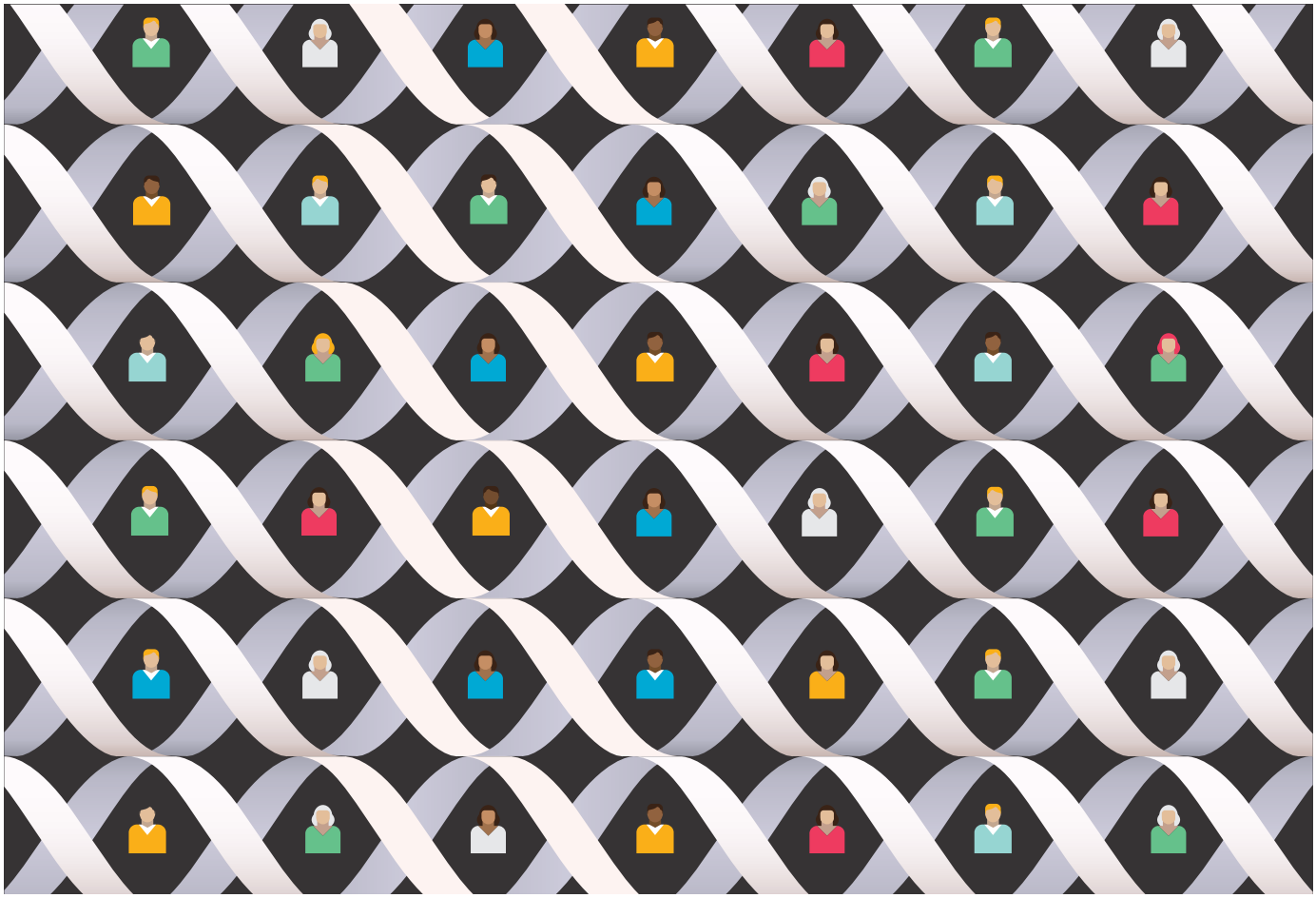
THOMAS POROSTOCKY



Lessons for big-data projects

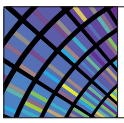
To be successful, consortia need clear management, codes of conduct and participants who are committed to working for the common good, says ENCODE lead analysis coordinator **Ewan Birney**.

The ENCODE consortium has for the past five years been building up an encyclopaedia of functional DNA elements to be used as a reference for the scientific community. Today it publishes 30 publicly accessible papers in three journals — and all are connected to the processed analysis and raw data. This scientific undertaking has inspired new publishing models, such as the interweaving of topic threads between papers in different journals, and will, I hope, have a large impact on biology.

The ENCODE project has delivered an incredible amount of information because of its sheer scale: more than 1,600 experiments on 147 cell types, including 235 antibodies or other assay protocols. The main paper has nearly 450 authors, working from more than 30 institutions.

Because of its complexity (see page 46), the project could not have worked in the same way as one involving just one or two laboratories. Typically, scientists try to do the best science they can, with a limited set of collaborators, to earn grants and publications to do what is best for science, their own careers and their own laboratories.

This mindset doesn't work in consortium science. Instead, researchers must focus on creating the best data set they can. Maybe they will use the data, maybe they won't. What is important is the community resource, not individual success. This requires a shift in perspective to a common goal of data output rather than publications. In turn, the success of consortium participants must be measured at least as much by how their data have enabled science as by the insights they have produced.



ENCODE
Encyclopedia of DNA Elements
nature.com/encode

SUPPORTING THE COMMUNITY

Big-biology consortia such as ENCODE, HapMap and the 1000 Genomes Project approach grand-scale work systematically. For example, they often take a 'catalogue' approach to create foundational resources rather than spotlighting areas of interest, and they use standardized methods, reagents and analysis schemes. The cost of these projects is justified by the breadth of science they support — from genome-wide analysis down to smaller-scale, hypothesis-driven studies.

Has the big project had its day in the current era of 'democratized' data gathering? Certainly the drop in the price of data gathering has changed the game for all biology groups — and nearly always for the better (although there are of course new challenges in how to handle this). But the cheapness of data just extends the reach of large-scale projects; it does not alter the need to create systematic reference data sets. It is hard, if not impossible, to combine smaller data sets into reference data sets — as demonstrated by the initial chromosome maps in the Human Genome Project or the attempt to patch together collections of microarray data into an atlas of gene expression.

Instead, a systematic data 'skeleton' is needed (for genomes, functional elements and variation, for example), around which smaller-scale experiments can add insight, colour and deeper understanding. ENCODE, BLUEPRINT and the 1000 Genomes Project are examples of such skeletons. The main products of ENCODE and similar projects are not just raw data, but also analysed intermediates that allow scientists to choose the level of detail at which they wish to start.

I have been involved in consortia at various levels since 1999. In 2004, I became the coordinator of the ENCODE analysis. I have learned that consortia are difficult to make successful, because they involve people who might be competing with one another in another context. Getting competitors to work openly together towards a shared goal is not trivial. It relies on the good will of all.

ENCODE has made it clear to me that effective consortium science requires all participants to buy into a structure, a code of conduct and the goal of high-quality data that are made accessible and usable to all scientists around the world.

CLEAR STRUCTURE

In my opinion, for large consortia to succeed, they need to create a structure that is transparent to everyone involved.

This structure cannot follow the classic model of a single institute with a fixed hierarchy, or even a single 'virtual' institute agreed on by multiple partners. Instead, as happened for ENCODE, an open, peer-reviewed process should select and evaluate the partners who are best suited to a self-organized structure. And the structure should be flexible enough to change over time and to encompass multiple sources of funding. Considering each partner as an individual — rather than regarding the consortium as a single group — allows the addition of innovative participants from outside the expected group. ENCODE probably would not have such a great depth of input from statistical groups had the project been funded by a single large grant.

A diverse collection of scientists keeps the ideas fresh and the technology agile. It prevents 'group think'. For example, when there is a shift in technology, labs differ in their uptake. It would be damaging if everyone either committed too early to a poorly performing technology, or delayed uptake of a successful one. Broad participation also connects the output to a much larger audience worldwide.

Large consortia do, however, need to avoid a common pitfall: sharing the responsibility between too many principal investigators and senior postdoctoral fellows. This renders decision-making difficult. Without a core structure, there is a risk that members will shift their focus to their own interest areas at the expense of the overall project. At the same time, these projects are too big and complex to be managed by one person, who is unlikely to have expertise in all the relevant areas. Initiatives that are piloted by one or a few principal investigators are more common in consortia working on diseases, and in my experience they often lack an operational project manager with a well-defined role.

"Consortium science involves interaction between humans, with all the social complications this entails."

The ENCODE consortium had an internal structure that I believe was instrumental to its success. It had a 'spine' of leadership comprising: scientifically aware project officers in the primary funding agency, the National Human Genome Research Institute at the US National Institutes of Health; a few leading scientists with goals aligned to the consortium; and one or two scientific project managers hired inside the consortium who had a detailed understanding of all the tasks and people involved. ENCODE's two key project coordinators (Ian Dunham and Anshul Kundaje) were funded for the lifetime of the project through a grant for which I was the principal investigator. Successful consortia tend to have similar core structures, suggesting that this is a natural and effective way to organize such projects.

The spine was able to resolve some of the most complex problems — both scientific and social — such as sorting out a quality-control disagreement between a data-production and data-analysis group. As in any endeavour that involves many individuals, communication channels are crucial for success. We should have explicitly broadcast the existence of this spine both to the group and externally, to provide more transparency with respect to how decisions were made.

I also think that funding agencies should become more involved in shaping consortia. They should be flexible enough to shift their support from one group to another as needed, with adequate warning, and to withdraw funding from poorly performing or uncooperative partners — again with warning and with real consequences. Funding agreements often include such terms and conditions, but they are rarely used, perhaps because the threat of action is enough. And perhaps funding agencies feel uncomfortable, understandably, taking on such a scientifically directive role. But the responsibility for the overall success of the project rests firmly with the funding agency, so it must feel empowered to intervene when necessary.

CODES OF CONDUCT

Consortium science involves interaction between humans, with all the social complications this entails. It happens across multiple sites and time zones, and the partners generally communicate electronically, rather than in person. Misunderstandings and clashes can arise because of cultural differences — at national, organizational and individual levels.

To ensure that things run smoothly, rules are essential. An agreed-upon, written and publicly accessible code of conduct is extremely beneficial to large consortia, particularly when they need to incorporate less-experienced partners. ENCODE had several written rules, on issues such as data release, and these were circulated internally.

Such rules help to ensure that partners work within the goals of the consortium and do not (consciously or unconsciously) form a cartel that controls access to the data and analysis. An advisory board should regularly scrutinize internal and external partners for scientific impact, capacity to deliver and ability to interact effectively. Although I am confident that ENCODE did not restrict access to data or analysis through the rules of the funding agency, outside groups occasionally had that impression, and that is a failing I deeply regret.

We should also have had written guidelines on how to transfer work between groups, how to assign credit when papers are published and how and when project officers should communicate, especially during times of conflict. Implicit rules of behaviour in consortia are often exploited by more experienced participants.

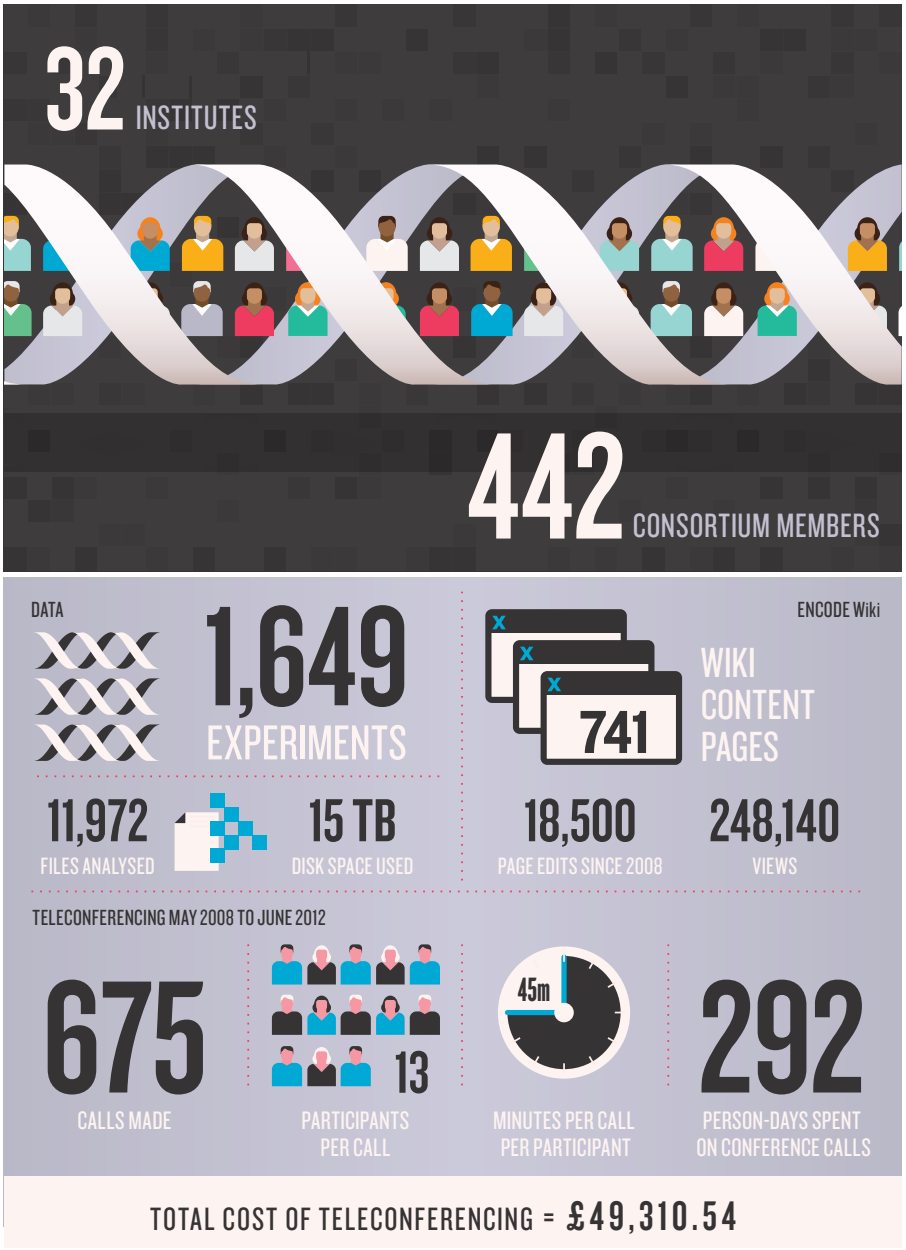
Large consortia clearly benefit from an open-door policy that allows new, unfunded analysts to participate. And when these individuals join the group or work with released consortium data, their analyses should be considered equally creditable and stigma-free relative to those performed by long-standing group members.

That brings us to error-catching. Big projects generate errors and have a range

THOMAS POROSTOCKY; SOURCE: MEETINGZONE

BY THE NUMBERS

The ENCODE project involved hundreds of people from around the world, and a lot of editing, disk space and phone calls.



of artefacts, so most researchers agree that data should be released to the larger community sooner rather than later. In ENCODE, we came to understand how time-consuming and involved quality control at scale is. It was not until around halfway through the process that we were able to assess the experiments retrospectively with a formalized, centralized, quality-control system. Most experiments were exemplary; some had to be redone. A few had to be left out.

The quality-control metrics and our final 'call' on whether a data set would be in or out is publicly accessible on the project website. Although important and biologically correct, some experiments scored low on quality-control metrics because they had, for example, very few true sites where a protein bound to DNA. Other sources of error, such as that from a cross-reactive antibody, generated excellent scores — the antibody 'worked' because it bound to a particular class of molecule, but it also bound to many others that were not predicted by the analysis. I wish now that we had accelerated the centralized quality-control process earlier, and been more open about this process.

Although most errors are caught within a consortium before they are released, new analysis of public data inevitably uncovers more, particularly early in data production. When analysing such early data, external groups should report such errors promptly and without rancour. Although funders need to measure data quality in a standardized way, during early data production consortia should really be judged not by absolute error rates, but by how quickly they can rectify reported errors.

Funders have considerable influence in how raw and analysed data are released, and should design policies that maximize reuse. Early data-release policies focused on how data should be shared before publication, with clumsy etiquette-based restrictions on the first publications of global analysis, such as waiting for the authors who generated the data to publish their analyses before others can publish on the entire data set. These agreements are starting to show their age and a lack of clarity.

The new era of analysis calls for a rethink, with more focus on the release of intermediate analysis throughout the project, so that the community can use the resource more fully during the project; the 1000 Genomes consortium has done well in this regard.

DOES IT DELIVER?

The overall importance of consortia science can not be assessed until years after the data are assembled. But reference data sets are repeatedly used by numerous scientists worldwide, often long after the consortium disbands. We already know of more than 100 non-consortium publications that make use of ENCODE data, and I expect many more in the forthcoming years.

Even if massive projects are successful, I feel strongly that the vast majority of funding should still go to smaller, more creative, hypothesis-led science.

For consortium participants, my call for more scrutiny, more clarity and more independent utilization of the data might seem restrictive, but I am confident that it will only benefit science and scientists in the long run. Even if large consortia receive only a small proportion of a discipline's funding, that can be a substantial amount when concentrated on a limited set of groups. If this is to continue, the entire community must be able to understand and use the resultant data.

ENCODE is a foundational data set for understanding the human genome. I am proud of what we have delivered, but there are things we could have done better. I hope that other groups can learn from our experience. ■

Ewan Birney *is lead ENCODE analysis coordinator and associate director of the European Molecular Biology Laboratory's European Bioinformatics Institute in Hinxton, UK.*
e-mail: birney@ebi.ac.uk