Gene Index analysis estimates the human genome contains 120,000 genes

F. Liang *et al*.

Nature Genet. **25**, 239–240 (2000).

In a recent Letter, we used the TIGR Human Gene Index (HGI, release 5.0) to estimate that the human genome contained 120,000 genes. The estimate of the number of genes, N_{genome}, relied on five parameters: the number N_{THCs} of tentative human consensus sequences (THCs) containing ESTs; the estimated redundancy R of the THC data set; the estimated number $N_{ESTgenes}$ of previously identified human genes represented by ESTs; the estimated total number N_{genes} of distinct known gene sequences within the available data set; and r, the 'enrichment factor' reflecting the representation of the data set relative to the genome average:

$$N_{genome} = \frac{1}{r}\,\frac{N_{THCs}}{R}\,\frac{N_{genes}}{N_{ESTgenes}}$$

A reassembly of HGI (release 6.0) allowed us to identify a source of error in our original estimate. HGI 6.0 contains 83,892 THCs constructed from 1,746,022 ESTs and 58,329 gene sequences. The latter number includes human transcripts (NP sequences) parsed from GenBank records plus curated expressed transcript (ET) sequences from the TIGR EGAD database. Of 83,892 THCs, 82,274 THCs contain at least 1 EST. Our analysis of HGI 6.0 indicates that the percentage of genes sampled by ESTs is much higher than the estimate based on HGI 5.0. Of 58,329 NP/ET sequences, 45,767 appear in 12,731 THCs containing ESTs; 5,838 appear in 1,618 THCs containing only NP/ET sequences; and 6,724 remain as singletons. A new detailed analysis of these singletons indicates that most encode immunoglobulin heavy/light chain variable regions. After collapsing immunoglobulin-like genes and other potential gene families and alternative transcripts through high-stringency clustering, the number of unique singleton genes drops to 1,216. This suggests that the gene sequence data represents 15,565 (12,731 + 1,618 + 1,216) unique genes, of which 12,731 (82%) have been sampled by EST sequencing projects. With N_{THCs}=82,274; R = 1.3[5]; $N_{ESTgenes}$=12,731, N_{genes}=15,565 and r = 1 (by definition as we are sampling the entire genome), we estimate the total number of human genes is N_{genome}=81,273, a value much lower than our previous estimate.

We also re-estimated the number of genes using a second, independent method. We mapped the THCs from HGI to chromosomes 21 and 22 as follows. First, we masked repeats in the published chromosome 21 and 22 sequence using RepeatMasker (A.F.A. Smit and P. Green, unpublished data; http://repeatmasker.genome.washington.edu) and then we searched HGI against the chromosomal sequence as described previously. THC sequences having ≥95% identity with the chromosomal sequence over >80% of their length were considered positives. A total of 2,232 THCs were mapped to chromosomes 21 and 22 with an average 99.2% identity. To account for possible redundancy in aligning THCs to the chromosome due alternative splicing and mispriming, we 'condensed' THCs that mapped to chromosomal locations within 5,000 bp of each other. This yielded 1,275 distinct THC hits. The total number of annotated genes plus pseudogenes on chromosomes 21 and 22 is 953 (770+183); based on these numbers, the publications describing chromosomes 21 and 22 estimated 40,000 and 45,000 total genes[3,6], giving an average extrapolated estimate of 42,500. If one assumes (conservatively) that the relative number of pseudogenes in the EST collection is the same as those predicted for chromosomes 21 and 22, then the revised estimate of the number of human genes should be: (42,500*1,275/953)=56,960. These improved estimates provide a lower bound of 56,960 and an upper bound of 81,273 genes in the human genome.

Mutations in *AXIN2* cause colorectal cancer with defective mismatch repair by activating β-catenin/TCF signalling

W Liu *et al*.

Nature Genet. **26**, 146–147 (2000).

The corrected list of authors is: Wanguo Liu, Xiangyang Dong, Ming Mai, Ratnam S. Seelan, Ken Taniguchi, Kausilia K. Krishnadath, Kevin C. Halling, Julie M. Cunningham, Lisa A. Boardman, Chiping Qian, Eric Christensen, Shauna J. Schmidt, Patrick C. Roche, David I. Smith & Stephen N. Thibodeau

Loss-of-function mutations in the EGF-CFC gene *CFC1* are associated with human left-right laterality defects

R N Bamford *et al.*

Nature Genet. **26**, 365–369 (2000).

The corrected list of authors is: Richard N. Bamford, Erich Roessler, Rebecca D. Burdine, Umay Saplakoglu, June dela Cruz, Miranda Splitt, Judith A. Goodship, Jeffrey Towbin, Peter Bowers, Giovanni B. Ferrero, Bruno Marino, Alexander F. Schier, Michael M. Shen, Maximilian Muenke & Brett Casey

J.A. Goodship's affiliation: Department of Human Genetics, University of Newcastle upon Tyne, Newcastle upon Tyne, UK. G.B. Ferrero's affiliation: Department of Pediatrics, University of Torino, Torino, Italy.

This work was supported by grants from the British Heart Foundation (to J.A.G.), and the Telethon (E.434) and MURST (9906314313; to G.B.F.).