

- [ScienceWatch Home](#)
- [Inside This Month...](#)
- [Interviews](#)

- [Featured Interviews](#)
- [Author Commentaries](#)
- [Institutional Interviews](#)
- [Journal Interviews](#)
- [Podcasts](#)

Analyses

- [Featured Analyses](#)
- [What's Hot In...](#)
- [Special Topics](#)

Data & Rankings

- [Sci-Bytes](#)
- [Fast Breaking Papers](#)
- [New Hot Papers](#)
- [Emerging Research Fronts](#)
- [Fast Moving Fronts](#)
- [Corporate Research Fronts](#)
- [Research Front Maps](#)
- [Current Classics](#)
- [Top Topics](#)
- [Rising Stars](#)
- [New Entrants](#)
- [Country Profiles](#)

About Science Watch

- [Methodology](#)
- [Archives](#)
- [Contact Us](#)
- [RSS Feeds](#)



[Interviews](#)

[Analyses](#)

[Data & Rankings](#)

2009 : January 2009 - Author Commentaries : EBI's Ewan Birney: Quest for the Genomic Dragons

AUTHOR COMMENTARIES - 2009

January/February 2009



[+enlarge image](#)

EBI's Ewan Birney: Quest for the Genomic Dragons

Science Watch® Newsletter Interview

When the human genome project was completed in 2003, it provided the DNA sequences of the 3 billion base pairs that make up the genome, but that was effectively all it did. Researchers were left to translate that sequence data into meaningful information about the protein-coding genes that have been the focus of molecular biology for the past 40-odd years. They were left to guess at what this meant for the region of the genome that has traditionally been written off as "junk"—the DNA that sits between genes and inside of genes but plays no apparent role in the coding of proteins.

In 2004 an international collaboration known as ENCODE (Encyclopedia of DNA Elements) set out to address these issues by defining all functional elements in a representative 1% of the human genome—whether in the protein-coding regions or in the erstwhile junk. The results were published in June of 2007 in a summary paper in *Nature* and in 23 more detailed articles in the journal *Genome Research*. The *Nature* article, "Identification and analysis of functional elements in 1% of the human genome," (447[7146]: 799-816, 2007) quickly became a Hot Paper, racking up nearly 300 citations in roughly a year and half and duly taking up residence in this publication's Top Ten in Biology, where it currently ranks at #3. In the process, the paper helped catapult Ewan Birney, a scientist at the European Bioinformatics Institute in Hinxton, U.K., and first author on the ENCODE paper, into the #9 position in the current **Thomson Reuter's Essential Science Indicators**SM ranking of the hottest researchers in molecular biology & genetics. Birney's elite selection of 42 *ESI*-covered papers in the field since 1998, representing just a portion of his career output of more than 100 reports, has collectively tallied over 13,000 citations—averaging an amazing 307 citations per paper.

Birney, at 36, is already a veteran of a host of genome projects, from the human, mice, and rat genomes to the platypus and *Anopheles gambiae*, the mosquito that carries malaria. He received his bachelor's degree in biochemistry from Balliol College Oxford in 1996. He then spent the next four years working with Richard Durbin at the Sanger Centre, where he received his Ph.D. in 2000. Since then, Birney has been a senior scientist at the EBI, where he heads the Nucleotide Data division.

Birney spoke to Science Watch from his office at the EBI, just outside Cambridge.

SW: You started young as a PI at the European Bioinformatics Institute. What was your first major project?

I co-founded, with Tim Hubbard and Michelle Clamp, a project called Ensembl, which was and still is one of the major resources for using genomic information on the web. It presents the human genome to researchers—it's a major access portal. I'm still one of the principal investigators for that project and it still dominates my life quite a lot.

SW: How did ENCODE get started, and what was the original plan?

ENCODE was one of a series of projects that followed on the human genome project. At the end of 2003, people knew the human genome was going to be finished; they knew the path forward for the mouse and rat genomes. There was a kind of logical follow-up for other important organisms. The question was, what else do we need to do to really enable genomics? One project, for instance, which is kind of orthogonal to ENCODE, is the HapMap project to discover how variations in the genome occur in different humans. That's one thing we need to know. There are lots and lots of subtle differences. What are they? Let's just build a catalog.

The other side was understanding the genome better. Although we have some reasonable appreciation of protein-coding genes, everything else was a here-be-dragons kind of thing.

"Dark matter of the genome' is a better term than 'junk DNA,'" says Ewan Birney of the European Bioinformatics Institute. "It implies that we don't know what this stuff does."

SW: Here be dragons?

You know those old maps, where the known world ends and they just write "here be dragons." We have these huge expanses of genome and we just don't know what's going on. This is the noncoding DNA, in the parlance. And we want to know what this stuff is doing.

SW: Is this "junk DNA" we're talking about?

Well, the phrase "junk DNA" has morphed over time. Back in the 1970s, with the discovery of introns, junk DNA implied these large chunks of DNA that got transcribed into RNA, cut out, and then seemingly thrown away. As people in the 1980s began to rather painstakingly put together big chunks of genomic DNA, they started to see the layout of genes in the genome, and one of the first things they realized was that the protein-coding stuff doesn't make for much of it. Even in the densest part of the genome, it makes up maybe 10%. On average it makes 2%. Then you also find these disperse repeats, these parasitic elements that are found in every large genome, whether plants, fish, or humans. They have their own set of specific genome parasites, copying themselves happily across genomes. About half of our genome comes from these repeats. The phrase "junk DNA" started to be used interchangeably to mean these disperse repeats selfishly copying themselves, as well as all these other parts of the genome that we didn't understand. The phrase started to have a life of its own. It's not a very scientific term, although many scientists use it.

SW: What phrase do you prefer?

"Dark matter of the genome" is a better term. It implies that we don't know what this stuff does.

SW: So back to ENCODE. How was the collaboration put together, and who decided what techniques to use to analyze the genome?

The project was saying, in effect, let's just throw the kitchen sink of experimental techniques at this problem of understanding what the non-coding DNA does. Dream up any kind of useful experiment, propose it to NHGRI; peer review then says yes or no, and off we go to basically chart the here-be-dragons part of the genome, to discover what's going on in the dark matter. When the project started in 2004, we really didn't know which experiments were going to work and which would be too expensive to do in a whole genome, so the pilot project focused on 1% of the genome, divided up into 44 distinct regions. That sounds like a small percentage, but it's a lot of DNA. It's like if you want to study the Atlantic Ocean and you study just 1%, you'll probably learn a lot about it. Beginning in 2005, a dozen experimental groups were funded to look at this 1%. And one of the key rules was that everybody had to use the same 1%. That was critical. If we didn't do that—if everybody had chosen their own bit—it would have been a disaster.

SW: How did you end up first author?

I am not first author—the first author is the "ENCODE Project Consortium." I am first in the list of equals after that. About halfway through the project—and this is typical in biology these days—the real headaches shift from the experiments to the bioinformatics. Once you collect the data, you have

to aggregate it and store it sensibly. That's mostly boring plumbing and engineering; it's very tedious, but you have to get it right. Then the far more interesting aspect of understanding the data—which again is mainly bioinformatics—starts. I was originally funded to do a very small bit in ENCODE, but when everybody else assumed that the integration of the data would just magically happen, some of us from different laboratories essentially put our hands up and said, yes, we will dedicate our own effort to making this work. I was one of them. It turned into a pretty painful two years of work, and by the end, I was the first amongst equals in the author list, though many, many people contributed to the paper. The thing to stress is that the paper did have 308 authors.

Highly Cited Papers by Ewan Birney and Colleagues, Published Since 2000

(Ranked by total citations)

| Rank | Papers | Cites |
|------|---|-------|
| 1 | E.S. Lander, <i>et al.</i> , "Initial sequencing and analysis of the human genome," <i>Nature</i> , 409 (6822): 860-921, 2001. | 6,409 |
| 2 | R.H. Waterston, <i>et al.</i> , "Initial sequencing and comparative analysis of the mouse genome," <i>Nature</i> , 420(6915): 520-62, 2002. | 2,158 |
| 3 | A. Bateman, <i>et al.</i> , "The Pfam protein families database," <i>Nucl. Acids Res.</i> , 30(1): 276-80, 2002. | 1,289 |
| 4 | A. Bateman, <i>et al.</i> , "The Pfam protein families database," <i>Nucl. Acids Res.</i> , 28(1): 263-6, 2000. | 836 |
| 5 | G.M. Rubin, <i>et al.</i> , "Comparative genomics of the eukaryotes," <i>Science</i> , 287(5461): 2204-15, 2000. | 811 |

SOURCE: Thomson Reuters *Web of Science*®

SW: So what did ENCODE tell us about the nature of the genome?

ENCODE didn't have the neatest take-home message, although we did have bullet points in the *Nature* paper explaining what we had learned. When I discussed all this with a U.K. science journalist, he ended up using the phrase "boffins are baffled," which about captures it.

The major thing we learned is that the genome is a very complex place, including all these here-be-dragons areas, this dark matter. It's complex and multilayered. We certainly removed the idea that the genome is a simple thing, with very discrete units of genes and discrete regulatory information not in the genes, all very neatly packaged. When you look at the ENCODE data, it doesn't hold up. Introns of many genes, for instance, are just alive with regulatory information. One of the more challenging things we discovered is that we see lots more RNA being transcribed than we ever expected, and we really don't have a proper way of understanding and classifying this RNA. So RNA is being transcribed not just in the traditional way, into messenger RNA that makes proteins, but it's also happening in bits between genes and, weirder still, it's crossing between gene boundaries, going from one region that we're pretty sure is not involved with protein coding, and then going into a gene or out of a gene.

SW: Can you describe that in a little more detail?

Some RNA transcripts start outside a gene and then go into the intron of another gene, but they don't actually seem to make an exon. Rather they overlap with the exon of that gene. Some transcripts go the other way: they start inside the intron of one gene, go through that gene a couple of introns, and then stop. Even more amazing, some transcripts start in the intron of one gene and go for one megabase, crossing four or five other genes, and then somehow, in a way we don't understand, start making exons in another gene. They join two genes which have four genes in between them at a very long genomic distance.

SW: Any chance this is being interpreted incorrectly?

When I started, I was one of the real arch skeptics of this RNA data. I now firmly believe that these RNAs exist. But what they do, if anything, is something I can get into very long debates about. So this will lead into what I think is the most interesting discovery from ENCODE.

SW: Which is?

Well, the genome is alive with what are called DNaseI hypersensitive sites, which are sites on the genome that bind transcription factors. This was not a surprise. It was completely expected that these sites are all over the genome, even in lots of places that didn't have a gene. The surprise is that only about half the sequences under these DNaseI hypersensitive regions are conserved across mammals. The other half are not. The weird RNA described above is even less conserved than this. That's pretty odd, because, for protein-coding genes, maybe 95% are conserved. So this is really weird. We just didn't

expect this. What on earth is going on?

SW: So what on earth is going on?

What I believe, and what we put this in the paper as the last of the bullet points, is that, in fact, many of these elements arise by chance, and once having done so, they're neither good nor bad for the organism. Then they just hang around until, by chance again, they disappear. So this is a very interesting idea, that there is this big pool of neutral stuff that's biochemically active but not actually important now for the organism.

SW: If you looked again in 10 million years, the 50% that's not conserved would look entirely different?

That would be the logic. Interestingly enough, very similar results have been seen in *Drosophila*. One interpretation of this is that maybe biology has tuned itself to create a pool of elements that could be used by evolution in the future. The trouble with that hypothesis, and the reason why theoreticians of evolutionary biology will poo-poo it, is that evolution isn't supposed to have foresight. It can't say, for example, that we'd better get evolving these opposable thumbs because we might need them in the future to hold a telephone. Evolution works on the instantaneous here and now. On the other hand, the counter argument to this is that if you take the hard line, you can't justify sex. You can't justify why an individual would throw away half its genes to combine with some other individual's half. Surely that's a bad idea from the gene's perspective, to roll the dice every time whether it will make it to the next generation. Surely it's not a good thing for the gene. We hint at this idea in the paper, again without getting stuck in it. But for me, this is the most interesting thing that has come out of ENCODE.

When people wrote about ENCODE, most of them picked up on the idea that we are redefining the word "gene," what with the RNA running in and out of genes. That was the easiest thing to understand and put across. This evolutionary stuff is genuinely harder to get your head around, but, from my point of view, it's the most important thing.

SW: What can we expect from ENCODE in the future?

It's being scaled up to cover 100% of the genome. The grants were awarded last year. One of the great things about genomics is that it will cost about the same to do this, perhaps even a little less, than it did to do 1% a few years ago. And we'll do it twice as fast. The reason is we now have all these ultra-high-throughput sequence machines. That's just completely changed the game and made these experiments really cost effective. ■

Related information: Ewan Birney talks with *ScienceWatch.com* and answers a few questions about his Fast Moving Front in the field of Computer Science in July 2006.

Keywords: Ewan Birney, EBI, European Bioinformatics Institute, ENCODE, junk DNA, introns, RNA transcription, DNaseI, Pfam, Genomics.



PDF

[back to top](#)

2009 : January 2009 - Author Commentaries : EBI's Ewan Birney: Quest for the Genomic Dragons