Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi

Ravi S. Kamath*†, Andrew G. Fraser*†§, Yan Dong*, Gino Poulin*, Richard Durbin‡, Monica Gotta*§, Alexander Kanapin||, Nathalie Le Bot*, Sergio Moreno*¶, Marc Sohrmann‡§, David P. Welchman*, Peder Zipperlen* & Julie Ahringer*

* *Wellcome Trust/Cancer Research UK Institute and Department of Genetics, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK*
‡ *Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK*
|| *EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK*
¶ *Centro de Investigacion del Cancer, CSIC / Univ. Salamanca, Campus Miguel de Unamuno, 37007 Salamanca, Spain*
† *These authors contributed equally to this work*

A principal challenge currently facing biologists is how to connect the complete DNA sequence of an organism to its development and behaviour. Large-scale targeted-deletions have been successful in defining gene functions in the single-celled yeast *Saccharomyces cerevisiae*, but comparable analyses have yet to be performed in an animal. Here we describe the use of RNA interference to inhibit the function of ~86% of the 19,427 predicted genes of *C. elegans*. We identified mutant phenotypes for 1,722 genes, about two-thirds of which were not previously associated with a phenotype. We find that genes of similar functions are clustered in distinct, multi-megabase regions of individual chromosomes; genes in these regions tend to share transcriptional profiles. Our resulting data set and reusable RNAi library of 16,757 bacterial clones will facilitate systematic analyses of the connections among gene sequence, chromosomal location and gene function in *C. elegans.*

The ability to inactivate a target gene transiently by RNAi[1] has greatly accelerated the analysis of loss-of-function phenotypes in *C. elegans* and other organisms. Although several large-scale RNAi-based screens have been used to study gene function in *C. elegans*[2–4], in total only about a third of the predicted genes have been analysed so far. Genome-wide RNAi analyses would not only provide a key resource for studying gene function in *C. elegans* but should also address important issues in functional genomics, such as the global organization of gene functions in a metazoan genome. In addition, because more than half of the genes in *C. elegans* have a human homologue, this kind of functional analysis in the worm should provide insights into human gene function.

Analysis of gene functions by RNAi

Loss-of-function RNAi phenotypes can be generated efficiently by feeding worms with bacteria expressing double-stranded RNA (dsRNA) that is homologous to a target gene[5–7]; we previously used this method to screen roughly 87% of predicted genes on chromosome I of *C. elegans* (ref. 2). To screen most of the predicted genes in *C. elegans* by RNAi, we constructed a library of bacterial strains, each capable of expressing dsRNA designed to correspond to a single gene. The library consists of 16,757 bacterial strains, which in total correspond to about 86% of the 19,427 current predicted genes in *C. elegans* with similar coverage across each chromosome (see Supplementary Tables 1 and 2). Using this library, we screened wild-type *C. elegans* hermaphrodites to identify genes for which RNAi reproducibly results in sterility, embryonic or larval lethality, slow post-embryonic growth, or a post-embryonic defect (Methods). Such phenotypes were obtained with 1,722 bacterial strains (10.3% of those analysed; Fig. 1a).

Many strains gave rise to several reproducible RNAi phenotypes, indicating that the targeted gene has many developmental roles. For example, RNAi against Y77E11A.13a (which encodes a homologue of the yeast Sec13p protein implicated in protein trafficking from the endoplasmic reticulum to the Golgi[8]) results in sterility, embryonic lethality or uncoordinated movement. To simplify subsequent genomic analyses, we defined three mutually exclusive phenotypic classes: the nonviable (Nonv) class, consisting of embryonic or larval lethality or sterility (with or without associated post-embryonic defects); the growth defects (Gro) class, consisting of slow or arrested post-embryonic growth; and the viable post-embryonic phenotype (Vpep) class, consisting of defects in post-embryonic development (for example, in movement or body shape) without any associated lethality or slowed growth. The RNAi phenotypes obtained on each chromosome are summarized in Fig. 1a, and a full list of phenotypes by gene is given in Supplementary Tables 2–4; these data are available publicly on Wormbase (http://www.wormbase.org).

To determine the effectiveness of the screen, we assessed our ability to identify correctly the known loss-of-function phenotypes for previously studied loci. Overall, we obtained RNAi phenotypes for 63.5% of 323 detectable loci; almost all of those detected (92%) produced an RNAi phenotype similar to the known mutant phenotype (see Supplementary Tables 5 and 6). More loci with a Nonv phenotype were detected (77.9%) than loci with a Vpep phenotype (42.2%). This difference is likely to arise because certain classes of gene with Vpep phenotypes (for example, neuronally expressed genes) are relatively resistant to RNAi[7,9] and because Vpep phenotypes are more difficult to detect in this screen (Methods). Notably, the estimated rate of false-positive RNAi phenotypes is very low (<1%; see Supplementary Fig. 1). In addition, our results correlate well with, and are as sensitive as, previous RNAi screens (refs 3, 4, and Supplementary Fig. 1), indicating that RNAi data are highly reproducible irrespective of the method used.

The most common RNAi phenotype is embryonic lethality, which was observed for 929 strains (5.5%). On the basis of our efficiency of detecting known embryonic lethal loci, this probably includes over 70% of embryonic lethal genes and thus will be an excellent starting point for more detailed analyses of the molecular

§ Present addresses: Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK (A.G.F.); Institute of Biochemistry ETH Hoenggerberg, 8093-Zurich, Switzerland (M.G., M.S.)..

mechanisms of embryogenesis in *C. elegans*. Of the post-embryonic phenotypes detected, the largest class was uncoordinated movement (Unc), which is typically indicative of a defect in the neuromuscular system. We also defined an RNAi phenotype for 33 close homologues (BlastP *E* values less than 10^{-6}) of human disease genes (Table 1). Notably, many of these genes had Vpep phenotypes (50% versus 16% among all genes with a phenotype), consistent with their embryonic viability in humans, and thus may be useful for establishing *C. elegans* models of some human diseases.

A small percentage of the bacterial strains were predicted to target more than one predicted gene. Before carrying out global analyses, we removed these ambiguous data to generate a set of 1,528 clones for which RNAi phenotypes could be attributed to a single predicted gene (Methods).

Conservation and gene function

We and others have previously found relationships between the RNAi phenotype of a gene and its degree of conservation and putative molecular function, using relatively small datasets[2–4,10]. Using the larger dataset obtained here, we have confirmed and extended those conclusions. We find that *C. elegans* genes with an orthologue in another eukaryote are much more likely to have a detectable RNAi phenotype than all other genes (21% versus 6%). In addition, highly conserved genes that are present as a single copy in the *C. elegans* genome are more than twice as likely to have an RNAi phenotype as those that are present in more than one copy (31% versus 12%); this suggests that many recently duplicated paralogues are at least partially functionally redundant or have specialized functions that are not detectable in this screen.

The highest cross-species conservation is seen among genes with a Nonv RNAi phenotype, of which 52% have an orthologue in another eukaryote; this shows that similar essential basal cellular machinery is common to all eukaryotes. Indeed, 51% of *C. elegans* orthologues of yeast essential genes[11] have a Nonv RNAi phenotype. Consistent with these findings, genes involved in the basic metabolism and maintenance of the cell are significantly enriched for having a Nonv RNAi phenotype (Fig. 2a); by contrast, genes involved in more complex processes that are expanded in metazoa, such as signal transduction and transcriptional regulation, are enriched for Vpep phenotypes (Fig. 2b).

Domain evolution and gene function

To study further the relationship between the sequence and function of a gene, we examined the domain composition of genes in each phenotypic class. Of the 200 most abundant InterPro domains[12] in the *C. elegans* genome, 28 show significant ($P < 0.05$) associations

a

	Chr I (2,445 clones)		Chr II (2,978 clones)		Chr III (2,132 clones)		Chr IV (2,693 clones)		Chr V (4,152 clones)		Chr X (2,357 clones)		Total (16,757 clones)	
Phe	Num	%	Num	%	Num	%	Num	%	Num	%	Num	%	Num	%
All	334	13.7	354	11.9	395	18.5	293	10.9	215	5.2	131	5.6	1722	10.3
Nonv	253	10.3	240	8.1	298	14.0	187	6.9	145	3.5	47	2.0	1170	7.0
Grow	37	1.5	41	1.4	59	2.8	70	2.6	46	1.1	23	1.0	276	1.6
Vpep	44	1.8	73	2.5	38	1.8	36	1.3	24	0.6	61	2.6	276	1.6
Emb	226	9.2	204	6.9	220	10.3	143	5.3	108	2.6	28	1.2	929	5.5
Ste	83	3.4	43	1.4	132	6.2	58	2.2	44	1.1	12	0.5	372	2.2
Stp	15	0.6	33	1.1	18	0.8	14	0.5	19	0.5	3	0.1	102	0.6
Gro/Lva	147	6.0	129	4.3	161	7.6	131	4.9	102	2.5	34	1.4	704	4.2
Lvl	38	1.6	61	2.0	35	1.6	18	0.7	20	0.5	24	1.0	196	1.2
Adl	3	0.1	19	0.6	34	1.6	7	0.3	7	0.2	16	0.7	86	0.5
Bli	4	0.2	3	0.1	1	0.0	1	0.0	0	0.0	2	0.1	11	0.1
Bmd	27	1.1	89	3.0	32	1.5	17	0.6	15	0.4	14	0.6	194	1.2
Clr	14	0.6	84	2.8	45	2.1	45	1.7	18	0.4	38	1.6	244	1.5
Dpy	19	0.8	39	1.3	16	0.8	20	0.7	7	0.2	16	0.7	117	0.7
Egl	6	0.2	29	1.0	18	0.8	13	0.5	8	0.2	21	0.9	95	0.6
Him	12	0.5	4	0.1	1	0.0	2	0.1	1	0.0	2	0.1	22	0.1
Lon	2	0.1	11	0.4	9	0.4	8	0.3	3	0.1	5	0.2	38	0.2
Mlt	8	0.3	8	0.3	4	0.2	2	0.1	2	0.0	13	0.6	37	0.2
Muv	2	0.1	4	0.1	5	0.2	2	0.1	0	0.0	1	0.0	14	0.1
Prz	18	0.7	41	1.4	15	0.7	11	0.4	9	0.2	25	1.1	119	0.7
Pvl	32	1.3	39	1.3	37	1.7	17	0.6	13	0.3	9	0.4	147	0.9
Rol	2	0.1	5	0.2	1	0.0	2	0.1	1	0.0	2	0.1	13	0.1
Rup	10	0.4	39	1.3	25	1.2	18	0.7	11	0.3	16	0.7	119	0.7
Sck	6	0.2	32	1.1	67	3.1	44	1.6	25	0.6	6	0.3	180	1.1
Unc	72	2.9	111	3.7	55	2.6	46	1.7	40	1.0	64	2.7	388	2.3

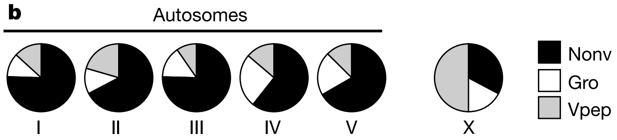

Figure 1 Summary of RNAi phenotypes. **a**, Number of bacterial strains associated with each RNAi phenotype. The Nonv (nonviable, including all phenotypic classes that result in lethality or sterility), Gro (growth defects, including slow post-embryonic growth or larval arrest) and Vpep (viable post-embryonic phenotype, including all other phenotypic classes) categories are mutually exclusive; however, many genes are associated with several specific RNAi phenotypes. Phenotypic classes are described in Methods. The percentages are out of the total number of clones screened per chromosome. **b**, Relative proportion of Nonv, Gro and Vpep phenotype on each chromosome.

with particular classes of RNAi phenotype (Table 2). Notably, of the seven InterPro domains that are significantly associated with Vpep RNAi phenotypes, most (six) are represented in the fly[13] and human[14,15] genomes but not in the genome of budding yeast[16] or *Arabidopsis*[17]. Genes with a Vpep phenotype by definition have no associated lethality but instead have a role in the multicellular animal (such as in movement or body shape); therefore, these data suggest that many of the 'animal-specific' functions encoded by genes with Vpep phenotypes may have arisen through the evolution of new domains.

To explore this idea further, we examined whether genes with animal-specific domains are, in general, more likely to have an 'animal-specific' function (that is, to have a Vpep RNAi phenotype). *C. elegans* genes encoding at least one identifiable domain were split into three groups: 'ancient', in which all encoded protein domains are found in yeast, *Arabidopsis*, *Drosophila* and humans; 'animal', in which at least one domain is found in *Drosophila* or humans but not in yeast or *Arabidopsis*; and 'worm', in which any domain is found only in *C. elegans* (37% are ancient, 8% are animal, 10% are worm and 46% have no identifiable domain).

Whereas genes with a Nonv RNAi phenotype are highly enriched for being in the ancient class (Fig. 3; 90% of those with an identifiable domain are 'ancient'), genes with a Vpep RNAi phenotype are enriched for being in the animal class (16% of Vpep genes but only 6% of Nonv genes are in the animal class). This supports the idea that the evolution of new domains has been important for the evolution of animal-specific gene functions. In addition, we found that almost none of the genes in the 'worm' class has an essential role in *C. elegans*, although many have a Vpep phenotype. This suggests that these genes have nematode-specific developmental functions and supports the view that the basal machinery of eukaryotes is shared and not phylum-specific.

The X chromosome

The *C. elegans* genome is organized into five autosomes and a sex chromosome (X)[18]. Sex in *C. elegans* is determined by the number of copies of the X chromosome: hermaphrodites have two copies of the X chromosome, each of which is partially transcriptionally silenced to ensure dosage compensation to and males have a single copy (reviewed in ref. 19). We explored whether there are functional differences between genes on the autosomes and the X chromosome. We found that whereas the autosomes each have a similar distribution of RNAi phenotypes, the distribution on the X chromosome is markedly different (Fig. 1b). This difference is due almost completely to a reduction in the percentage of genes with a Nonv phenotype (Fig. 1a), an effect previously reported by other groups using smaller datasets[3,10]. Thus, there has been strong selection against the encoding of essential functions on the X chromosome.

Previous studies have shown that X-linked genes are transcriptionally silenced in the germ line during mitosis and early meiosis[20,21]. Genes required for the basic cellular processes that are essential for the viability of all cells (including those in the germ line) might thus be expected to be absent from the X chromosome; many such genes have Nonv RNAi phenotypes. We indeed found that genes in the functional classes enriched for Nonv phenotypes (such as protein synthesis) are highly underrepresented on the X chromosome (Fig. 2c and Supplementary Fig. 2). The reduction in the number of essential functions encoded on the X chromosome therefore seems to be related to the transcriptional repression of X-linked genes in the germ line. Differential expression of X-linked genes does not explain the entire difference, however, because X-linked and autosomal genes with similar germline expression profiles have very different roles. For example, although genes with oocyte-enriched expression are found in similar numbers on the X chromosome and the autosomes[20], none of the X-linked oocyte-enriched genes have a Nonv RNAi phenotype, whereas 19% of the autosomal oocyte-enriched genes are essential.

A second, more intriguing property of the X chromosome is that it is enriched for genes with Vpep phenotypes ($P < 0.01$; chromo-

Table 1 **Thirty-three human disease gene homologues with an RNAi phenotype**

Predicted gene	*C. elegans* locus	Human disease	Human gene	BlastP *E* value	RNAi phenotype
B0035.5		G6PD deficiency	*G6PD*	1×10^{-176}	Emb, Clr, Gro
B0350.2A	*unc-44*	Hereditary spherocytosis	*ANK1*	0.00	Slu
C01G6.8	*cam-1/kin-8*	Insulin-resistant diabetes mellitus	*INSR*	6×10^{-55}	Unc, Pvl, clear patch
C01G8.5A		Neurofibromatosis	*NF2*	1×10^{-123}	Unc, Lvl, Gro
C06A1.1		Zellweger syndrome	*PEX1*	3×10^{-67}	Emb, Bmd, Sck, Gro
C07H6.7	*lin-39*	MODY, type IV	*IPF1*	5×10^{-14}	Egl, Vul, Muv
C17E4.5		Oculopharyngeal muscular dystrophy	*PABPN1*	3×10^{-41}	Emb, Unc, Lva
C29A12.3	*lig-1*	DNA ligase I deficiency	*DNA ligase1*	1×10^{-167}	Emb
C48A7.1	*egl-19*	Long QT syndrome 3	*SCN5A*	2×10^{-64}	Egl, Clr
C50H2.1		Leydig cell hypoplasia	*LHCGR*	9×10^{-76}	Gro
D2045.1		Spinocerebellar ataxia 2	*SCA2*	7×10^{-09}	Emb
F01G10.1		Wernicke–Korsakoff syndrome	*TKT*	0.00	Emb, Clr, Gro
F07A5.7	*unc-15*	Tuberous sclerosis	*TSC1*	1×10^{-07}	Unc, Prz, Egl
F11C1.6	*nhr-25*	Pseudohyperaldosteronism	*NR3C2*	7×10^{-24}	Unc, Prz, Clr, Egl
F11H8.4	*cyk-1*	Nonsyndromic sensorineural deafness	*DFNA1*	9×10^{-49}	Emb, Adl, Rup, Clr
F20B6.2	*vha-12*	Renal tubular acidosis	*ATP6B1*	0.00	Emb, Ste, Adl, Lvl, Prz
F54D8.1		Ehlers–Danlos syndrome, type IV	*COL3A1*	1×10^{-06}	Dpy
F53G12.3		Chronic Granulomatous Disease	*X-CGD*	3×10^{-34}	Bli, Mlt, Lvl
F58A3.2A	*egl-15*	Multiple venous malformations	*VMCM*	1×10^{-62}	Egl
K04G2.8A	*apr-1*	Adenomatous polyposis of the colon	*APC*	9×10^{-34}	Unc, Bmd, Lvl
K07A1.12	*rba-2*	Cockayne syndrome	*CKN1*	6×10^{-13}	Emb, Pvl, Lvl
K08A8.2		Gonadal dysgenesis	*SRY*	3×10^{-31}	Unc, Egl
K08C7.3	*epi-1*	Usher syndrome 2a	*USH2A*	1×10^{-112}	Ste, Unc, Muv, Dpy, Pvl, Rup
K11D9.2A		Darier–White disease	*SERCA*	0.00	Ste, Sck
M02A10.2		Hyperinsulinism	*KCNJ11*	4×10^{-78}	Unc
R107.8	*lin-12*	Alagille syndrome	*JAG1*	2×10^{-90}	Egl
R12B2.1	*sma-4*	Pancreatic carcinoma	*MADH4*	2×10^{-39}	Sma, Dpy
T03F6.5	*lis-1*	Miller–Dieker lissencephaly syndrome	*PAF*	1×10^{-148}	Emb
W05E10.3	*ceh-32*	Holoprosencephaly	*SIX3*	1×10^{-69}	Unc
W10G6.3	*ifa-2*	Keratoderma	*KRT9*	7×10^{-26}	Unc, Lvl, Mlt
Y47D3A.6A	*tra-1*	Grieg cephalopolysyndactyly syndrome	*GLI*	6×10^{-58}	Rup, clear patch
Y76A2A.2		Menkes disease	*ATP7A*	0.00	Prz, Adl, Unc
ZC506.4	*mgl-1*	Hypercalcemia	*CASR*	2×10^{-77}	Gro

C. elegans genes with a human disease gene homologue are defined as those with a BlastP *E* value less than 1.0×10^{-6}, taken from refs 38, 39. Shown are those with an RNAi phenotype. The phenotypes are defined in Methods. MODY, maturity onset diabetes of the young. G6PD, glucose-6-phosphate dehydrogenase.

some II is also enriched for Vpep genes). In addition, significantly ($P < 0.01$) more X-linked genes than autosomal genes encode components of signalling pathways and transcription factors; these genes are enriched for Vpep phenotypes. This concentration of Vpep genes on the X chromosome may have evolutionary benefits. Whereas a hermaphrodite worm that is heterozygous for a mutant allele of an X-linked gene is likely to be phenotypically wild type, a (hemizygous) male inheriting the mutant allele will be mutant. Hermaphrodites could thus act as wild-type repositories for mutant alleles of genes affecting the patterning, structure or behaviour of worms; these alleles could then be selected for or against in a dominant manner in the hemizygous male animal. Because the number of males spontaneously arising from hermaphrodites through meiotic non-disjunction events increases markedly under stressful conditions (such as increased temperature), this haploselection for relatively subtle phenotypic changes might be a powerful mechanism by which to adapt to a changing environment.

Large-scale functional gene clustering

Our RNAi experiments targeted most of the genes in *C. elegans*, with similar proportions of genes covered along each chromosome. Using these data, we examined whether genes of similar function cluster in specific regions of chromosomes. Unlike most animals, *C. elegans* has holocentric chromosomes that lack a localized centromeric region. The five autosomes have a central 'cluster', where rates of recombination are low and where most studied genetic loci are found, which is flanked by chromosome 'arms', where recombination rates are more than tenfold higher[22]. These clusters have characteristic features on all autosomes: lower repeat content, greater conservation and greater representation by expressed sequence tags (ESTs)[18]. By contrast, the X chromosome does not have a defined cluster region.

In agreement with data derived from classical genetics, we found that genes with RNAi phenotypes are enriched twofold in the cluster regions relative to the arms (7.6% of genes on arms have an RNAi phenotype versus 14.9% in the cluster regions; Fig. 4a). We next examined the distribution of the Nonv, Gro and Vpep genes in the genome (Methods). Notably, genes with a Nonv RNAi phenotype are strongly enriched in large regions of the clusters of chromosomes I, II and III ($P < 0.01$; Fig. 4b): 36% of the Nonv genes lie in

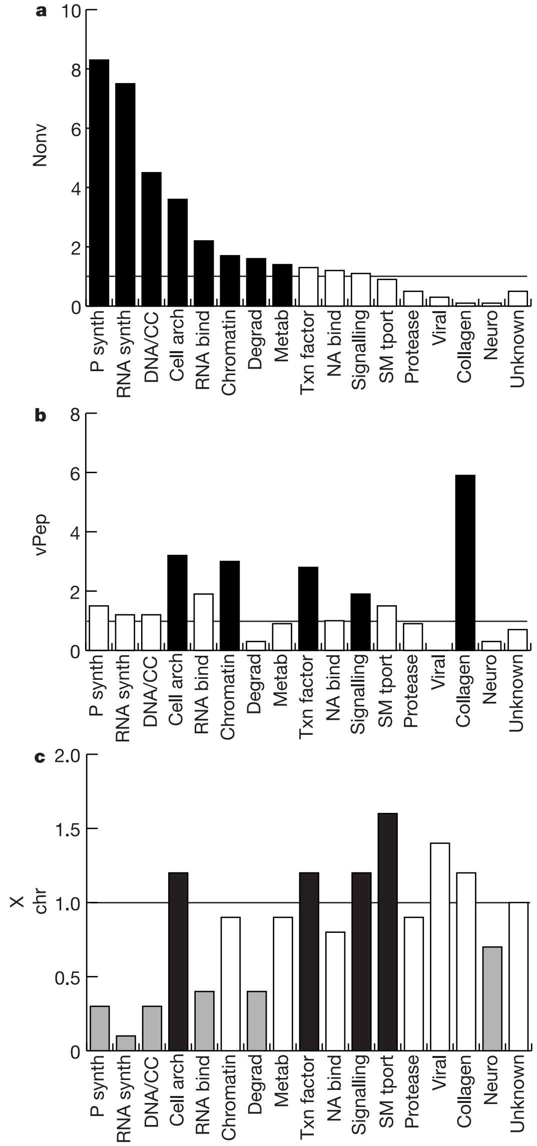

Figure 2 Relative enrichment of Nonv, Vpep and X chromosome genes for different functional classes. The functional classes are protein synthesis (P synth), RNA synthesis (RNA synth), DNA synthesis and repair/cell cycle (DNA/CC), cellular architecture (Cell arch), RNA binding (RNA bind), chromatin regulation (Chromatin), protein degradation (Degrad), energy and intermediary metabolism (Metab), transcription factors (Txn factor), nucleic-acid binding (NA bind), signal transduction (Signalling), small-molecule transport (SM tport), specific proteases (Protease), retroviral- and transposon-derived sequences (Viral), collagens (Collagen), genes with neuronal functions (Neuro), and Unknown. Shown are the levels of enrichment among genes in each functional class for Nonv phenotypes (**a**), Vpep phenotypes (**b**) or genes on the X chromosome (**c**); bars in black denote a statistically significant overenrichment ($P < 0.01$). The grey bars in **c** represent an underenrichment ($P < 0.01$). For reference, a line is drawn at a relative representation of 1.0.

Table 2 **InterPro domains associated with RNAi phenotypes**

Nonv only
Elongation factor, GTP-binding
Cyclin
Ubiquitin domain
TPR repeat
Zinc-finger, CCHC type
Myb DNA-binding domain
Laminin-type EGF-like domain
DEAD/DEAH box helicase
Ubiquitin-associated domain
Zinc-finger, C_2H_2 type
Mitochondrial substrate carrier
Protein kinase C, phorbol ester/DAG binding
Gro only
Glycosyl transferase, family 2
Zinc-finger, RING
Phosphotyrosine interaction domain
Proline-rich extensin
Nonv and Gro
G-protein β-subunit WD40 repeat
AAA ATPase
KH domain
Zinc-finger, C-X_8-C-X_5-C-X_3-H type
RNA-binding region RNP-1 (RNA recognition)
Vpep
Immunoglobulin/major histocompatibility complex
Collagen triple helix repeat
Immunoglobulin-like
EGF-like calcium-binding
Aspartic acid and asparagine hydroxylation site
Fibronectin, type III
Worm-specific repeat type 1

We examined the phenotypes of genes containing any of the 200 most abundant InterPro[12] domains in the *C. elegans* genome; genes containing the listed domains were significantly enriched ($P < 0.05$) for the indicated phenotypes, in order of decreasing significance. DAG, diacylglycerol; EGF, epidermal growth factor.

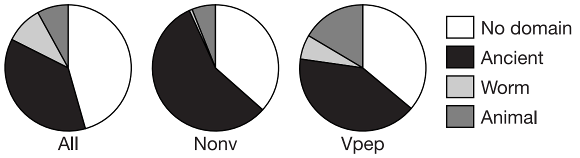

Figure 3 Conservation of domains in genes with different RNAi phenotypes. All predicted genes were placed into one of four mutually exclusive classes on the basis of their InterPro domain content. The 'ancient' class comprises genes for which all predicted domains are also encoded in the *S. cerevisiae*, *A. thaliana*, *D. melanogaster* and *H. sapiens* genomes; the 'animal' class comprises genes that contain any domain present in the *D. melanogaster* or *H. sapiens* genomes, but not in *S. cerevisiae* or *A. thaliana*, and the 'worm' class comprises genes containing any domain present in the *C. elegans* genome, but not in the other four. The proportions of All, Nonv and Vpep genes that fall into each class are shown.

these enriched regions, which represent about 13% of the genome. By contrast, Nonv genes are underenriched on the autosomal arms and the whole of the X chromosome. Functional redundancy among paralogous genes might explain some of the underenrichment, because these regions frequently overlap those areas of the autosomes with increased gene duplication (Fig. 4b).

Genes with Vpep and Gro phenotypes are enriched in different regions of the genome from those showing enrichment for Nonv genes. Notably, genes with a Vpep phenotype are enriched significantly in the centre of the X chromosome, despite the absence of a recombinationally defined cluster[22]. This suggests that the X chromosome, like the autosomes, has a central accumulation of genes with nonredundant functions; on the X chromosome, however, these genes are not required for viability, but rather for worm behaviour or morphology. These findings suggest that in *C. elegans* there is selective pressure for genes with similar organismal functions to be colocalized in large domains of the genome.

How such domains are maintained and what they represent mechanistically are unclear. A possible hypothesis is that, perhaps as a consequence of long-range chromatin regulation, genes in these domains are transcriptionally co-regulated. To investigate this possibility, we examined sets or "mounts"[23] of *C. elegans* genes identified by microarray analysis to share expression profiles; we found that genes in each mount are enriched in distinct regions of the chromosomes (Supplementary Fig. 3). Such large-scale clustering has also been observed in both humans[24] and *Drosophila*[25].

Notably, genes in mounts 7 and 11 are significantly enriched in the same regions of the genome as are the Nonv genes (Fig. 4b and Supplementary Fig. 3); in addition, these mounts are enriched for genes with Nonv RNAi phenotypes. This suggests that in regions of the genome that have concentrations of genes of similar functions, there is large-scale broad transcriptional co-regulation. The scale of these regions (over 1 megabase) indicates that this mode of regulation is clearly distinct from that previously reported in yeast[26] and in *C. elegans*[27], in which small clusters of nearly adjacent genes are likely to be co-regulated, perhaps as a consequence of open loops of chromatin[26,28]. When an assembled genome sequence is available for the nematode *Caenorhabditis briggsae*, which is closely related to *C. elegans*, it will be intriguing to see whether these functional domains are maintained as syntenic regions.

In summary, we note that there are differences in gene function between the X chromosome and the autosomes, as well as functional clustering in different regions of the genome. Each chromosome has unique features—for example, chromosome V has few essential genes relative to the other autosomes and has a high degree of gene duplications, whereas chromosome III is enriched for Nonv genes, and chromosome II is enriched for Vpep genes. These data suggest that different chromosomes and regions of the genome may be specialized for particular functions.

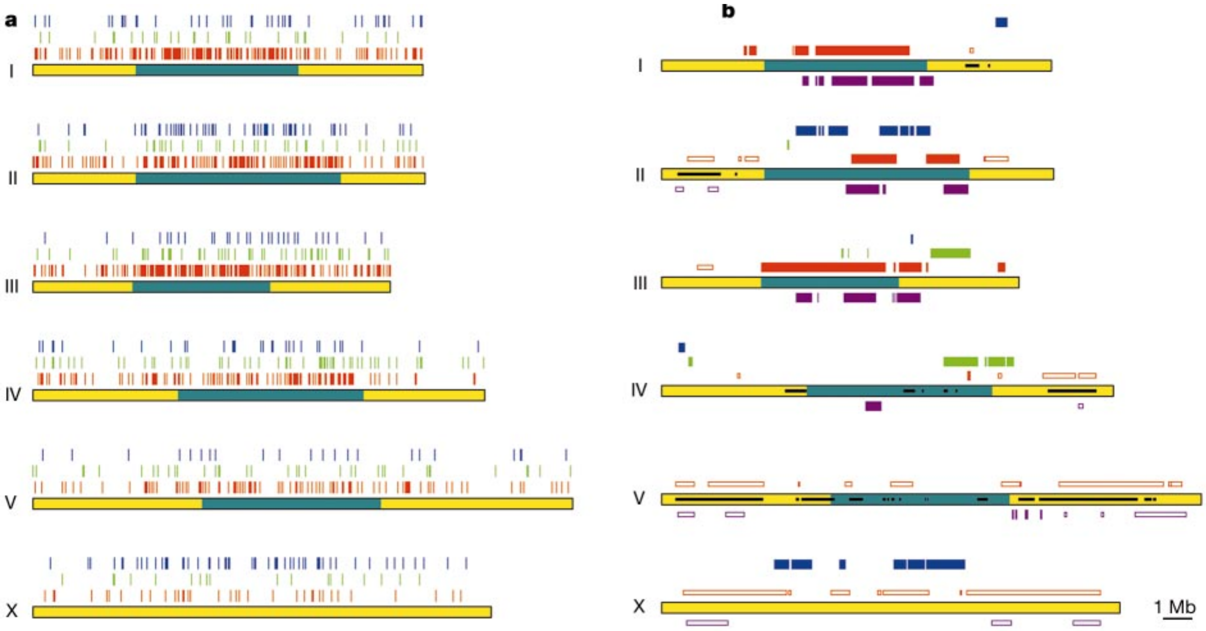

Figure 4 Distribution of RNAi phenotypes across the *C. elegans* chromosomes. **a**, Genomic locations of genes with RNAi phenotypes. Horizontal yellow (arm regions) and blue-green (cluster regions) bars represent *C. elegans* chromosomes; black bars indicate regions enriched for duplicated genes (that is, those with a *C. elegans* homologue). Each RNAi phenotype is represented by a single red (Nonv), green (Gro) or blue (Vpep) line above the chromosomes. **b**, Chomosomal enrichment of genes with different RNAi phenotypes. Overenrichment is indicated by filled boxes, underenrichment by open boxes. No windows could be significantly underenriched for Gro or Vpep phenotypes owing to the smaller sample sizes. The purple bars below the chromosomes represent regions that are significantly ($P < 0.01$) over- or underenriched for genes in mount 11 (ref. 23). In the enriched regions, 36% of Nonv genes lie in 13% of the genome, 11.6% of Gro genes lie in 3.9% of the genome, and 23.9% of Vpep genes lie in 7.8% of the genome.

Conclusion

We have used RNAi to examine the loss-of-function phenotypes of about 86% of predicted genes in *C. elegans*. To our knowledge, this is the first systematic functional analysis of a metazoan genome. Of the 1,528 genes for which we could assign an RNAi phenotype, over two-thirds had not been previously associated with a biological function *in vivo*. In addition, we have created an RNAi feeding library of bacterial clones that can be replicated and reused for an unlimited number of future genome-wide RNAi screens in *C. elegans*.

Much as the genome sequence has provided an invaluable platform for investigating *C. elegans* biology, these data and the availability of this library will form a useful tool for functional genomic studies in *C. elegans*. In the future, an analogous genome-wide RNAi library approach could be extended to mammalian cells by capitalizing on techniques using DNA constructs to encode hairpin RNAs[29–34]. We anticipate that in the coming years the quantity of functional data derived from RNAi-based screens in *C. elegans* and in other organisms will greatly expand our understanding of how genes function to bring about the phenotype of an organism. □

Methods

Generation of bacterial feeding library

Polymerase chain reaction (PCR) products were generated using the Research Genetics *C. elegans* GenePairs primer set of 19,213 primer pairs. The set of predicted genes used includes only those genes thought to encode proteins. Primer sequences are listed on the Kim Lab website at Stanford University (http://cmgm.stanford.edu/~kimlab/primers.12-22-99.html). Current alignments of predicted GenePair PCR products on the *C. elegans* genome are available at WormBase (http://www.wormbase.org). We generated PCR products and constructed bacterial strains as described[2]. Inserts were checked for the correct size and confirmed by PCR using the original GenePair oligomers. The whole-genome library consists of 16,757 clones, which represent 87.2% of the GenePairs set and are predicted to correspond to 86.3% of *C. elegans* predicted genes[18], exclusive of cross-RNAi interactions (see below). To assess the quality of the cloning procedure, we sequenced 100 random clones and found all of them to be correct. For the 13% of GenePairs for which no bacterial strain was made, either the GenePair failed to generate a PCR product or the generated product could not be cloned into the T-tailed vector; up to three cloning attempts were made for each GenePair. Supplementary Table 2 gives the complete list of GenePairs and RNAi phenotype class, and indicates whether a clone is available.

Screening using RNAi by feeding

We carried out RNAi as described[2,7]. Embryonic lethality was defined as $>$10% dead embryos, and sterility required a brood size of $<$10 among fed worms (Ste) or their progeny (Stp); wild-type worms under similar conditions typically have $>$100 progeny. Each post-embryonic phenotype was required to be present among at least 10% of analysed worms; the phenotypes assayed were Emb (embryonic lethal), Ste (sterile), Stp (sterile progeny), Gro (slow post-embryonic growth), Lva (larval arrest), Lvl (larval lethality), Adl (adult lethal), Bli (blistering of cuticle), Bmd (body morphological defects), Clr (clear), Dpy (dumpy), Egl (egg-laying defective), Him (high incidence of males), Lon (long), Mlt (moult defects), Muv (multivulva), Prz (paralysed), Pvl (protruding vulva), Rol (roller), Rup (ruptured), Sck (sick) and Unc (uncoordinated). Phenotypes expressed in adults (such as Egl) were difficult to score in this screen because food became limiting at this time point; some of the late expressing phenotypes will therefore have been missed. Detailed listings of GenePairs with corresponding RNAi phenotypes are given in Supplementary Tables 3 and 4 and are available at WormBase (http://www.wormbase.org).

Bioinformatic analyses

We carried out BlastP[35] analyses for all *C. elegans* predicted genes against similar databases (downloaded on 13 Feb 2002) for *S. cerevisiae* (6,183 entries), *Arabidopsis* (25,813 entries), *Drosophila* (13,957 entries) and *Homo sapiens* (36,493 entries), or against *C. elegans* itself. *C. elegans* genes with orthologues were defined as those with BlastP *E* values of less than 10^{-10} with conservation extending over at least 80% of matched protein lengths; 21% of predicted genes in *C. elegans* have such conservation. Predicted gene products were placed into functional classes by manual inspection, primarily using data from Proteome, InterPro release 4.0 (ref. 12) and BLAST analysis[35,36]. We could place 41% of all predicted genes into 1 of 16 functional classes (Supplementary Table 2), with the remaining 59% having unknown function.

Predicted genes targeted by a given bacterial clone were determined by comparing electronic PCR (ePCR) products corresponding to the bacterial clone insert (ftp://ftp.ncbi.nlm.nih.gov/pub/schuler/e-PCR)[37] obtained using chromosome DNA files from the WS61 release of Wormbase (ftp://ftp.sanger.ac.uk/pub/wormbase) to gene predictions from the same database. Roughly 94% of bacterial strains tested correspond to a single predicted gene. To identify genes elsewhere in the genome that might be targeted by cross-RNAi owing to strong homology of part of the gene to the ePCR product, we found genes having $>$80% identity over a region of at least 200 nucleotides for each ePCR product by parsing BlastN results against Wormpep release 71. In total, 1,528 clones with RNAi phenotypes could be assigned directly to a single *C. elegans* predicted gene; these are listed in Supplementary Table 3. By contrast, 194 clones with RNAi phenotypes could not be assigned definitively to a single predicted gene; these are listed in Supplementary Table 4 and include GenePairs with either no or multiple ePCR products or for which the ePCR product is not predicted to overlap any coding sequence.

We found chromosomal regions of significant over- or underrepresentation by considering moving windows of 250 consecutive genes along the chromosomes, and by examining whether the number of genes showing a particular phenotype or in a particular expression cluster within a window was significantly different from that expected according to the genomic mean, using a 1% significance level in a two-tailed test using the binomial distribution. Figure 4b and Supplementary Fig. 3 show continuous significant windows, from the midpoint of the leftmost to the midpoint of the rightmost window. Gene positions were taken from the predicted gene set from Wormbase release WS61.

Received 2 July; accepted 29 October 2002; doi:10.1038/nature01278.

1. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391,** 806–811 (1998).
2. Fraser, A. G. *et al.* Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* **408,** 325–330 (2000).
3. Maeda, I., Kohara, Y., Yamamoto, M. & Sugimoto, A. Large-scale analysis of gene function in *Caenorhabditis elegans* by high- throughput RNAi. *Curr Biol* **11,** 171–176 (2001).
4. Gonczy, P. *et al.* Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. *Nature* **408,** 331–336 (2000).
5. Timmons, L. & Fire, A. Specific interference by ingested dsRNA. *Nature* **395,** 854 (1998).
6. Timmons, L., Court, D. L. & Fire, A. Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in *Caenorhabditis elegans*. *Gene* **263,** 103–112 (2001).
7. Kamath, R. K., Martinez-Campos, M., Zipperlen, P., Fraser, A. G. & Ahringer, J. Effectiveness of specific RNA-mediated interference through ingested double-stranded RNA in *C. elegans*. *Genome Biol.* **2,** 1–10 (2001).
8. Pryer, N. K., Salama, N. R., Schekman, R. & Kaiser, C. A. Cytosolic Sec13p complex is required for vesicle formation from the endoplasmic reticulum in vitro. *J. Cell Biol.* **120,** 865–875 (1993).
9. Tavernarakis, N., Wang, S. L., Dorovkov, M., Ryazanov, A. & Driscoll, M. Heritable and inducible genetic interference by double-stranded RNA encoded by transgenes. *Nature Genet.* **24,** 180–183 (2000).
10. Piano, F., Schetter, A. J., Mangone, M., Stein, L. & Kemphues, K. J. RNAi analysis of genes expressed in the ovary of *Caenorhabditis elegans*. *Curr. Biol.* **10,** 1619–1622 (2000).
11. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **28,** 37–40 (2000).
12. Apweiler, R. *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29,** 37–40 (2001).
13. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
14. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
15. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
16. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 563–567 (1996) 546.
17. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
18. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018 (1998).
19. Kuwabara, P. E. Developmental genetics of *Caenorhabditis elegans* sex determination. *Curr. Top. Dev. Biol.* **41,** 99–132 (1999).
20. Reinke, V. *et al.* A global profile of germline gene expression in *C. elegans*. *Mol. Cell* **6,** 605–616 (2000).
21. Kelly, W. G. *et al.* X-chromosome silencing in the germline of *C. elegans*. *Development* **129,** 479–492 (2002).
22. Barnes, T. M., Kohara, Y., Coulson, A. & Hekimi, S. Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. *Genetics* **141,** 159–179 (1995).
23. Kim, S. K. *et al.* A gene expression map for *Caenorhabditis elegans*. *Science* **293,** 2087–2092 (2001).
24. Lercher, M. J., Urrutia, A. O. & Hurst, L. D. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genet.* **31,** 180–183 (2002).
25. Spellman, P. T. & Rubin, G. M. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J. Biol.* **1**(1), paper no. 5 ⟨http://jbiol.com/content/1/1/5⟩ (2002).
26. Cohen, B. A., Mitra, R. D., Hughes, J. D. & Church, G. M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genet.* **26,** 183–186 (2000).
27. Roy, P. J., Stuart, J., Lund, J. & Kim, S. K. Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. *Nature* **418,** 975–9 (2002).
28. Kruglyak, S. & Tang, H. Regulation of adjacent yeast genes. *Trends Genet.* **16,** 109–111 (2000).
29. Paddison, P. J., Caudy, A. A., Bernstein, E., Hannon, G. J. & Conklin, D. S. Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. *Genes Dev.* **16,** 948–958 (2002).
30. Donze, O. & Picard, D. RNA interference in mammalian cells using siRNAs synthesized with T7 RNA polymerase. *Nucleic Acids Res.* **30,** e46 ⟨http://nar.oupjournals.org/cgi/content/full/30/10/e46⟩ (2002).
31. Elbashir, S. M., Harborth, J., Weber, K. & Tuschl, T. Analysis of gene function in somatic mammalian cells using small interfering RNAs. *Methods* **26,** 199–213 (2002).
32. Paul, C. P., Good, P. D., Winer, I. & Engelke, D. R. Effective expression of small interfering RNA in human cells. *Nature Biotechnol.* **20,** 505–8 (2002).

33. Miyagishi, M. & Taira, K. U6 promoter driven siRNAs with four uridine 3′ overhangs efficiently suppress targeted gene expression in mammalian cells. *Nature Biotechnol.* **20,** 497–500 (2002).
34. Sui, G. *et al.* A DNA vector-based RNAi technology to suppress gene expression in mammalian cells. *Proc. Natl Acad. Sci. USA* **99,** 5515–5520 (2002).
35. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
36. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
37. Schuler, G. D. Sequence mapping by electronic PCR. *Genome Res.* **7,** 541–550 (1997).
38. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215 (2000).
39. Wood, V. *et al.* The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415,** 871–880 (2002).

Supplementary Information accompanies the paper on *Nature*'s website (➧ **http://www.nature.com/nature**).

Acknowledgements We thank T. Blumenthal, B. Kelly, S. Kim and V. Reinke for discussions and sharing data before publication; B. Shumacher for assistance early in the project; and G. Chalklin, J. Peacock, J. Abbott and K. Rossiter for preparing the media. R.S.K. was supported by a Howard Hughes Medical Institute Predoctoral Fellowship; A.G.F. by a US Army Breast Cancer Research Fellowship; Y.D., R.D., M.G., D.P.W. and P.Z. by the Wellcome Trust; G.P. by the Canadian Institute of Health Research and the Wellcome Trust; A.K. by the European Molecular Biology Laboratory; N.L.B. by the European Molecular Biology Organization; S.M. by the Centro de Investigacion del Cancer; M.S. by a Swiss National Science Foundation fellowship and J.A. by a Wellcome Trust Senior Research Fellowship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.A.
(e-mail: jaa@mole.bio.cam.ac.uk).