Single Nucleotide Polymorphisms in Wild Isolates of *Caenorhabditis elegans*

Romke Koch, Henri G.A.M. van Luenen, Marieke van der Horst, et al.

Genome Res. 2000 10: 1690-1696 originally published online November 8, 2000
Access the most recent version at doi:10.1101/gr.GR-1471R

Single Nucleotide Polymorphisms in Wild Isolates of *Caenorhabditis elegans*

Romke Koch,[1] Henri G.A.M. van Luenen,[2] Marieke van der Horst,[1] Karen L. Thijssen,[1] and Ronald H.A. Plasterk[1,3]

[1]*The Hubrecht Laboratory, Centre for Biomedical Genetics, 3584 CT Utrecht, Netherlands;* [2]*The Netherlands Cancer Institute, 1066 CX Amsterdam, Netherlands*

Caenorhabditis elegans (isolate N2 from Bristol, UK) is the first animal of which the complete genome sequence was available. We sampled genomic DNA of natural isolates of *C. elegans* from four different locations (Australia, Germany, California, and Wisconsin) and found single nucleotide polymorphisms (SNPs) by comparing with the Bristol strain. SNPs are under-represented in coding regions, and many were found to be third base silent codon mutations. We tested 19 additional natural isolates for the presence and distribution of SNPs originally found in one of the four strains. Most SNPs are present in isolates from around the globe and thus are older than the latest contact between these strains. An exception is formed by an isolate from an island (Hawaii) that contains many unique SNPs, absent in the tested isolates from the rest of the world. It has been noticed previously that conserved genes (as defined by homology to genes in *Saccharomyces cerevisiae*) cluster in the chromosome centers. We found that the SNP frequency outside these regions is 4.5 times higher, supporting the notion of a higher rate of evolution of genes on the chromosome arms.

Caenorhabditis elegans is the first animal of which the genome was sequenced (The *C. elegans* Sequencing Consortium 1998). Recently, the genome sequence of *Drosophila* has also become available (Adams et al. 2000). *C. elegans* is a sexually-reproducing animal, but the egg-laying animals are actually hermaphrodites: They produce some sperm that they can use to self-fertilize. Self-fertilization quickly results in inbred lines. Although the generation time of *C. elegans* is ~3–4 days, it is likely that in the wild the average time of clonal expansion without male–female mating is much longer. The strain Bristol N2, of which the genome sequence was determined, was isolated from mushroom compost in Bristol, UK, before 1956 (Nicholas et al. 1959; Fatt and Dougherty 1963) and frozen by John Sulston in 1969 (Brenner 1974). This animal occurs worldwide; isolates have been found on all continents except Antarctica (Hodgkin and Doniach 1997). Based on restriction fragment length polymorphisms (RFLPs) associated with Tc*1* transposons, at least 20 races were defined. Previous research has indicated that spontaneous mutation rates in *C. elegans* are low (Anderson 1995), except for transposon insertions in strains that show germ-line transposition. Most strains have been stored frozen since their isolation from nature (Hodgkin and Doniach 1997). For this reason we consider it likely that the single nucleotide polymorphism (SNP) pattern we observe in the strains is identical to that of the original isolate. In this paper we sampled the genome of different natural isolates of *C. elegans* for SNPs. We investigated the nature of the polymorphisms and determined how they are distributed over the chromosomes and whether we could see differences between coding and noncoding regions. We also investigated how SNPS are distributed over natural isolates from over the globe, and we used this to infer relationships among them. We found that SNP patterns can be shared between strains and that SNP levels are elevated on the chromosome arms.

[3] Corresponding author.
E-MAIL plasterk@niob.knaw.nl; FAX 31302516554.
Article published online before print: *Genome Res.*, 10.1101/gr.147100. Article and publication are at www.genome.org/cgi/doi/10.1101/gr.147100.

RESULTS

SNP Frequencies

We performed shotgun sequence analysis of 970 random clones from several natural isolates (AB1, CB4857, RC301, and TR403), which resulted in ~730 kb of sequence information (Table 1), and we searched for SNPs by comparing with the Bristol N2 sequence. In total we found 366 SNPs. SNPs were defined as small substitutions, deletions, or insertions, mostly of 1–3 nucleotides (Jakubowski and Kornfeld 1999). TR403 has the lowest frequency of SNPs (on average 1 in 8750 bp; Table 1) and CB4857 the highest (1 in 1445 bp). Table 2 shows the source of all natural isolates used in this study.

The majority (90%) are point mutations of one nucleotide, but we also encounter small deletions or insertions, and substitutions of 2 bp (Fig. 1A). The data set does not permit a determination of which sequence was the ancestral and which the derived sequence.

10:1690–1696 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

Table 1. Sequence Statistics

Strain	Clones sequenced	Base pairs sequenced	Clones with SNPs	Number of SNPs	SNPs/bp
CB4857	472	372,926	108	258	1:1445
AB1	311	221,192	42	91	1:2430
RC301	137	100,580	12	13	1:7740
TR403	50	35,000	4	4	1:8750
Total	970	729,700	166	366	*

Therefore, the Bristol N2 sequence was taken as ancestral. We found transitions to be over-represented: Sixty-one percent of all single base pair substitutions are transitions. This is also the case for SNPs recently described for human (Hacia et al. 1999) and *Drosophila* (Petrov and Hartl 1999). We analyzed the distribution of SNPs over coding versus noncoding DNA. For the classification we used the Genefinder predictions (unpublished software developed by P. Green and L. Hillier). We find 15% of the SNPs in exons, 85% of the SNPs in non-exon DNA. Twenty-seven percent of the genome is exonic (The *C. elegans* Sequencing Consortium 1998); so there is a twofold under-representation of SNPs in exons. Of the SNPs in coding regions, 53% do not change the coded amino acid; there is a clear bias for SNPs in third base positions (Fig. 1B). As expected, selective removal of deleterious mutations has played an important part in the generation of the SNP pattern as we observe it today (Stenico et al. 1994; Shabalina and Kondrashov 1999).

SNPs Are Shared

We initially found each SNP in one natural isolate. To check whether they also occurred in other natural isolates, the study was broadened by inclusion of a set of six additional natural isolates. We analyzed all SNPs that change the recognition site of a restriction enzyme, and in addition we sequenced all remaining SNPs on autosomes I and III. We found that most SNPs are not unique to the isolate they were detected in and probably preceded the latest contact between the strains; of 109 SNPs that were tested in other strains even within this small set of natural isolates, only 25 were found to be unique (Fig. 2A). Isolates from the same geographical region are not necessarily similar. For example, the two Australian strains AB1 and AB4 are not at all identical (one has its chromosome II pattern largely in common with Bristol N2, the other with CB4857 and KR314); nevertheless the X chromosomes are almost identical for the SNPs tested.

To further investigate the diversity of strain variants at one geographical location, we also sampled a collection of strains that had all been isolated in California. Some were isolated at several time points from the same vegetable garden or flower bed (Hodgkin and Doniach 1997). These strains were tested for all SNPs that can be visualized by RFLP. We found strains to have essentially three different SNP patterns (Fig. 2B). Some were largely similar to the English N2 strain, others to CB4857 from Claremont, but an intermediate type was also observed. Some phylogenies have been proposed for natural isolates of *C. elegans*, based on phenotypical traits (Dion and Brun 1971; Egilmez et al. 1995; Abdul Kader and Cote 1996; Hodgkin and Doni-

Table 2. Strains Analyzed in This Study

Strain	Source	Strain	Source
N2	Bristol, UK	DR1348	Altadena, CA
AB1	Adelaide, Australia	DR1350	Pasadena, CA
AB4	Adelaide, Australia	GA3	Altadena, CA
CB4856	Hawaii	GA4	Altadena, CA
CB4857	Claremont, CA	GA13	Altadena, CA
CB3191	Altadena, CA	GA23	Altadena, CA
CB4555	Pasadena, CA	KR314	Vancouver, Canada
CB4853	Altadena, CA	PA3	Pasadena, CA
CB4854	Altadena, CA	RC301	Freiburg, Germany
CB4858	Pasadena, CA	RW7000	Bergerac, France
DH424	El Prieto, CA	TR389	Madison, WI
DR1347	Altadena, CA	TR403	Madison, WI

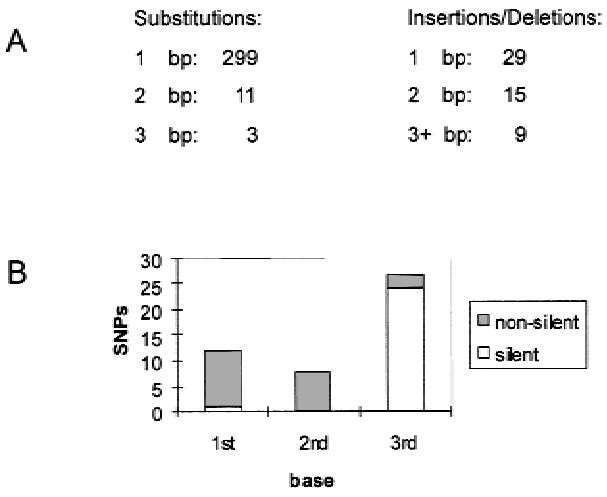

Figure 1 Nature of the SNPs. (*A*) The number of base pairs involved in substitutions, insertions, and deletions is indicated. Note that we find on average one SNP per 2000 bp. After checking the Bristol N2 sequence for 63 potential SNPs, we found 3 to be a sequence error in N2. Thus, we estimate the error rate of the *C. elegans* genome sequence to be ~1 in 42,000. (*B*) The number of SNPs involved in the first, second, or third base of a codon in coding sequences is plotted. It also shows whether the SNP alters the amino acid. Coding sequences were predicted by Genefinder.

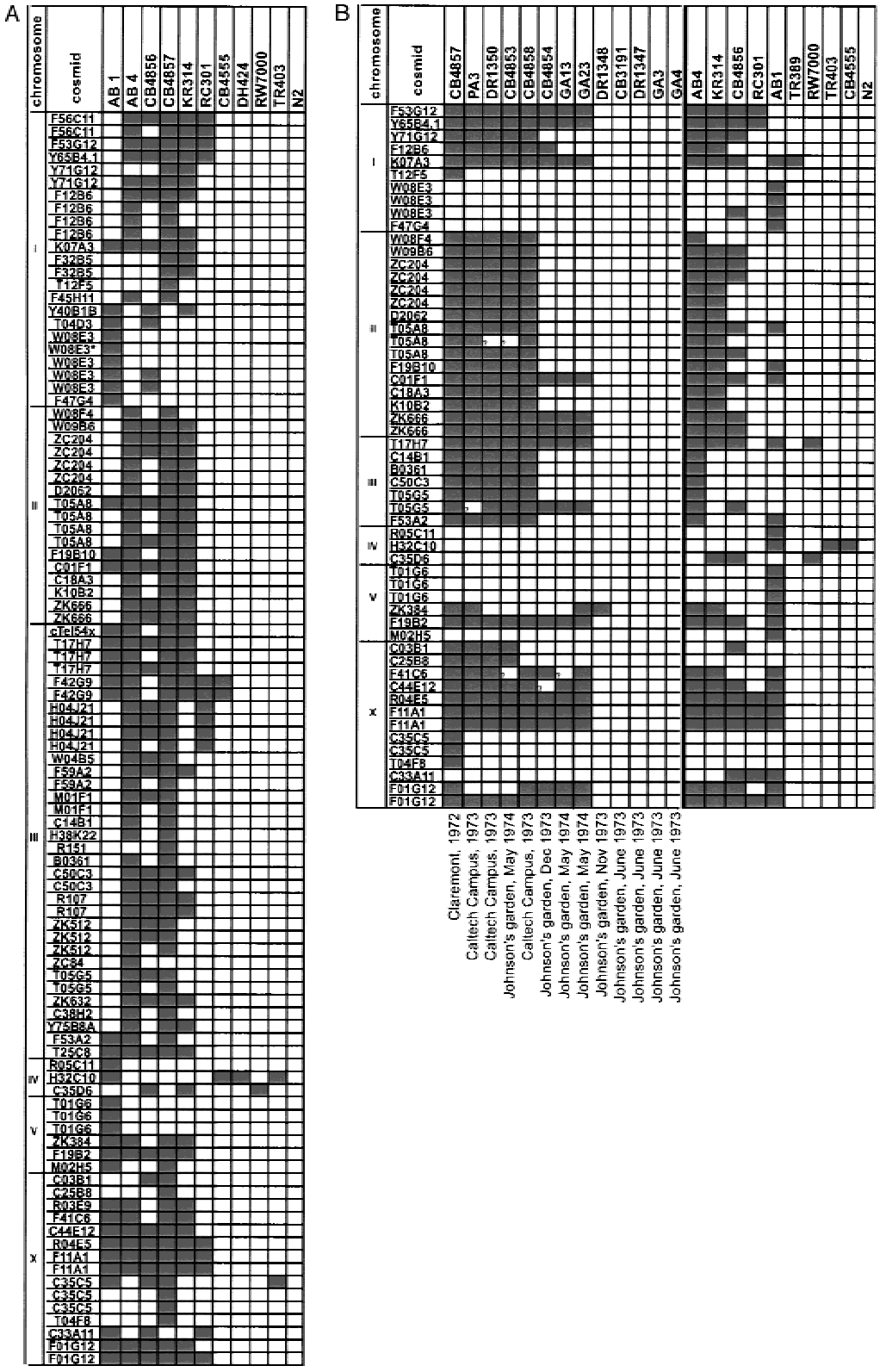

Figure 2 Comparison of SNPs among different natural isolates. (*A*) The presence of SNPs originally identified in one strain was tested in nine other strains and Bristol N2 by RFLP analysis or sequencing. SNP-containing clones are shown vertically along the chromosomes using their physical map position. Some clones contain multiple SNPs. Eleven unique SNPs in clone W08E3* are not shown in this figure. (*B*) Local variation was tested by comparison of strains isolated in California for SNPs that can be visualized by RFLP analysis. Most were isolated from flower beds on Caltech campus and from a vegetable garden in Altadena. The presence of a SNP is indicated with a shaded square. Isolates from other locations are plotted for reference.

ach 1997) as well as genome typing (Egilmez et al. 1995; Hodgkin and Doniach 1997). Figure 2 indicates that one cannot speak of lineages in the strictest sense, because the similarity between strains is different for separate regions of the genome. No clear correlation between DNA variation and the geographical origin was seen. Analysis of genetic diversity within and between *Arabidopsis thaliana* ecotypes resulted in similar findings (Innan et al. 1997; Breyne et al. 1999).

We investigated at a microlevel how SNPs were distributed between the strains; we sequenced, now in a directed fashion, the environment of two regions that seemed highly polymorphic based on the number of SNPs found in shotgun clones (H04J21 for CB4857; W08E3 for AB1). Within these regions we found the level of polymorphism with Bristol N2 to be 1 in 200 bp for CB4857 (117 SNPs in 23 kb) and 1 in 170 bp for AB1 (86 SNPs in 14.5 kb), compared with 1 in 1800 bp for the whole SNP set, indicating the presence of highly polymorphic regions. The presence of SNPs in the other natural isolates was tested by sequencing 1-kb regions and also stretches 5 kb and 50 kb on each side of these regions. We find that the SNPs in the natural isolates do occur in tracts (Fig. 3). Presumably *C. elegans* lineages have been reproductively isolated for prolonged times, accumulated many mutations, and then came into contact again with other populations, resulting in polymorphic tracts. Many SNPs of the strain from Hawaii (CB4856) were absent in all other tested isolates. This is also true for a large set of SNPs from many different regions of the CB4856 genome (R. Waterston, pers. comm.), which were tested for their presence in the other strains: SNPs (63 of 90) of CB4856 were absent in the other nine strains (S. Wicks et al., in prep). This suggests that this strain has been reproductively isolated and diverged significantly. From the strains studied in this paper, this is the only one from an isolated island.

Elevated SNP Levels Suggest Higher Evolutionary Rates

It was found previously that the rate of meiotic recombination is more than five times higher on the chromosome arms than in the central clusters (Barnes et al. 1995). The genome sequence revealed that genes with similarity to the yeast genome were more frequent on the autosome centers than on the arms (Fig. 4A), whereas inverted and tandem repeats clustered mainly on the arms (The *C. elegans* Sequencing Consortium

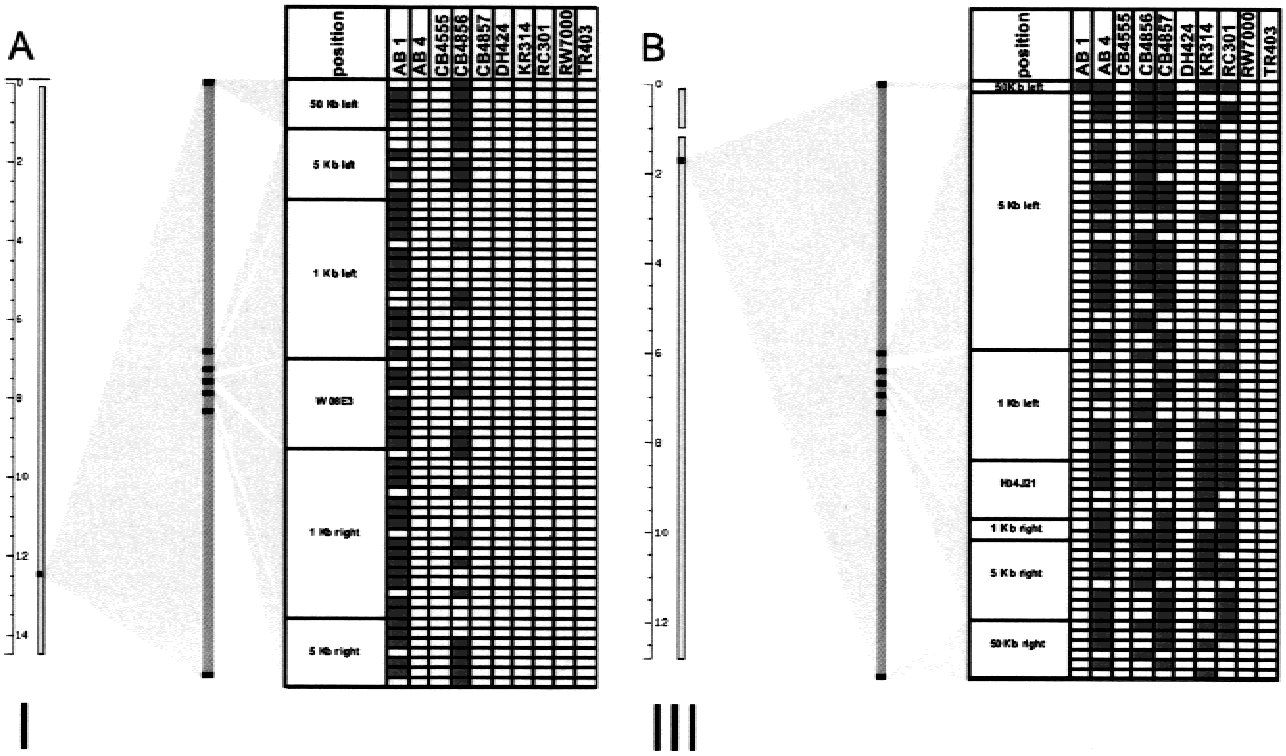

Figure 3 Detailed comparison of SNPs among isolates in SNP-rich regions near clone W08E3 on chromosome I (*A*) and clone H04J21 on chromosome III (*B*). Regions of ~1 kb flanking the highly polymorphic clones were analyzed by directed sequencing at 1, 5, and 50 kb on each side in 10 natural isolates and N2. Polymorphisms within the tracts are shown vertically. The presence of a SNP is marked with a shaded square.

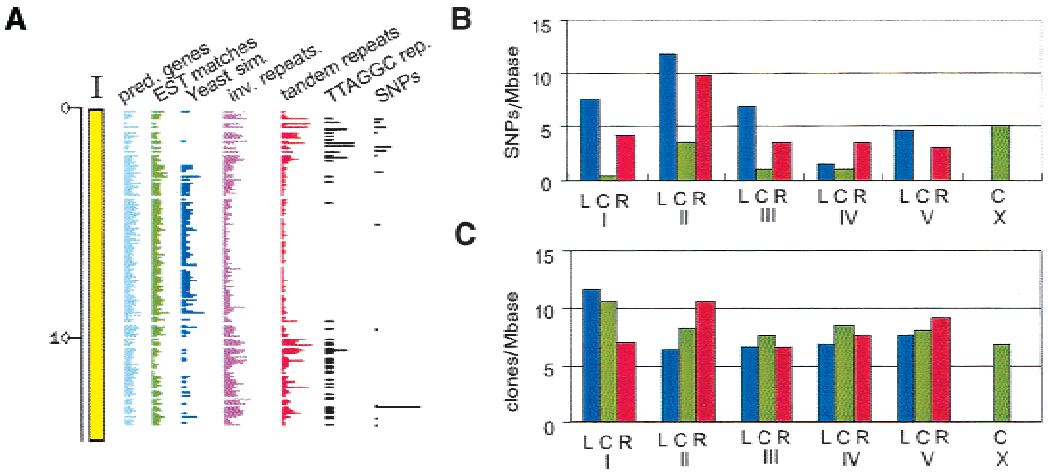

Figure 4 SNP distribution on the chromosomes. (*A*) DNA repeats and SNPs on chromosome I are found mainly on the chromosome arms, whereas similarity with yeast genes is highest on the autosome centers (figure adapted from The *C. elegans* Sequencing Consortium 1998, p. 2016). The number of SNPs/Mb (*B*) and the number of sequenced clones/Mb (*C*) are plotted for the six chromosomes. Autosomes are divided into genetically defined compartments of the left arm (L), the central cluster (C), and the right arm (R) (Barnes et al. 1995; The *C. elegans* Sequencing Consortium 1998). SNPs occur more frequently on the arms than on the central regions. The X chromosome does not show a higher SNP frequency on the arms.

1998; Surzycki and Belknap 2000). The authors suggested that possibly there was a higher evolutionary rate on the chromosome arms. Can we still find indications for a differential evolution rate in these segments of the genome? Figure 4A shows the distribution of the SNPs we found on chromosome I: Polymorphism levels are elevated on the arms. A systematic analysis of the SNP density on all the autosomes revealed that the SNPs were not distributed randomly ($\chi^2 = 40.28$, $P < 0.001$) but were elevated on the autosome arms. We found SNPs to be 4.5 times more abundant on the arms (L and R; Fig. 4B) of the autosomes than on the central region (C; Fig. 4B), whereas the shotgun clones were distributed uniformly (Fig. 4C). These data support the notion of more rapid evolvement of DNA on the autosome arms in the current *C. elegans* species.

DISCUSSION

We have characterized DNA polymorphisms in four *C. elegans* strains that were isolated from diverse geographical locations. We checked SNPs in the strain in which they were found and also checked the original Bristol N2 sequence by analyzing PCR products derived from N2 DNA. Any mutation that occurred during the 10 or more years of lab culturing of Bristol N2 would show up as a SNP that was unique for Bristol N2. We found none of those, suggesting that spontaneous mutations in Bristol N2 are extremely rare and that, by analogy, the SNPs we detect in natural isolates existed before the strains were isolated from nature.

Most polymorphisms do not alter the coding amino acid; they are found in introns or on the third base of an exon, supporting the idea of gene conservation. Furthermore, higher levels of polymorphism are found on the autosome arms than on the centers. In combination with the finding that more conserved sequences are found mainly in the chromosome centers, this suggests a higher tolerance of polymorphisms on the arms. It remains to be explained which mechanisms are responsible for the intriguing differences between the arms and the central clusters.

Analysis of the 24 isolates suggests that SNPs are shared between different strains. Most SNPs have probably occurred by mutagenic events before the last contact between the different continents. At one geographical location, worm types can be found that are similar to worms from several other locations (i.e., Californian strains). This could be explained by the idea of long range dispersal, spreading of worms as dauer larvae in soil adhering to birds or other animals (Hodgkin and Doniach 1997). The SNP patterns described here are in agreement with the classification of worm races made by Hodgkin and Doniach (1997). For example, based on their Tc*1* pattern, plug formation, and clumping phenotype, these authors suggested that CB4858 was similar to AB4. AB4 and CB4858 show almost the same SNP pattern; only one SNP on chromosome V is not present in CB4858.

The SNPs found in this study can also be used efficiently as a tool for gene mapping. SNP markers can easily be detected by sequencing or restriction digestion, and they do not interfere with subtle phenotypes.

A simple cross can be used to determine linkage to a chromosome (Jakubowski and Kornfeld 1999; S. Wicks, in prep.).

Analysis of SNP patterns can be used to further characterize the natural history of the worm against the background of a sequenced genome. The analysis is probably facilitated by the hermaphrodite lifestyle of *C. elegans*, which results in inbreeding and, thus, extended conservation of haplotypes. With the exception of the Hawaiian CB4856 strain, continent-specific genotypes were not recognized, and not many SNPs were unique and unshared between isolates from different regions of the world. SNPs were searched in a limited set of strains. It cannot be excluded that analysis of other natural isolates might reveal a clear subspecies that diverged significantly from all other known isolates. The analysis of SNP patterns within a global species of which the genome sequence is known such as *C. elegans* can provide a new perspective on many aspects of population biology and evolution.

METHODS

Strains and Genomic DNA Isolation

Natural isolates of *C. elegans* were obtained from the *Caenorhabditis elegans* Genetics Center (University of Minnesota, St. Paul). Initially, polymorphisms to Bristol N2 were searched in AB1 from Australia, CB4857 from California, RC301 from Germany, and TR403 from Wisconsin. Later, SNPs were also verified in other natural isolates (Table 2). The origin of these strains is extensively described in a paper by Hodgkin and Doniach (1997). Genomic DNA of *C. elegans* strains N2, CB4857, AB1, RC301, and TR403 was isolated as described by Sulston and Hodgkin (1988).

Cloning of Genomic DNA and SNP Detection

Genomic DNA (~20 µg) was partially digested with 10 units of *Sau*3A1 (Roche Molecular Biochemicals) in a 50-µL reaction containing 1× SuRECut buffer A for 5 min at room temperature and loaded on a 1% 1× TAE–agarose gel. After electrophoresis, fragments between 1000 bp and 1500 bp were purified from the gel by freezing the excized bands in liquid nitrogen in separate tubes and centrifuging 10 min at maximum speed. Supernatants were extracted twice with phenol–chloroform, and DNA was precipitated with 0.1 volume of NaAc (pH 5.2) and 2.5 volumes of ethanol. After centrifugation, the pellet was redissolved in 50 µL H_20. To create an overhang at the 3′ end for more efficient cloning in a pGEM-T vector (Promega), DNA was incubated with 5 units of *Taq* polymerase (GIBCO BRL) for 20 min at 72°C in 1× PCR buffer with 0.2 mM dNTPS. Fragments were subsequently ligated into the vector and transformed into DH5α cells. Transformants were grown on ampicillin selective plates and used for sequencing with SP6 and T7 primers on an ABI 377 sequencer. Sequence traces were aligned to the Bristol N2 sequence using the *C. elegans* BLAST server. All sequence differences with N2 were confirmed by visually analyzing the raw data to exclude mistakes in base calling. Clones with similarity to repetitive sequences were discarded. For confirmation, oligonucleotides were designed to amplify the genomic region containing the SNP. After amplification of 1 µL of genomic DNA (20 ng/µL) and 20 µL of PCR mix [4.4 pmoles of each oligonucleotide, 0.5 unit of*Taq* polymerase (GIBCO BRL), 2 µm of each dNTP in 50 mM KCl, 20 mM Tris-HCl (pH 8.3), 1.5 mM $MgCl_2$] with 35 cycles (1 min at 95°C, 1 min at 52°C, and 1 min at 72°C), the presence of a SNP was confirmed. This was done either by sequencing of the PCR product or by digestion of the product with a restriction enzyme. SNPs will be submitted to the SNP database of the Sanger Centre.

ACKNOWLEDGMENTS

We thank Amanda McMurray and Jane Rogers of the Sanger Centre and Roelof Pruntel for help in sequencing, Dr. R.H. Waterston for sharing unpublished data, Dr. J. Hodgkin and the Caenorhabditis Genetics Center for strains, and Dr. Stephen Wicks for critical reading of the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Abdul Kader, N. and Cote, M.G. 1996. Isolation, identification and characterization of some strains of *Caenorhabditis elegans* (Maupas, 1900) from Quebec. *Fund. App. Nemat.* **19:** 381–389.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Anderson, P. 1995. Mutagenesis. In *Methods in cell biology. Caenorhabditis elegans*: Modern biological analysis of an organism (eds. H.F. Epstein and D.C. Shakes), pp. 31–48. Academic Press, San Diego, CA.

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.

Brenner, S. 1974. The genetics of *Caenorhabditis elegans*. *Genetics* **77:** 71–94.

Breyne, P., Rombaut, D., Van Gysel, A., Van Montagu, M., and Gerats, T. 1999. AFLP analysis of genetic diversity within and between *Arabidopsis thaliana* ecotypes. *Mol. & Gen. Genet.* **261:** 627–634.

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

Dion, M. and Brun, J.L. 1971. Genetic mapping of the free-living nematode *Caenorhabditis elegans* Maupas 1900, var. Bergerac. I. Study of two dwarf mutants. *Mol. & Gen. Genet.* **112:** 133–151.

Egilmez, N.K., Ebert II, R.H., and Shmookler Reis, R.J. 1995. Strain evolution in Caenorhabditis elegans: Transposable elements as markers of interstrain evolutionary history. *J. Mol. Evol.* **40:** 372–381.

Fatt, H.V. and Dougherty, E.C. 1963. Genetic control of differential heat tolerance in two strains of the nematode *Caenorhabditis elegans*. *Science* **141:** 266–267.

Hacia, J.G., Fan, J.B., Ryder, O., Jin, L., Edgemon, K., Ghandour, G., Mayer, R.A., Sun, B., Hsie, L., Robbins, C.M., et al. 1999. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat. Genet.* **22:** 164–167.

Hodgkin, J. and Doniach, T. 1997. Natural variation and copulatory plug formation in *Caenorhabditis elegans*. *Genetics* **146:** 149–164.

Innan, H., Terauchi, R., and Miyashita, N.T. 1997. Microsatellite

polymorphism in natural populations of the wild plant *Arabidopsis thaliana*. *Genetics* **146:** 1441–1452.

Jakubowski, J. and Kornfeld, K. 1999. A local, high-density, single-nucleotide polymorphism map used to clone *Caenorhabditis elegans* cdf-1. *Genetics* **153:** 743–752.

Nicholas, W.L., Dougherty, E.C., and Hansen, E.L. 1959. Axenic cultivation of *C. briggsae* (Nematoda: Rhabditidae) with chemically undefined supplements; comparative studies with related nematodes. *Ann. N.Y. Acad. Sci.* **77:** 218–236.

Petrov, D.A. and Hartl, D.L. 1999. Patterns of nucleotide substitution in Drosophila and mammalian genomes. *Proc. Natl. Acad. Sci.* **96:** 1475–1479.

Shabalina, S.A. and Kondrashov, A.S. 1999. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. *Genet. Res.* **74:** 23–30.

Stenico, M., Lloyd, A.T., and Sharp, P.M. 1994. Codon usage in *Caenorhabditis elegans*: Delineation of translational selection and mutational biases. *Nucleic Acids Res.* **22:** 2437–2446.

Sulston, J. and Hodgkin, J. 1988. Methods. In *The nematode Caenorhabditis elegans* (ed. W.B. Wood), pp. 587–606. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Surzycki, S.A. and Belknap, W.R. 2000. Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. *Proc. Natl. Acad. Sci.* **97:** 245–249.

Received May 9, 2000; accepted in revised form July 27, 2000.