

## GENOMICS

# Unknown polypeptides galore

**A bounty of short peptides found in cells points to a larger, more mysterious proteome.**

Proteins are known movers and shakers in cells, but the function of one set of their relatives, so-called SEPs—which are short polypeptides that are genetically encoded from short open reading frames (sORFs)—is poorly understood.

Genetic screens have revealed sORF-encoded polypeptides in bacteria, viruses, plants, yeast, worms, insects and humans, in which they control processes as varied as cell death and glucose uptake. Discovery of a new crop of SEPs shows they are more plentiful than previously assumed. The work also reveals a weak spot in gene-finding algorithms and indicates these molecules might also be pervasive in disease.

Peptides in this size range are hard to identify using mass spectrometry, which relies on protein databases full of annotated

genes. But by linking high-throughout RNA sequencing with peptide profiling using mass spec, a team of researchers at Harvard University and the Broad Institute identified 90 human SEPs in a human leukemia cell line called K562. Practically all of the SEPs, which are between 18 and 149 amino acids long, have not been described before and have unknown function.

One resource that helped the scientists characterize these SEPs was a custom proteomics database with RNA-seq data from K562 cells, says Harvard University biochemist Alan Saghatelian, who led the work. “This custom database contained every possible polypeptide that could be generated from K562 RNA, which enabled the discovery of unannotated polypeptides.”

The team believes their results show that the human proteome is larger than previously believed. The identification of additional sORFs “underscores the limitations of current

gene-finding algorithms for sORF identification,” Saghatelian says. For example, fewer than half of the genes use ATG as a start codon, which provides “a potential explanation for why some of these sORFs may have been missed.”

Another important find is that regions of RNA considered to be noncoding and a small fraction of noncoding RNAs also produce SEPs, both of which shows that the understanding of protein translation is still incomplete, Saghatelian says.

Next on the agenda is the development of new strategies to characterize the biology of these polypeptides and to then discern which SEPs are associated with disease.

**Vivien Marx**

#### RESEARCH PAPERS

Slavoff, S.A. *et al.* Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nat. Chem. Biol.* advance online publication (18 November 2012).