examine the quaternary complex, *LexA–AP3* and *PI* were expressed on the bait vector, and *GAL4 AD-AG* and/or *SEP3-MIK* were expressed on the prey vector. When two genes were expressed on the same vector, they were both driven by ADH1 promoters. Amino-acid residues 1–167 and 1–171 were used for the truncated AP1-MIK and SEP3-MIK proteins, respectively. Other processes and the colony-lift β-gal assays were performed in accordance with the manufacturer's instructions (Clontech).

Immunoprecipitation

For immunoprecipitation experiments, radiolabelled AP1 or SEP3 were mixed with haemagglutinin (HA)-tagged proteins and precipitated with anti-HA antibody. Precipitated AP1 and SEP3 were separated by SDS–PAGE and detected by radio-imaging analyser, BAS2000 (Fujifilm). Other procedures were done as described[7,12].

Transactivation assay

For yeast, MADS proteins cDNAs were fused in-frame to GAL4 DNA-binding domain on pAS2-1 (Clontech) and transformed into the yeast strain YRG-2 (*UAS::lacZ*, Stratagene). AP1-K2C (residues 125–256) and SEP3-K2C (128–257) were used as truncated MADS proteins. Yeast cells were grown at 22 °C overnight, and the β-gal activity was assayed at 30 °C using *o*-nitrophenyl-β-D-galactopyranoside.

For onion epidermal cells, 35S promoter-driven MADS cDNAs that express native MADS proteins (effector) and *CArG::LUC* (reporter) were co-transfected into onion epidermal cells by using a particle delivery system (Bio-Rad). *CArG::LUC* has seven repeats of MADS protein binding consensus sequence[29], 5′-GGGGTGGCTTTCCTTTTTTGGTAAATTTTGGATCC-3′ (CArG box is underlined), upstream of the 35S minimal promoter (–30). 35S::*Renilla* luciferase (RLUC) was used for the internal control. LUC assays were conducted using Dual-luciferase reporter system (Promega). Other procedures were done as described[30].

Plant material

Arabidopsis Columbia ecotype was used for *Agrobacterium*-mediated vacuum transformation[31]. Plant crossing was carried out by manual cross-pollination. The presence of the transgenes was confirmed by PCR. *AP3::GUS* plants have a 600-base-pair region of the *AP3* promoter[16]. Staining for GUS activity was done as described[16].

Cryo-scanning electron micrograph

We used a Hitachi S-3500N scanning electron microscope equipped with a cryo-stage. For observation and photography, the stage was chilled at –20 °C and the natural scanning electron microscopy (SEM) mode (70 Pa) was used with a 25-kV accelerating voltage.

Received 2 October; accepted 6 November 2000.

1. Coen, E. S. & Meyerowitz, E. M. The war of the whorls: genetic interactions controlling flower development. *Nature* **353,** 31–37 (1991).
2. Bowman, J. L., Smyth, D. R. & Meyerowitz, E. M. Genetic interactions among floral homeotic genes of *Arabidopsis*. *Development* **112,** 1–20 (1991).
3. Mizukami, Y. & Ma, H. Ectopic expression of the floral homeotic gene AGAMOUS in transgenic *Arabidopsis* plants alters floral organ identity. *Cell* **71,** 119-131 (1992).
4. Krizek, B. A. & Meyerowitz, E. M. The *Arabidopsis* homeotic genes APETALA3 and PISTILLATA are sufficient to provide the B class organ identity function. *Development* **122,** 11–22 (1996).
5. Pelaz, S., Ditta, G. S., Baumann, E., Wisman, E. & Yanofsky, M. F. B and C floral organ identity functions require SEPALLATA MADS-box genes. *Nature* **405,** 200–203 (2000).
6. Mandel, M. A., Gustafson-Brown, C., Savidge, B. & Yanofsky, M. F. Molecular characterization of the *Arabidopsis* floral homeotic gene *APETALA1*. *Nature* **360,** 273–277 (1992).
7. Goto, K. & Meyerowitz, E. M. Function and regulation of the *Arabidopsis* floral homeotic gene PISTILLATA. *Genes Dev.* **8,** 1548–1560 (1994).
8. Jack, T., Brockman, L. L. & Meyerowitz, E. M. The homeotic gene *APETALA3* of *Arabidopsis thaliana* encodes a MADS box and is expressed in petals and stamens. *Cell* **68,** 683-697 (1992).
9. Yanofsky, M. F. *et al.* The protein encoded by the *Arabidopsis* homeotic gene *agamous* resembles transcription factors. *Nature* **346,** 35–39 (1990).
10. Schwarz-Sommer, Z., Huijser, P., Nacken, W., Saedler, H. & Sommer, H. Genetic control of flower development: homeotic genes in *Antirrhinum majus*. *Science* **250,** 931–936 (1990).
11. Ma, H., Yanofsky, M. F. & Meyerowitz, E. M. *AGL1-AGL6*, an *Arabidopsis* gene family with similarity to floral homeotic and transcription factor genes. *Genes Dev.* **5,** 484–495 (1991).
12. Riechmann, J. L., Krizek, B. A. & Meyerowitz, E. M. Dimerization specificity of *Arabidopsis* MADS domain homeotic proteins APETALA1, APETALA3, PISTILLATA, and AGAMOUS. *Proc. Natl Acad. Sci. USA* **93,** 4793–4798 (1996).
13. Herskowitz, I. A regulatory hierarchy for cell specialization in yeast. *Nature* **342,** 749–757 (1989).
14. Tilly, J. J., Allen, D. W. & Jack, T. The CArG boxes in the promoter of the *Arabidopsis* floral organ identity gene APETALA3 mediate diverse regulatory effects. *Development* **125,** 1647–1657 (1998).
15. Hill, T. A., Day, C. D., Zondlo, S. C., Thackeray, A. G. & Irish, V. F. Discrete spatial and temporal cis-acting elements regulate transcription of the *Arabidopsis* floral homeotic gene APETALA3. *Development* **125,** 1711–1721 (1998).
16. Honma, T. & Goto, K. The *Arabidopsis* floral homeotic gene PISTILLATA is regulated by discrete cis-elements responsive to induction and maintenance signals. *Development* **127,** 2021–2030 (2000).
17. Sadowski, I., Ma, J., Triezenberg, S. & Ptashne, M. GAL4-VP16 is an unusually potent transcriptional activator. *Nature* **335,** 563–564 (1988).
18. Mandel, M. A. & Yanofsky, M. F. The *Arabidopsis AGL9* MADS box gene is expressed in young flower primordia. *Sex Plant Reprod.* **11,** 22–28 (1998).
19. Rubinelli, P., Hu, Y. & Ma, H. Identification, sequence analysis and expression studies of novel anther-specific genes of *Arabidopsis thaliana*. *Plant Mol. Biol.* **37,** 607–619 (1998).
20. Fan, H. -Y., Hu, Y., Tudor, M. & Ma, H. Specific interactions between the K domains of AG and AGLs, members of the MADS domain family of DNA binding proteins. *Plant J.* **12,** 999–1010 (1997).
21. Cho, S. *et al.* Analysis of the C-terminal region of *Arabidopsis thaliana* APETALA1 as a transcription activation domain. *Plant Mol. Biol.* **40,** 419–429 (1999).
22. Riechmann, J. L. & Meyerowitz, E. M. MADS domain proteins in plant development. *J. Biol. Chem.* **378,** 1079–1101 (1997).
23. Egea-Cortines, M., Saedler, H. & Sommer, H. Ternary complex formation between the MADS-box proteins SQUAMOSA, DEFICIENS and GLOBOSA is involved in the control of floral architecture in *Antirrhinum majus*. *EMBO J.* **18,** 5370–5379 (1999).
24. Davies, B., Egea-Cortines, M., de Andrade Silva, E., Saedler, H. & Sommer, H. Multiple interactions amongst floral homeotic MADS box proteins. *EMBO J.* **15,** 4330-4343 (1996).
25. Rounsley, S. D., Ditta, G. S. & Yanofsky, M. F. Diverse roles for MADS box genes in *Arabidopsis* development. *Plant Cell* **7,** 1259–1269 (1995).
26. Smyth, D. A reverse trend—MADS functions revealed. *Trends Plant Sci.* **5,** 315–317 (2000).
27. Parcy, F., Nilsson, O., Busch, M. A., Lee, I. & Weigel, D. A genetic framework for floral patterning. *Nature* **395,** 561–566 (1998).
28. Bartel, P. L., Chien, C., Sternglanz, R. & Fields, S. in *Cellular Interactions in Development: a Practical Approach.* (ed. Hartley, D. A.) 153–179 (IRL Press, Oxford, 1993).
29. Shiraishi, H., Okada, K. & Shimura, Y. Nucleotide sequences recognized by the AGAMOUS MADS domain of *Arabidopsis thaliana* in vitro. *Plant J.* **4,** 385–398 (1993).
30. Pan, S., Sehnke, P. C., Ferl, R. J. & Gurley, W. B. Specific interactions with TBP and TFIIB in vitro suggest that 14-3-3 proteins may participate in the regulation of transcription when part of a DNA binding complex. *Plant Cell* **11,** 1591–1602 (1999).
31. Bechtold, N., Ellis, J. & Pelletier, G. *In planta Agrobacterium* mediated gene transfer by infiltration of adult *Arabidopsis* plants. *C. R. Acad. Sci. Paris* **316,** 1194–1199 (1993).

Acknowledgements

We are grateful to M. Yanofsky for communicating data before publication, and to D. Weigel for providing the cDNA library. We also thank J. Bowman, T. Ito and H. Tsukaya for critical reading of the manuscript. This work was supported by grants from the Monbusho and JSPS.

Correspondence and requests for materials should be addressed to K.G. (e-mail: kgoto@v004.vaio.ne.jp).

……

Genome sequence of enterohaemorrhagic *Escherichia coli* 0157:H7

Nicole T. Perna[*,†], Guy Plunkett III[‡], Valerie Burland[‡], Bob Mau[‡], Jeremy D. Glasner[‡], Debra J. Rose[‡], George F. Mayhew[‡], Peter S. Evans[‡], Jason Gregor[‡], Heather A. Kirkpatrick[‡], György Pósfai[§], Jeremiah Hackett[‡], Sara Klink[‡], Adam Boutin[‡], Ying Shao[‡], Leslie Miller[‡], Erik J. Grotbeck[‡], N. Wayne Davis[‡], Alex Lim[‖], Eileen T. Dimalanta[‖], Konstantinos D. Potamousis[‡,‖], Jennifer Apodaca[‡,‖], Thomas S. Anantharaman[¶], Jieyi Lin[#], Galex Yen[*], David C. Schwartz[*,‡,‖], Rodney A. Welch[☆] & Frederick R. Blattner[*,‡]

[*] *Genome Center of Wisconsin,* [†] *Department of Animal Health and Biomedical Sciences,* [‡] *Laboratory of Genetics,* [‖] *Department of Chemistry,* [¶] *Department of Biostatistics, and* [☆] *Department of Medical Microbiology and Immunology, University of Wisconsin, Madison, Wisconsin 53706, USA*
[§] *Institute of Biochemistry, Biological Research Center, H-6701 Szeged, Hungary*
[#] *Cereon Genomics, LLC, 45 Sidney Street, Cambridge, Massachusetts 02139, USA*

The bacterium *Escherichia coli* O157:H7 is a worldwide threat to public health and has been implicated in many outbreaks of haemorrhagic colitis, some of which included fatalities caused by haemolytic uraemic syndrome[1,2]. Close to 75,000 cases of O157:H7 infection are now estimated to occur annually in the United States[3]. The severity of disease, the lack of effective treatment and the potential for large-scale outbreaks from contaminated food supplies have propelled intensive research on the pathogenesis and detection of *E. coli* O157:H7 (ref. 4). Here we have sequenced the genome of *E. coli* O157:H7 to identify candidate genes responsible for pathogenesis, to develop better methods of strain detection and to advance our understanding of

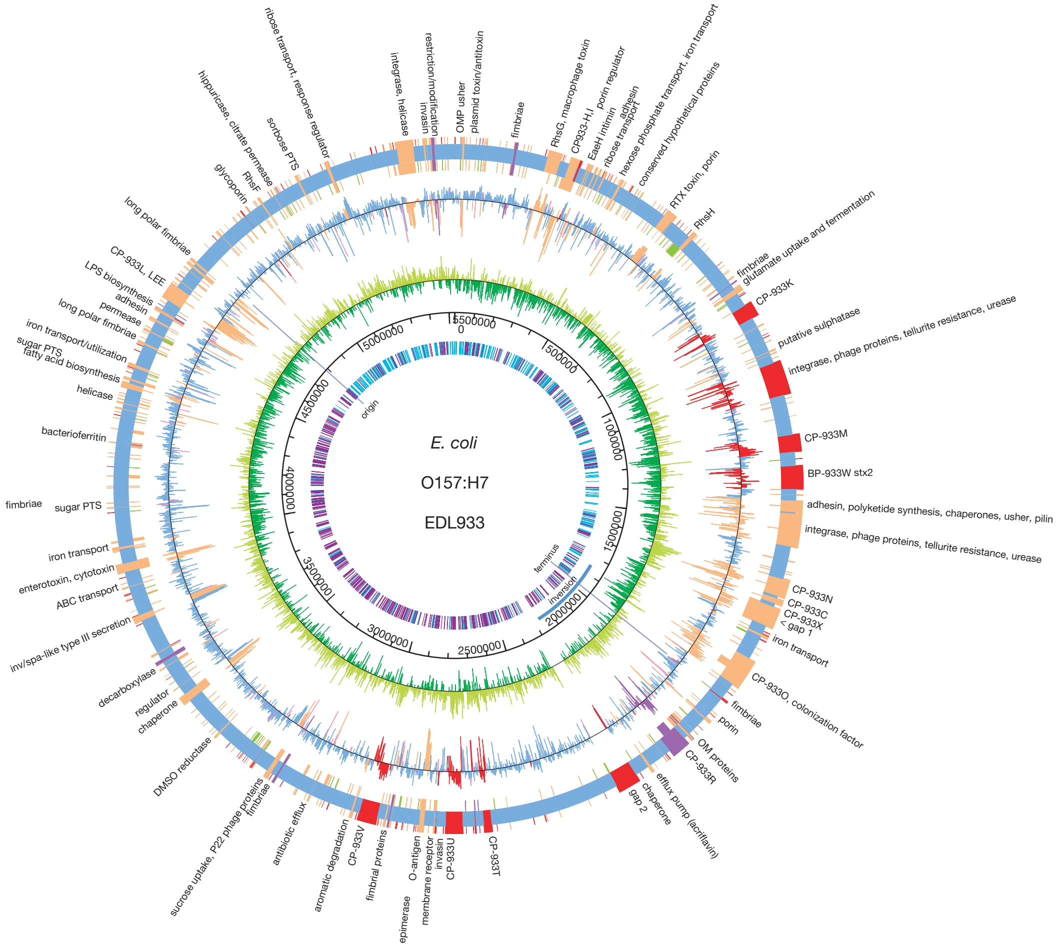

Figure 1 Circular genome map of EDL933 compared with MG1655. Outer circle shows the distribution of islands: shared co-linear backbone (blue); position of EDL933-specific sequences (O-islands) (red); MG1655-specific sequences (K-islands) (green); O-islands and K-islands at the same locations in the backbone (tan); hypervariable (purple). Second circle shows the G+C content calculated for each gene longer than 100 amino acids, plotted around the mean value for the whole genome, colour-coded like outer circle. Third circle shows the GC skew for third-codon position, calculated for each gene longer than 100 amino acids: positive values, lime; negative values, dark green. Fourth circle gives the scale in base pairs. Fifth circle shows the distribution of the highly skewed octamer Chi (GCTGGTGG), where bright blue and purple indicate the two DNA strands. The origin and terminus of replication, the chromosomal inversion and the locations of the sequence gaps are indicated. Figure created by Genvision from DNASTAR.

the evolution of *E. coli*, through comparison with the genome of the non-pathogenic laboratory strain *E. coli* K-12 (ref. 5). We find that lateral gene transfer is far more extensive than previously anticipated. In fact, 1,387 new genes encoded in strain-specific clusters of diverse sizes were found in O157:H7. These include candidate virulence factors, alternative metabolic capacities, several prophages and other new functions—all of which could be targets for surveillance.

Escherichia coli O157:H7 was first associated with human disease after a multi-state outbreak in 1982 involving contaminated hamburgers[1]. The strain EDL933 that we sequenced was isolated from Michigan ground beef linked to this incident, and has been studied as a reference strain for O157:H7. Figures 1 and 2 show the gene content and organization of the EDL933 genome, and compare it with the chromosome of the K-12 laboratory strain MG1655 (ref. 5). These strains last shared a common ancestor about 4.5 million years ago[6]. The two *E. coli* genomes revealed an unexpectedly complex segmented relationship, even in a preliminary examination[7]. They share a common 'backbone' sequence which is co-linear except for one 422-kilobase (kb) inversion spanning the replication terminus (Fig. 1). Homology is punctuated by hundreds of islands of apparently introgressed DNA—numbered and designated 'K-islands' (KI) or 'O-islands' (OI) in Fig. 2, where K-islands are DNA segments present in MG1655 but not in EDL933, and O-islands are unique segments present in EDL933.

The backbone comprises 4.1 megabases (Mb), which are clearly homologous between the two *E. coli* genomes. O-islands total 1.34 Mb of DNA and K-islands total 0.53 Mb. These lineage-specific segments are found throughout both genomes in clusters of up to 88 kb. There are 177 O-islands and 234 K-islands greater than 50 bp in length. Histograms (Fig. 3) show more intermediate and large islands in EDL933 than in MG1655. Only 14.7% (26/177) of the O-islands correspond entirely to intergene regions. The two largest are identical copies of a 106-gene island, both in the same orientation and adjacent to genes encoding identical transfer RNAs.

Figure 2 Detailed comparative map of the EDL933 and MG1655 genomes. The upper double bar in each tier shows the genome comparison in EDL933 coordinates, with segments shown in detail and colour coded as in Fig. 1. Segments shown below the blue bar represent K-islands (MG1655-specific sequence). Segments extending above the blue backbone bar represent O-islands (EDL933-specific sequence). Unique identifying names (KI and OI numbers) were assigned to all segments of more than 50 bp. Unnamed vertical black lines across the blue bar indicate segments of less than 50 bp. In the lower line of each tier, EDL933 genes are presented showing orientation, and are coloured by segment type. Genes spanning segment junctions are shown in pink. Some gene names are given to provide landmarks in the backbone regions, and the sequence gaps are indicated. The scale in base pairs marks the base of each tier. Map created by Genvision from DNASTAR.

Labelling lineage-specific segments 'islands' is an extension of the term 'pathogenicity island' now in common, albeit varied, use. The original term arose from observations that virulence determinants are often clustered in large genomic segments showing hallmarks of horizontal transfer[8]. However, we found K- and O-islands of all sizes with no obvious association with pathogenicity; conversely, genes probably associated with virulence are not limited to the largest islands.

Roughly 26% of the EDL933 genes (1,387/5,416) lie completely within O-islands. In 189 cases, backbone-island junctions are within predicted genes. We classified the EDL933 genes into the functional groups reported for the MG1655 genome[5] and this is included in the annotation. Of the O-island genes, 40% (561) can be assigned a function. Another 338 EDL933 genes marked as unknowns lie within phage-related clusters and are probably remnants of phage genomes. About 33% (59/177) of the O-islands contain only genes of unknown function. Many classifiable proteins are related to known virulence-associated proteins from other *E. coli* strains or related enterobacteria.

Nine large O-islands ($>$15 kb) encode putative virulence factors: a macrophage toxin and ClpB-like chaperone (OI#7); a RTX-toxin-like exoprotein and transport system (OI#28); two urease gene clusters (OI#43 and #48); an adhesin and polyketide or fatty-acid biosynthesis system (OI#47); a type III secretion system and secreted proteins similar to the *Salmonella–Shigella inv–spa* host-cell invasion genes (OI#115); two toxins and a PagC-like virulence factor (OI#122); a fatty-acid biosynthesis system (OI#138); and the previously described locus of enterocyte effacement (OI#148)[9]. Among the large islands, four include a P4-family integrase and are directly adjacent to tRNAs (OI#43-*serW*, #48-*serX*, #122-*pheV* and #148-*selC*). Only the locus of enterocyte effacement and two of the lambdoid phages (see below) have as yet been experimentally associated with virulence in animal models.

Smaller islands that may be involved in virulence contain fimbrial biosynthesis systems, iron uptake and utilization clusters, and putative non-fimbrial adhesins. Many clusters have no obvious role in virulence, but may confer strain-specific abilities to survive in different niches. Examples include candidates for transporting diverse carbohydrates, antibiotic efflux, aromatic compound degradation, tellurite resistance and glutamate fermentation. Not all islands are expected to be adaptive. Some may represent neutral variation between strains. Still others may be deleterious but either have not yet been eliminated by selection or cannot be eliminated because of linkage constraints.

We identified 18 multigenic regions of the EDL933 chromosome related to known bacteriophages. Only one, the Stx2 Shiga toxin-converting phage BP-933W, is known to be capable of lytic growth and production of infectious particles[10]. We named the other EDL933 prophages cryptic prophage (CP) to indicate that they probably lack a full complement of functional phage genes. They vary in size from 7.5 kb (CP-933L) to 61.6 kb (BP-933W) and consist of a mosaic of segments similar to various bacteriophages, recalling the 'modular' phage genome hypothesis[11]. The two remaining physical gaps in the genome sequence correspond to prophage-related regions, and resolution of the sequence is complicated by extensive similarities to other prophage within this genome. The gap sizes and positions (4 kb and 54 kb) were determined from optical restriction maps. With only one exception, the EDL933 prophages and the eight cryptic prophages of MG1655 are all lineage-specific. Prophage Rac (MG1655) and CP-933R are similarly located in the backbone, and are sufficiently related to suggest a common prophage ancestor at the time that the strains diverged.

Subunits Stx1A and Stx1B of the second Shiga toxin of EDL933 (ref. 12) are encoded in the newly identified CP-933V. The position of the *stx1* genes in a putative Q antiterminator-dependent transcript is analogous to the placement of the *stx2* genes in BP-933W, although there are no tRNAs adjacent to *stx1AB*. Genes in this position should be expressed maximally during lytic growth. The relationship between Stx toxin expression and phage induction is important, because treatment of O157:H7 with macrolide and quinolone antibiotics increase expression of the toxins[13,14]. Clinical decisions regarding drug therapy are complicated by strain-specific variation in this response[15], and reports in the literature (for example, refs 6 and 12) taken together suggest that the Stx phage status is variable among O157:H7 strains. Given the potential for recombination among the prophage reported here, this does not seem surprising. In addition, the *stx* locus in *Shigella* is known to lie within a cryptic prophage, inserted at a site different from either *stx* phage of EDL933 (ref. 16).

The MG1655 genome contains 528 genes (528/4,405 = 12%) not found in EDL933. About 57% (303) of these were classified into known functional groups and include genes, such as for ferric citrate utilization, that would suggest a role in virulence if identified in a pathogen. It is unclear whether these are remnants of a recent pathogenic ancestor, steps along a path leading to evolution of a new pathogen, indicators that K-12 strains may be pathogenic for non-human hosts, or completely unrelated to pathogenicity. There are 106 examples of O-islands and K-islands present at the same locations relative to the conserved chromosomal backbone. The two replichores in each strain are nearly equal in length despite the large number of insertion/deletion events necessary to generate the observed segmented structure between strains. Only a subset of islands is associated with elements likely to be autonomously mobile.

Each island might be ancestral and lost from the reciprocal genome; however, atypical base composition suggests that most islands are horizontal transfers of relatively recent origin from a donor species with a different intrinsic base composition. Restricting analysis to the 108 O-islands greater than 1 kb, 94% (101/108) are significantly different ($\chi^2 > 7.815$, $P < 0.05$) from the average base composition of shared backbone regions in the same replichore. The percentage drops very little with a Bonferroni correction for multiple tests (91/108; $\chi^2 > 17.892$, $P < 0.05$). Similar results are obtained for analysis of the third-codon position composition (Fig. 1). Still more islands may have originated as horizontal transfers but have been resident in genomes with a spectrum of mutation similar enough to *E. coli* to have obtained equilibrium

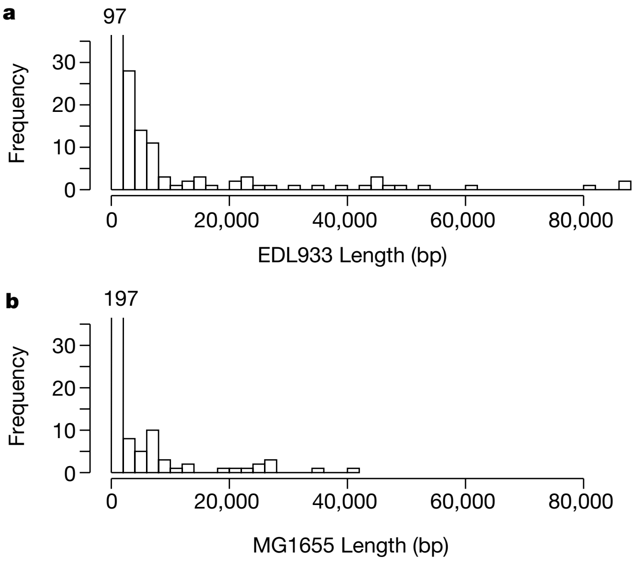

Figure 3 Histograms of lineage-specific segment lengths. **a**, EDL933; **b**, MG1655. The frequencies for the smallest length class are truncated to emphasize the distribution of longer clusters.

nucleotide frequencies or at least obscure statistical significance[17]. Still other gene clusters may be horizontal transfers that predate the divergence of MG1655 and EDL933.

Single nucleotide polymorphisms (75,168 differences) are distributed throughout the homologous backbone. There are 3,574 protein-coding genes encoded in backbone, and the average nucleotide identity for orthologous genes is 98.4%. Many orthologues (3,181/3,574 = 89%) are of equal length in the two genomes, but only 25% (911) encode identical proteins. Table 1 shows the number of each type of polymorphism observed by codon position. As expected, most differences are synonymous changes at third-codon positions. Multiple mutations at the same site should be infrequent at this low level of divergence. Thus the co-occurrence matrix provides insight into the substitution pattern, despite uncertainty of the ancestral state. The overall ratio of transitions to transversions is close to 3:1. A bias towards a greater number of T↔C than A↔G transitions on the coding strand previously attributed to transcription-coupled repair is evident[18]. An additional bias was observed at third-codon positions. Thymidines are more frequently involved in tranversions than cytosines, and G↔T are the most frequent transversions for the coding strand. The reciprocal polymorphisms, C↔A, are not over-represented. This bias is consistent for genes on both the leading and lagging strands (data not shown) and is therefore not related to asymmetries in the replication process. One possible explanation is transcription-coupled repair of damage associated with oxidative stress. Oxidized products of guanine (2,2,4-triaminooxazolone and 7,8-dihydro-8-oxoguanine) lead to G→T transversions by mispairing with A, and two DNA glycosylases (MutY and Fpg) are responsible for mismatch resolution[19]. Preferential repair of these lesions on the transcribed strand has been observed in humans[20], and a similar mechanism could account for the observed transversion bias on the coding strand in *E. coli*.

Some chromosomal regions are more divergent ('hypervariable') than the average homologous segment but encode a comparable set of proteins at the same relative chromosomal position. In the most extreme case (YadC), the MG1655 and EDL933 proteins exhibit only 34% identity. Four such loci encode known or putative fimbrial biosynthesis operons. Another encodes a restriction/modification system. Elevated divergence has been associated with positive selection at both these types of loci and among proteins that interact directly with the host[9,21,22]. Alternatively, hypervariable genes may result from locally elevated mutation rates or differential paralogue retention from an ancient tandem duplication.

Table 1 Frequency of each type of single nucleotide polymorphism by codon position for 3,181 genes of equal length in EDL933 and MG1655

First-codon position

	Base in MG1655				EDL933 totals
Base in EDL933	G	A	T	C	
G	–	865	154	137	1,156
A	924	–	129	286	1,339
T	170	143	–	1,260	1,573
C	124	284	1,156	–	1,564
MG1655 totals	1,218	1,292	1,439	1,683	5,632

Second-codon position

	Base in MG1655				EDL933 totals
Base in EDL933	G	A	T	C	
G	–	410	66	118	594
A	393	–	159	166	718
T	67	147	–	464	678
C	107	176	394	–	677
MG1655 totals	567	733	619	748	2,667

Third-codon position

	Base in MG1655				EDL933 totals
Base in EDL933	G	A	T	C	
G	–	6,021	1,562	1,024	8,607
A	6,107	–	1,242	1,124	8,473
T	1,619	1,228	–	8,538	11,385
C	1,049	1,010	8,307	–	10,366
MG1655 totals	8,775	8,259	11,111	10,686	38,831

Comparison of our observations with other genome-scale analyses of closely related strains or species supports the idea that enterobacterial genomes are particularly subject to recombinational evolution. Two *Helicobacter pylori* strains exhibited only 6–7% differential coding capacity despite showing less identity among orthologues (92.6%) than observed among these *E. coli*. Furthermore, almost half of the lineage-specific *Helicobacter* genes are clustered in a single region referred to as the plasticity zone[23]. Analyses of four *Chlamydia* genomes with orthologues that differ by as much as 19.5% show little evidence of horizontal transfer, and this is attributed to the inherent isolation of an obligate intracellular parasite[24]. Most lineage-specific genes are expansions of paralogous gene families. As in *Helicobacter*, many of the *Chlamydia* lineage-specific elements are clustered in a plasticity zone. Continuing genome projects will elucidate the generality of observations made from these comparisons of closely related organisms.

Together, our findings reveal a surprising level of diversity between two members of the species *E. coli*. Most differences in overall gene content are attributable to horizontal transfer, and offer a wealth of candidate genes that may be involved in pathogenesis. Base substitution has introduced variation into most gene products even among conserved regions of the two strains. Many of these differences can be exploited for development of highly sensitive diagnostic tools; but diagnostic utility will require a clearer understanding of the distribution of genetic elements in *E. coli* species as a whole. An independently isolated O157:H7 strain showed differences from EDL933 by restriction mapping[25]. Additional genome sequence data from other *E. coli* strains as well as functional characterization of gene products is necessary before the complex relationship between *E. coli* genotypes and phenotypes can be understood. Showing that disease-related traits are associated with predicted genes will require many areas of study including extensive testing in animal models that mimic symptoms of human infections, but the genome sequence offers a unique resource to help meet the challenge. □

Methods

Clones and sequencing

EDL933 was kindly provided by C. Kaspar, who obtained it from the American Type Culture Collection (ATCC 43895). The sequenced isolate has been redeposited at the ATCC and is available as ATCC 700927. Whole-genome libraries in M13Janus and pBluescript were prepared from genomic DNA as described for genome segments used in the K-12 genome project[26]. Random clones were sequenced using dye-terminator chemistry and data were collected on ABI377 and 3700 automated sequencers. Sequence data were assembled by Seqman II (DNASTAR). Finishing used sequencing of opposite ends of linking clones, several PCR-based techniques and primer walking. Whole-genome optical maps for restriction enzymes *Nhe*I and *Xho*I were prepared[27] so that the ordering of contigs during assembly could be confirmed. Two gaps remain in the genome sequence. Extended exact matches pose a significant assembly challenge. The final determination of sequence for the 100-kb duplicated region was based on clones that span the junction between unique flanking sequences and the ends of the duplicated island, concordance of the two regions in optical restriction maps, excess random sequence coverage in the duplicated region, lack of polymorphism and confirmation of duplication of an internal segment by Southern blotting (data not shown).

Sequence features and database searches

Potential open reading frames (ORFs) were defined by GeneMark.hmm[28]. The GenPept118 protein and MG1655 protein and DNA databases were searched by each ORF using BLAST[29]. Annotations were created from the search output in which each gene was inspected, assigned a unique identifier, and its product classified by functional group[5]. Alternative start sites were chosen to conform to the annotated MG1655 sequence. Orthology was inferred when matches for EDL933 genes in the MG1655 database exceeded 90% nucleotide identity, alignments included at least 90% of both genes, and the MG1655 gene did not have an equivalent match elsewhere in the EDL933 genome. This list was supplemented by manual inspection of the protein-level matches in the complete GenPept database to include genes with lower similarities if they occurred within co-linear regions of the genomes. The genome sequence was compared with that of MG1655 by the maximal exact match (MEM) alignment utility, (B.M., manuscript in preparation) an adaptation of MUMmer[30]. This program was based on suffix arrays rather than suffix trees, and exact rather than unique matches, coupled with a custom anchored-alignment algorithm that extends sequence homology into the regions separating contiguous co-linear exact matches. Inferences on biases in polymorphism patterns are based on χ^2 goodness-of-fit tests of a nested sequence of multinomial log–linear models. These predict symmetric elevated levels of A↔G, T↔C and G↔T polymorphisms, above a quasi-independent baseline generated from marginal frequencies in the co-occurrence matrix of synonymous third-codon differences. Further information may be found at our Website http://www.genome.wisc.edu/, including a Genome Browser displaying a comparative map of EDL933 and K-12.

Received 24 July; accepted 6 November 2000.

1. Riley, L. W. *et al.* Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. *N. Engl. J. Med.* **308,** 681–685 (1983).
2. Karmali, M. A., Steele, B. T., Petric, M. & Lim, C. Sporadic cases of haemolytic–uraemic syndrome associated with faecal cytotoxin and cytotoxin-producing *Escherichia coli* in stools. *Lancet* **ii,** 619–620 (1983).
3. Mead, P. S. *et al.* Food-related illness and death in the United States. *Emerg. Infect. Dis.* **5,** 607–625 (1999).
4. Su, C. & Brandt, L. J. *Escherichia coli* O157:H7 infection in humans. *Ann. Intern. Med.* **123,** 698–714 (1995).
5. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453–1474 (1997).
6. Reid, S. D., Herbelin, C. J., Bumbaugh, A. C., Selander, R. K. & Whittam, T. S. Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* **406,** 64–67 (2000).
7. Blattner, F. R. *et al.* Comparative genome sequencing of *E. coli* O157:H7 versus *E. coli* K 12. *Microb. Compar. Genom.* **2,** 174 (1997).
8. Hacker, J. *et al.* Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal *Escherichia coli* isolates. *Microb. Pathog.* **8,** 213-225 (1990).
9. Perna, N. T. *et al.* Molecular evolution of a pathogenicity island from enterohemorrhagic *Escherichia coli* O157:H7. *Infect. Immun.* **66,** 3810–3817 (1998).
10. Plunkett, G. III, Rose, D. J., Durfee, T. J. & Blattner, F. R. Sequence of Shiga toxin 2 phage 933W from *Escherichia coli* O157:H7: Shiga toxin as a phage late-gene product. *J. Bacteriol.* **181,** 1767–1778 (1999).
11. Campbell, A. & Botstein, D. in *Lambda II* (eds Hendrix, R. W., Roberts, J. W., Stahl, F. W. & Weisberg, R. A.) 365–380 (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1983).
12. O'Brien, A. D. *et al.* Shiga-like toxin-converting phages from *Escherichia coli* strains that cause hemorrhagic colitis or infantile diarrhea. *Science* **226,** 694–696 (1984).
13. Walterspiel, J. N., Ashkenazi, S., Morrow, A. L. & Cleary, T. G. Effect of subinhibitory concentrations of antibiotics on extracellular Shiga-like toxin I. *Infection* **20,** 25–29 (1992).
14. Neely, M. N. & Friedman, D. I. Functional and genetic analysis of regulatory regions of coliphage H-19B: location of shiga-like toxin and lysis genes suggest a role for phage functions in toxin release. *Mol. Microbiol.* **28,** 1255–1267 (1998).
15. Grif, K., Dierich, M. P., Karch, H. & Allerberger, F. Strain-specific differences in the amount of Shiga toxin released from enterohemorrhagic *Escherichia coli* O157 following exposure to subinhibitory concentrations of antimicrobial agents. *Eur. J. Clin. Microbiol. Infect. Dis.* **17,** 761–766 (1998).
16. McDonough, M. A. & Butterton, J. R. Spontaneous tandem amplification and deletion of the shiga toxin operon in *Shigella dysenteriae* 1. *Mol. Microbiol.* **34,** 1058–1069 (1999).
17. Lawrence, J. G. & Ochman, H. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* **44,** 383–397 (1997).
18. Francino, M. P., Chao, L., Riley, M. A. & Ochman, H. Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science* **272,** 107–109 (1996).
19. Blaisdell, J. O., Hatahet, Z. & Wallace, S. S. A novel role for *Escherichia coli* endonuclease VIII in prevention of spontaneous G→T transversions. *J. Bacteriol.* **181,** 6396–6402 (1999).
20. Le Page, F. *et al.* Transcription-coupled repair of 8-oxoguanine: requirement for XPG, TFIIH, and CSB and implications for Cockayne syndrome. *Cell* **101,** 159–171 (2000).
21. Boyd, E. F., Li, J., Ochman, H. & Selander, R. K. Comparative genetics of the *inv–spa* invasion gene complex of *Salmonella enterica*. *J. Bacteriol.* **179,** 1985–1991 (1997).
22. Sharp, P. M., Kelleher, J. E., Daniel, A. S., Cowan, G. M. & Murray, N. E. Roles of selection and recombination in the evolution of type I restriction-modification systems in enterobacteria. *Proc. Natl Acad. Sci. USA* **89,** 9836–9840 (1992).
23. Alm, R. A. & Trust, T. J. Analysis of the genetic diversity of *Helicobacter pylori*: the tale of two genomes. *J. Mol. Med.* **77,** 834–846 (1999).
24. Read, T. D. *et al.* Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.* **28,** 1397–1406 (2000).
25. Ohnishi, M. *et al.* Chromosome of the enterohemorrhagic *Escherichia coli* O157:H7: comparative analysis with K-12 MG1655 revealed the acquisition of a large amount of foreign DNAs. *DNA Res.* **6,** 361–368 (1999).
26. Mahillon, J. *et al.* Subdivision of *Escherichia coli* K-12 genome for sequencing: manipulation and DNA sequence of transposable elements introducing unique restriction sites. *Gene* **223,** 47–54 (1998).
27. Lin, J. *et al.* Whole-genome shotgun optical mapping of *Deinococcus radiodurans*. *Science* **285,** 1558–1562 (1999).
28. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26,** 1107–1115 (1998).
29. 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).
30. Delcher, A. L. *et al.* Alignment of whole genomes. *Nucleic Acids Res.* **27,** 2369–2376 (1999).

Acknowledgements

We thank T. Forsythe, M. Goeden, H. Kijenski, B. Leininger, J. McHugh, B. Peterson, G. Peyrot, D. Sands, P. Soni, E. Travanty and other members of the University of Wisconsin genomics team for their expert technical assistance. This work was funded by grants from the NIH (NIAID and NCHGR), the University of Wisconsin Graduate School and the RMHC to F.R.B., the NIH (NCHGR, NIAID) to D.C.S., HHMI/OTKA to G.P., an Alfred P. Sloan/DOE Fellowship to B.M., a CDC/APHL Fellowship to P.S.E., and an Alfred P. Sloan/NSF Fellowship to N.T.P. Sixteen University of Wisconsin undergraduates participated in this work and particular thanks are due to A. Byrnes for web-site development to complement this project, and to A. Darling for programming.

Correspondence and requests for materials should be addressed to N.T.P. (e-mail: perna@ahabs.wisc.edu). The GenBank accession number for the annotated sequence is AE00517H.

……

Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF

Vishwanath R. Iyer*†‡, Christine E. Horak§‡, Charles S. Scafe‖¶, David Botstein‖, Michael Snyder§ & Patrick O. Brown*#

** Department of Biochemistry and # Howard Hughes Medical Institute, Stanford University Medical Center, Stanford, California 94305, USA*
‖ Department of Genetics, Stanford University Medical Center, Stanford, California 94305, USA
§ Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA
† These authors contributed equally to this work

Proteins interact with genomic DNA to bring the genome to life; and these interactions also define many functional features of the genome. SBF and MBF are sequence-specific transcription factors that activate gene expression during the G1/S transition of the cell cycle in yeast[1,2]. SBF is a heterodimer of Swi4 and Swi6, and MBF is a heterodimer of Mbp1 and Swi6 (refs 1, 3). The related Swi4 and Mbp1 proteins are the DNA-binding components of the respective factors, and Swi6 may have a regulatory function[4,5]. A small number of SBF and MBF target genes have been identified[3,6–10]. Here we define the genomic binding sites of the SBF and MBF transcription factors *in vivo*, by using DNA microarrays. In addition to the previously characterized targets, we have identified about 200 new putative targets. Our results support the hypothesis that SBF activated genes are predominantly involved in budding, and in membrane and cell-wall biosynthesis, whereas DNA replication and repair are the dominant functions among MBF activated genes[6,11]. The functional specialization of these factors may provide a mechanism for independent regulation of distinct molecular processes that normally occur in synchrony during the mitotic cell cycle.

To identify the targets of SBF and MBF, we combined chromatin immunoprecipitation and microarray hybridization (Fig. 1). Proteins were crosslinked with formaldehyde to their target sites *in vivo*. DNA that was specifically crosslinked to either of the transcription factors was purified by immunoprecipitation using an antibody against either the native protein or an epitope tag that was fused to the protein. Polymerase chain reaction (PCR) analysis of immunoprecipitated DNA confirmed the specific association of Swi4, Swi6 and Mbp1 with several known target promoters, and other target promoters that are identified here (see Supplementary Information). After reversal of the crosslinks, immunoprecipitated DNA was amplified and fluorescently labelled with the Cy5 fluoro-

---

‡ Present address: Institute of Molecular and Cellular Biology, University of Texas at Austin, Austin, Texas 78712, USA.

¶ Present address: Applied Biosystems, Foster City, California 94404, USA.

is less than 100%). Models for the additional oligonucleotide, GTP molecules and Mg^{2+} ions, have been fitted into electron density maps and refinement of these oligo–Mn^{2+}–polymerase and oligo–GTP–Mg–Mn–polymerase complexes against their data sets, imposing strict threefold NCS constraints, resulted in models with *R* factors of 23.7 and 21.4%, respectively, and good stereochemistry (Table 1).

Figures

Unless otherwise stated figures were drawn using BOBSCRIPT[26] and rendered with RASTER3D[27].

Received 2 August; accepted 28 December 2000.

1. Reinisch, K. M., Nibert, M. L. & Harrison, S. C. Structure of the reovirus core at 3. 6 Å resolution. *Nature* **404,** 960–967 (2000).
2. Grimes, J. M. *et al.* The atomic structure of the bluetongue virus core. *Nature* **395,** 470–478 (1998).
3. Makeyev, E. V. & Bamford, D. H. Replicase activity of purified recombinant protein P2 of double-stranded RNA bacteriophage ϕ6. *EMBO J.* **19,** 124–133 (2000).
4. Butcher, S. J., Makeyev, E. V., Grimes, J. M., Stuart, D. I. & Bamford, D. H. Crystallization and preliminary X-ray crystallographic studies on the bacteriophage ϕ6 RNA-dependent RNA polymerase. *Acta Crystallogr. D* **56,** 1473–1475 (2000).
5. Mindich, L. Reverse genetics of dsRNA bacteriophage ϕ6. *Adv. Virus Res.* **53,** 341–353 (1999).
6. Gottlieb, P., Strassman, J., Quao, X., Frucht, A. & Mindich, L. In vitro replication, packaging, and transcription of the segmented, double-stranded RNA genome of bacteriophage ϕ6: studies with procapsids assembled from plasmid-encoded proteins. *J. Bacteriol.* **172,** 5774–5782 (1990).
7. Mindich, L. Precise packaging of the three genomic segments of the double-stranded-RNA bacteriophage ϕ6. *Microbiol. Mol. Biol. Rev.* **63,** 149–160 (1999).
8. Makeyev, E. V. & Bamford, D. H. The polymerase subunit of a dsRNA virus plays a central role in the regulation of viral RNA metabolism. *EMBO J.* **19,** 124–133 (2000).
9. Ollis, D. L., Kline, C. & Steitz, T. A. Domain of E. coli DNA polymerase I showing sequence homology to T7 DNA polymerase. *Nature* **313,** 818–819 (1985).
10. Delarue, M., Poch, O., Tordo, N., Moras, D. & Argos, P. An attempt to unify the structure of polymerases. *Protein Eng.* **3,** 461–467 (1990).
11. Lesburg, C. A. *et al.* Crystal structure of the RNA-dependent RNA polymerase from hepatitis C virus reveals a fully encircled active site. *Nature Struct. Biol.* **6,** 937–943 (1999).
12. Ago, H. *et al.* Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Struct. Fold. Des.* **7,** 1417–1426 (1999).
13. Bressanelli, S. *et al.* Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Proc. Natl Acad. Sci. USA* **96,** 13034–13039 (1999).
14. Stuart, D. I., Levine, M., Muirhead, H. & Stammers, D. K. Crystal structure of cat muscle pyruvate kinase at resolution of 2. 6Å. *J. Mol. Biol.* **134,** 109–142 (1979).
15. Oh, J. W., Ito, T. & Lai, M. M. A recombinant hepatitis C virus RNA-dependent RNA polymerase capable of copying the full-length viral RNA. *J. Virol.* **73,** 7694–7702 (1999).
16. Lohmann, V., Overton, H. & Bartenschlager, R. Selective stimulation of hepatitis C virus and pestivirus NS5B RNA polymerase activity by GTP. *J. Biol. Chem.* **274,** 10807–10815 (1999).
17. Frilander, M., Poranen, M. & Bamford, D. H. The large genome segment of dsRNA bacteriophage ϕ6 is the key regulator in the in vitro minus and plus strand synthesis. *RNA* **1,** 510–518 (1995).
18. van Dijk, A. A., Frilander, M. & Bamford, D. H. Differentiation between minus- and plus-strand synthesis: polymerase activity of dsRNA bacteriophage ϕ6 in an in vitro packaging and replication system. *Virology* **211,** 320–323 (1995).
19. Huang, H., Chopra, R., Verdine, G. L. & Harrison, S. C. Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. *Science* **282,** 1669–1675 (1998).
20. Zhong, W., Uss, A. S., Ferrari, E., Lau, J. Y. & Hong, Z. De novo initiation of RNA synthesis by hepatitis C virus nonstructural protein 5B polymerase. *J. Virol.* **74,** 2017–2022 (2000).
21. Yazaki, K. & Miura, K. Relation of the structure of cytoplasmic polyhedrosis virus and the synthesis of its messenger RNA. *Virology* **105,** 467–479 (1980).
22. Hendrickson, W. A. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254,** 51–58 (1991).
23. Brunger, A. T. *et al.* Crystallography and NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
24. Navaza, J. AMoRe: an automated package for molecular replacement. *Acta Crystallogr. A* **50,** 164–182 (1994).
25. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26,** 283–291 (1993).
26. Esnouf, R. M. An extensively modified version of MolScript that includes greatly enhanced colouring capabilities. *J. Mol. Graph.* **15,** 132–134 (1997).
27. Merritt, E. A. & Bacon, D. J. in *Macromolecular Crystallography* (eds Carter, J. W. Jr & Sweet, R. M.) 505–524 (Academic, San Diego, 1997).
28. Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11,** 281–296 (1991).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

J. Diprose and G. Sutton helped with synchrotron data collection; J. Diprose and S. Ikemizu with calculations; and R. Esnouf and K. Harlos with computing and in-house data collection. We thank the staff at the beamlines of the ESRF, SRS and APS, in particular Sergey Korolev at the APS for help with the MAD experiment. S.J.B. is a Marie Curie Fellow. J.M.G. is funded by the Royal Society and D.I.S. by the Medical Research Council. The work was supported by the Academy of Finland, the Medical Research Council and the European Union.

Correspondence and requests for materials should be addressed to D.I.S. (e-mail: dave@strubi.ox.ac.uk). Coordinates have been deposited in the RCSB Protein database under accesssion codes: 1HHS, 1HHT, 1HI0, 1HI1, 1HI8.

---

correction

Improved estimates of global ocean circulation, heat transport and mixing from hydrographic data

Alexandra Ganachaud & Carl Wunsch

Nature **408,** 453–457 (2000).

In this first paragraph of this paper, the uncertainty on the net deep-water production rates in the North Atlantic Ocean was given incorrectly. The correct value should have been $(15 \pm 2) \times 10^6\,\mathrm{m}^3\,\mathrm{s}^{-1}$. ☐

---

errata

Changes in Greenland ice sheet elevation attributed primarily to snow accumulation variability

J. R. McConnell, R. J. Arthern, E. Mosley-Thompson, C. H. Davis, R. C. Bales, R. Thomas, J. F. Burkhard & J. D. Kyne

Nature **406,** 877–879 (2000).

As the result of an editing error, the 1993–1998 aircraft-based altimetry surveys of the southern Greenland ice sheet reported by Krabill *et al.* (1999) were erroneously described as satellite-based. ☐

Genome sequence of enterohaemorrhagic *Escherichia coli* 0157:H7

Nicole T. Perna, Guy Plunkett III, Valerie Burland, Bob Mau, Jeremy D. Glasner, Debra J. Rose, George F. Mayhew, Peter S. Evans, Jason Gregor, Heather A. Kirkpatrick, György Pósfai, Jeremiah Hackett, Sara Klink, Adam Boutin, Ying Shao, Leslie Miller, Erik J. Grotbeck, N. Wayne Davis, Alex Lim, Eileen T. Dimalanta, Konstantinos D. Potamousis, Jennifer Apodaca, Thomas S. Anantharaman, Jieyi Lin, Galex Yen, David C. Schwartz, Rodney A. Welch & Frederick R. Blattner

Nature **409,** 529–533 (2001).

The Genbank accession number for the annotated sequence given in this paper was typeset incorrectly. The correct accession number is AE005174. ☐