Scaling up genomics

Scale can be as problematic in genetics as it is in microscopy or astronomy. Luckily, pan-genomics is here to tackle the complexity of genetics on the large scale.

The twentieth century produced two great theories in physics, relativity and quantum mechanics. Between them they deal with the four fundamental forces that underpin all phenomena in the known universe: electromagnetism, strong and weak nuclear forces, and gravity. Quantum mechanics deals with the first three of these, which operate on atomic and subatomic scales. Relativity deals with gravity, which is the weakest of the four forces but operates over astronomical distances and, on such scales, is capable of warping the fabric of space–time itself. But relativity and quantum mechanics are fundamentally incompatible. The grand unifying ‘Theory of Everything’ that has long been the goal of physicists, including the late Stephen Hawking, remains a near mythical beast. The small scale and the large cannot be reconciled.

In biology, something similar occurs with inheritance and evolution. Mendel and Darwin had essentially no knowledge of the functioning of life at a cellular level, leaving them free to postulate the existence of ‘units of inheritance’ without worrying too much about what they might be. Through the work of Hugo de Vries and August Weismann, the concept of ‘genes’ located in a cell’s nucleus was established, and at the turn of the twentieth century, Theodor Boveri tracked them to chromosomes.

It wasn’t until half a century later that DNA was fully established as the physical structure from which genes are transcribed, but such a materialistic definition is not always helpful for understanding evolution. It is a perfectly logical approach to say that as the sequence of base pairs in DNA is the only thing inherited, such sequences must be the units on which selection applies its pressure. Darwin’s drivers of evolution, by acting on this information, would lead to the survival of the fittest sequence of DNA, regardless of the organism in which it is located. The resulting ‘selfish gene’ hypothesis, popularized by geneticists such as Richard Dawkins in the 1970s, is mathematically rigorous but requires complex interactions between genes to produce the emergence of cellular behaviours and some leaps of faith to account for those of organisms.

Evolution by natural selection requires the existence of an excess of individuals, of which only a small fraction survive to reproduce. It is the whole organism that dies,

so perhaps selection acts on entire organisms and thus entire genomes. But genomes are shared, at least partially, between an individual’s kin, giving evolutionary incentives to aiding the survival and reproduction of close relatives, even at the expense of oneself: a phenomenon known as kin selection. At an even larger scale, genome sequences are common, more or less, to all members of a species; perhaps survival of the species is what counts, and individual survival is irrelevant. Such ‘group selection’ was widely discredited in the 1970s but has been partially resurrected in the last decade by researchers such as Edward O. Wilson as a way to understand the behaviour of eusocial animals such as ants and bees.

Genomic research is currently grappling with its own problem of scale. With sequencing of genomes becoming routine, it is increasingly obvious just how flexible they are not only to single nucleotide polymorphisms, but to larger-scale insertions, deletions and rearrangements. Furthermore, the ability of plants to hybridize and accommodate varying degrees of polyploidy can strain the conventional concepts of species to the point where any individual will not contain the full roster of genes available to the species to which it belongs.

A ‘pan-genome’ is the sum of all possible sequences and sequence variations that can be found in individual members of a species. The concept originally arose from studies on the pathogenic bacterium *Streptococcus agalactiae* (*Proc. Natl Acad. Sci. USA* **102**, 13950–13955; 2005) but rapidly became adopted by plant and animal geneticists (*Nat. Rev. Genet.* **21**, 243–254; 2020). Indeed, even if they were not using the term, researchers who were interested in the multiplicity of *Arabidopsis* ecotypes were well aware that a single variety was not telling the whole story of the species. A pan-genome includes a core genome, consisting of essential genes present in all individuals, and a dispensable genome, including genes present in only a subset of strains within the species. All this additional information over and above a single reference genome make pan-genomes a window into the diversity, conservation, function and evolutionary significance of DNA elements.

The increasing availability of long-read sequencing technologies will soon make the presentation of individual genomes

seem horribly old-fashioned. *Nature Plants* has itself published the pan-genomes of *Brassica napus* (*Nat. Plants* **6**, 34–45; 2020) and sunflower (*Nat. Plants* **5**, 54–62; 2019) in the last 18 months, while a recent paper on hornworts (*Nat. Plants* **6**, 259–272; 2020) attempted an initial construction of the pan-genome of *Anthoceros agrestis*, albeit based on only two strains. For crops, selective breeding directed by robust genomic data has the ability to create specific selections from the available variations present in the species’ pan-genome, and there are pan-genomes available for wheat, maize, soybean and rice among others. These pan-genomes — important for directing breeding programs — also bear the history of evolution under artificial selection: gene losses, gains and rearrangements that accumulate during domestication and improvement.

As yet, we have not seen attempts to discuss a variation of group selection at the pan-genome level, nor the assembly of pan-genomes of multi-species genera. But it can only be a matter of time.

COVID-19

The COVID-19 epidemic has been causing significant disruption to research in China for some months, and now that it has expanded into a global pandemic, there can be very few, if any, researchers untouched. We are very aware that many in our community will have difficulty in completing steps associated with peer review as quickly as under normal circumstances. Reviewing and revising manuscripts will take longer, and it may be impossible for authors to perform experiments requested by referees until access to their labs, samples and plants is restored. We completely understand this and will attempt to accommodate such added difficulties. We hope our authors will also understand that decision times may be slower over the coming months.

For the time being, we will attempt to stick to our usual timelines and our systems will continue to remind authors and reviewers of our standard deadlines, but we will be as flexible as we possibly can be. So please let us know if you need additional time, or if what we have asked of you is impractical. And, most importantly, stay safe and healthy. □

Published online: 15 April 2020
<https://doi.org/10.1038/s41477-020-0652-0>