REVIEW: CELL BIOLOGY

Conservation and Novelty in the Evolution of Cell Adhesion and Extracellular Matrix Genes

Harald Hutter,[1]* Bruce E. Vogel,[2] John D. Plenefisch,[3] Carolyn R. Norris,[2] Rui B. Proenca,[2] John Spieth,[4] Chaobo Guo,[2] Surjeet Mastwal,[2] Xiaoping Zhu,[2]† Jochen Scheel,[5] Edward M. Hedgecock[2]

New proteins and modules have been invented throughout evolution. Gene "birth dates" in *Caenorhabditis elegans* range from the origins of cellular life through adaptation to a soil habitat. Possibly half are "metazoan" genes, having arisen sometime between the yeast-metazoan and nematode-chordate separations. These include basement membrane and cell adhesion molecules implicated in tissue organization. By contrast, epithelial surfaces facing the environment have specialized components invented within the nematode lineage. Moreover, interstitial matrices were likely elaborated within the vertebrate lineage. A strategy for concerted evolution of new gene families, as well as conservation of adaptive genes, may underlie the differences between heterochromatin and euchromatin.

The genome of the nematode *Caenorhabditis elegans*, now fully sequenced, affords remarkable insights into the origin and nature of multicellular life (*1*). Moreover, it raises challenging, often unforeseen, questions about the molecular processes and evolutionary consequences of genome change. Some 20% of *C. elegans* genes have orthologs in the budding yeast *Saccharomyces cerevisiae* (*2*) that function in cellular processes common to all eukarya. Beyond those shared with yeast, about 30% of *C. elegans* genes have known orthologs in insects or vertebrates that are involved in developmental and physiological processes common to all higher animals (*3–5*). The remaining genes are thus far found only in nematodes (*6*). About half are single-copy genes and could represent ancient genes not yet discovered in other phyla. If so, as many as 50% of all *C. elegans* genes arose sometime between the radiations of cellular eukarya [about 2 gigayears ago (Gya)] and metazoa (about 0.8 Gya), and are therefore expected to be found in all higher animals. *C. elegans* is an excellent experimental model for studying conserved functions of these inherently metazoan genes. Finally, comparison of the *C. elegans* genome with other nematodes, and with itself, reveal robust, ongoing processes of gene invention (*7–9*).

We examined the evolution of extracellular matrix and cell adhesion molecules, protein classes that frequently overlap in structure or interact molecularly. To identify candidate genes, we used representative insect and mammalian proteins, or their fragments, as queries for BLAST searches of Wormpep (*10*, *11*). From these initial hits, we performed reciprocal BLAST searches to identify potential insect or mammalian orthologs in GenBank, and to expand the sample of nematode proteins. Direct searches against Wormpep allowed identification of nematode-specific protein domains and families (*12*). For all protein domains summarized in Web table 1 (*13*) and discussed below (*12*, *14*), this search cycle proved a sensitive means of detection with no false-negatives to the best of our knowledge. To confirm known protein domains and to count tandem repeats, we used Pfam profiling with the hidden Markov model algorithm, HMMer (*15*); profile parameters were set to their most sensitive value, allowing for module fragments due, for example, to imperfect GENEFINDER predictions. We used manual sequence alignment assisted by CLUSTAL to define new protein domains. Potential signal sequences, transmembrane helices, or glycosyl-phosphatidylinositol (GPI)-anchoring signals, were identified by PSORTII (*16*). Finally, these genes were sorted by chromosome position with known genetic loci. Proteins with orthologs in all eukarya, for example, ribosomal proteins, histones, and tubulins, were included for comparison. Genes with known mutations were identified from a *C. elegans* database (ACEDB); cDNA matches were identified from BLASTN searches, or counted from online lists, correcting for duplicate entries of 5′ and 3′ expressed sequence tags (ESTs) from a single cDNA clone (*17*). Genes are identified below by their Wormpep accession numbers (*11*). In addition, where available, protein and gene names are appended to these Wormpep accession numbers using colons and parentheses, respectively. Where a single protein apparently comprises two or more separate database entries, we list NH_2-terminal fragments first, e.g., ZK944.4/ZK944.3. Further information regarding our analysis is described online at www.mpimf-heidelberg.mpg.de/ewgdn/genome_paper/.

Basement Membrane Proteins and Receptors

Basement membranes are polymeric sheets of laminin, collagen IV, and associated proteins found on the basal surfaces of epithelia and condensed mesenchyma that provide a substratum for attachment and present a barrier to cell mixing during development (*18*). Basement membrane components are among the oldest and most conserved extracellular matrix proteins (*19*). In *C. elegans*, two distinct laminin molecules, designated $\alpha_A\beta\gamma$ and $\alpha_B\beta\gamma$, arise from four laminin chain genes, α_A::T22A3.8, α_B::K08C7.3 (*epi-1*), β::W03F8.5 (*lam-1*), and γ::C54D1.5 (Fig. 1). Comparison of these four genes suggests that exchange between two genes of protolaminin (with subunit composition δεε) resulted in two parental, $\delta^N\delta^C$ and $\varepsilon^N\varepsilon^C$, and two recombinant, $\delta^N\varepsilon^C$ and $\varepsilon^N\delta^C$, chains, seen in nematodes today. Laminins have duplicated further within the vertebrate lineage. Thus, α_A and α_B branches split into α1/α2 and α3/α4/α5 chains, respectively (*20*).

Basement membrane collagens are encoded by three genes, α1(IV)::K04H4.1 (*emb-9*), α2(IV)::F01G12.5 (*let-2*), and α1(XV/XVIII):: F39H11.4. Other basement membrane proteins have unique representatives, for example, agrin::F41G3.8, fibulin::F56H11.1, Kallmann-syndrome protein::K03D10.1, nidogen:: F54F3.1 (*nid-1*), osteonectin:: C44B12.2 (*ost-1*), and perlecan::ZC101.2 (*unc-52*). Several, thus far novel, matrix proteins are required for cellular attachments of mechanosensory neurons and other tissues. This category includes hemicentin::F15G9.4 (*him-4*),

[1]Max-Planck-Institute for Medical Research, Jahnstrasse 29, 69120 Heidelberg, Germany. [2]Department of Biology, Johns Hopkins University, Baltimore, MD 21218, USA. [3]Department of Biology, University of Toledo, Toledo, OH 43606, USA. [4]The Washington University Genome Sequencing Center, St. Louis, MO 63108, USA. [5]Max-Planck-Institute for Developmental Biology, Spemannstrasse 35, 72076 Tübingen, Germany.

*To whom correspondence should be addressed. E-mail: hutter@mpimf-heidelberg.mpg.de

†Present address: Laboratory of Molecular Biology, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA.

MEC-5::E03G2.3, and MEC-9::C50H2.3a (*21*). Adamalysin, astacin, matrixin, and neprilysin, which are among the families of metalloproteases observed, are implicated in extracellular matrix interactions or neuropeptide processing (*22*).

Integrins couple assembly of various extracellular matrix and cytoskeletal polymers (*23*). In *C. elegans*, there are two α chains, INA-1::F54F2.1 and PAT-2::F54G8.3, representing the laminin-binding and RGD-binding branches of this family, respectively, and two β chains, INB-1::C05D9.3, PAT-3::ZK1058.2, which allows four possible heterodimers. The patterning of these receptors, together with the two laminins, may specify where apposing basement membranes should fuse together or remain separate.

The dystroglycan-dystrophin complex couples laminin and agrin to the membrane cytoskeleton in skeletal muscle and other tissues (*24*). In humans, mutations in various components of this complex, including dystrophin and sarcoglycans, as well as laminin α2 itself, cause muscular dystrophies. Components of the dystroglycan complex have conserved *C. elegans* orthologs, for example, dystroglycan::T21B6.1, α/ε-sarcoglycan::H22K11.4, β-sarcoglycan::K01A2.1 and γ/δ-sarcoglycan::F07H5.2, dystrophin::F15D3.9/F32B4.3 (*dys-1*), and various syntrophins. Family T07D3.4 encodes homologs of the secreted protein fukutin, implicated in Fukuyama-type congenital muscular dystrophy (*25*).

What extracellular matrix proteins are not found in nematodes? Like connective tissue itself, interstitial matrix polymers, such as elastin, fibrillar collagen, and fibronectin, are generally absent in *C. elegans*, suggesting that these genes were elaborated within the vertebrate lineage. Fibrillar collagen genes occur in other invertebrates, however, indicating that this and perhaps other matrix components have been lost in the nematode lineage (*26*).

Cell Adhesion Molecules

We classified cell adhesion molecules and other receptors into superfamilies on the basis of their NH_2-terminal domains, which are largely responsible for ligand recognition. The five largest superfamilies (*27*) begin with CA, EG, IG, LA, or LR repeats [(*14*), Web table 1 and Web figure 1 (*13*)]. For convenience, we refer to these proteins collectively as cadherins, EgfCAMs, IgCAMs, LdlCAMs, and LrrCAMs, respectively. Finally, several well-known cell adhesion molecules that defy easy classification have apparent orthologs, for example, chondroitin sulfate proteoglycan NG2::C48E7.6, polycystins::ZK945.10/.9 (*lov-1*) and Y73F8A.B, selectin::C54G4.4, and syndecan::F57C7.3.

IgCAMs. Twenty-six genes encode predicted transmembrane or GPI-anchored proteins with extracellular IG modules (Web figure 1). With the exception of Lrr(Ig) CAMs (discussed below), the IG domain is located at the NH_2-terminus. Thirteen proteins, Ig(F3)CAMs, have one or more fibronectin III (F3) modules after their IG domain, whereas one protein, UNC-5::B0273.4, has thrombospondin 1 (T1) modules instead. Two neuregulin-like proteins, F28E10.2 and F48C5.1, contain an EG module after their IG domain. The intracellular domains of the transmembrane IgCAMs have characteristic enzymatic and binding functions. Six additional genes encode extracellular matrix proteins with IG domains.

The IgCAMs are principal mediators of cell recognition and adhesion in the developing nervous system and other tissues (*28*). These proteins act combinatorially among themselves, and with receptors from other structural families, to pattern cell movements and attachment. In particular, GPI-anchored, or small soluble IgCAMs (discussed below), might directly modulate the ligand specificities of transmembrane IgCAMs. Whereas IG and F3 motifs are found in a few yeast proteins (*2*), the combination of these motifs and their recruitment for cell adhesion occurred within the metazoan lineage. The immunoglobulin superfamily has remained static in the nematode lineage, Nine of the thirteen Ig(F3)CAMs in *C. elegans* have known orthologs in insects and vertebrates. We predict the various new IgCAMs identified in this species have yet undiscovered orthologs in other animals.

An ancient mechanism evolved for adhesive recognition, IgCAMs and their effectors were adapted for antigen recognition in the vertebrate immune system. In lymphocytes, antigen receptors signal through *fyn* and related *src* family protein tyrosine kinases to elicit cellular responses. In neurons, L1 and neural cell adhesion molecule (NCAM) can signal through these same nonreceptor tyrosine kinases (*29*). Covalent association of antibody heavy and light chains contributes to the specificity and versatility of antigen recognition. Reminiscent of these light chains, *C. elegans* has seven secreted proteins comprising just two tandem IG modules. Conceivably, these chains complex with membrane-associated IgCAMs to form functional receptors.

Semaphorins and plexins. Five genes encode proteins with a semaphorin domain at their NH_2-terminus including three semaphorins, semaphorin I::Y54E5B.1, semaphorin II::Y71G12A_05.G and D1037.2, and two plexins, K04B12.1 and Y55F3B_45.A

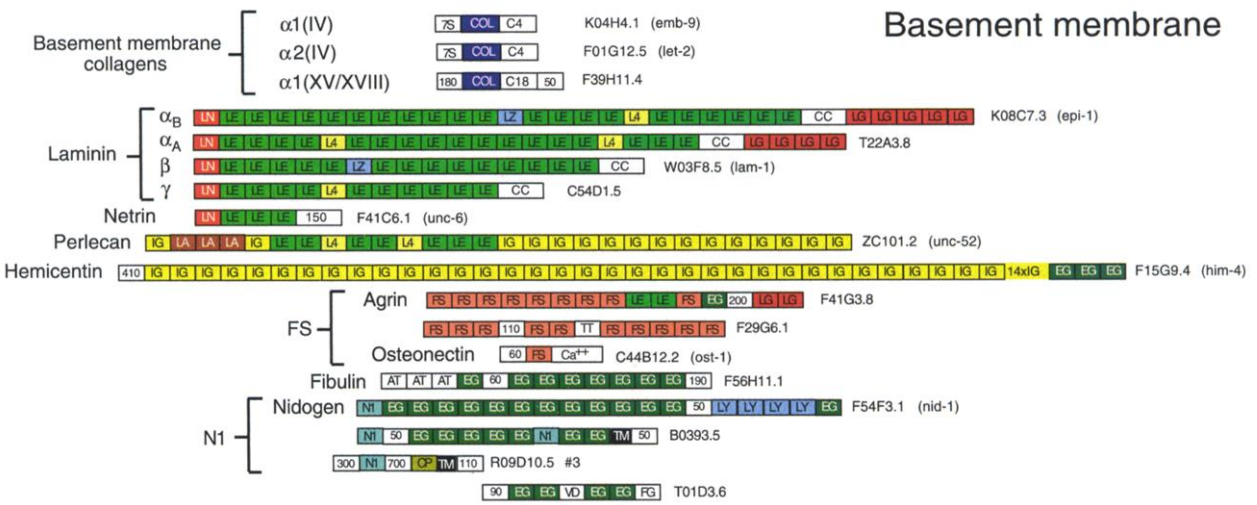

Fig. 1. Organization of selected basement membrane molecules. Lettered boxes represent identified protein motifs; numerals indicate the lengths of uncharacterized sequences (not drawn to scale). Extracellular modules and catalytic domains are described online (*56*). Other features, including intracellular domains, are abbreviated: C4 (collagen IV COOH-terminus), C18 (collagen XV/XVIII COOH-terminus), Ca^{++} (calcium-binding domain), CC (coiled-coil domain), and 7S (collagen IV NH_2-terminus). The suffix "#n" indicates the number of genes with this same organization. Cell adhesion molecules are available in Web figure 1 (*13*).

(Web figure 1). Semaphorins are guidance cues in the developing nervous system, whereas plexins, acting as semaphorin receptors, mediate growth cone collapse (*30*).

LrrCAMs. Twenty-three genes encode proteins with an extracellular LR-repeat domain at their NH_2-terminus including apparent orthologs of slit::C26G2.C/F40E10.4, peroxidasin::K09C8.5 and ZK944.4/.3, chaoptin::C56E6.6, 18-wheeler::T05A1.3, and FSH/TSH-receptor::C50H2.1 (Fig. 1). Three of these proteins, designated Lrr(Ig) CAMs, have one or more IG modules following their LR-repeat domain, for example, GAC1::F20D1.7 and LIG-1::T21D12.9. Several LrrCAMs have been implicated in adhesive recognition in the nervous system, including regulation of synapse formation (*31*). Additional genes encode proteins with intracellular LR-repeat domains, which are possibly unrelated motifs that converged onto a similar protein fold (*32*).

Cadherins, latrophilins, and neurexins. Ten genes encode classical cadherins with a CA repeat domain at their NH_2-terminus; three more genes encode FAT-related cadherins where a crumbs-like region of alternating EG and laminin G (LG) modules follows the CA domain. Classical and FAT-related cadherins are implicated in both general adhesion between cells and specialized junctions, for example, adherens, desmosomes, and synapses (*33*). One additional cadherin, CELSR1::F15B9.7, has a FAT-related NH_2-terminus followed by laminin epidermal growth factor–like (LE) repeats and a latrophilin-related COOH-terminus (Web figure 1); a mammalian ortholog is expressed in the nervous system (*34*).

Latrophilin and neurexin are presynaptic membrane proteins identified as receptors for latrotoxin, a neurotoxin from black widow spider venom that triggers massive, unregulated exocytosis of synaptic vesicles from nerve terminals (*35*). Like CELSR1, latrophilins are members of the secretin receptor family, an ancient branch of serpentine receptors implicated in secretory coupling. Two genes encode latrophilins, B0286.2 and B0457.1, and three more genes encode secretin receptor–related proteins without large extracellular domains. These proteins could play roles in synapse formation, triggering exocytosis in response to potential synaptic targets, or synapse maintenance (*36*). Finally, five genes, including axotactin::W03D8.6, crumbs::F11C7.4, neurexin I/II/III::C29A12.4 and neurexin IV::F20B10.1, encode crumbs-like receptors with alternating EG and LG repeats.

New Proteins Combine Novel Modules and Select Old Parts

New genes arise from specific, often novel, feedstock, which changes over time, suggesting that not all regions of eukaryotic genomes are equally available for gene invention. Some motifs used for gene invention in early metazoans seem inert today, whereas once minor or entirely novel sequences have became important within the nematode or chordate lineages. We identified more than 40 ancient protein motifs [(*14*) and Web table 1], clearly predating the metazoan radiation, found in the extracellular domains of *C. elegans* proteins (*27*). By far the most promiscuous extracellular module, epidermal growth factor (EG) appears in 30 distinct structural contexts. At the other extreme, the LN motif, implicated in polymerization, occurs at the NH_2-terminus of laminin chains and netrin but nowhere else. Remarkably, most ancient motifs occur in a stable set of contexts from nematodes through chordates. However, some have been used for new "gene shuffling" within specific lineages. For example, CK, FC, FS, KR, SR, and VD motifs are more promiscuous in vertebrates than nematodes (*37*). By parsimony, those contexts shared with nematodes likely reflect the ancestral functions of these domains. Conversely, CL and KU modules have been recruited for many novel contexts in nematodes (Web figure 1).

Some extracellular motifs present in vertebrates are apparently absent in *C. elegans*, and vice versa, suggesting new protein modules have been invented more or less continuously throughout eukaryote evolution (*2*, *6*, *38*). Twenty nematode-specific protein motifs were found in the extracellular domains of *C. elegans* proteins (*12*). Most of these motifs are present in only one or two structural contexts and may have duplicated quite recently. However, DC, SX, and CT modules are more promiscuous and presumably expanded early within the nematode lineage. The DC module, a 45-residue motif with six conserved cysteines, occurs in more than 60 secreted or membrane proteins representing nine distinct structural contexts (Web figure 1). It is interesting that these proteins contain various ancient modules, i.e., EG, IG, F3, KU, TY, and WA, intermixed with apparently nematode-specific motifs. Although these observations suggest a relatively ancient origin, the DC module is not currently represented in any human gene or EST sequence. By inference, this module, rare or absent in our common nematode-chordate ancestors, expanded greatly in early nematodes. Secreted proteins with SX (SXC) modules include nematode surface coat components and several enzymes possibly involved in cuticle maturation (*6*). Finally, the CT (cuticulin) module occurs in proteins found at the apical surface of nematode epidermis and mucosa, as well as a transmembrane protein from *Drosophila* epidermis (*39*). Expression and phenotype studies suggest a role in epithelial morphogenesis.

New Genes Arise in Specific Regions

Many new genes have arisen within the *C. elegans* lineage since the metazoan radiation. Some arose through duplication of known genes; others were apparently invented within the nematode lineage itself. We examined gene families and superfamilies of various ages to learn whether new genes arise evenly throughout the genome, and to gain insight into possible mechanisms. The immunoglobulin superfamily, which has remained remarkably static within the nematode lineage, is dispersed throughout the genome as single genes, or rarely, pairs, in regions overall enriched in adaptive, often highly expressed, genes (Fig. 2). The younger superfamilies DC and CT, which we suggest expanded comparatively early within the nematode lineage, have a similar genomic organization. The SX superfamily comprises both dispersed genes, mostly encoding proteins with catalytic domains, and several local gene clusters (discussed below) encoding simpler proteins with SX modules alone (*6*).

Remarkably, a majority of potentially nematode-specific genes occur in large families, some with over 200 members in *C. elegans* (*1*, *3*–*9*). Several gene families are implicated in structures and processes important to all nematodes, for example, collagenous cuticle. Although these families clearly expanded within the nematode lineage, until other complete genomes are available for comparison, it remains possible that their founding members originated earlier. Indeed, several of the largest families in *C. elegans* are evidently recent expansions of individual members of more ancient families, for example, chitinase, glutathionine-S-transferase, nuclear receptor, SCP (TPX), serpentine receptor, and UDP-glucuronyl transferase.

Cuticle collagens form one of the largest, and possibly oldest, nematode-specific gene families (*40*). Most of these 160 genes are represented in *C. elegans* EST databases and many have known mutations affecting body morphology. The family is dispersed throughout the genome, no cluster larger than four genes, at 128 sites. The flanking regions are enriched in highly expressed genes and mutant loci (Fig. 2). Why so many genes? Many isoforms are expressed in characteristic order during each molt cycle to create a layered cuticle; others provide stage- or region-specific modifications. Requirements for rapid, synchronous synthesis of large amounts of mRNA may select for increased copy number. Why are the gene products so similar? Requirements of triple-helix formation and polymerization may impose structural constraints on these chains that belie their true age. Like the DC and CT superfamilies, we suggest the cuticle collagen gene family expanded early in nematodes and is now maintained largely by independent selection on each member.

Most multigene families in *C. elegans* are strongly clustered within the genome. Selecting highly similar gene pairs, Semple and Wolfe (*9*) compared the relative spacing and orientation for 2929 duplicated genes representing 655 families in Wormpep release 12. Local gene clusters with mixing and inversion (discussed below), not pure tandem repeats or unlinked duplications, dominate the aggregate distribution of gene families in this large sample. Using dot-matrix and BLAST comparisons, we examined several large gene clusters in detail, finding frequent examples of recent gene duplication or conversion. Two representative gene families, C01B7.7 and M176.8 (chitinase), are summarized in Figs. 2 and 3. Compared with cuticle collagens and various superfamilies, these apparently younger families are less evenly dispersed through the genome, occurring in a few large clusters, often with some isolated members (*41*). Within a cluster, repeated genes tend to have common orientation and regular spacing, but frequently, this pattern is disrupted by partial gene duplications and inversions. Sequence comparisons indicate these families expanded primarily through some mechanism of local duplication, i.e., adjacent genes were generally more similar than distant pairs, but sequences sometimes move to farther sites or even separate clusters. Often two or more unrelated gene families are intermixed within a single cluster. Examples of very recent duplications or conversion suggest genes but not intergenic regions were moved. Robertson (*8*) found a very similar pattern of gene expansion and movement, supported by comparisons with *Caenorhabditis briggsae*, in a study of serpentine receptor families.

What molecular processes and selective forces shape the evolution of gene families? Contrary to previous belief, random gene duplication followed by independent, divergent evolution of the copies cannot explain the distribution of gene families and superfamilies in *C. elegans*. As they obtain more chances for duplication and divergence, this model predicts older gene families should tend to be larger, more divergent in structure and function, and more dispersed in the genome, than younger families. Eventually, such processes would produce protein superfamilies sharing only limited regions of homology. Contrary to these predictions, many old superfamilies appear relatively static, whereas large gene families are often young and dynamic.

What mechanisms are responsible for clustering of young gene families? Unequal crossovers and sequence drift could create tandem duplications where adjacent repeats are more similar than distant sequences (*42*). Occasional duplication or conversion to distant sites might drive concerted evolution of an entire family (*43*). We favor a role for mRNA intermediates. Gene clusters would expand through integration of cDNAs made from nascent transcripts in the same region. Similar chromosome-associated reactions, or RNA-mediated integration, have been proposed for retrotransposition of non–long terminal repeat (non-LTR) retrotransposable elements, short interspersed nuclear elements (SINEs) and processed pseudogenes in other eukaryotes (*44*). Frequent precise loss of individual introns during gene duplication (*8*), could be explained by limited processing of mRNAs before reverse transcription. Intermixing of gene families, inversions, and movements to farther sites might occur if the coupling of transcription and integration sites were relaxed. Indeed non-LTR retrotransposons can mediate gene movement and exon shuffling in cultured somatic cells (*45*). Finally, our model, which explains why gene duplications and conversions rarely extend into intergenic regions (*9*), suggests that transcribed and regulatory sequences generally have independent origins (*46*).

Is It Heterochromatin?

Eukaryote genomes are generally packaged into euchromatin and heterochromatin where the former regions are enriched in expressed genes and contain most known mutant loci (*47*). Can we identify bona fide euchromatin in the *C. elegans* genome sequence? Most members of the immunoglobulin superfamily have a single representative in *C. elegans*. In many cases, these proteins have been shown to be adaptive (conferring increased fitness) in insects or vertebrates, if not nematodes themselves. By inference, the *C. elegans* orthologs must be located in transcriptionally active regions of the genome, presumably euchromatin. Inspection of the regions flanking these genes reveals an assortment of structurally unrelated, often unique, genes of comparatively ancient origin (*48*). Consistent with the notion that these regions represent euchromatin, many of the flanking genes are themselves highly expressed, or else known through mutation to be adaptive (Fig. 2). Extrapolating to similar regions, most cuticle collagen genes are likely contained in euchromatin, and similarly for the DC superfamily, although no mutants have been found in the latter.

The fraction of predicted protein genes on each chromosome with known visible muta-

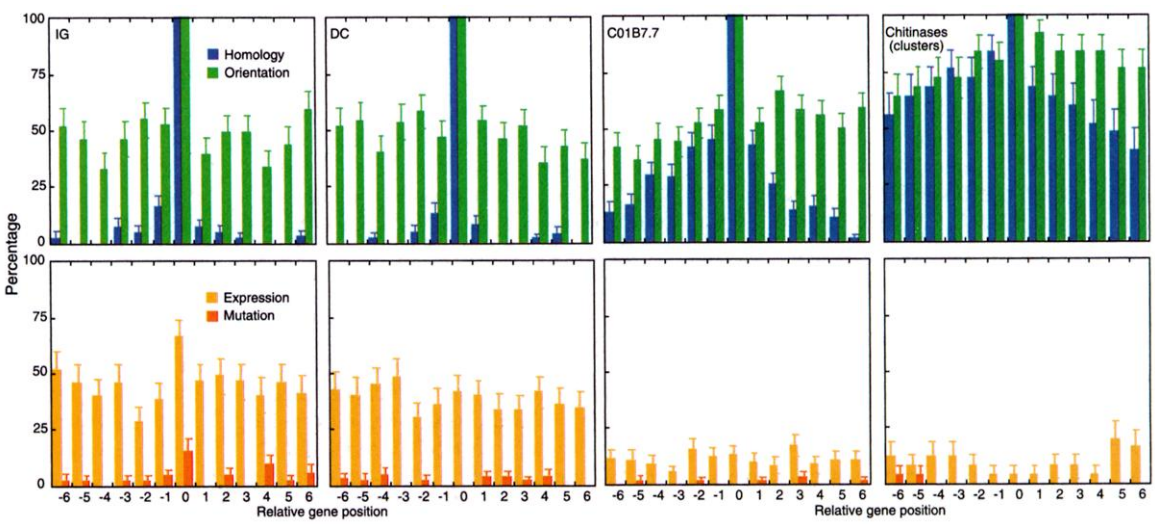

Fig. 2. Genomic regions of selected gene families and superfamilies illustrating an inverse relation between gene expression and clustering. Regions from gene families and superfamilies representative of varying ages were compared by using a window of 13 genes centered on the target gene in 5′ to 3′ orientation. Intracellular members of the IG superfamily are not included. Plots for cuticle collagens, as well as chitinases outside the local gene clusters (*41*), are available in (*57*) and Web figure 2 (*13*), with the mean numbers of cDNAs in the current EST databases (± SD) for all gene families and superfamilies. For each family, bars on the upper panel indicate the percentage of genes at relative positions "– 6" to "+6" matching the target gene at position "0" by structural family or orientation, respectively. Bars on the lower panel indicate the percentage of genes with one or more cDNAs in the current EST databases (*17*), or known phenotypic alleles, respectively. The error symbols above each bar indicate the SD for binomial sampling. For IG superfamily, $n = 45$; for DC superfamily, $n = 50$; for the C01B7.7 family, $n = 61$; for the chitinases (clusters), $n = 25$.

tions correlates strongly, but negatively, with the fraction of genes in multigene families (*9, 49*). Inspection of these families reveals that clustered genes, which are found rarely, if at all, among characterized *C. elegans* mutants, account for this bias (Fig. 2). Moreover, they are highly underrepresented in the EST databases (*1, 8, 17*); this effect is not absolute as nearly all clusters examined have occasional EST hits (Fig. 3). The simplest explanation for these observations is that most local gene clusters are transcriptionally silent, but these data do not preclude significant levels of gene expression in a few cell types (*50*), or under unusual conditions, combined with functional redundancy among the gene products. Regardless, expression and selection of genes evolving in clusters must be qualitatively different from "typical" adaptive genes as described above.

Heterochromatin was first described cytologically as regions of late replicating DNA that remain condensed during interphase (*47*). Genetic studies indicated these same regions were impoverished in adaptive genes and undergo little recombination during meiosis. Early in situ hybridization studies revealed that heterochromatin is often enriched in simple repeated sequences, or "satellite" DNA. These observations lead to hypotheses that all heterochromatin might have a common, rather simple, sequence organization, and moreover, specific sequence repeats might themselves direct heterochromatin formation. However, subsequent studies, including recent analyses of long, representative genomic sequences, shown that heterochromatic regions are highly dynamic and remarkably heterogeneous in sequence. Several classes of transposable elements, including non-LTR retrotransposons and related SINEs, occur preferentially in heterochromatin (*51*). Moreover, clusters of recently duplicated genes or pseudogenes have been found in pericentromeric and subtelomeric heterochromatin of human chromosomes (*52*); it is unclear whether these duplicated genes are generally expressed or adaptive. Finally, juxtaposition or insertion into heterochromatin can silence otherwise active genes. In both insects and mammals, local duplication of transgenes or endogenous chromosomal sequences can itself cause heterochromatin formation and gene silencing (*53*).

Interphase nuclei in *C. elegans* have numerous regions of condensed heterochromatin, but little is known about their chromosomal arrangement. Does the genome sequence provide clues to chromatin organization at this level? In this species, spindle microtubules tether along the length of the chromosome during mitosis, rather than at a localized kinetochore (*54*); this distribution of kinetochore function could reflect a dispersal of centromeric heterochromatin along the chromosome. Unexpectedly, local gene clusters have several characteristics better ascribed to heterochromatin than euchromatin. Unlike dispersed gene families and superfamilies, most clustered genes predicted by genomic sequencing potentially fail two important criteria for adaptive genes, namely, expression of RNA products and observable phenotypes. Averaging only 2 to 3 kb in length, these repeated sequences generally form complex mixed arrays at multiple chromosomal sites (Figs. 2 and 3). Nonhomologous exchanges between inverted or unlinked duplications common in local gene clusters are a potential source of chromosome rearrangement. Although no studies have measured recombination rates within gene clusters, the stability of *C. elegans* chromosomes might be explained by the suppression of all recombination within these regions.

Like heterochromatin, local gene clusters are dynamic with frequent sequence movement through duplication or conversion (*8*). By contrast, the local arrangement of genes appears relatively stable in other regions of the *C. elegans* genome (*55*). Regions of synteny with *C. briggsae*, which separated 10 to 100 million years ago (*6*), have genome organization we ascribe to euchromatin, i.e., an assortment of structurally unrelated genes of comparatively ancient origin, many of which are highly expressed, or known through mutation to be adaptive. The *unc-4 egl-43* region, shown in Fig. 3, illustrates these features. Whereas this region has undergone large rearrangements, including inversion and translocation, there has been little local movement or duplication of sequences.

Conclusion

The *C. elegans* genome contains both ancient regions enriched in adaptive genes and more dynamic regions associated with emerging gene families. The expansion and collapse of complex gene clusters could reflect an ancient evolutionary process for the invention of new, potentially adaptive genes. Can concerted evolution speed acquisition and fixation of adaptive alleles or the elimination of useless members? Despite considerable interest, the molecular processes and selective forces underlying concerted evolution remain

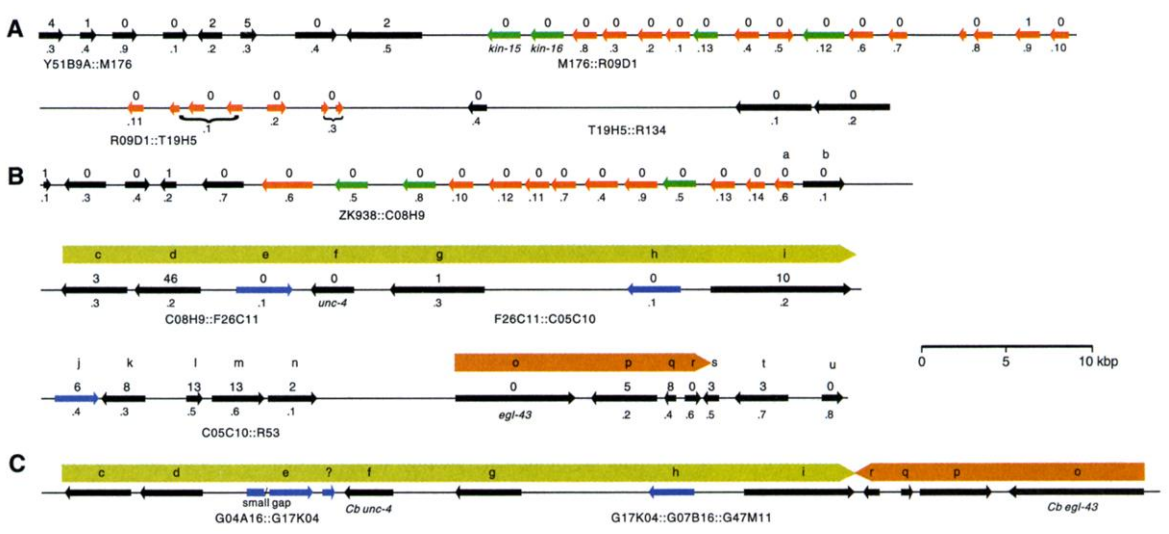

Fig. 3. Chitinase gene clusters. (**A** and **B**) In *C. elegans*, two local gene clusters on chromosome II, separated by 320 kilobase pairs, contain 25 chitinase genes or pseudogenes of the M176.8 family (red arrows), intermixed with seven members of the *kin-15* protein tyrosine kinase family (green arrows). Arrows show gene orientation and extent of the predicted protein coding sequence; gene names and number of reported cDNAs in EST databases are shown below and above these arrows, respectively. The 3′ portions of R09D1.6 and R09D1.8 differ at just one nucleotide among 1497 base pairs, suggesting local sequence movement, possibly gene conversion, is an ongoing evolutionary process within these clusters. (B) The *unc-4 egl-43* region, presumptive euchromatin immediately downstream of the chitinase clusters, encodes three acid phosphatases (blue arrows) and other, structurally unrelated genes (black arrows), many of which are highly expressed. (**C**) *C. briggsae* orthologs of *unc-4*, *egl-43*, and nine other genes from this region, lettered for purpose of comparison, are contained in the genomic contig G04A16::G17K04::C07B16::G47M11. Despite a large inversion, the order, orientation and spacing of genes labeled "c to i" and "o to r" have been conserved between these species (large arrows).

uncertain (*43*). The *C. briggsae* genome sequence, when completed, should help the interpretation of recent genome changes, including mutational mechanisms. We must also learn more about gene expression and function to understand the selective forces on genes evolving in families. Only selection on expressed sequences, whether direct or indirect, allows conservation of genes and exons.

We propose a simple, testable model for gene invention, namely, that heterochromatin is continually expanding through incorporation of cDNAs, creating local gene clusters, tandem protein repeats, and sometimes new exon combinations. At the euchromatin boundary, these sequences must succeed as adaptive genes, or more often, disappear completely. If heterochromatin is the primary site of gene invention today, was it always an organelle of chromosome growth? An attractive hypothesis is that heterochromatin arose during the transition from RNA- to DNA-based life as a mechanism for incorporating cDNA into chromosomes (*46*). In primitive chromosomes, the chromatin structure and enzymatic activities needed for converting RNA-to-DNA and incorporating the product were concentrated at specialized regions that persist today, near telomeres and centromeres, as heterochromatin (*56*).

References and Notes

1. The *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
2. S. A. Chervitz *et al.*, *Science* **282**, 2022 (1998). Orthology and paralogy were introduced over 30 years ago as phylogenetic terms to distinguish gene duplication from speciation. Interspecies orthologs were defined as any genes that diverged strictly after the speciation; all other interspecies homologs, and vacuously intraspecies homologs, are termed paralogs. Purely a cladistic relation, orthology says nothing per se about the degree of structural or functional conservation postspeciation. In particular, because protein genes, and even regions within them, diverge at widely different rates, orthologs may differ from negligibly to significantly in sequence and modular organization.
3. C. I. Bargmann, *Science* **282**, 2028 (1998).
4. G. Ruvkun and O. Hobert, *Science* **282**, 2033 (1998).
5. N. D. Clarke and J. M. Berg, *Science* **282**, 2018 (1998).
6. M. Blaxter, *Science* **282**, 2041 (1998).
7. E. L. L. Sonnhammer and R. Durbin, *Genomics* **46**, 200 (1997).
8. H. M. Robertson, Genome Research **8**, 449 (1998).
9. C. Semple and K. H. Wolfe, *J. Mol. Evol.* **48**, 555 (1999).
10. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
11. Sequences in the Wormpep database (release 17), provided online at www.sanger.ac.uk/Projects/C_elegans/wormpep/, are uniquely identified by clone name and extension, e.g., K08C7.3.
12. Nematode-specific motifs: DC, SX (SXC), CT, VE, FE, CX, DB, EA, IY (worm_family_2), CZ, T4 (worm_family_8), CN, EB, DD, MC, MD, DX, ES, and ET.
13. Supplementary material in the form of a table and expanded figures 1 and 2 is available to *Science* Online subscribers at www.scienceonline/feature/data/1036101.shl
14. Ancient motifs originating in eukarya: IG, immunoglobulin; VA, von Willebrand factor A; F3, fibronectin III; LR, leucine-rich repeat; and PER, peroxidase. Ancient motifs originating in metazoa: AD, ADAM Zn-metalloprotease; ASC, amiloride-sensitive channel; AT, anaphylatoxin; CA, cadherin; CK, COOH-terminal cystine knot; CL, C-type lectin; CP, Sushi/CCP/SCR; CU, CUB; CY, cystatin; DI, disintegrin; DS, delta serrate; EG, epidermal growth factor; FC, coagulation factor 5/8 C; FG, fibrinogen β/γ COOH-terminal; FS, follistatin/Kazal; GL, galaptin/S-type lectin; HX, hemopexin; KR, kringle; KU, Kunitz/BPTI; L4, laminin IV (B-type); LA, LDL-receptor A; LE, laminin EGF-like; LG, laminin G; LN, laminin NH_2-terminus; LY, LDL-receptor YWTD; LZ, laminin IV'; MP, Zn-metalloprotease, astacin type; MX, matrixin; N1, nidogen NH_2-terminus; NL, Notch/LIN-12; PD, P-type/trefoil; PX, plexin; SE, SEA; SM, semaphorin; SR, scavenger receptor C-rich; T1, thrombospondin 1; TY, thyroglobulin I; UL, sea urchin egg lectin; VC, von Willebrand factor C; VD, von Willebrand factor D; and WA, WAP (4-disulfide core). Note: the solidus is used to conform with the annotation system of Pfam.
15. E. L. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, R. Durbin, *Nucleic Acids Res.* **26**, 320 (1998).
16. K. Nakai and M. Kanehisa, *Genomics* **14**, 897 (1992).
17. Y. Kohara, *Tanpakushitsu Kakusan Koso* **41**, 715 (1996). This EST sample has 40,379 clones representing 40% (7432) of the predicted protein-coding genes in Wormpep.
18. R. Timpl, *Curr. Opin. Cell Biol.* **8**, 618 (1996); R. E. Burgeson and A. M. Christiano, *Curr. Opin. Cell Biol.* **9**, 651 (1997).
19. J. M. Kramer, in C. elegans *II*, D. L. Riddle, T. Blumenthal, B. J. Meyer, J. R. Priess, Eds. (Monograph 33, Cold Spring Harbor Press, Cold Spring Harbor, NY, 1997), pp. 471–500.
20. J. H. Miner , R. M. Lewis, J. R. Sanes, *J. Biol. Chem.* **270**, 28523 (1995).
21. H. Du, G. Gu, C. M. William, M. Chalfie, *Neuron* **16**, 183 (1996); B. Vogel and E. Hedgecock, unpublished observation.
22. C. B. Basebaum and Z. Werb, *Curr. Opin. Cell Biol.* **8**, 731 (1996); A. J. Turner and K. Tanzawa, *FASEB J.* **11**, 355 (1997).
23. A. Howe, A. E. Aplin, S. K. Alahari, R. L. Juliano, *Curr. Opin. Cell Biol.* **10**, 220 (1998).
24. E. Ozawa, S. Noguchi, Y. Mizumo, Y. Hagiwara, M. Yoshida, *Muscle Nerve* **21**, 421 (1998).
25. K. Kobayashi *et al.*, *Nature* **394**, 388 (1998).
26. F. X. Sicot *et al.*, *Eur. J. Biochem.* **246**, 50 (1997).
27. Superfamilies comprise proteins with a specified region of homology, for example, proteins belong to the IG superfamily if they have one or more IG modules. Families comprise proteins with more-or-less identical domain organization, for example, the Notch family.
28. T. Brummendorf and F. G. Rathjen, *Curr. Opin. Neurobiol.* **6**, 584 (1996); F. S. Walsh and P. Doherty, *Annu. Rev. Cell Dev. Biol.* **13**, 425 (1997).
29. P. F. Maness, H. E. Beggs, S. G. Klinz, W. R. Morse, *Perspect. Dev. Neurobiol.* **4**, 169 (1996).
30. M. L. Winberg *et al.*, *Cell* **95**, 903 (1998).
31. Y. Suzuki, N. Sato, M. Tohyama, A. Wanaka, T. Takagi, *J. Biol. Chem.* **271**, 22522 (1996); D. Rose, X. Zhu, H. Kose, B. Hoang, J. Cho, A. Chiba, *Development* **124**, 1561 (1997); A. Almeida *et al.*, *Oncogene* **16**, 2997 (1998); E. Shishido *et al.*, *Science* **280**, 2118 (1998).
32. A. V. Kajava, *J. Mol. Biol.* **277**, 519 (1998).
33. A. S. Yap, W. M. Brieher, B. M. Gumbiner, *Annu. Rev. Cell Dev. Biol.* **13**, 119 (1997); T. Uemura, *Cell* **93**,1095 (1998).
34. B. A. Davletov *et al.*, *EMBO J.* **17**, 3909 (1998).
35. A. K. Hadjantonakis *et al.*, *Genomics* **45**, 97 (1997).
36. C. Sala, J. S. Andreose, G. Fumagalli, T. Lomo, *J. Neurosci.* **15**, 520 (1995).
37. Conservatively, CK (Pfam: cys_knot), FC (Pfam: F5_F8_type_c), FS (Pfam: Kazal), KR (Pfam: kringle), SR (Pfam: SRCR), and VD (Pfam: vwd) modules occur in at least 6, 13, 6, 5, 7, and 7 distinct structural contexts, respectively, among known vertebrate proteins.
38. Conceivably, some protein modules have been lost in particular lineages (discussed below); however, differential loss without any invention cannot satisfactorily account for the observed distribution of protein families and superfamilies among species.
39. R. F. Jackson and L. M. Newby, *Comp. Biochem. Physiol.* **104**, 749 (1993); F. Lassandro *et al.*, *Mol. Biochem. Parasitol.* **65**, 147 (1994); S. L. Jones and D. L. Baillie, *Mol. Gen. Genet.* **48**, 719 (1995); M. G. Buechner, D. H. Hall, E. M. Hedgecock, *Dev. Biol.* **214**, 227 (1999).
40. I. L. Johnstone, *Bioessays* **16**, 171 (1994); I. L. Johnstone and J. D. Barry, *EMBO J.* **15**, 3633 (1996); J. S. Gilleard, D. K. Henderson and N. Ulla, *Gene* **193**, 181 (1997); C. A. Peixoto, J. M. Kramer, W. de Souza, *J. Parasitol.* **83**, 368 (1997).
41. In the M176.8 (chitinase) family, 25 genes or pseudogenes are clustered at two sites on chromosome II intermixed with seven members of the *kin-15* family (Fig. 3). Another 11 chitinase genes are found, singly or as tandem pairs, at nine separate sites throughout the genome. In the C01B7.7 family, 64 genes or pseudogenes are spread along chromosome V at 14 separate sites; 4 sites contain single genes, 9 sites contain 2 to 5 genes, and 1 site contains 29 genes of this family intermixed with at least seven other, locally duplicated gene families, in the interval from C50H11 to T20D4 (ca. 350 kbp). Another 12 genes in the C01B7.7 family are found at 7 separate sites on other chromosomes, including one site with 4 genes.
42. G. P. Smith, *Science* **191**, 528 (1976); R. M. Harding, A. J. Boyce and J. B. Clegg, *Genetics* **132**, 847 (1992).
43. G. Dover, *Curr. Biol.* **4**, 1165 (1994); K. J. Fryxell, *Trends Genet.* **12**, 356 (1996); D. Liao, T. Pavelitz, J. R. Kidd, K. K. Kidd, A. M. Weiner, *EMBO J.* **16**, 588 (1997); M. Nei, X. Gu, T. Sitnikova, *Proc. Natl. Acad. Sci. USA* **94**, 7799 (1997).
44. D. D. Luan *et al.*, *Cell* **72**, 595 (1993); J. Jurka, *Proc. Natl. Acad. Sci. USA* **94**, 1872 (1997).
45. J. V. Moran, R. J. DeBerardinis, H. H. Kazazian, *Science* **283**, 1530 (1999).
46. P. Nouvel, *Genetica* **93**, 191 (1994); J. Brosius and H. Tiedge, *Virus Genes* **11**, 163 (1995); A. J. Flavell, *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **110**, 3 (1995).
47. D. R. Dorer and S. Henikoff, *Cell* **77**, 993 (1994); M. H. Le, D. Duricka and G. H. Karpen, *Genetics* **141**, 283 (1995); S. C. R. Elgin, *Curr. Opin. Genet. Dev.* **6**, 193 (1996); L. L. Wallrath, *Curr. Opin. Genet. Dev.* **8**, 147 (1998).
48. H. Hutter and E. Hedgecock, unpublished information.
49. Excluding loci identified using reverse genetics and *let* genes identified using regional balancers, the fraction of protein genes on each chromosome with known visible mutations falls from 6.4% (161/2508), 6.0% (167/2803), 5.5% (146/2631), 4.1% (133/3259), 4.0% (124/3094) to 3.0% (124/4082), as the proportions of genes on these chromosomes that are duplicated rises from 0.32, 0.31, 0.36, 0.39, 0.45 to 0.51, for chromosomes III, I, X, II, IV, and V, respectively.
50. In a survey of presumptive olfactory receptors, E. R. Troemel *et al.* [*Cell* **83**, 207 (1995)] reported that about half of the genes sampled, whether isolated or clustered, had upstream sequences capable of driving transgene expression in specific chemosensory neurons. It is uncertain whether this reflects gene expression in situ or just latent potentials of these sequences.
51. S. Pimpinelli *et al.*, *Proc. Natl. Acad. Sci. USA* **92**, 3804 (1995); P. Dimitri, *Genetica* **100**, 85 (1997).
52. E. E. Eichler, *Hum. Mol. Genet.* **8**, 151 (1999).
53. S. Henikoff, *Bioessays* **20**, 532 (1998); D. Garrick, S. Fiering, D. I. Martin, E. Whitelaw, *Nature Genet.* **18**, 56 (1998).
54. D. G. Albertson, A. M. Rose, A. M. Villeneuve, in C. elegans *II*, D. L. Riddle, T. Blumenthal, B. J. Meyer, J. R. Priess, Eds. (Monograph 33, Cold Spring Harbor Press, Cold Spring Harbor, NY, 1997), pp. 47–78.
55. P. E. Kuwabara and S. Shah, *Nucleic Acids Res.* **22**, 4414 (1994).
56. Modern bacteria have evidently lost this and other capacities for RNA metabolism important for RNA life or the transition from RNA to DNA life.
57. Further information is online at www.mpimf-heidelberg.mpg.de/ewgdn/genome_paper/.

58. We wish to thank C. Bargmann, R. Barstead, T. Cebula, C. Dunn, E. Kipreos, D. Moerman, S. Nowak, G. Ruvkun, and W. Wadsworth for helpful discussions. H.H. received fellowships from the Deutsche Forschungsgemeinschaft and Max-Planck-Gesellschaft. This paper was supported by a grant from the NIH and originally submitted 7 October 1998.