$ATR^{flox/-}$ cells, which already had reduced ATRIP expression. Transfection of siRNA in HCT116 cells effectively reduced ATRIP expression (*17*). Transfection of control siRNAs in $ATR^{flox/-}$ cells, Ad-Cre infection of $ATR^{+/+}$ cells, or Ad-GFP infection of $ATR^{flox/-}$ cells had no effect on the ability of these cells to delay entry into mitosis after ionizing radiation. However, transfection of siRNAs against ATRIP yielded a profound γ-irradiation–induced G_2-M checkpoint defect that was similar to that seen in the $ATR^{flox/-}$ cells treated with Ad-Cre (Fig. 4D). About 40% of Cre-infected or ATRIP siRNA-transfected $ATR^{flox/-}$ cells enter mitosis 16 hours after irradiation compared with 20% of control cells. These results are consistent with checkpoint defects of cells overexpressing catalytically inactive ATR protein (*6*). Thus, ATR and ATRIP are essential for a normal DNA-damage–induced delay of mitosis initiated by ionizing radiation.

These data strongly suggest that ATRIP is the functional human homolog of the Rad26 family of genes. ATRIP associates with ATR, is a substrate of ATR in vitro and a phosphoprotein in vivo, and colocalizes with ATR to sites of DNA synthesis and repair after treatment of cells with DNA-damaging agents or replication inhibitors. Furthermore, interference with ATRIP function generates the same G_2-M checkpoint defect as observed after deletion of ATR. ATRIP expression is dependent on ATR, and ATR expression is dependent on ATRIP. This mutual dependency for expression suggests that the amount of ATR and ATRIP in cells is tightly coordinated and may indicate that these proteins form a stable complex with each other at a fixed stoichiometry.

ATR function is required for the viability of undamaged, proliferating cells and in cells exposed to DNA-damaging agents. In this respect, ATR is similar to MEC1, which is essential for viability because of difficulties in the proper coordination of DNA replication (*19*, *20*). We did observe an increase in the percentage of S phase cells after Ad-Cre infection of the $ATR^{flox/-}$ cells (see Fig. 3C), perhaps reflecting a requirement for ATR and ATRIP signaling to ensure successful DNA replication.

References and Notes

1. D. Durocher, S. P. Jackson, *Curr. Opin. Cell Biol.* **13**, 225 (2001).
2. Y. Shiloh, *Curr. Opin. Genet. Dev.* **11**, 71 (2001).
3. N. J. Bentley *et al.*, *EMBO J.* **15**, 6641 (1996).
4. K. A. Cimprich, T. B. Shin, C. T. Keith, S. L. Schreiber, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 2850 (1996).
5. J. A. Wright *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 7445 (1998).
6. W. A. Cliby *et al.*, *EMBO J.* **17**, 159 (1998).
7. R. S. Tibbetts *et al.*, *Genes Dev.* **13**, 152 (1999).
8. R. S. Tibbetts *et al.*, *Genes Dev.* **14**, 2989 (2000).
9. E. J. Brown, D. Baltimore, *Genes Dev.* **14**, 397 (2000).
10. A. de Klein *et al.*, *Curr. Biol.* **10**, 479 (2000).
11. R. J. Edwards, N. J. Bentley, A. M. Carr, *Nature Cell Biol.* **1**, 393 (1999).
12. V. Paciotti, M. Clerici, G. Lucchini, M. P. Longhese, *Genes Dev.* **14**, 2046 (2000).
13. J. Rouse, S. P. Jackson, *EMBO J.* **19**, 5801 (2000).
14. T. Wakayama, T. Kondo, S. Ando, K. Matsumoto, K. Sugimoto, *Mol. Cell. Biol.* **21**, 755 (2001).
15. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
16. M. H. Brodsky, J. J. Sekelsky, G. Tsang, R. S. Hawley, G. M. Rubin, *Genes Dev.* **14**, 666 (2000).
17. D. Cortez, unpublished data.
18. See supplemental Web information available on *Science* Online at www.sciencemag.org/cgi/content/full/294/5547/1713/DC1.
19. B. A. Desany, A. A. Alcasabas, J. B. Bachant, S. J. Elledge, *Genes Dev.* **12**, 2956 (1998).
20. X. Zhao, E. G. Muller, R. Rothstein, *Mol. Cell* **2**, 329 (1998).
21. C. E. Canman *et al.*, *Science* **281**, 1677 (1998).
22. Y. Wang *et al.*, *Genes Dev.* **14**, 927 (2000).
23. S. M. Elbashir *et al.*, *Nature* **411**, 494 (2001).
24. The siRNA duplexes were 21 base pairs including a 2–base pair deoxynucleotide overhang. The coding strands of the three ATRIP siRNAs were GGUCCACAGAUUAUUAGA*UTT*, AGAGGAACAGAGAAGAUCAC*A*, and GAAGAGGCCCAGAAAAGCU*TT*. The two control siRNAs used were GACCCGCGCCGAGGUGAAGUU and UGGCUUUCUGUAGAGGACAUC*TT*. Italics indicate deoxynucleotides.
25. B. Xu, S. Kim, M. B. Kastan, *Mol. Cell. Biol.* **21**, 3445 (2001).

26. We thank T. Lee for technical assistance. D.C. is a Fellow of the Jane Coffin Childs Memorial Fund for Medical Research. S.J.E. is an Investigator with the Howard Hughes Medical Institute.

20 August 2001; accepted 9 October 2001

Two Essential DNA Polymerases at the Bacterial Replication Fork

Etienne Dervyn,[1] Catherine Suski,[1] Richard Daniel,[2] Claude Bruand,[1] Jérome Chapuis,[1] Jeff Errington,[2] Laurent Jannière,[1] S. Dusko Ehrlich[1]*

DNA replication in bacteria is carried out by a multiprotein complex, which is thought to contain only one essential DNA polymerase, specified by the *dnaE* gene in *Escherichia coli* and the *polC* gene in *Bacillus subtilis*. *Bacillus subtilis* genome analysis has revealed another DNA polymerase gene, $dnaE_{BS}$, which is homologous to *dnaE*. We show that, in *B. subtilis*, $dnaE_{BS}$ is essential for cell viability and for the elongation step of DNA replication, as is *polC*, and we conclude that there are two different essential DNA polymerases at the replication fork of *B. subtilis*, as was previously observed in eukaryotes. $dnaE_{BS}$ appears to be involved in the synthesis of the lagging DNA strand and to be associated with the replication factory, which suggests that two different polymerases carry out synthesis of the two DNA strands in *B. subtilis* and in many other bacteria that contain both *polC* and *dnaE* genes.

The paradigm of the bacterial replication fork is that of *Escherichia coli* (*1*). The fork includes the DNA polymerase III holoenzyme, which contains 10 different subunits. One of the subunits, α, is the catalytic DNA polymerase. There are two α molecules in the holoenzyme, each of which copies a different DNA strand. The holoenzyme from *B. subtilis* is not as well characterized, but was reported to contain a similar number of polypeptides, including a catalytic DNA polymerase subunit (*2*). The *E. coli* and *B. subtilis* catalytic subunits of the holoenzyme, encoded by *dnaE* and *polC* genes, respectively, are prototypes of two different DNA polymerase classes within the so-called family C, which groups replicate DNA polymerases from eubacteria (*3*). The *B. subtilis* genome sequence (*4*) indicated the existence of two other genes encoding family C DNA polymerases, in addition to *polC*, both of the *E. coli* class. One of these genes, *yorI*, is carried on a prophage that can be eliminated from *B. subtilis* (*5*) and thus is not essential for cell growth. The other gene, referred to here as $dnaE_{BS}$, is encoded chromosomally. Genes homologous to *polC* and $dnaE_{BS}$ are present in all fully sequenced genomes of bacteria belonging to the *Bacillus/Clostridium* group and in the genome of a thermophilic microorganism species called *Thermotoga maritima*. $dnaE_{BS}$ encodes a protein with DNA polymerase activity (*6*), as does its homolog from *Streptococcus pyogenes* (*7*).

The *polC* gene is essential for *B. subtilis* cell growth (*8*). We show here that $dnaE_{BS}$ is also essential. First, inactivation of the gene by recombination with an insertional plasmid vector, pMUTIN (*9*), carrying an internal segment of the gene, was not successful. The failure was not due to a polar effect of insertion, which we know because a downstream

[1]Génétique Microbienne, Institut National de la Recherche Agronomique, Jouy-en-Josas, 78352 Cedex, France. [2]Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford, OX1 3RE, UK.

*To whom correspondence should be addressed. E-mail: ehrlich@jouy.mra.fr

gene (*ytsJ*; Fig. 1, top) could be inactivated. Similar results were obtained with the *polC* gene and the downstream *ylxS* gene (Fig. 1, top). Second, it was possible to place the $dnaE_{BS}$ gene under the control of the P_{spac} promoter, which is regulated by the LacI repressor and is therefore induced by isopropyl-β-D-thiogalactopyranoside (IPTG), by insertion of pMUTIN carrying a DNA segment overlapping the 5′-end of the gene (Fig. 1, top). However, the resulting strain required IPTG for growth (Fig. 1, middle) (*10*). Analogous observations were obtained with the *polC* gene (Fig. 1, middle). Third, we isolated 11 different thermosensitive $dnaE_{BS}$ mutants (*10*).

To test whether $dnaE_{BS}$ is required for DNA synthesis, we cultivated cells having the gene under P_{spac} control, with or without IPTG, and monitored incorporation of [^{3}H]-thymidine into acid-insoluble material (*11*). DNA synthesis was inhibited by depletion of $DnaE_{BS}$ whereas RNA and protein syntheses were not (Fig. 1, bottom). Analogous results were obtained for the *polC* gene (Fig. 1, bottom). We conclude that because both genes are required for DNA synthesis, their products must not have the same role in this process.

Because PolC is required for elongation (*8*), it seemed possible that $DnaE_{BS}$ might be required for initiation. Two lines of evidence show that $DnaE_{BS}$ is essential for elongation. First, incorporation of labeled thymidine was arrested immediately upon a shift of a thermosensitive $dnaE_{BS}$ mutant from 30° to 47°C (Fig. 2, left). Second, the relative abundance of chromosomal markers close to the origin and terminus of replication did not change upon depletion of DnaE or PolC (Fig. 2, right). The origin-to-terminus ratio is between 2:1 and 4:1 in *B. subtilis* cells during exponential growth (*12*). We would expect this ratio to remain constant if the elongation of replication were arrested, whereas it would change to 1:1 if the elongation continued but the initiation ceased. We derived two conclusions from these observations. First, depletion of the unaltered catalytic subunit of the *B. subtilis* holoenzyme PolC interferes with elongation of DNA replication. This suggests that the polymerase is frequently lost from and then reloaded onto the progressing replication fork, which might explain the presence of high PolC levels in *B. subtilis*, revealed by the analysis of cells carrying the protein fused to green fluorescent protein (GFP) (*13*). Second, depletion of $DnaE_{BS}$ also arrests elongation, indicating that $DnaE_{BS}$ is required for this replication step.

Because both PolC and $DnaE_{BS}$ are required for the elongation step of DNA replication, they must have different roles in this process. We considered that each might be involved in the synthesis of a different DNA strand, and so we used a pAMβ1-derived plasmid to test this possibility. Replication of pAMβ1 is initiated by DNA polymerase I, which progresses for about 200 base pairs, thus generating a D loop intermediate (*14*). A PriA-dependent primosome then assembles at a specific site present within the single-stranded part of the D loop (*15–17*) and promotes assembly of the holoenzyme, which completes the replication of the plasmid. This process mimics the restart of the arrested replication fork. pAMβ1 replication is unidirectional, and the leading and lagging strands are thus known (*18*). A derivative plasmid was introduced into $dnaE_{BS}$ or *polC* thermosensitive mutants, and the resulting strains were propagated at 37°C, the temperature at which we speculated the polymerase activity might already be affected. Total cell DNA was prepared and hybridized with probes specific for one or the other of the plasmid strands (Fig. 3). As expected, only the double-stranded plasmid DNA was detected in the control wild-type strain. In contrast, in the $dnaE_{BS}$ mutant, the probe complementary to the leading strand revealed, in addition to double-stranded DNA, a fast-migrating DNA form. This form did not hybridize with the probe specific for the other strand, which shows that it was single stranded and corresponded to only one DNA strand. The amount of single-stranded DNA was relatively low (about 3%), possibly for two reasons. One is that inactivation of $DnaE_{BS}$ at 37°C was incomplete, because sufficient polymerase activity had to be maintained to allow cell viability, and thus the uncoupling of the synthesis of the two strands may have taken place in only a fraction of molecules. The other is that the conversion of plasmid single-stranded DNA into a double-stranded form is efficient in *B. subtilis*. For instance, rolling circle plasmids, in which leading and lagging strand synthesis are fully uncoupled, often accumulate no more than 10% of single-

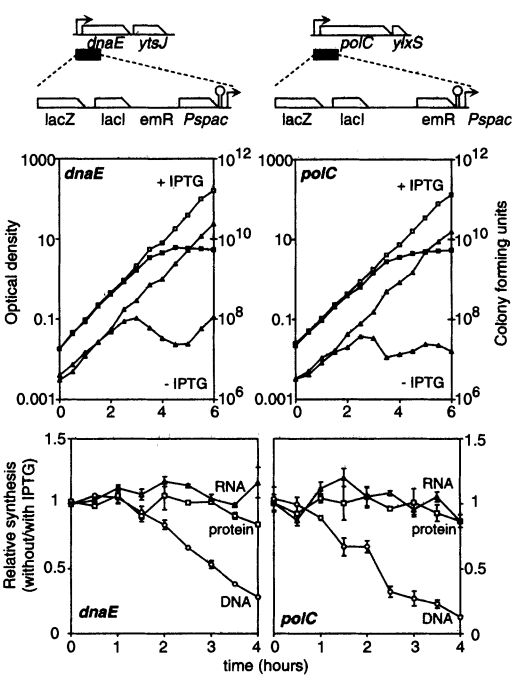

Fig. 1. $dnaE_{BS}$ and *polC* genes are essential for *B. subtilis* growth and DNA replication. (**Top**) $dnaE_{BS}$ and *polC* regions in *B. subtilis*. DNA segments overlapping the 5′ end of each gene (thick lines) were used for single cross-over integration of pMUTIN2 (depicted as a thin line; boxes, lollipops, and bent arrows indicate genes, transcription terminators, and promoters, respectively; the *emR* gene derives from plasmid pE194) to obtain the strains in which the *dnaE* or *polC* genes are placed under the control of the P_{spac} promoter and were designated HVS614 and HVS609, respectively. (**Middle**) Growth of HVS614 (left) and HVS609 (right). Cells were grown in Luria broth (LB) medium supplemented or not supplemented with IPTG (open and solid symbols, respectively) and were diluted periodically to maintain exponential growth. Optical density (squares) and colony-forming units (on IPTG-supplemented plates, triangles) were determined. (**Bottom**) Macromolecular synthesis in HVS609 and HVS614 grown with or without IPTG, measured as described previously (*11*). Average values and dispersion from four experiments for DNA (circles) and two experiments for RNA (triangles) or protein (squares) are shown.

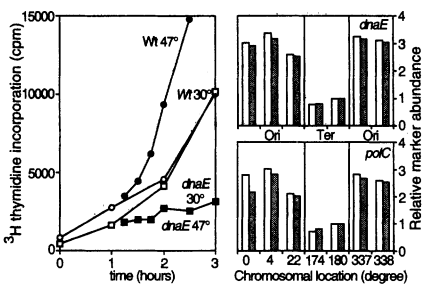

Fig. 2. $DnaE_{BS}$ is required for elongation. (**Left**) The $dnaE_{BS}$ thermosensitive strain EDJ51 and the parental strain 168 were grown at 30°C in the presence of [^{3}H]-thymidine. At the 1-hour point, half of each culture was transferred to 47°C and the [^{3}H]-thymidine incorporation was determined at 30° and 47°C (open and solid symbols). Squares and circles refer to the mutant (*dnaE*) and parental (Wt) strains, respectively. (**Right**) Exponentially growing cells having $dnaE_{BS}$ or *polC* genes under the control of P_{spac} (HVS614 or HVS609, respectively) were inoculated into the rich LB medium with or without IPTG and grown for 3 hours. Their DNA was extracted, cleaved with Eco*RV*, and analyzed by Southern blotting with probes from seven different chromosomal regions (positions are indicated in degrees) (*10*). Gray and white bars refer to samples with and without IPTG.

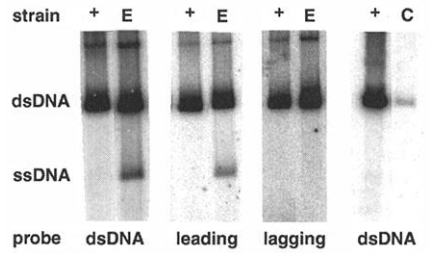

Fig. 3. Synthesis of plasmid pAMβ1 lagging strand is deficient in *dnaE*$_{BS}$ mutant cells. *Bacillus subtilis* cells harboring plasmid pIL252p1 (*30*), were grown exponentially at 30°C and shifted to 37°C for 90 min. Their total DNA was prepared and hybridized with probes homologous to the plasmid. +, E, and C refer to the wild-type strain, *dnaE*$_{BS}$ thermosensitive strain EDJ51, and *polC* thermosensitive mutant *dnaF33*, respectively. Probes were double-stranded (ds) plasmid DNA or oligonucleotides complementary to the leading or lagging plasmid DNA strand. Positions of ds and single-stranded (ss) DNA are indicated.

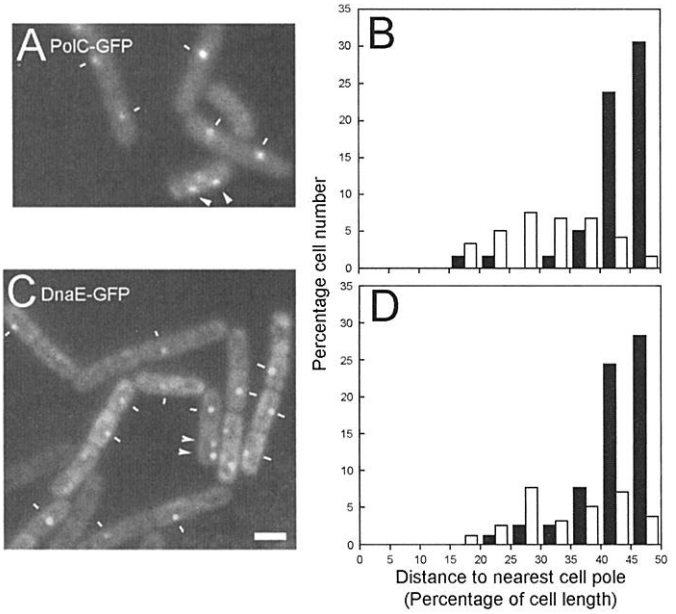

Fig. 4. Subcellular localization of PolC-GFP and DnaE-GFP in *B. subtilis*. Strains carrying *gfp* fusions to the *polC* (**A** and **B**) or *dnaE* (**C** and **D**) genes were grown in TS medium (*22*) to exponential phase, and samples were examined by microscopy (*10*). (A and C) GFP fluorescence. White lines show examples of cells with a single central focus; arrowheads point to the foci in cells with two foci. Scale bar, 2 μm. (B and D) Quantitative analysis of the positions of PolC-GFP (B) and DnaE-GFP (D) foci relative to cell length. The position of each focus was measured relative to the nearest cell pole, and the distance was expressed as a percentage of the length of the cell. Results for cells with a single focus are shown in black and results for cells with two foci are in white. Results are plotted as the percentage of total cells counted, based on measurements of (B) 65 and (D) 80 cells.

stranded DNA (*19*). We conclude that synthesis of the lagging strand of the plasmid is affected in the *dnaE*$_{BS}$ mutant. A different plasmid phenotype was observed in the *polC* mutant, in which the plasmid copy number was drastically reduced (Fig. 3). A low level of single-stranded plasmid DNA could have therefore escaped detection. However, inefficient synthesis of the leading DNA strand could reduce the plasmid copy number, without any accumulation of single-stranded DNA. A simple explanation of these observations is that the plasmid replication fork normally contains two polymerases, PolC and DnaE$_{BS}$, each of which synthesizes one DNA strand. We suggest that an incomplete fork can assemble without DnaE$_{BS}$ and progress far enough to allow formation of a full plasmid single-stranded DNA (about 5 kb). Formation of a fork containing only one catalytic polymerase subunit has been reported for *E. coli* (*20*).

Is the fork that assembles on plasmid pAMβ1, in a process that mimics the restart of replication, identical to that which normally replicates the *B. subtilis* chromosome? An argument in favor of this hypothesis is that, in *B. subtilis,* the process that leads to holoenzyme assembly during restart of the replication fork and initiation of chromosomal replication involves the same four proteins (DnaB, DnaC, DnaD, and DnaI) and only one specific protein [PriA, involved in restart, or DnaA, in initiation at the origin, (*16*, *17*, *21*)]. It has been predicted, then, that PolC and DnaE should both be present within the replication factory (*13*). We therefore examined the localization of the two proteins in the cell, using functional polymerase-GFP protein fusions (*10*). As shown in Fig. 4, A and B, cells containing a PolC-GFP fusion protein usually exhibited a single discrete focus localized at about mid-cell (lines in Fig. 4A), as reported previously (*13*). The foci were frequently offset from the central long axis of the cell, which is suggestive of a submembrane localization. In some cells (33%), generally the longer ones, there were two foci, one on either side of mid-cell (arrowheads in Fig. 4A). Under the conditions used, most cells would be expected to have a single ongoing round of DNA replication, with initiation of new rounds occurring just before cell division (*22*). The localization of a DnaE-GFP fusion was similar to that of a PolC-GFP fusion (Fig. 4, C and D). Most cells showed a single mid-cell focus (lines), with a minor proportion (30%) showing two foci (arrowheads). Again, the foci were often offset from the central long axis of the cell, and the dual foci were spaced out on either side of mid-cell. The low levels of fluorescence associated with the fusions prevented us from testing for colocalization directly. However, the striking similarity of the distributions of the two fusion proteins indicate that the PolC and DnaE proteins colocalize in the replication factory.

The *B. subtilis* replication fork thus appears to contain two different polymerases. This is similar to the situation regarding the eukaryotic replication fork, which also contains two polymerases—delta and epsilon—which have been proposed to be involved in the synthesis of different DNA strands (*23*, *24*). However, the catalytic domain of the latter polymerase is not essential, indicating that, if missing, it can be replaced by another polymerase, whereas a noncatalytic domain has an essential structural role (*25*). Further work should clarify whether an analogous situation holds true for *B. subtilis*. It was recently reported that the *E. coli* holoenzyme is asymmetric, with distinguishable leading and lagging polymerases (*26*). The need for the difference between the leading and lagging polymerase may thus be universal, the latter being more amenable to recycling than the former (*26*); and only the means to achieve the difference may vary, relying either on holoenzyme properties, as in *E. coli*, or on two different enzymes, as appears to be the case in *B. subtilis*, phylogenetically related bacteria, and eukaryotes.

Our results do not rule out additional nonessential roles of DnaE$_{BS}$ in the cell, such as mismatch correction or DNA repair [*B. subtilis* mutants strongly affected in these processes are viable (*27*)]. In particular, the *B. subtilis* genome sequence indicates that DNA polymerase II is absent, and DnaE$_{BS}$ could possibly fulfill the role this enzyme plays in *E. coli* (*28*).

References and Notes

1. A. Kornberg, T. A. Baker, *DNA Replication* (Freeman, New York, ed. 2, 1992).
2. A. T. Ganesan, T. Townsend, *Genet. Biotechnol. Bacilli* **2**, 317 (1988).
3. J. Ito, D. K. Braithwaite, *Nucleic Acids Res.* **19**, 4045 (1991).
4. F. Kunst *et al.*, *Nature* **390**, 249 (1997).
5. R. E. Yasbin, P. I. Fields, B. J. Andersen, *Gene* **12**, 155 (1980).
6. E. d'Alençon, E. LeChatelier, unpublished data.
7. I. Bruck, M. O'Donnell, *J. Biol. Chem.* **275**, 28971 (2000).
8. G. W. Bazill, J. D. Gross, *Nature New Biol.* **243**, 241 (1973).
9. V. Vagner, E. Dervyn, S. D. Ehrlich, *Microbiology* **144**, 3097 (1998).
10. See supplementary Web materials available on *Science* Online at www.sciencemag.org/cgi/content/full/294/5547/1716/DC1 for experimental details (*29*).
11. M.-A. Petit *et al.*, *Mol. Microbiol.* **29**, 261 (1998).

12. H. Yoshikawa, N. Sueoka, *Proc. Natl. Acad. Sci. U.S.A.* **49**, 559 (1963).
13. K. P. Lemon, A. D. Grossman, *Science* **282**, 1516 (1998).
14. V. Bidnenko, S. D. Ehrlich, L. Jannière, *Mol. Microbiol.* **28**, 1005 (1998).
15. C. Bruand, S. D. Ehrlich, L. Jannière, *EMBO J.* **14**, 2642 (1995)
16. C. Bruand, M. Farache, S. McGovern, S. D. Ehrlich, P. Polard, *Mol. Microbiol.* **42**, 245 (2001).
17. C. Bruand, P. Polard, unpublished data.
18. C. Bruand, S. D. Ehrlich, L. Jannière, *EMBO J.* **10**, 2171 (1991).
19. L. Boe, M. F. Gros, H. te Riele, S. D. Ehrlich, A. Gruss, *J. Bacteriol.* **171**, 3366 (1989).
20. A. E. Pritchard, H. G. Dallmann, B. P. Glover, C. S. McHenry, *EMBO J.* **19**, 6536 (2000).
21. S. Moriya, Y. Imai, A. K. Hassan, N. Ogasawara, *Plasmid* **41**, 17 (1999).
22. M. E. Sharpe, P. M. Hauser, R. G. Sharpe, J. Errington, *J. Bacteriol.* **180**, 547 (1998).
23. H. Araki, P. A. Ropp, A. L. Johnson, L. H. Johnston, A. Morrison, A. Sugino, *EMBO J.* **11**, 733 (1992).
24. R. Karthikeyan, E. J. Vonarx, A. F. Straffon, M. Simon, G. Faye, B. A. Kunz, *J. Mol. Biol.* **299**, 405 (2000).
25. R. Dua, D. L. Levy, J. Campbell, *J. Biol. Chem.* **274**, 22283 (1999).
26. B. P. Glover, C. S. McHenry, *Cell* **105**, 925 (2001).
27. R. E. Yasbin, D. Cheo, D. Bol, in B. subtilis *and Other Gram-Positive Bacteria*, A. Sonenshein, J. A. Hoch, R. Losick, Eds., (American Society for Microbiology, Washington DC, 1993), pp. 529–537.
28. M. F. Goodman, *Trends Biol. Sci.* **25**, 189 (2000).
29. P. J. Lewis, S. Thaker, J. Errington, *EMBO J.* **19**, 710 (2000).
30. E. LeChatelier, S. D. Ehrlich, L. Jannière, *Mol. Microbiol.* **20**, 1099 (1996).
31. We thank N. Givernaud for the construction of the IPTG-controlled *dnaE* mutant and S. Séror for the choice of probes to measure the *ori:ter* ratio. Supported in part by grant BIO4 CT95–0278 from the European Commission (S.D.E.) and by the UK Biotechnology and Biological Sciences Research Council (J.E.).

18 September 2001; accepted 11 October 2001

Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21

Nila Patil, Anthony J. Berno, David A. Hinds, Wade A. Barrett, Jigna M. Doshi, Coleen R. Hacker, Curtis R. Kautzer, Danny H. Lee, Claire Marjoribanks, David P. McDonough, Bich T. N. Nguyen, Michael C. Norris, John B. Sheehan, Naiping Shen, David Stern, Renee P. Stokowski, Daryl J. Thomas, Mark O. Trulson, Kanan R. Vyas, Kelly A. Frazer, Stephen P. A. Fodor, David R. Cox*

Global patterns of human DNA sequence variation (haplotypes) defined by common single nucleotide polymorphisms (SNPs) have important implications for identifying disease associations and human traits. We have used high-density oligonucleotide arrays, in combination with somatic cell genetics, to identify a large fraction of all common human chromosome 21 SNPs and to directly observe the haplotype structure defined by these SNPs. This structure reveals blocks of limited haplotype diversity in which more than 80% of a global human sample can typically be characterized by only three common haplotypes.

Human DNA sequence variation accounts for a large fraction of observed differences between individuals, including susceptibility to disease. The majority of human sequence variation is due to substitutions that occurred once in the history of mankind at individual base pairs, called single nucleotide polymorphisms (SNPs) (*1–3*). Although most of these biallelic SNPs are rare, it has been estimated that 5.3 million common SNPs, each with a frequency of 10 to 50%, account for the bulk of the DNA sequence difference between humans. Such SNPs are present in the human genome once every 600 base pairs (*4*). Alleles making up blocks of such SNPs in close physical proximity are often correlated, resulting in reduced genetic variability and defining a limited number of "SNP haplotypes," each of which reflects descent from a single, ancient ancestral chromosome (*5*). Although a block of N independent biallelic SNPs could in theory generate 2^N different haplotypes, in the absence of recurrent mutation and/or recombination the number of observed haplotypes should be no greater than (N + 1). The complexity of local haplotype structure in the human genome and the distance over which individual haplotypes extend is poorly defined. Empirical studies investigating different segments of the human genome in different populations have revealed tremendous variability in local haplotype structure. These studies indicate that the relative contributions of mutation, recombination, selection, population history, and stochastic events to haplotype structure vary in an unpredictable manner, resulting in some haplotypes that extend for only a few kilobases (kb) and others that extend for greater than 100 kb (*6–8*). These findings suggest that any comprehensive description of the haplotype structure of the human genome, defined by common SNPs, will require empirical analysis of a dense set of SNPs in many independent copies of the human genome. As a first step toward achieving this goal, we have used high-density oligonucleotide arrays, in combination with somatic cell genetics, to identify a large fraction of all common human chromosome 21 SNPs and to directly observe the haplotype structure they define.

SNPs were discovered by using a publicly available panel of 24 ethnically diverse individuals (*9*). We physically separated the two copies of chromosome 21 from each individual using a rodent-human somatic cell hybrid technique (*10*). Twenty independent copies of chromosome 21, representing African, Asian, and Caucasian chromosomes, were analyzed for SNP discovery and haplotype structure. Finished human chromosome 21 genomic DNA sequence consisting of 32,397,439 bases was masked for repetitive sequences and the resulting 21,676,868 bases (67%) of unique sequence were assayed for variation with high-density oligonucleotide arrays (*11*). In total, we synthesized 3.4×10^9 oligonucleotides on 160 wafers to scan 20 independent copies of human chromosome 21 for DNA sequence variation. Each unique chromosome 21 was amplified from a rodent-human hybrid cell line by using long range–polymerase chain reaction (LR-PCR) (*12*). We designed unique oligonucleotides to generate 3253 minimally overlapping LR-PCR products of 10-kb average length spanning 32.4 Mb of contiguous chromosome 21 DNA. LR-PCR products corresponding to the bases present on a single wafer were pooled and hybridized to the wafer as a single reaction (*13*). SNPs were detected as altered hybridization by using a pattern recognition algorithm (*14*). In total, we identified 35,989 SNPs in our sample of 20 chromosomes. The position and sequence of these human polymorphisms have been deposited in the National Center for Biotechnology Information (NCBI) dbSNP database, accession numbers ss#3995623 through ss#4020948. We used dideoxy sequencing to assess a random sample of 227 of these SNPs in the original DNA samples and confirmed 220 (97%) of the SNPs assayed. In order to achieve this low rate of 3% false-positive SNPs, we

Perlegen Sciences, Inc., 2021 Stierlin Court, Mountain View, CA 94043, USA.

*To whom correspondence should be addressed. E-mail: david_cox@perlegen.com