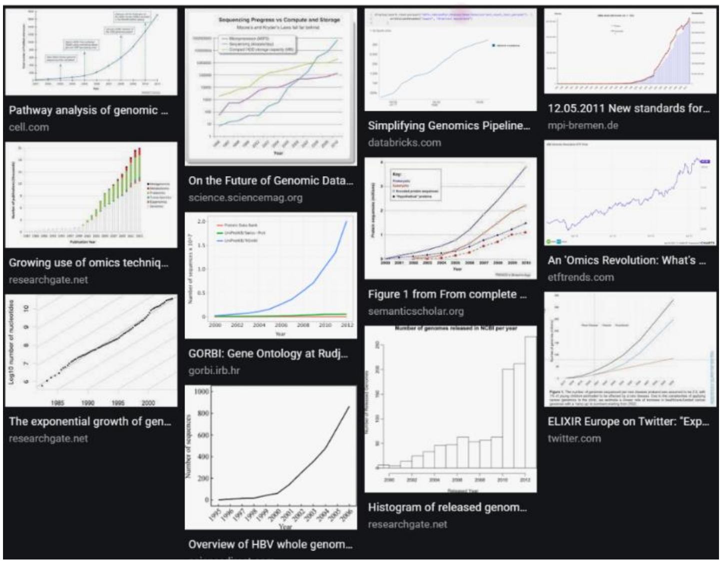
FindingPheno

July 02, 2021

Big data is getting bigger

Updated: Jan 2, 2024

At the FindingPheno Kickoff Meeting, I spoke about the data tsunami. What is this? Why is it happening? Why should we care?



Pathway analysis of genomic ...
cell.com
Growing use of omics techniq...
researchgate.net
The exponential growth of gen...
researchgate.net
On the Future of Genomic Data...
science.sciencemag.org
GORBI: Gene Ontology at Rudj...
gorbi.irb.hr
Overview of HBV whole genom...
Simplifying Genomics Pipeline...
databricks.com
Figure 1 from From complete ...
semanticscholar.org
Histogram of released genom...
researchgate.net
12.05.2011 New standards for...
mpi-bremen.de
An 'Omics Revolution: What's ...
etftrends.com
ELIXIR Europe on Twitter: "Exp...
twitter.com

Biological research is a data-intensive endeavor

A look across the trends in genomics and related areas show the same thing, the amount of data generated and published increases year on year. The adjacent figure is a Google image search illustrating this idea, with different omics data graphs increasing exponentially. This is true on the commercial side too, with bioinformatics and genomics market reports predicting double-digit CAGRs driven by falling prices and rising demand. Meanwhile, omics technology continues improving, giving better, more detailed, and more complex measurements, often in higher throughput to give more data points each time. Biological insights and knowledge gained from these experiments feed into the system, allowing us to ask –and answer– more sophisticated questions as the field evolves.

While new data generation is increasing, the old data does not disappear. Journals and funding bodies generally require that data sets are uploaded into pubic databases after publication meaning that most of this data are available for public use. Even after being analyzed, data remains in its original form, retaining its biological value. This value may even increase over time as we compile different types of data for a specific organism (i.e. the plant or animal being studied), allowing new combinations that give more complete information than is found in one experiment alone.

All together this result in a tsunami of data, a gathering wave of biological measurements with rich potential for new ideas, new knowledge, and new answers to our pressing problems.

Data integration and analysis across molecular layers is difficult

So the data accumulates. How to make the most of this potential? The ability to analyse different types of omics data from the same sample all at once can improve our understanding of the underlying interactions and regulation within and between the different molecular layers. We can then use this understanding to predict how a plant or animal will change in response to outside interventions allowing us to optimize their growth and wellbeing. However, this is not an easy task, with several major challenges standing in the way.

The first problem is harmonizing all data produced when measuring different chemicals, i.e. DNA, mRNA, proteins, and metabolites, into a single multi-omics package ready for biological interrogation. Their different physical and chemical properties influence the types of measurements taken. Different measurement platforms or software packages may produce similar data in different formats. This causes interoperability problems where data created by one system is not readable or understood by another. Include differences in data cleaning, normalization, and transformation, and combining all data in one place can be a major hurdle.

Researchers need to integrate different data types into one analysis framework that can identify true causal interactions and associations. For example, while RNA-seq or microarrays measure tens of thousands of transcripts with high coverage, mass-spec or similar may only profile a few hundred or thousand proteins or metabolites from the same sample. This can introduce annotation bias, where the transcript data has more weight simply because there is more of it even though the proteins may be more functionally meaningful. Additionally, small changes in low-concentration proteins or metabolites (e.g. cytokines, eicosanoids) may be lost among the transcript data even when they have strong effects on the plant or animal. On the microbiome side, a key challenge is how to move from species or strain abundances (i.e. which microbes are there) to modelling the metabolic capacity of the whole microbiome, since it can be argued that the latter is what causes a phenotype. This becomes more complex when including host metabolism, due to likely strong feedback loops in the whole system.

So what do we do about it?

Overall, data integration must be approached in a clever way to reduce noise, bias, and computational burden by focusing in on only what is important or meaningful, without losing useful biological signals along the way. FindingPheno takes a range of approaches to this problem, including biology-agnostic unsupervised machine learning methods, biology-informed hierarchical modelling, and time course or spatial dynamic modelling, with the overall aim of creating a holistic statistical model that can finally tame the tsunami.

Written: Shelley Edmunds
Updated: Marie Sorivelle