Supplementary Information

Methods

We have surveyed 82 prokaryotes and 19 eukaryotes and illustrated in the Supplementary Table the list of genome sequence size (N) in kilo-base-pair (kb), the number of predicted genes (n), the ratio of coding sequence over the genome sequence (r) for each of these species and validation information about the eukaryotes' genome coding sequences together with the key reference from which the information was collected. From the data, we calculated the mean length of genic coding sequence (MLGCS) within each of the species as $L = N \times r / n$, and regressed the estimates of L on the size of genome coding sequence ($N \times r$) for each of two species groups.

Supplementary Table 1. Genome Sequence Data and Data Validation Information for 19 Eukaryotes

Species	**Estimated genome size[a] (N, kb)**	**Estimated gene number[a] (n)**	**Estimated ratio of coding sequence [a] (r)**	**Data validation Information[b]**
A. gambiae (mosquito)	278244.06	13683	0.070	Reliability of gene prediction was tested by EST and homology comparison and at least 70% accuracy was guaranteed.
A. thaliana	115409.95	25498	0.288	As high as 80 percent of gene structure predicted by three independent centres involved were completely consistent, 93% of ESTs matched gene models.
A. gossypii (ascomycete)	9200.00	5006	0.812	More than 90% of *A. gossypii* genes show both homology and a particular pattern of synteny with *S. cerevisiae,* which has been adequately annotated.

C. elegans (Nematode)	97000	19320	0.258	92% of the predicted introns have an exact match to the experimentally confirmed ones, and approximately 76% of predicted proteins have other nematode proteins or out nematode proteins homologues.
C. Intestinalis (Ascidians)	160000	16000	0.132	More than 75% of the predicted genes have cDNA matches.
C. merolae (Red alga)	16520	5331	0.449	All genes were identified from mapping 99.85% ESTs onto the genome sequence and 86.3% of the predicted genes have corresponding ESTs.
D. discoideum	8100	2872	0.563	EST, protein, and/or InterPro matches provide support for approximately 70% of predicted genes.
D. melanogaster (Fruit fly)	120000	13600	0.150	There are EST matches for 65% of the predicted genes.
F. rubripes (Torafugu)	365000	32000	0.115	Approximately 75% of Fugu genes predicted has a homologue in the human genome by tblastx.
H. sapiens (Human)	3000000	35000	0.015	Supporting evidence for predicted human genes includes known genes, genes with good protein or EST homology. Around 81% of the genes in the RIKEN mouse set show sequence similarity to the human genome sequence, whereas 69% show sequence similarity to the IGI/IPI. This suggests a sensitivity of 85% (69/81). About 69% of the IGI (human) matches the RIKEN cDNA set (rodents).
M. musculus (Mouse)	2500000	35000	0.020	Approximately 99% of mouse genes annotated have a homologue in the human genome. For 80% of mouse genes, the best match in the human genome in turn has its best match against that same mouse gene in the conserved syntenic interval.
N. crassa (fungus)	40000	10082	0.376	Predicted genes are validated against ESTs aligned to the genome.
O. sativa (chr1)	45746	6756	0.183	A test made on chromosome I showed that 71% of predicted genes have homologue to a domain, a functional site, a cereal EST or a protein.
R. norvegicus (Rat)	2750000	33000	0.017	90% of rat genes possess strict orthologues in both mouse and human genomes.
S. cerevisiae (yeast)	12100	5726	0.705	Comparative analysis with three different yeast genomes indicates that *S. cerevisiae* contains 5,726 genes.
G. gallus (Chicken)	1063000	23000	0.031	Among the genes predicted, 80% had cDNA data support and covered more than 80% coding sequence in the genome.
C. briggsae (nematode)	104000	19500	0.232	Approximately 96% of predicted genes had orthologues to those in *C. elegans*.
P. chrysosporium (fungus)	30000	11777	0.450	72.1% of predicted genes have sequence similarity to GenBank proteins, and other 6.4% can be aligned to conserved protein domains under InterPro scanning.
T. nigroviridis (teleost fish)	340000	27918	0.101	There were 90.5% orthologous proteins detected between Tetraodon and Takifugu.

References:

1 a: Holt, R.A. et al. Science 2002. 298: 129-149. b: Evgeny M. Zdobnov, et al. *Science* 2002. 298: 149-159.
2 a,b: Arabidopsis Genome Initiative. *Nature* 2000. 408: 796-815.
3 a,b: Fred S Dietrich, et al. *Science* 2004. 304: 304-307. Supporting online material.
4 a,b: C. Elegans Sequencing Consortium. *Science* 1998. 282: 2012-2018.
5 a,b: Dehal, P. et al. *Science* 2002. 298: 2157-2167. a: Simmen, M.W. et al. *Proc Natl Acad Sci U S A* 1998. 95: 4437-4440.
6 a,b: Motomichi Matsuzaki, et al. *Nature* 2004. 428: 653-657.
7 a,b: Glockner, G. et al. *Nature* 2002. 418: 79-85.
8 a,b : Adams, M.D. et al. *Science* 2000. 287: 2185-2270.
9 a,b: Aparicio, S. et al. *Science* 2002. 297: 1283-1285. a: Jun Yu, et al. *Science* 2002. 297: 1301-1310. Web Supplement 1
10 a,b: Lander, E.S. et al. *Nature* 2001. 409: 860-921. b: Venter, J.C. et al. *Science* 2001. 291: 1304-1351.
11 a,b: Waterston, R.H. et al. *Nature* 2002. 420: 520-562. a: Alexander E. Vinogradov. *J Mol Evol* 1999. 49: 376-384.
12 a,b: Galagan, J.E. et al. *Nature* 2003. 422: 859-868.
13 a,b: Takuji Sasaki, et al. *Nature* 2002. 420: 312-316.
14 a,b: Rat Genome Sequencing Project Consortium. *Nature* 2004. 428: 493-521.
15 a: Goffeau, A. et al. *Science* 1996. 274: 546,563-567. a:Wood, V. et al. *Nature* 2002. 415: 871-880. a,b: Manolis,K. et al. Nature 2003. 423: 241-254.
16 a,b: International Chicken Genome Sequencing Consortium. *Nature* 432, 695-716(2004). a: Wallis1, J.W. et al. *Nature* 2004. 432: 761-764. a: Jun Yu, et al. *Science* 2002. 297: 1301-1310. Web Supplement 1
17 a,b: Lincoln.D.S, et al. *PLoS Biology* 2003. 1: 166-192.
18 a,b: Diego M, et al. *Nature Biotechnology* 2004. 22: 695-700.
19 a,b: Jaillon1 O, et al. *Nature* 2004. 431: 946-957.