmap project mark the birth of a new era in livestock and agricultural research. The size and scope of the project (US\$52M, 25 countries, >300 collaborators) is truly impressive. One of the unique aspects of this undertaking was the community based gene annotation that involved hundreds of individual researchers who volunteered their time and expertise, which made human curation of genes possible for an agricultural genome. The agricultural applications of this research are likely to be profound for both the dairy and beef cattle industries.

In spite of this success, such mega-dollar projects for single organism genome sequencing are probably headed for extinction, as the price of DNA sequencing continues to plummet and the cost of sequencing an animal or plant genome is now becoming affordable within the context of an individual research grant.

Another reason for this is that the process of puzzling together the raw data obtained from sequencing into the context of a whole genome is rapidly becoming easier.

The exponential growth of genome sequencing projects in the past 3-4 years has largely been driven by microorganisms and other life forms with small genomes (Figure 1). One of the key barriers to the adoption of next-generation DNA sequencing for large genomes is the difficulty of genome assembly from short reads. Significant improvements in assembly algorithms and software are making this less of a problem and it is only a matter of time before this barrier crumbles.

been able to do with other new techniques such as gene expression arrays. This has implications for how research grants are defined and funded; should wet bench projects that generate large data sets only be funded if their analysis projects/consortia are funded as well? These will be moot points if there are insufficient computational biologists and bioinformaticians to carry out the analyses.

This problem will not be restricted to DNA sequence datasets. Advances in proteomics and metabolomics instruments are lowering costs and will create additional complexity as their data are integrated with DNA sequence and expression data in what is currently called systems biology.



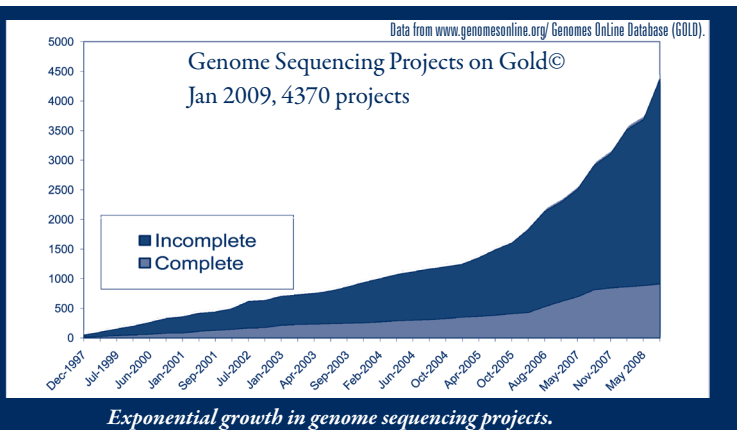
...it is clear that the analysis shortfall for biology is a global phenomenon, so if we want to remain competitive we have to generate 'in house' analytical know how and not just rely on overseas educational systems.

What to do? The short-term answer will be to engage more with the international research community in the form of research consortia in order to remain at the cutting edge. However, it is clear that the analysis shortfall for biology is a global phenomenon, so if we want to remain competitive we have to generate 'in house' analytical know how and not just rely on overseas educational systems.

It is therefore imperative that Australia allocate sufficient funds for bioinformatics.

One particular concern is that at present, bioinformatics and computational biology are not explicitly targeted as national technology funding priorities by either the ARC or NHMRC. Furthermore, funding a few more Centres of Excellence will not be the solution, because the analytical shortfall will be uniform and pervasive. Right now any bright person with an internet connection and a reasonably powerful desktop machine can become a computational/systems biologist, but they will not do so if there are no career paths and if training is hard to come by. At present there are few true interdisciplinary training programs in Australia that can attract and train the next generation of analytical biologists, and this needs to be addressed as a matter of urgency.

We live in exiting times; fifteen years ago I never dreamed that I would be involved in livestock genome sequencing efforts, but today I am an analysis team leader in two such consortia. While it's hard to predict where we might be in another fifteen years, it is clear that at present, the 'omics playing field has been leveled, and Australia needs to take advantage of this state of affairs to field a competitive team, as is done for the Olympic Games. The benefits of this investment will be twofold; it will be easier to forge strong collaborative links with overseas institutions that are already at the leading edge in order to participate in future multi-megadollar projects and it will ensure the continued high quality of local research and development in the biological sciences and the agricultural sector.



This new era of cheap, whole genome sequencing will revolutionize the way we do biology. But revolutions are never easy and in a Darwinian sense will impose strong selective pressure on labs and institutions. This selective pressure will be exerted at the level of data analysis, as the genomics bottleneck shifts downstream to the computational biologists. The bovine genome project was a success because it enlisted the bovine research community as its analytical engine.

However, it still required about 15 analysis team leaders and numerous additional personnel to transform the raw data into knowledge.

It is perhaps this requirement that explains the fact that the number of completed genome sequences has flattened out in the past couple of years, while the number of incomplete projects is continuously increasing. Individual labs are not able to tackle all of the analyses required to characterise a genome sequence on their own, in contrast to what they have