GENOMICS

Too much information? Not for long

The far-reaching, National Human Genome Research Institute backed ENCODE project hopes to advance the state-of-the-art of genomic analysis and derive the definitive functional index of the human genome.

Sequencing the human genome was a landmark achievement, yet the raw information alone is limited in usefulness, like a massive directory with only telephone numbers but no names or addresses. Genomic data are also available for several animal species, but these also underline just how little is truly known. "Only 5% of [human] sequence is conserved in mammals, and only 1.5% seems to be coding sequence out of that conserved 5%," says Elise Feingold of the National Human Genome Research Institute (NHGRI). "So what is that other three and a half percent doing, and what is the rest of the genome doing?"

Feingold is on the Scientific Management team for the ENCODE (ENCyclopedia Of DNA Elements) project, a multi-institutional and multinational initiative to develop a comprehensive directory of functional elements contained within the human genome, including protein-coding and non-protein-coding expressed sequences, regulatory elements and so forth.

The ENCODE project has begun to address this daunting challenge with two initial phases. The 'pilot' phase entails the rigorous analysis of 1% (30 megabases) of the human genome by a broad range of existing technologies. This 1% includes sequence from several non-contiguous regions, both closely characterized segments and others selected at random. To account for differences between individuals, sequence variations in conserved regions will be determined from the 48 samples being used by the HapMap consortium. Meanwhile, groups involved in the 'technology development' phase are working on innovative technologies to enhance the pursuit of high-quality data. The outcome of these two phases will ultimately determine the conduct of the monumental 'production' phase, wherein the other 99% of the genome will be studied with equal rigor.

As ENCODE project data are verified, they will be made freely available via the UCSC Genome Browser and other databases, and some of the new techniques being developed by consortium members are already making their way into publication (see Dorschner *et al.*, pp. 219–225). Feingold says funding for the initial phases currently covers three years, and she anticipates a far longer road beyond that—but she is already greatly encouraged by the findings and collaborations that have emerged to date, indicating that getting there may be much more than half the fun.

Michael Eisenstein

RESEARCH PAPERS
ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).
WEB SITES
The ENCODE project homepage: http://www.genome.gov/ENCODE
The UCSC genome browser: http://genome.cse.ucsc.edu/ENCODE

GENOMICS

HIGH-THROUGHPUT MATCHMAKING

By combining the yeast one-hybrid assay with Gateway cloning techniques, a research group at the University of Massachusetts set up a screen in Caenorhabditis elegans with high-throughput potential for matching transcription factors and promoters.

If genes could post personal ads, they would read something like this: "Good-looking promoter seeking transcription factor for binding experience with promise of regulated gene expression." Who wouldn't be interested in the reply? In the absence of such correspondence, biochemical screens have to reveal matches made between transcription factors (TFs) and promoters. In a recent article in *Genome Research*, Marian Walhout from the University of Massachusetts introduced a modification to the yeast one-hybrid (Y1H) assay, which opens the door to the high-throughput characterization of TF-promoter interactions (Deplancke *et al.*, 2004).

In a traditional Y1H assay, a yeast strain is transformed with a bait vector containing multiple copies of short *cis*-regulatory elements followed by a reporter gene and a prey vector that expresses the TF. If the TF binds the regulatory sequence, expression of the reporter gene is initiated, and its product confers a growth advantage to the yeast strain. Using the Gateway system, which is based on homologous recombination rather than restriction enzyme digests, Walhout's team was able to greatly increase the speed of cloning so that they can now generate TF prey libraries and large numbers of bait clones. Also, instead of using *cis*-regulatory elements, their baits consist of whole promoter sequences. As Walhout points out, "This opens up a whole new level of possibility for systems biology.…You can just take a promoter without knowing anything about it and try to identify transcription factors that can bind to that promoter."

She chose *C. elegans* as a model system because the genome is very well annotated, but she sees no reason why this approach could not be used in other organisms as well. Initially, Walhout screened only four promoters against a TF library, thus validating the technique, but in the near future her group will apply this method to high-throughput analysis of many promoters. When asked for the motivation behind the study, she replied, "We want to understand how differential gene expression is controlled in space and time during development and, long term, also during homeostasis and in disease." A tall order, but with the new Y1H assay it is certainly doable.

Nicole Rusk

RESEARCH PAPERS
Deplancke, B. *et al*. A Gateway-compatible yeast one-hybrid system. *Genome Res*. **14**, 2093–2101 (2004).