Washington University School of Medicine
Digital Commons@Becker

Open Access Publications

2005

Overview of gene structure

John Spieth
Washington University School of Medicine in St. Louis

Daniel Lawson
Sanger Institute

Recommended Citation

Spieth, John and Lawson, Daniel, ,"Overview of gene structure." WormBook.,. 1-10. (2005).
http://digitalcommons.wustl.edu/open_access_pubs/3466

Overview of gene structure[*]

John Spieth[§], Genome Sequencing Center, Washington University School of Medicine, St. Louis, MO 63108 USA

Daniel Lawson, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Sanger Institute, Hinxton, Cambridge, CB10 1SA UK

Table of Contents

Abstract

Throughout the *C. elegans* sequencing project Genefinder was the primary protein-coding gene prediction program. These initial predictions were manually reviewed by curators as part of a “first-pass annotation” and are actively curated by WormBase staff using a variety of data and information. In the WormBase data release WS133 there are 22,227 protein-coding gene, including 2,575 alternatively-spliced forms. Twenty-eight percent of these have every base of every exon confirmed by transcription evidence while an additional 51% have some bases confirmed. Most of the genes are relatively small covering a genomic region of about 3 kb. The average gene contains 6.4 coding exons accounting for about 26% of the genome. Most exons are small and separated by small introns. The median size of exons is 123 bases, while the most common size for introns is 47 bases. Protein-coding genes are denser on the autosomes than on chromosome X, and denser in the central region of the autosomes than on the arms. There are only 561 annotated pseudogenes but estimates but several estimates put this much higher.

[*] Edited by Philip Anderson and Jonathan Hodgkin. Last revised March 30, 2005. Published January 18, 2006. This chapter should be cited as: Spieth, J. and Lawson, D. Overview of gene structure (January 18, 2006), *WormBook*, ed. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.65.1, http://www.wormbook.org.

[§] To whom correspondence should be addressed. E-mail: jspieth@watson.wustl.edu

1. What is a gene?

Sidney Brenner, the founder of modern worm biology, once said, “Old geneticists knew what they were talking about when they used the term ‘gene’, but it seems to have become corrupted by modern genomics to mean any piece of expressed sequence…” (Brenner, 2000). Sidney's lament serves to illustrate two points. The first is that the concept of a gene can mean different things to different people in different contexts. The second is that the concept of a gene has been evolving, not only in the modern genomic era, but ever since it first appeared in the early 1900s as a term to conceptualize the particulate basis of heritable physical traits (Snyder and Gerstein, 2003). Therefore, in a review of gene structure in *C. elegans* it seems prudent to define what we mean by a gene.

Our definition of a gene is necessarily heavily influenced by modern genomics, but we prefer to think it has not been corrupted by it. We define a gene as “the complete sequence region necessary for generating a functional product”. This encompasses promoters and control regions necessary for the transcription, processing and if applicable, translation of a gene. Hence, we include not only protein-coding genes (genes that encode polypeptides), but also non-coding RNA genes (ribosomal RNA, transfer RNA, micro RNA and small nuclear RNA genes). One additional type of gene we will briefly discuss is the pseudogene, though usually these are not considered to be functional.

The full extent of a gene, or the “complete sequence region” is not known for most *C. elegans* genes because promoters remain, for the most part, incompletely defined. Even the full extent of the primary transcript is frequently not known because a majority (70%) of protein-coding genes are rapidly modified by trans-splicing, which involves the addition of a short 22 nt exogenous RNA species to the 5’ end of a transcript (Zorio et al., 1994). Recently, it has been shown that some non-coding RNA genes are also trans-spliced; a precursor of the microRNA *let-7* was identified with a trans-splice leader sequence (Bracht et al., 2004).

2. Protein-coding genes

Protein-coding genes are the largest class of genes in the *C. elegans* genome, and probably the most interesting to the majority of people, so we will cover these genes first.

2.1. Prediction and curation

Throughout the *C. elegans* sequencing project Genefinder (Green and Hillier, unpublished software) was the gene prediction program of choice. Genefinder is an *ab initio* predictor and requires only a genomic DNA sequence and parameters based on a training set of confirmed coding sequences. Note that Genefinder, like most other gene prediction tools, is actually a coding-sequence (CDS) predictor and does not attempt to locate promoters, or untranslated regions (UTRs). All Genefinder predictions were appraised by human curators as part of the ‘first-pass annotation’ prior to the publication of the genome in 1998 (The *C. elegans* Sequencing Consortium, 1997). The quality of CDS predictions has improved over the course of the sequencing project as better training sets of confirmed CDS were generated and improved versions of Genefinder became available.

With the completion of the *C. elegans* genome sequence and funding for WormBase (NHGRI and MRC sources), a renewed effort to improve the gene predictions was initiated. These ‘curated’ gene structures are available through WormBase (http://www.wormbase.org/) and public nucleotide and protein databases (GenBank/EMBL/DDBJ/UniProt).

The generation and maintenance of a gene prediction data set is always a ‘work in progress’. WormBase has invested heavily in correlating transcript data (experimentally confirmed coding sequences) with gene predictions. Hence, all messenger RNA (mRNA), Expressed Sequence Tag (EST) (Yugi Kohara, unpublished; http://nematode.lab.nig.ac.jp/) and Orfeome Sequence Tag (OST) (Reboul et al., 2003; Vaglio et al., 2003) sequences are routinely mapped to the *C. elegans* genome and compared to the current set of gene predictions. Annotators then modify the exon/intron structure of the prediction to accommodate any changes highlighted by new transcript data. Other types of information used in gene curation include protein alignment data (with special weighting to matches within *C. elegans* and related nematode species such as *C. briggsae* (Stein et al., 2003), repeat sequences and sequence features such as trans-splice leaders and poly(A) sites.

Curation endeavors to continually improve the accuracy of the gene structures. How accurate are the gene structures? In WormBase release WS133 of September 24, 2004, 6,202 CDS predictions (28%) have every base of every exon confirmed by some type of transcription evidence, showing the gene is real and the structure correct. An additional 11,459 (51%) have at least one base of an exon confirmed by transcript data, indicating the gene is real and part of the structure is correct. The remaining 21% of the CDS predictions currently have no EST or mRNA support but could have underlying protein alignments or strong sequence conservation with *C. briggsae* (Stein et al., 2003).

2.2. Gene number and sizes

There were 22,227 protein-coding genes on Sept 24, 2004 (WormBase data release WS133), including 2,575 alternatively-spliced forms. This is up from 19,099 in 1998 when the genome was declared essentially complete (The *C. elegans* Sequencing Consortium, 1997), primarily due to new transcript data indicating the existence of previously undetected new genes, the splitting of existing genes, or the detection of conserved sequences between the *C. elegans* and *C. briggsae* genomes that are predicted to be coding (Stein et al., 2003).

How many additional genes might be missing from WormBase? It's hard to know for certain, but an estimated 1,119 new genes come from 2,228 gene predictions made by TWINSCAN (Korf et al., 2001); these gene predictions do not overlap existing WormBase genes and an RT-PCR success rate of 55% confirms a subset of the novel genes (Wei et al., 2005, in press).

The average *C. elegans* protein-coding gene is compact in comparison to vertebrate genes. Most *C. elegans* genes are relatively small (Figure 1), covering a genomic region of approximately 3 kb (from start to stop codon including introns), however there are some very large genes, which skew the average. The median size is only 1,956 bases with a range from 48 bases (Y108G3AL.6, confirmed by transcript data) to 80,957 bases (W06H8.8g, the *C. elegans* titin gene).

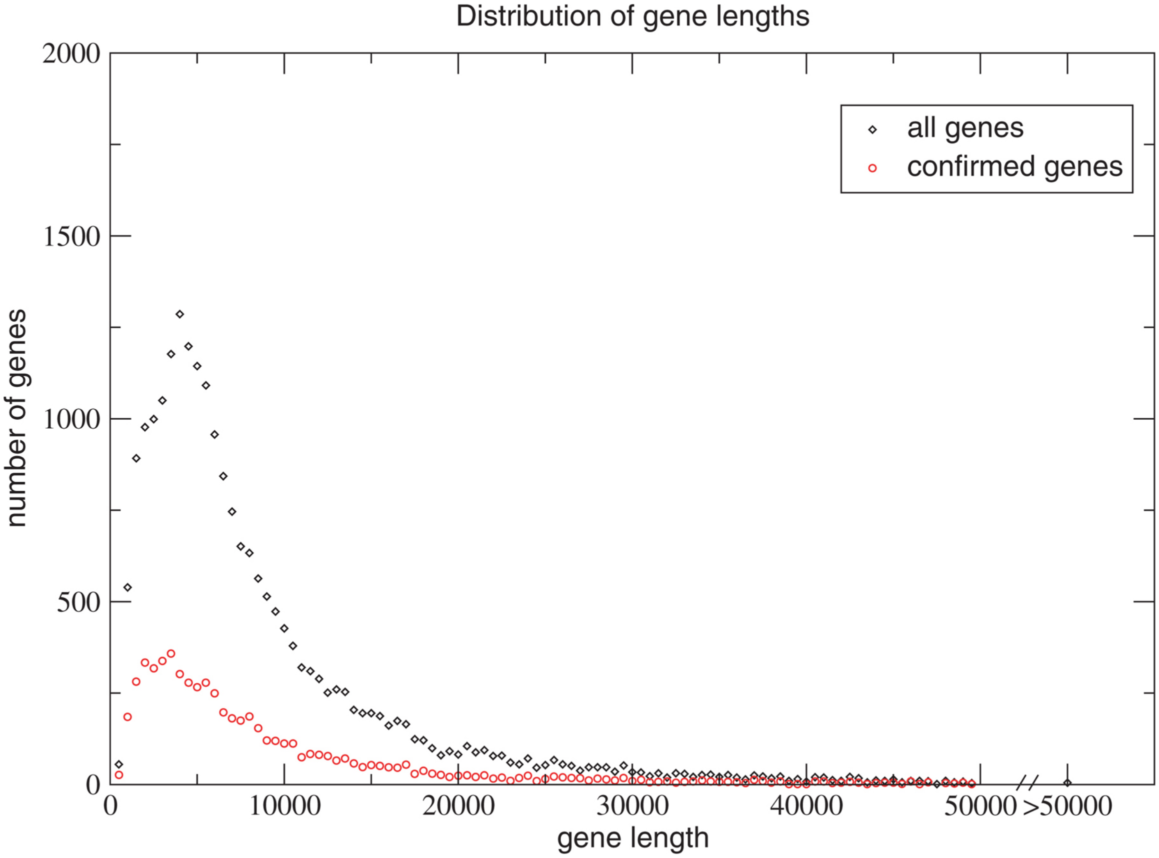

Figure 1. Distribution of gene lengths for confirmed and all genes. Data was obtained from Wormbase data release WS133. Each data point represents the number of genes in each size class in increments of 500 bases: 1–500, 501–1000, etc.

The distribution of gene sizes for confirmed genes is nearly identical to that for all genes (Figure 1) suggesting that Genefinder does not significantly over-or under-predict the size of genes.

2.3. Exons

C. elegans genes, like most eukaryotic protein-coding genes, contain exons separated by introns. There are 126,477 predicted unique, coding exons (the same exon used in alternatively-spliced isoforms of the same gene is considered as one unique exon) in the WS133 protein-coding gene set, which account for 25.55% of the genome, considerably more than the 1.5% estimated for the human genome (The International Human Genome Sequencing Consortium, 2001).

The average gene contains 6.4 coding exons, 6.0 if only genes that are confirmed by ESTs or mRNAs are considered; however, there are a few genes with a large number of exons (Figure 2). W06H8.8g, an isoform of the titin gene, has 66 coding exons. If only confirmed genes are considered then the gene with the largest number of exons (62) is F15G9.4b, an isoform of *him-4*. There are also a few single exon genes (570 in WS133) amounting to about 3% of total genes. Almost 60% of these are supported by EST or mRNA data.

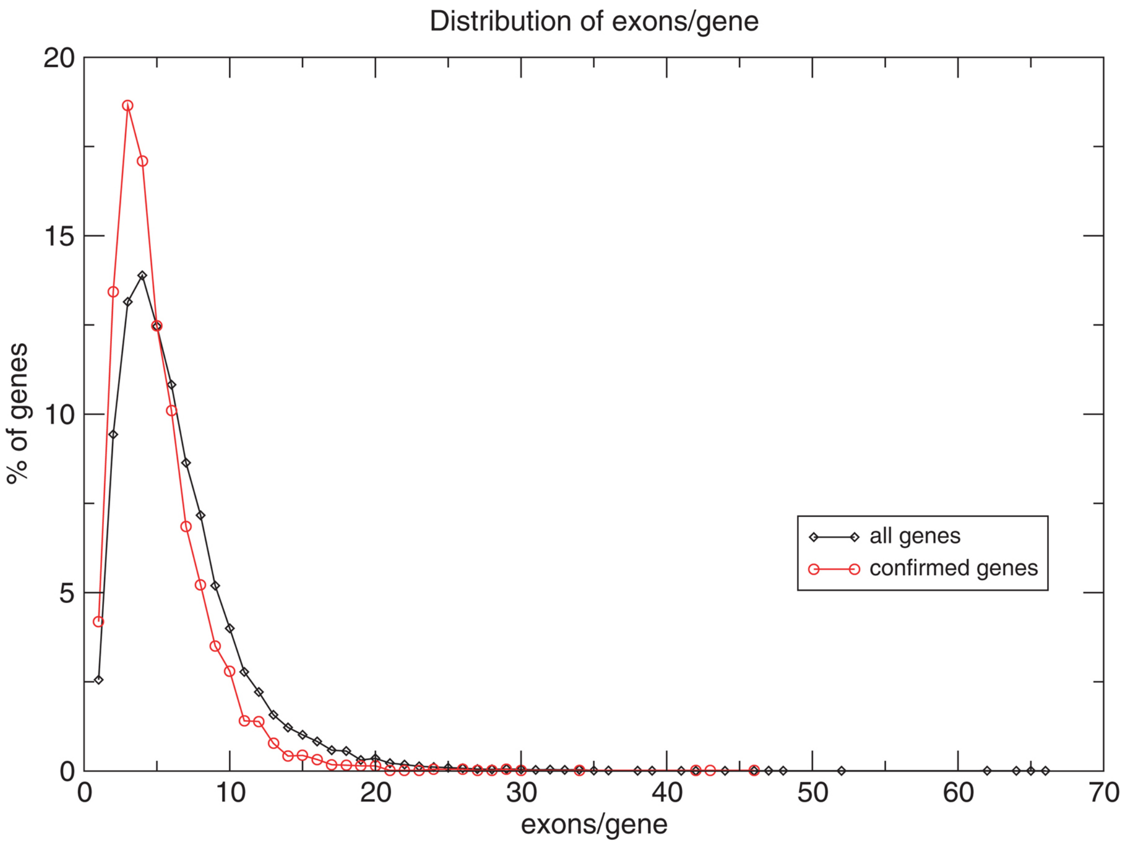

Figure 2. Distribution of the number of exons per gene in all and confirmed genes. Data was obtained from Wormbase data release WS133. Each data point represents the percent of genes having a specific number of exons.

The average size of unique exons in all protein-coding genes is 208 bases, but there are a small number of very large exons. Again, as with gene size, these few large exons skew the average. The median size is only 123 bases, thus exons are similar in size to exons in human and fly genes (The International Human Genome Sequencing Consortium, 2001). The average size of unique exons in confirmed genes (201 bases), the median size (144 bases) and the distribution (Figure 2) is very similar to that in all worm genes, and to an earlier study based on 862 *C. elegans* exons in GenBank entries (Blumenthal and Steward, 1997). The largest exon in a confirmed gene is 7,569 bases found in *pqn-43* (F54E2.3b). The largest exon in all genes is 14,975 bases, found in an isoform of the *unc-44* gene. The smallest confirmed coding exon is 3 bases, found in F54C1.3b, an isoform of *mes-3*.

2.4. Introns

There are 106,909 predicted unique introns (the same intron used in different isoforms or spliced variants is considered as one unique intron) in all of the protein-coding genes of *C. elegans* (WS133 release). Some of these are probably not real introns or have incorrect boundaries because they are either predicted only by Genefinder or based on imperfect alignments of cDNA or single-pass EST reads. Of these, 824 are less than 30 bases, almost all of which probably result from erroneous EST alignments in WormBase. 67,833 introns are considered confirmed because there is EST or cDNA sequences spanning the intron boundaries. The most common size of confirmed introns is 47 bases with the median size being 65 bases. The smallest confirmed intron is only 10 bases. It is found in the 3’ UTR of *mag-1* (R09B3.5), has good splice acceptor and donor sites (5’-CAAAAA/gtacagttag/AAAAG-3’) and is supported by an mRNA sequence and 3 separate EST clones. The largest confirmed intron is 21,230 bases, found in *kin-1* (ZK909.2). It is confirmed by a single EST clone containing the SL1 trans-splice leader sequence on its 5’ end. Interestingly, intron size in *C. elegans* appears to be positively correlated with local recombination rates (Prachumwat et al., 2004) and short introns are preferentially found in highly expressed genes (Castillo-Davis et al., 2002).

The introns of *C. elegans* have always been considered small, but as more genomes are being sequenced and annotated it is becoming evident that they are not distinctly smaller than those of most eukaryotes. The most common size for fly introns is only 59 bases (The International Human Genome Sequencing Consortium, 2001), as compared to 47 bases for the worm. The average size of introns on the largest, somatic, macronuclear chromosome of Paramecium is only 25 bases (Zagulski et al., 2004). Fungal introns are also small; Neurospora introns average 134 bases (Galagan et al., 2003), *S. macrospora* 106 bases (Nowrousian et al., 2004), and *C. neoformans* 67 bases (Loftus et al., 2005). Even in humans the most common intron size is only 87 bases, but there are also some very large introns, shifting the mean sized to more than 3,300 bases genes (The International Human Genome Sequencing Consortium, 2001).

C. elegans introns follow the GU-AG splice site rule, although GC is a rare 5’ splice site variant (Blumenthal and Stewart, 1997). From their analysis of 669 introns Blumenthal and Steward found that *C. elegans* has a highly conserved and extended 3’ splice site (UUUCAG) and no obvious polypyrimidine track other than the 3’ splice site consensus. They suggest that the 3’ intron boundary may be more important in *C. elegans* intron recognition than in other organisms.

In addition to splicing information, some *C. elegans* introns contain sequences involved in the regulation of gene expression (Zhang and Emmons, 2000).

2.5. Alternative splicing

Alternative splicing will be covered in detail in another chapter (see Alternative splicing in *C. elegans*) so here we will just mention the topic with reference to WormBase. Alternative splice forms are only annotated when there is direct transcript or literature citation evidence for the alternative form. In WormBase release WS133 there are 1,834 genes that have a total of 4,407 alternatively-spliced forms. The number of alternatively-spliced forms per gene tends to be small. Over 90% have either one (1,375) or two (302) alternatively-spliced forms. Many of these alternative forms show only minor changes to the CDS with modulo 3 (i.e., 3,6,9 base differences) to the splice donor or acceptor.

3. Pseudogenes

Processed pseudogenes, which are created by reverse transcription of mRNA into cDNA followed by reintegration into the genome, are fairly easy to detect because they lack introns. These are rare in *C. elegans* (Harrison et al., 2001). Unprocessed pseudogenes arise by duplication of a gene, which is subsequently disabled by random mutation. Unprocessed pseudogenes usually have features that aid in their identification, such as frameshifts, premature stops, insertions and truncations compared to their functional homolog, or a ratio of non-synonymous vs. synonymous nucleotide substitution rates indicating a lack of purifying selection. These features are probably valid indicators for most pseudogenes, assuming that the function of the gene is at the protein level. Some pseudogenes may even be expressed, but mRNAs containing premature stops are usually subject to rapid, nonsense-mediated decay (NMD), making them difficult to detect.

The *uaf-1* gene is an interesting example of how features indicating a pseudogene should be viewed with caution. Uaf-1 encodes the essential splicing factor U2AF^{65} (Zorio et al., 1997) and produces several classes of mRNA, including a 1.7 kb mRNA that encodes a functional U2AF^{65} and a slightly larger mRNA with an extra exon, which inserts an in-frame stop (MacMorris et al., 1999). The premature stop in the latter isoform suggests that this form is non-functional and should be degraded. However, the larger mRNA remains in the nucleus, thus escaping nonsense mediated decay, probably because the extra exon contains multiple copies of a 3' splice-site consensus sequence, which can bind U2AF^{65} (Zorio and Blumenthal, 1999). The likely function of this alternatively-spliced form of *uaf-1*, which is retained in the nucleus, is to down-regulate levels of *uaf-1* when the need for splicing is reduced and to retain free splicing factors in the nucleus where they can be made quickly available when the need for splicing increases. So even though this alternatively-spliced form of *uaf-1* is non-functional at the protein level, and would appear to be a pseudogene version of *uaf-1*, it does function at the RNA level and therefore is not a pseudogene.

Even before large-scale sequencing of the *C. elegans* genome commenced, pseudogenes were identified in the major sperm protein (MSP; Ward et al., 1988) and heat-shock protein gene families (Heschl and Baillie, 1989). Genefinder predictions associated with the genome sequencing project did not attempt to predict pseudogenes. The first genome-wide analysis of pseudogenes was done in 2001 (Harrison et al., 2001). Analyzing Wormpep release 18 and the corresponding version of the genomic sequence (which had only 332 annotated pseudogenes), the authors found 2,168 pseudogenes, or 11.7% of the annotated genes. Most of these were unprocessed pseudogenes, with only 208 designated as processed pseudogenes. They found that pseudogenes are unevenly distributed across the genome with a disproportionate number on chromosome IV; the density was also higher on chromosome arms than in the central regions. Looking at the distribution of pseudogenes among gene families they found that the number of pseudogenes is not correlated with the size of the gene family, but several families were associated with large numbers of pseudogenes. One of these families was the 7-TM receptor family, a finding supported by Robertson's characterization of chemoreceptor gene families (Robertson, 1998, Robertson, 2000, Robertson, 2002).

A higher estimate of the number of pseudogenes comes from an analysis of reporter gene fusions (Mounsey et al., 2002). Extrapolating from the number of 364 randomly selected reporter gene fusions that showed no expression, the authors estimate that 20% of the annotated *C. elegans* genes may be pseudogenes. Furthermore, they found that pseudogenes were enriched for genes that had been recently duplicated.

WormBase release WS133 contains only 561 annotated pseudogenes, far fewer than either of the above estimates. Half of these are located on chromosome V (Table 1), reflecting the curation of chemoreceptor genes, which are located primarily on chromosome V (Robertson, 1998, Robertson, 2000, Robertson, 2002; see Putative chemoreceptor families of *C. elegans*). It seems likely that the number of annotated pseudogenes in WormBase is too low and that other gene families need to be scrutinized for them in the same way the chemoreceptor gene family has.

Table 1. Chromosomal distributions of protein-coding genes, pseudogenes and tRNAs[1]

Chromosome[2]	**Size (Mb)**	**Protein-coding genes**	**Density (genes/Mb)**	**Annotated non-tRNA pseudogenes**	**tRNAs**
I	15.08	3260	216	49	66
Left	4.00	685	171	4	6
Center	6.26	1573	251	6	28
Right	4.82	1002	202	39	32
II	15.28	3874	253	77	56
Left	5.90	1648	279	43	26
Center	5.44	1435	263	23	19
Right	3.94	791	201	11	11
III	13.76	3103	225	30	64
Left	4.80	972	202	10	18
Center	4.29	1199	279	3	29
Right	4.68	932	199	17	17

Chromosome[2]	**Size (Mb)**	**Protein-coding genes**	**Density (genes/Mb)**	**Annotated non-tRNA pseudogenes**	**tRNAs**
IV	17.49	3606	206	100	70
Left	6.74	1339	198	23	26
Center	5.08	1321	260	29	6
Right	5.67	946	167	48	38
V	20.92	5256	251	284	78
Left	6.51	1615	248	63	11
Center	6.99	1880	269	59	25
Right	7.42	1761	237	162	42
X	17.72	3186	180	21	274

[1] Data taken from WormBase data release WS133.

[2] Chromosomal cluster boundaries were taken from Barnes et al. (1995).

4. non-coding RNA genes

4.1. transfer RNA genes

There are 608 nuclear and 22 mitochondrial tRNA genes in *C. elegans*. Seven nuclear genes have been identified as suppressor tRNAs (*sup-5*, *sup-7*, *sup-21*, *sup-24*, *sup-29*, *sup-28* and *sup-33*) and two are likely to be pseudogenes (*rtw-5* and *rtw-6*). The remainder are predicted by tRNAscan-SE (Lowe and Eddy, 1997). The tRNA genes range in size from 64 to 122 bases with 72% having 72 or 73 bases. 29 of the 608 are genes with two exons and the remainder have a single exon. tRNAscan also predicts that there are 213 tRNA pseudogenes.

Nuclear tRNA genes are over-represented on the X chromosome with 45% residing there (Table 1). The other 55% are distributed uniformly over the autosomes; however, there is a slight enrichment on chromosome III and a lower density in the central region of chromosome IV and on the left arm of chromosome I.

4.2. ribosomal RNA genes

The genes for the 18S, 5.8S and 26S ribosomal RNAs, first sequenced and characterized by Ellis et al. (1986), are found in a large tandem-repeat of 100–150 copies on the right-end of chromosome I. Each repeat contains one copy each of the 18S, 5.8S and 26S genes. The 5S ribosomal RNA genes are found in a tandem-repeat of an estimated 100 copies on chromosome V. Each copy of the 5S gene is interspersed with one SL1 splice leader gene.

5. Genomic organization

Protein-coding genes are found equally on either strand of DNA and are fairly uniformly distributed throughout the genome. They are slightly denser on autosomes than on chromosome X (Table 1) and, in general, the central regions of the autosomes are denser than the arms. The left arm of chromosome II is an exception.

Genes in general do not overlap one another, that is to say, their exons do not overlap, but there are numerous examples of genes that fall within introns of another gene, either on the same or the opposite strand. F10F2.2 contains 5 genes on the opposite strand in 2 large exons. A rare and unusual example of overlapping genes can be found with *unc-17* and *cha-1*. These two genes share a common promoter and a first, non-coding exon. The rest of the coding exons do not overlap, so the two genes encode different proteins with mutationally separable functions. *unc-17* encodes a synaptic vesicle acetylcholine transporter, while *cha-1* encodes a choline acetyltransferase (Alfonso et al., 1994).

An unusual and interesting feature of the worm genome is the existence of genes organized into operons. These polycistronic gene clusters contain two or more closely spaced genes, which are oriented in a head to tail direction. They are transcribed as a single polycistronic mRNA and separated into individual mRNAs by the process of trans-splicing (Spieth et al., 1993). These topics are covered in detail in the chapter Trans-splicing and operons.

6. Acknowledgments

We thank Darin Blasiar for help in compiling Table 1. This work was supported by grants from the National Human Genome Research Institute (HG002042, principal investigator, R.K.W.; and P41-HG002223, principal investigator, P.W.S.) and the British Medical Research Council.

7. References

Alfonso, A., Grundahl, K., McManus, J.R., Asbury, J.M., Rand, and J.B. (1994). Alternative splicing leads to two cholinergic proteins in *Caenorhabditis elegans*. J. Mol. Biol. *24*, 627–630. Abstract Article

Barnes, T.M., Kohara, Y., Coulson, A., and Hekimi, S. (1995). Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. Genetics. *141*, 159–179. Abstract

Blumenthal, T., and Steward, K. (1997). RNA processing and gene structure. In: *C. elegans II*, D.L. Riddle, T. Blumenthal, B.J. Meyer, J.R. Priess, eds (Plainview, New York: Cold Spring Harbor Laboratory Press), pp. 117–145.

Bracht, J., Hunter, S., Eachus, R., Weeks, P., and Pasquinelli, A.E. (2004). Trans-splicing and polyadenylation of *let-7* microRNA primary transcripts. RNA. *10*, 1586–1594. Abstract Article

Brenner, S. (2000). The end of the beginning. Science *287*, 2173–2174. Abstract Article

The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: A platform for investigating biology. Science *282*, 2012–2018. Article

Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V., and Kondrashov, F.A. (2002). Selection for short introns in highly expressed genes. Nat Genet. *31*, 415–418. Abstract Article

Ellis, R.E., Sulston, J.E., and Coulson, A.R. (1986). The rDNA of *C. elegans*: sequence and structure. Nucleic Acids Res. *11*, 2345–2364. Abstract

Galagan, J.E., Calvo, S.E., Borkovich, K.A., Selker, E.U., Read, N.D., Jaffe, D., FitzHugh, W., Ma, L.J., Smirnov, S., Purcell, S., Rehman, B., Elkins, T., Engels, R., Wang, S., Nielsen, C.B., Butler, J., Endrizzi, M., Qui, D., Ianakiev, P., Bell-Pedersen, D., Nelson, M.A., Werner-Washburne, M., Selitrennikoff, C.P., Kinsey, J.A., Braun, E.L., Zelter, A., Schulte, U., Kothe, G.O., Jedd, G., Mewes, W., Staben, C., Marcotte, E., Greenberg, D., Roy, A., Foley, K., Naylor, J., Stange-Thomann, N., Barrett, R., Gnerre, S., Kamal, M., Kamvysselis, M., Mauceli, E., Bielke, C., Rudd, S., Frishman, D., Krystofova, S., Rasmussen, C., Metzenberg, R.L., Perkins, D.D., Kroken, S., Cogoni, C., Macino, G., Catcheside, D., Li, W., Pratt, R.J., Osmani, S.A., DeSouza, C.P., Glass, L., Orbach, M.J., Berglund, J.A., Voelker, R., Yarden, O., Plamann, M., Seiler, S., Dunlap, J., Radford, A., Aramayo, R., Natvig, D.O., Alex, L.A., Mannhaupt, G., Ebbole, D.J., Freitag, M., Paulsen, I., Sachs, M.S., Lander, E.S., Nusbaum, C., and Birren, B. (2003). The genome sequence of the filamentous fungus Neurospora crassa. Nature *422*, 859–868. Abstract Article

Harris, T.W., Chen, N., Cunningham, F., Tello-Ruiz, M., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bradnam, K., and Chan, J. (2004). WormBase: a multi-species resource for nematode biology and genomics. Nucleic Acids Res. *32*, D411–D417. Abstract Article

Harrison, P.M., Echols, N., and Gerstein, M.B. (2001). Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. Nucleic Acids Res. *29*, 818–830. Abstract Article

Heschl, M.F., and Baillie, D.L. (1989). Identification of a heat-shock pseudogene from *Caenorhabditis elegans*. Genome *32*, 190–1955. Abstract

The International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. Abstract

Korf, I., Flicek, P., Duan, D., and Brent, M.R. (2001). Integrating genomic homology into gene structure prediction. Bioinformatics *17(Suppl 1)*, S140–S148 Abstract

Loftus, B.J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I.J., Fraser, J.A., Allen, J.E., Bosdet, I.E., Brent, M.R., Chiu, R., Doering, T.L., Donlin, M.J., D'Souza, C.A., Fox, D.S., Grinberg, V., Fu, J., Fukushima, M., Haas, B.J., Huang, J.C., Janbon, G., Jones, S.J., Koo, H.L., Krzywinski, M.I., Kwon-Chung, J.K., Lengeler, K.B., Maiti, R., Marra, M.A., Marra, R.E., Mathewson, C.A., Mitchell, T.G., Pertea, M., Riggs, F.R., Salzberg, S.L., Schein, J.E., Shvartsbeyn, A., Shin, H., Shumway, M., Specht, C.A., Suh, B.B., Tenney, A., Utterback, T.R., Wickes, B.L., Wortman, J.R., Wye, N.H., Kronstad, J.W., Lodge, J.K., Heitman, J., Davis, R.W., Fraser, C.M., and Hyman, R.W. (2005). The genome of the *basidiomycetous* yeast and human pathogen *Cryptococcus neoformans*. Science *307*, 1321–1324. Abstract Article

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955–964. Abstract Article

MacMorris, M.A., Zorio, D.A., and Blumenthal, T. (1999). An exon that prevents transport of a mature mRNA. Proc. Natl. Acad. Sci. USA *96*, 3813–3818. Abstract Article

Mounsey, A., Bauer, P., and Hope, I.A. (2002). Evidence suggesting that a fifth of annotated *Caenorhabditis elegans* genes may be pseudogenes. Genome Res. *12*, 770–775. Abstract Article

Nowrousian, M., Wurtz, C., Poggeler, S., and Kuck, U. (2004). Comparative sequence analysis of *Sordaria macrospora* and *Neurospora crassa* as a means to improve genome annotation. Fungal Genet. Biol. *41*, 285–292. Abstract Article

Prachumwat, A., DeVincentis, L., and Palopoli, M.F. (2004). Intron size correlates positively with recombination rate in *Caenorhabditis elegans*. Genetics *163*, 1585–1590. Abstract Article

Reboul, J., Vaglio, P., Rual, J.F., Lamesch, P., Martinez, M., Armstrong, C.M., Li, S., Jacotot, L., Bertin, N., Janky, R., Moore, T., Hudson, J.R. Jr, Hartley, J.L., Brasch, M.A., Vandenhaute, J., Boulton, S., Endress, G.A., Jenna, S., Chevet, E., Papasotiropoulos, V., Tolias, P.P., Ptacek, J., Snyder, M., Huang, R., Chance, M.R., Lee, H., Doucette-Stamm, L., Hill, D.E., and Vidal, M. (2003). *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. Nat. Genet. *34*, 35–41. Abstract Article

Robertson, H.M. (1998). Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss. Genome Res. *8*, 449–463. Abstract

Robertson, H.M. (2000). The large srh family of chemoreceptor genes in *Caenorhabditis* nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. Genome Res. *10*, 192–203. Abstract Article

Robertson, H.M. (2002). Updating the *str* and *srj (stl)* families of chemoreceptors in *Caenorhabditis* nematodes reveals frequent gene movement within and between chromosomes. Chem. Senses *26*, 151–159. Article

Spieth, J., Brooke, G., Kuersten, S., Lea, K., and Blumenthal, T. (1993). Operons in *C. elegans*: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. Cell *73*, 521–532. Abstract Article

Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A. et al. (2003). The Genome Sequence of *Caenorhabditis briggsae*: A Platform for Comparative Genomics. PLoS Biol. *1*, 166–192. Abstract Article

Synder, M., and Gerstein, M (2003). Defining genes in the genomics era. Science *300*, 258–260 Abstract Article

Vaglio, P., Lamesch, P., Reboul, J., Rual, J.F., Martinez, M., Hill, D., and Vidal, M. (2003). WorfDB: the *Caenorhabditis elegans* ORFeome Database. Nucleic Acids Res. *31*, 237–240. Abstract Article

Ward, S., Burke, D.J., Sulston, J.E., Coulson, A.R., Albertson, D.G., Ammons, D., Klass, M., and Hogan, E. (1988). Genomic organization of major sperm protein genes and pseudogenes in the nematode *Caenorhabditis elegans*. J Mol Biol.*199*, 1–13. Abstract Article

Wei, C., Lamesch, P., Arumugam, M., Rosenberg, J., Hu, P., Vidal, M., and Brent, M.R. (2005). Genome Res. in press.

Zhang, H., and Emmons, S.W. (2000). A mediator protein confers regulatory selectivity on lineage-specific expression of a transcription factor gene. Genes Dev. *14*, 2161–2172. Abstract Article

Zorio, D.A., Blumenthal, T. (1999). Both subunits of U2AF recognize the 3' splice site in *Caenorhabditis elegans*. Nature *402*, 835–832. Abstract Article

Zorio, D.A., Cheng, N.N., Blumenthal, T., and Spieth, J. (1994). Operons as a common form of chromosomal organization in *C. elegans*. Nature *372*, 270–272. Abstract Article

Zorio, D.A., Lea, K., and Blumenthal, T. (1997). Cloning of *Caenorhabditis* U2AF65: an alternatively spliced RNA containing a novel exon. Mol. Cell Biol. *17*, 946–953. Abstract

Zagulski, M., Nowak, J.K., Le Mouel, A., Nowacki, M., Migdalski, A., Gromadka, R., Noel, B., Blanc, I., Dessen, P., Wincker, P., Keller, A.M., Cohen, J., Meyer, E., and Sperling, L. (2004). High coding density on the largest *Paramecium tetraurelia* somatic chromosome. Curr. Biol. *14*, 1397–1404. Abstract Article