Ubx-homeodomain electron density, and polyalanine helices were placed into the Exd-homeodomain density. The MIRAS map was further improved by cycles of solvent flattening with program DM[24]. Using these and $F_o - F_c$ and $2F_o - F_c$ maps, interspersed with positional and individual B-factor refinement using X-PLOR[26], the model was rebuilt, side chains were added, and the Exd loops and N-terminal arms were built with the program O (ref. 27). The first four residues of Exd, the residues from −7 to 4 of Ubx and the first 6 residues of Ubx were disordered. There was clear density for the YPWM motif, but it was not readily interpretable for residues other than the tryptophan side chain. To obtain a more interpretable map, we refined and improved the MIRAS phases by solvent flattening and extended them to 2.8 Å using the program SHARP[28]. This map was greatly improved and showed us how to fit the YPWM motif. The YPWM fit was further verified by an anomalous-difference Fourier map calculated with data measured from selenomethionine (SeMet)-substituted protein; this map showed the positions of the three substituted seleniums, including the one in the YPWM motif. The refinement of the structure was extended to 2.4 Å resolution using the native 1 data, and the structure verified through extensive simulated annealing omit maps. Finally, 110 water molecules were added from the inspection of $F_o - F_c$ maps. The final refined structure has good stereochemistry, with the Ramachandran plot showing 84.5% of the residues in the core allowed regions and no residues in the disallowed or generously allowed regions.

Received 18 December 1998; accepted 3 February 1999.

1. Lewis, E. B. A gene complex controlling segmentation in *Drosophila*. *Nature* **276,** 565–570 (1978).
2. McGinnis, W. & Krumlauf, R. Homeobox genes and axial patterning. *Cell* **68,** 283–302 (1992).
3. Mann, R. S. & Chan, S.-K. Extra specificity from *extradenticle*: the partnership between HOX and exd/pbx homeodomain proteins. *Trends Genet.* **12,** 258–262 (1996).
4. Mann, R. S. The specificity of homeotic gene function. *Bioessays* **17,** 855–863 (1995).
5. Gehring, W. J., Affolter, M. & Burglin, T. Homeodomain proteins. *Annu. Rev. Biochem.* **63,** 487–526 (1994).
6. Wolberger, C. Homeodomain interactions. *Curr. Opin. Struct. Biol.* **6,** 62–68 (1996).
7. Lu, Q. & Kamps, M. Structural determinants of Pbx1 mediating cooperative DNA-binding with pentapeptide-containing HOX proteins. *Mol. Cell. Biol.* **16,** 1632–1640 (1996).
8. Chan, S.-K. & Mann, R. S. A structural model for an extradenticle-HOX-DNA complex accounts for the choice of HOX protein in the heterodimer. *Proc. Natl Acad. Sci. USA* **93,** 5223–5228 (1996).
9. Chang, C. P., Brocchieri, L., Shen, W. F., Largman, C. & Clearly, M. L. Pbx modulation of Hox homeodomain amino-terminal arms establishes different DNA-binding specificities across the Hox locus. *Mol. Cell. Biol.* **16,** 1734–1745 (1996).
10. Li, T., Stark, M. R., Johnson, A. D. & Wolberger, C. Crystal structure of the MATa1/MAT alpha 2 homeodomain heterodimer bound to DNA. *Science* **270,** 262–269 (1995).
11. Chan, S. K., Jaffe, L., Capovilla, M., Botas, J. & Mann, R. S. The DNA binding specificity of Ultrabithorax is modulated by cooperative interactions with extradenticle, another homeoprotein. *Cell* **78,** 603–615 (1994).
12. Fraenkel, E. & Pabo, C. O. Comparison of X-ray and NMR structures for the Antennapedia homeodomain–DNA complex. *Nature Struct. Biol.* **5,** 692–697 (1998).
13. Izpisua-Belmonte, J. C., Falkenstein, H., Dolle, P., Renucci, A. & Duboule, D. Murine genes related to teh *Drosophila* AbdB homeotic gene are sequentially expressed during development of the posterior part of the body. *EMBO J.* **10,** 2279–2289 (1991).
14. Johnson, F. B., Parker, E. & Krasnow, M. A. Extradenticle protein is a selective cofactor for the *Drosophila* homeotics: role of the homeodomain and YPWM amino acid motif in the interaction. *Proc. Natl Acad. Sci. USA* **92,** 739–743 (1995).
15. Chan, S.-K., Pöpperl, H., Krumlauf, R. & Mann, R. S. An extradenticle-induced conformational change in a HOX protein overcomes an inhibitory function of the conserved hexapeptide motif. *EMBO J.* **15,** 2477–2488 (1996).
16. Rauskolb, C., Smith, K., Peifer, M. & Wieschaus, E. Extradenticle determines segmental identities throughout development. *Development* **121,** 3663–3671 (1995).
17. Klemm, J. D. & Pabo, C. O. Oct-1 POU domain-DNA interactions: cooperative binding of isolated subdomains and effects of covalent linkage. *Genes Dev.* **10,** 27–36 (1996).
18. Phelan, M. L. & Featherstone, M. S. Distinct HOX N-terminal arm residues are responsible for specificity of DNA recognition by HOX monomers and HOX-PBX heterodimers. *J. Biol. Chem.* **272,** 8635–8643 (1997).
19. Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M. & Harrison, S. C. Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science* **242,** 899–907 (1988).
20. Otwinowski, Z. *et al*. Crystal structure of trp repressor/operator complex at atomic resolution. *Nature* **335,** 321–329 (1988).
21. Tan, S. & Richmond, T. J. Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. *Nature* **391,** 660–666 (1998).
22. Janin, J., Miller, S. & Chotia, C. Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.* **204,** 155–164 (1988).
23. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
24. Collaborative Computational Project, N. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
25. Hirsch, J. A. & Aggarwal, A. K. Structure of the even-skipped homeodomain complexed to AT-rich DNA: new perspectives on homeodomain specificity. *EMBO J.* **14,** 6280–6291 (1995).
26. Brunger, A. T. *XPLOR Version 3.1 Manual* (Yale Univ., New Haven, 1993).
27. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
28. de La Fortelle, E. & Bricogne, G. Maximum-likelihood heavy atom parameter refinement for the mutliple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol.* **276,** 472–494 (1997).
29. Evans, S. V. Setor: hardware lighted three-dimensional solid model representations of macromolecules. *J. Mol. Graph.* **11,** 134–138 (1993).
30. 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).

Acknowledgements. We thank the staff at CHESS for help with data collection; C. Escalante for advice on protein purification; L. Shapiro for help with map calculations; and T. Jessell and L. Shapiro for comments on this manuscript. J.M.P. thanks the Possen family for their hospitality during trips to CHESS. This work was supported by NIH grants to A.K.A. and R.S.M. R.S.M. is a Scholar of the Leukemia Society of America.

Correspondence and requests for material should be addressed to A.K.A. (e-mail: aggarwal@inka.mssm.edu). Coordinates have been deposited in the Brookhaven Protein Database (accession number 1B81).

erratum

Structural basis for activation of the titin kinase domain during myofibrillogenesis

Olga Mayans, Peter F. M. van der Ven, Matthias Wilm, Alexander Mues, Paul Young, Dieter O. Fürst, Matthias Wilmanns & Mathias Gautel

Nature **395,** 863–869 (1998)

D.O.F.'s full address should have included the Institut für Zoophysiologie und Zellbiologie (University of Potsdam, Lennéstrasse 7a, 14471 Potsdam, Germany); lane 1 of Fig. 5a referred to transfected kin4 (not in4); and in Fig. 6b, the panels referred to as "top" and "bottom" should have been left and right panels, respectively. □

correction

Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*

Richard A. Alm, Lo-See L. Ling, Donald T. Moir, Benjamin L. King, Eric D. Brown, Peter C. Doig, Douglas R. Smith, Brian Noonan, Braydon C. Guild, Boudewijn L. deJonge, Gilles Carmel, Peter J. Tummino, Anthony Caruso, Maria Uria-Nickelsen, Debra M. Mills, Cameron Ives, Rene Gibson, David Merberg, Scott D. Mills, Qin Jiang, Diane E. Taylor, Gerald F. Vovis & Trevor J. Trust

Nature **397,** 176–180 (1999)

Typographical errors in Table 1 caused the transposition of some numbers between the *H. pylori* 26695 and J99 columns. The affected rows should read as follows.

vacA genotype:	26695, *sla/ml*; J99, *slb/ml*
Functionally classified ORFs:	26695, 895; J99, 874
Conserved with no function ORFs:	26695, 290; J99, 275
H. pylori-specific ORFs:	26695, 367; J99, 346

In the table footnote, the coordinates of the second 26695 23S rRNA sequence should read 1,473,499–1,476,836. □

Structural basis for activation of the titin kinase domain during myofibrillogenesis

Olga Mayans*, Peter F. M. van der Ven†, Matthias Wilm‡, Alexander Mues‡, Paul Young‡, Dieter O. Fürst†, Matthias Wilmanns* & Mathias Gautel‡

* *European Molecular Biology Laboratory, Hamburg Outstation c/o DESY, Notkestrasse 85, D-22603 Hamburg, Germany*
† *Department for Cell Biology, University of Potsdam, Lenneéstrasse 7a, D-14471 Potsdam, Germany*
‡ *European Molecular Biology Laboratory Heidelberg, Meyerhofstrasse 1, Postfach 10 22 09, D-69012 Heidelberg, Germany*

The giant muscle protein titin (connectin) is essential in the temporal and spatial control of the assembly of the highly ordered sarcomeres (contractile units) of striated muscle. Here we present the crystal structure of titin's only catalytic domain, an autoregulated serine kinase (titin kinase). The structure shows how the active site is inhibited by a tyrosine of the kinase domain. We describe a dual mechanism of activation of titin kinase that consists of phosphorylation of this tyrosine and binding of calcium/calmodulin to the regulatory tail. The serine kinase domain of titin is the first known non-arginine–aspartate kinase to be activated by phosphorylation. The phosphorylated tyrosine is not located in the activation segment, as in other kinases, but in the P + 1 loop, indicating that this tyrosine is a binding partner of the titin kinase substrate. Titin kinase phosphorylates the muscle protein telethonin in early differentiating myocytes, indicating that this kinase may act in myofibrillogenesis.

Protein kinases are important in controlling cell proliferation and differentiation and they require specific mechanisms of regulation and substrate recognition[1–3]. Tight control of the enzymatic activity of protein kinases is achieved by phosphorylation of specific residues in the activation segment of the catalytic domain, sometimes combined with reversible conformational changes in carboxy-terminal autoregulatory tails, induced by effector molecules such as Ca^{2+}/calmodulin. In all available structures of protein kinases regulated by phosphorylation, one or more phosphorylated residues are always located in the activation segment (Fig. 1). These kinases contain a strictly conserved arginine preceding the catalytic aspartate, and are hence known as RD kinases. The arginine interacts with a phosphorylated residue from the activation segment[2]. In the non-RD kinases of the myosin-light-chain kinase (MLCK) family, this pattern is not found (Fig. 1), and activation of these kinases by phosphorylation has not been predicted.

The assembly of striated myofibrils in differentiating myocytes involves the controlled integration of hundreds of proteins into the highly ordered macromolecular complex of the sarcomere. The giant protein titin extends over one half of the sarcomeric unit and is crucial in this control[4,5]. Close to its C terminus, titin contains a MLCK-like kinase domain. Although sequence similarity of the catalytic domain of titin to that of its invertebrate analogue, the serine/threonine kinase domain of twitchin[6], has indicated that titin and twitchin may have similar function, the differential localization of the kinase domains[7,8], their distinct C-terminal regulatory tails[9,10] and the presence of several titin-specific residues in the active site[11]

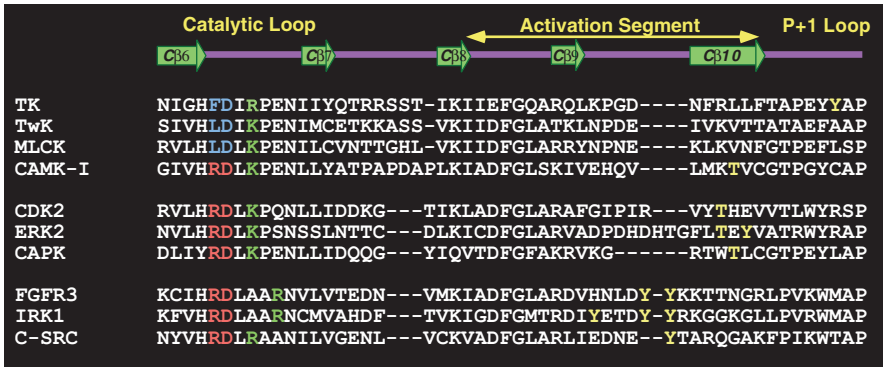

Figure 1 Sequence alignment of active-site regions. Serine/threonine kinases activated by calmodulin or by other calcium-binding proteins: human cardiac titin kinase (TK), twitchin (TwK) from *Aplysia californica*, rat skeletal myosin-light-chain kinase (MLCK), and human calcium/calmodulin-dependent protein kinase-I (CAMK-I); serine/threonine kinases activated by phosphorylation in the activation segment: human cell-division-protein kinase 2 (CDK2), human extracellular-signal-regulated kinase 2 (ERK2), and human cyclic AMP-dependent protein kinase α-catalytic subunit (cAPK); tyrosine kinases: fibroblast growth factor receptor kinase 3 (FGFR3), insulin-receptor inase (IRK3), and proto-oncogene tyrosine protein kinase c-Src (C-SRC). Yellow, identified phosphorylation sites; red, RD motif; cyan, deviations from the RD motif; green, conserved basic residues (K, R) in the +2 or +4 position relative to the RD motif. The secondary-structure elements shown at the top relate to the three-dimensional structure of titin kinase. The limits of the activation segment and the P + 1 loop are defined by the sequence motifs DFG (EFG in titin kinase) to APE[2].

hint at a different function for titin. Twitchin is a Ca^{2+}/S100A1-regulated MLCK[12]. Despite biochemical evidence for Ca^{2+}/calmodulin binding[9], however, the activation process and function of titin kinase have remained unknown.

The crystal structure of autoinhibited titin kinase shows how a tyrosine of the P + 1 loop inhibits the active site. The titin kinase structure has provided the basis for determining its dual activation mechanism, which comprises phosphorylation of Y170 and Ca^{2+}/calmodulin binding. Titin kinase is activated in differentiating myocytes, where it phosphorylates the muscle protein telethonin, indicating its probable importance in myofibrillogenesis.

Unusual autoinhibited conformation

The complete autoinhibited form of titin kinase, including the catalytic domain and the regulatory tail (kin1; Fig. 2), could be grown in thin-plate crystals <5 μm thick. We used these crystals to determine the atomic structure of titin kinase at 2.0 Å resolution from synchrotron radiation X-ray data. The overall fold shows the catalytic domain and the autoregulatory C-terminal tail, which wraps around the lower lobe and the active site of the catalytic domain (Fig. 3a, b). The amino-terminal helix of this tail (αR1) is in a similar location to the equivalent helices in twitchin[10] and calcium/calmodulin-dependent kinase-I (CaMK-I)[13]. The C terminus of this tail superimposes with that of twitchin (Fig. 3c) but not with the tail of CaMK-I. The second helix of the tail, αR2, binds into the ATP-binding site and is followed by a β-strand (βR1) which forms an antiparallel β-sheet with strands βC10 and βC11 in the lower lobe of the catalytic domain. Specific interactions to the catalytic domain are mostly restricted to the two termini, αR1 and βR1, of the regulatory tail (Fig. 3c). The C-terminal regulatory tail of titin kinase provides the calmodulin-binding site, which mapped to a segment that covers helix αR1 (ref. 9). The surface of this helix exhibits a cluster of basic and hydrophobic residues that is characteristic of Ca^{2+}/calmodulin-binding sequences[14]. The analogous segment is also involved in Ca^{2+}/calmodulin binding of other MLCK-like kinases[13,15,16].

Unlike other kinases, in which the activation segment undergoes a switch from a 'closed' to an 'open' conformation after serine and/or tyrosine phosphorylation[2], the activation segment of titin kinase is in an open conformation in the autoinhibited structure (Fig. 4). In the active site, however, the carboxylate group of the invariant catalytic base D127 is involved in a hydrogen-bonding network to R129, Q150 and Y170 (Fig. 4a, d), blocking the active site from access of its protein substrate and thus inhibiting catalysis. In turn, the hydroxyl group of Y170 is embedded in a network of hydrogen bonds to residues D127, R129 and the main-chain carbonyl group of F126. The important residue in this network is R129, a conserved lysine in all other MLCK-like kinases, which acts as a bridge between D127 and Y170. Steric inhibition of the catalytic base by a tyrosine in its vicinity is reminiscent of the inhibition of the inactive forms of insulin-receptor kinase (IRK)[17] (Fig. 4c) and the MAP kinase ERK2 (ref. 18). These kinases are activated by phosphorylation of at least one tyrosine residue in the activation segment, indicating that tyrosine phosphorylation may be involved in activation of titin kinase. In contrast to ERK2 and IRK, the important inhibitory residue of titin kinase, Y170, is located in the P + 1 loop that connects βC10 and αC4 in the lower lobe of the kinase. This region forms the pocket that accommodates the P + 1 position of the substrate in other kinase peptide-substrate structures[19,20].

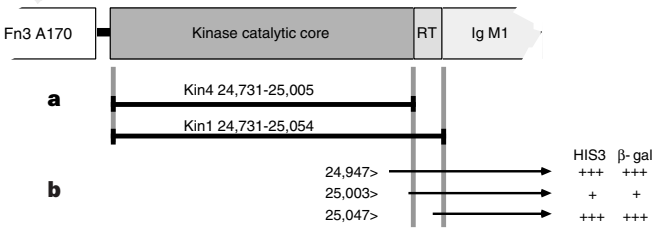

Figure 2 Titin kinase constructs used in this study. The domain pattern of the kinase region of human titin and the residues of the respective constructs are shown. Fn3, fibronectin-III-like domain; Ig, immunoglobulin like domain. **a**, Kin4 lacks the regulatory tail (RT), leaving a constitutively active catalytic core. The kin1 construct includes the regulatory tail. The N-terminal phasing is based on the N termini of DAP kinase[31] and of *Dictyostelium discoideum* MLCK[42]. **b**, Two-hybrid screens of human skeletal and cardiac complementary DNA libraries with kin4 and kin4(K36A) yield multiple overlapping clones from the regulatory tail, showing that there is a functional binding site in the active site of kin4. The strength of interaction is given by HIS3 signals and by β-galactosidase activity[30]. The first residue in the cardiac titin sequence of three representative clones is shown at the left.

Tyrosine phosphorylation of titin kinase

To determine whether Y170 is a site of phosphorylation in titin kinase, we made a truncated titin kinase construct (kin4) lacking the regulatory C-terminal tail. Kin4 is predicted to be constitutively active, like C-terminal-truncated twitchin[15]. In a yeast two-hybrid analysis, kin4 interacts specifically with overlapping peptides from the regulatory region of titin kinase (Fig. 2b), thus showing the structural integrity of this construct. We introduced a mutation of Y170 (Y170E) into kin4 to abolish the inhibitory tyrosine and, at the same time, to mimic the phosphorylated state[18]. The K36A mutation is predicted to disrupt catalytic activity, by analogy to K72A in cAMP-dependent protein kinase[21]. When expressed in C2C12 myocytes, phosphotyrosine was detected in the wild-type kin4 and in the K36A mutant. No phosphotyrosine signal could be detected for the Y170E mutant (Fig. 5a). Similarly, the full-length kinase (kin1) was tyrosine-phosphorylated by extracts from differentiating myocytes, but not by extracts from adult muscle. Again, tyrosine phosphorylation in the mutant kin1(Y170E) was markedly reduced (Fig. 5b). These results indicate that titin kinase is trans-phosphorylated by kinase activities in differentiating muscle at an unusual site in the P + 1 loop, where Y170 is the major phosphorylation site.

Y170 is buried in the autoinhibited, unphosphorylated structure of titin kinase (Fig. 4) like in ERK2 (ref. 18) in which the residues phosphorylated upon activation are equally inaccessible. To make the peptide-binding site accessible, major structural rearrangements are predicted to take place during, or following, tyrosine phosphorylation in both ERK2 (ref. 18) and titin kinase. We immunoprecipitated the native, *in vitro* phosphorylated titin kinase with anti-phosphotyrosine antibodies (Fig. 5c). This indicates that the P + 1 loop may undergo major structural rearrangements involving the exposure to solvent of phosphorylated Y170.

Activated titin kinase phosphorylates telethonin

To study kinase activity of titin kinase, we expressed the truncated titin kinase construct, kin4, in C2C12 myocytes. When we used kin4 in phosphorylation assays with different substrates, we found no myosin light-chain-kinase activity. We detected specifically increased phosphorylation of a protein of relative molecular mass 22,500 (M_r 22K) in day 2 myocyte extracts in the presence of wild-type kin4, but not in the presence of the catalytically inactive kin4 mutant K36A (Fig. 6a). We used mass-spectrometrical microsequencing to determine that this band was telethonin, a protein of cardiac and skeletal muscle[22]. By using phosphorylation assays with recombinant subfragments of telethonin and subsequent sequencing by tandem mass spectrometry, we identified a single phosphorylated serine in the recognition sequence ^{153}RRSLS(phospho)RSMSQEAQRG^{167}, close to the C terminus of telethonin. The identification of telethonin as a sarcomeric Z-disk protein[23] indicates that its phosphorylation may be involved in the control of myofibrillogenesis. This is supported by the observation that C2C12

cells expressing the constitutively active kin4 show a breakdown of normal cytoskeletal architecture (Fig. 6b). In the differentiated myofibril, titin kinase is located ~1 μm from the Z-disk, at the edge of the M-band[8]. However, both the titin C terminus and telethonin can be detected by immunofluorescence in dot-like aggregates on stress-fibre-like structures in day 2 myocytes (Fig. 6c). These data indicate that the activation of titin kinase in differentiating myocytes and the resulting phosphorylation of telethonin are involved in the reorganization of the cytoskeleton during myofibrillogenesis. During these events, the titin C terminus is transiently in close proximity to the titin kinase substrate telethonin.

Activation of titin kinase

To further characterize the mechanism of activation of titin kinase, we used purified, autoregulated kin1 (Fig. 1)[9]. Kin1 showed only weak basal kinase activity towards the substrate telethonin (residues 104–167) (Fig. 7a). Addition of the muscle calcium-binding proteins calmodulin, $S100A1_2$, CACY or CAPL[24] did not increase kinase activity significantly (Fig. 7a). In contrast, the basal activity

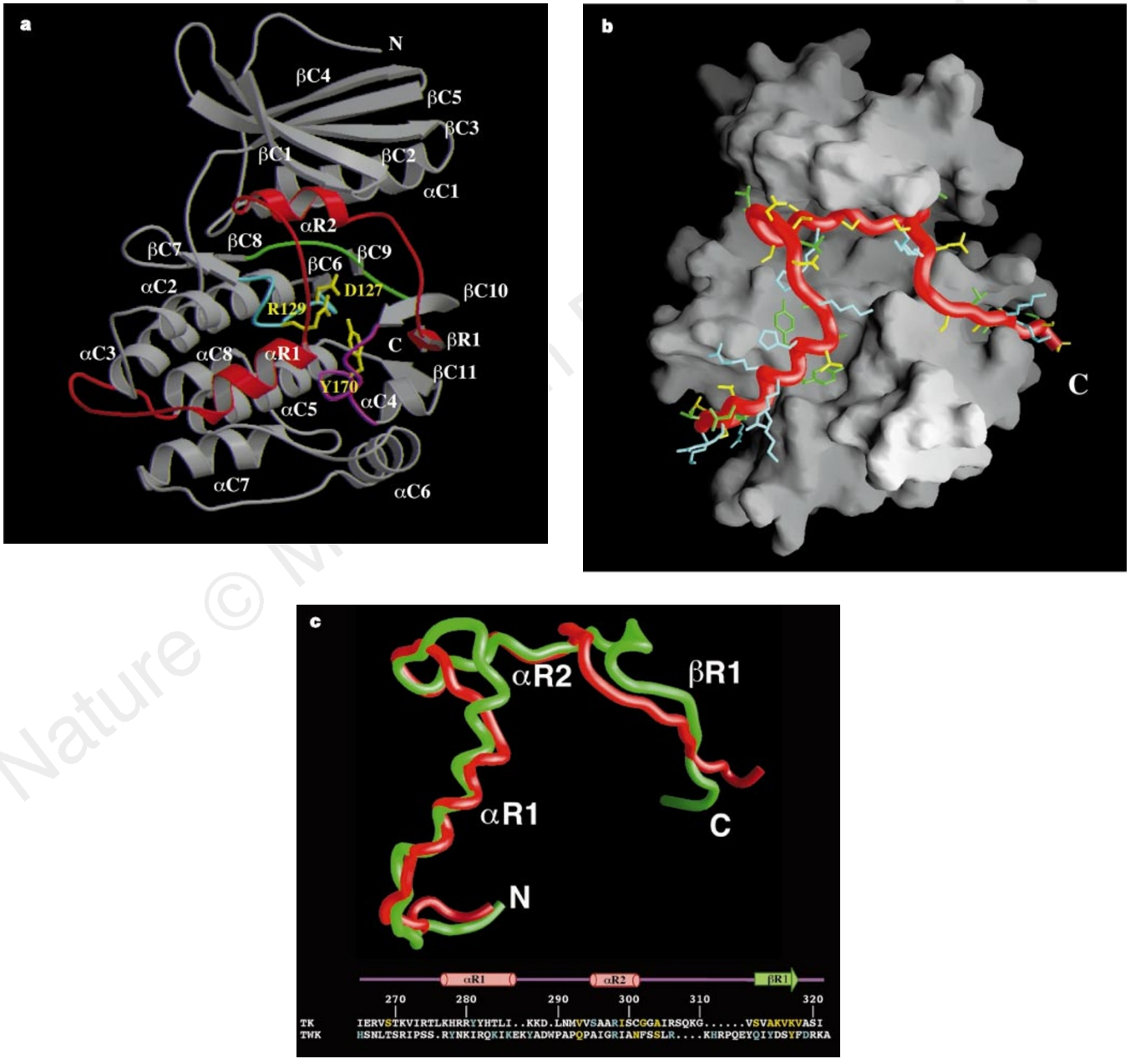

Figure 3 The three-dimensional structure of the autoinhibited form of titin kinase. **a**, Ribbon diagram. Red, regulatory tail; cyan, catalytic loop; green, activation segment; magenta, P + 1 loop. The side chains of residues D127, R129 and Y170 are highlighted in yellow. Y170 is phosphorylated upon activation. The secondary-structural elements are numbered in sequential order. They are labelled with 'C' and 'R', according to their location in the catalytic domain or regulatory tail, respectively. **b**, Surface presentation of the catalytic domain; the regulatory tail is shown as a tube. The side chains of the residues of the regulatory tail are coloured in cyan (basic residues), green (hydrophobic residues) and yellow (other residues). The central part of the regulatory tail blocks the ATP-binding site. **c**, Top, superposition of the regulatory tails of titin kinase (red) and twitchin[10] (green); bottom, a structure-based sequence alignment using the program ALIGN_FRG[43]. Residues that are involved in polar interactions to the catalytic domain through their main chains and side chains are coloured in yellow and cyan, respectively. Secondary-structure elements of the regulatory tail of the titin-kinase structure are shown schematically. The residue numbers correspond to the numbering convention used to describe the titin-kinase structure. **a** was prepared with MOLSCRIPT[44] and **b**, **c** with GRASP[45].

of the phosphate-mimickry mutant kin1(Y170E) towards telethonin was more than tenfold higher than that of wild-type and was stimulated by up to 100-fold by Ca^{2+}/calmodulin (Fig. 7b). None of the S100 proteins (S100A1, CACY or CAPL) activated the mutant titin kinase (Fig. 7b). These results show that two different autoinhibitory mechanism must be overcome to fully activate titin kinase: the P + 1 loop must be released from the substrate-binding site by tyrosine phosphorylation, mimicked in kin1(Y170E), and the conformation of the C-terminal tail must change as a result of binding of Ca^{2+}/calmodulin. This novel dual autoregulatory mechanism is likely to provide tight control for titin kinase activity during muscle differentiation.

Substrate recognition by titin kinase

In contrast to the RD kinases[2], the active site of titin kinase does not contain a pocket composed of basic residues that could accommodate a phosphorylated tyrosine. Replacement of the phosphorylated tyrosine of the kinase with substrate, which occurs in IRK[17,25], is unlikely because of the catalytic specificity of titin kinase for serine. None of the other MLCK-like kinases contains a tyrosine in the P + 1 loop (Fig. 1), excluding the possibility of a related activation mechanism. The most significant difference in the phosphorylation site on telethonin, as compared with the sites of phosphorylation of substrates for MLCK and twitchin[26] is the presence of an arginine in the P + 1 position. A potential function of the phosphorylated Y170 could, therefore, be to directly bind this arginine, explaining the unusual presence of arginine in the P + 1 position of the titin kinase substrate and the location of the titin kinase phosphorylation site, Y170, in the P + 1 loop. Furthermore, Q150 in the activation segment of titin kinase is replaced by a hydrophobic residue in other MLCK-like kinases (Fig. 1). This glutamine is located in the pocket that accommodates the P + 1 position of the substrate in known kinase peptide-substrate structures[27]. Therefore, Q150 might be another ligand for the P + 1 arginine of telethonin. The only other known protein kinase that shows the same pattern of a preferred arginine in the substrate P + 1 position and a tyrosine in the kinase P + 1 loop is NIMA[19]. The significance of this observation is unknown.

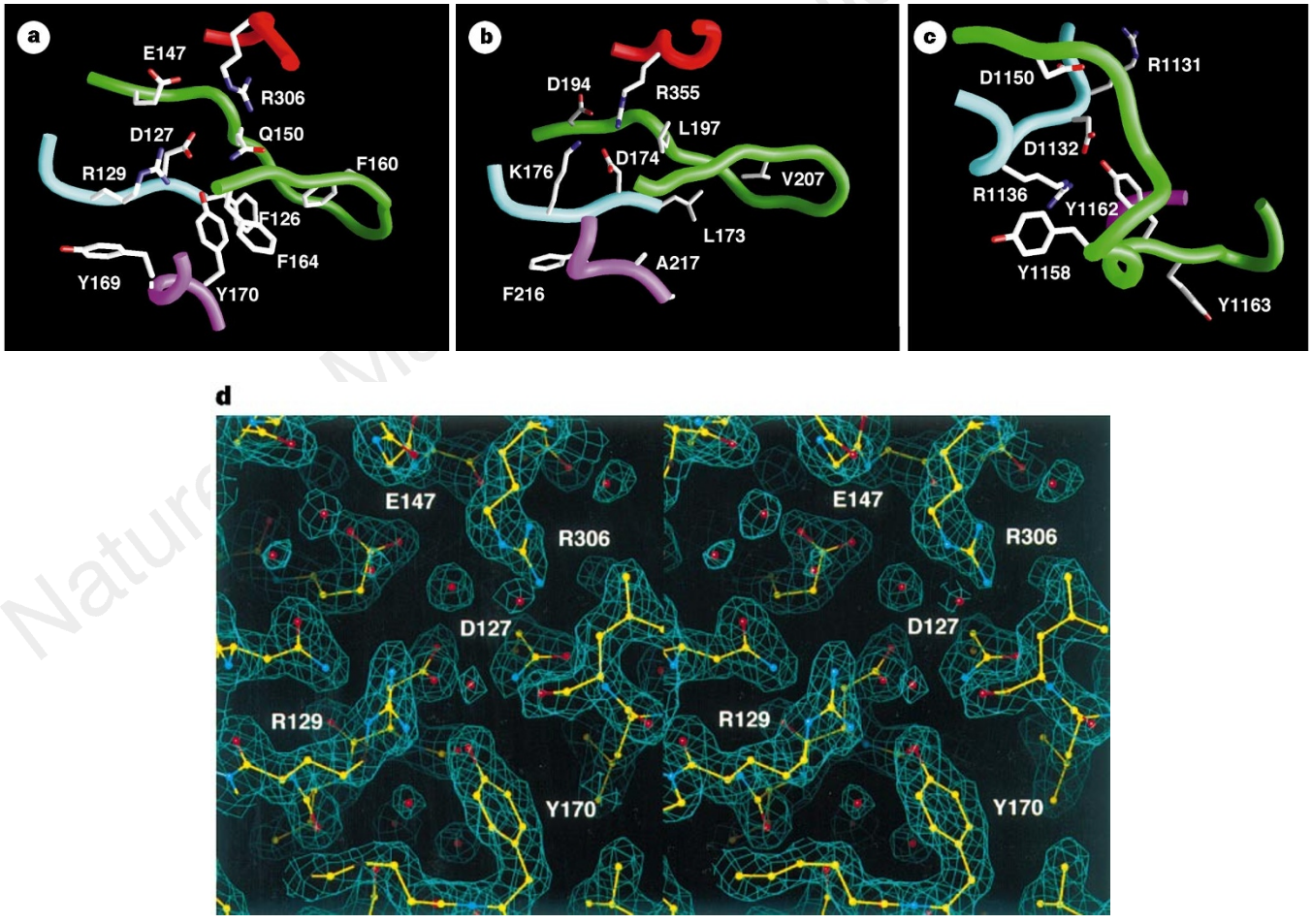

Figure 4 Active-site conformation of the autoinhibited forms of titin kinase, twitchin and IRK. **a**, The active site of titin kinase. The guanidinium group of R129 forms short hydrogen bonds with the side chains of D127 and Y170 from the P + 1 loop. There is a weak direct hydrogen bond between D127 and Y170 (3.1 Å in length). D127 forms further hydrogen bonds with Q150. **b**, Twitchin active site[10]. The catalytic aspartate, D174 forms hydrogen bonds with K176, Q200 and R355 from the regulatory tail. At the position of Y170 in titin kinase, there is an alanine in twitchin. In the autoinhibited twitchin structure[10], the catalytic aspartate is blocked by a salt bridge with an arginine (R355 in twitchin) of the regulatory tail, suggesting a different activation mechanism than for titin kinase. In titin kinase, the equivalent arginine, R306, does not interact with the catalytic aspartate. **c**, Active site of the autoinhibited form of IRK[17]. The catalytic aspartate, D1132, is bound to Y1162. This bond is disrupted after phosphorylation of Y1162, accompanied by phosphorylation of two other tyrosines and induces a conformational change of the activation segment from a closed to an open conformation[25]. The colour codes of the tubes are as in Fig. 3a. **d**, Stereo view of a $2F_o - F_c$ electron-density map, using phases of the final model, contoured at 1.3 σ. The electron density shown covers several active-site residues and solvent molecules. Some titin-kinase residues are labelled. **a**–**c** were prepared with GRASP[45] and **d** with program O (ref. 46).

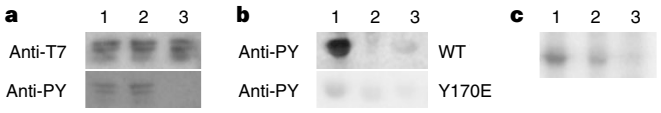

Figure 5 Titin kinase is tyrosine-phosphorylated on Y170 in differentiating C2C12 cells. **a**, Transfected in4 (lane 1), kin4(K36A) (lane 2) and kin4(Y170E) (lane 3) were immunoprecipitated with a rabbit antibody against titin kinase (α-TK-ra; ref. 9) and phosphotyrosine was detected on blots of the immunoprecipitates with the monoclonal antibody 4G10 (anti-PY). The immunoprecipitated kinase was detected with the anti-T7 tag antibody (anti-T7). The absence of a phosphotyrosine signal in kin4(Y170E) indicates that Y170 is the major tyrosine phosphorylation site. The phosphorylation of the catalytically inactive kin4(K36A) indicates that titin kinase is transphosphorylated by upstream kinase activities rather than intramolecularly autophosphorylated. **b**, Similarly, the recombinant full-length kinase kin1 (WT) is markedly tyrosine-phosphorylated *in vitro* by cytosolic extracts from differentiating day 2 C2C12 cells (lane 1) but only weakly from adult psoas muscle extracts (lane 3) as visualized in blots with 4G10. The phosphotyrosine signal of the kin1(Y170E) mutant is markedly reduced (lane 1, bottom), indicating that Y170 is the major phosphorylation site in the autoinhibited form of titin kinase also. Lanes 2 and 3, bottom: control without cell extracts; 1 μg of enzyme was loaded per lane. **c**, the phosphorylated tyrosine in titin kinase is accessible to anti-phosphotyrosine antibodies in the native state of the enzyme, indicating that it is exposed to solvent and that major structural rearrangements of the P + 1 loop occur in the activated kinase. Kin1 was phosphorylated as in **a** in the presence of [γ-^{32}P]ATP (lane 1) and immunoprecipitated with the 4G10 antibody. The autoradiograph shows the presence of phosphorylated titin kinase in the immunoprecipitated fraction (lane 2) and its near absence in the unbound supernatant (lane 3).

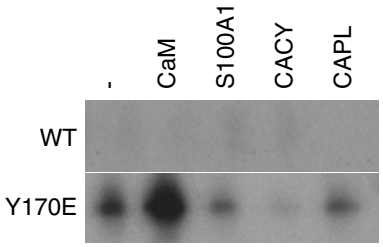

Figure 7 Full activation of titin kinase requires both tyrosine phosphorylation and Ca^{2+}/calmodulin. Wild-type (WT) kin1 shows very low basal activity towards telethonin(104–167) (lane 1); this activity is not stimulated by Ca^{2+}/calmodulin (CaM), Ca^{2+}/S100A1 (S100), Ca^{2+}/CACY (CACY), or Ca^{2+}/CAPL (CAPL). In these assays, we used 0.1 μg purified enzyme. Autoradiographs were exposed for 12 h. The phosphate-mimicry mutant kin1(Y170E) shows elevated basal activity (lane 1) which is stimulated markedly by Ca^{2+}/calmodulin. The addition of Ca^{2+}/S100A1, Ca^{2+}/CACY or Ca^{2+}/CAPL has no activating effect; the presence of Ca^{2+}/CACY suppresses basal titin-kinase activity by about tenfold, indicating that S100 proteins indeed modulate the activity of titin kinase. The addition of 10 μM Zn^{2+} in the assays using S100 proteins did not increase kinase stimulation but inhibited kin1(Y170E) as it inhibited *C. elegans* twitchin[26].

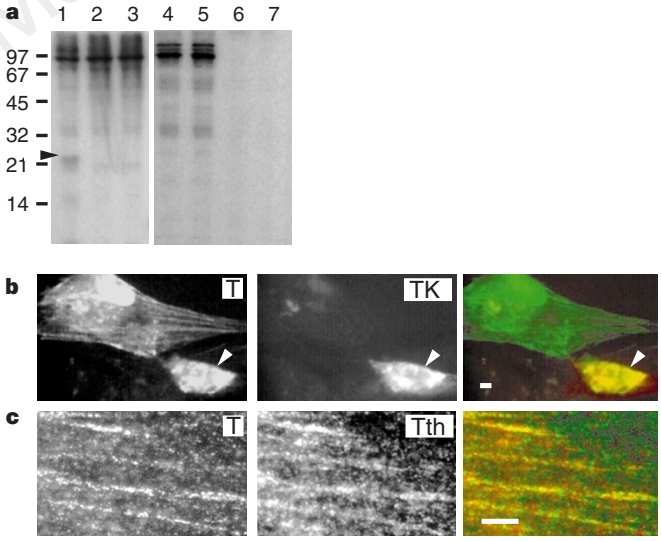

Figure 6 Titin kinase phosphorylates the muscle protein telethonin. **a**, Phosphorylation of muscle proteins by constitutively active kin4 was done using tyrosine-phosphorylated enzyme from transfected C2C12 cells. To suppress background calcium-activated kinases in the myocyte extracts, EGTA was used. In extracts of day 2 myocytes, increased phosphorylation of a protein of M_r~22K occurs in the presence of kin4 (lane 1). In the presence of kin4(K36A) (lane 2) or in blank assays (lane 3), phosphorylation of this band is not increased. Kin4 shows no discernible activity towards proteins in adult psoas cytosol (lane 4) or myofibrils (lane 6) compared with controls (lanes 5 and 7), indicating that the activity of titin kinase is directed towards proteins in differentiating myocytes. The constitutively active enzyme has no myosin-light-chain kinase activity (lane 6). **b**, the presence of constitutively active kin4 in myogenic cells disrupts myofibrillogenesis. Non-transfected cells at day 2 show the typical alignment of titin (top panel, visualized by anti-Z1Z2-ra staining[30]) along stress-fibre-like structures. In the transfected cells (arrowhead in all three panels), the alignment of titin along the actin cytoskeleton is disrupted and titin (green) and kin4 (red, visualized with the anti-T7-tag antibody in the bottom panel) are randomly distributed. **c**, The C terminus of titin loops back onto the stress-fibre-associated N-terminal portion, and localizes with telethonin. In day 2 myocytes, C-terminal epitopes of titin (visualized with the anti-titin antibody T30 (ref. 47) localize with telethonin on stress-fibre-like structures in dot-like aggregates. T, Titin; Tth, telethonin; TK, transfected titin kinase. Scale bars: 10 μm.

Discussion

The titin kinase structure exhibits dual inhibition of the active site: the catalytic aspartate is blocked by Y170 from the P + 1 loop, and the ATP-binding site is blocked by helix αR2 from the regulatory tail. Inhibition is removed by a new, dual-activation process, involving phosphorylation of Y170 by an unknown kinase and potential co-factors and the binding of Ca^{2+}/calmodulin. Titin kinase is, to our knowledge, the only kinase known to be activated by phosphorylation of a tyrosine from the P + 1 loop and to lack the RD motif. We propose that, on activation of titin kinase, the central part of the regulatory tail is expelled from the active site whereas the flanking interactions of αR1 and βR1 remain in contact with the catalytic domain. It remains to be determined whether activation by phosphorylation in the P + 1 loop will effect substrate specificity as a general mechanism for kinase regulation, as in protein kinase C[28]. Although all members of the MLCK family bind Ca^{2+}/calmodulin, other calcium-binding proteins are involved in activation; for example, twitchin is activated by the dimeric S100A1 protein[12,16]. It will be important to determine the specificity and complementarity of different calcium-binding proteins in the activation of these protein kinases, which appears to be a multistep process in most cases. □

Methods

Plasmids and two-hybrid screens. The titin kinase fragments kin1 (EMBL accession number X90568; amino acids (aa) 24,731–25,054) and kin4 (aa 24,731–25,005) and their mutants were cloned into a modified pCMV5 vector[29] with an N-terminal T7-tag (Invitrogen) sequence (MTGGQQMGR). The N-terminal phasing was based on MLCK-like kinases without extra N-terminal domains, for reasons of expression stability. For two-hybrid screens, we inserted kin4 and its mutants into a modified pLexA plasmid[30]. Two-hybrid screens and analysis were done as described[30]. All cloning and mutagenesis steps followed standard procedures.

Cell culture. We cultured C2C12 cells in DMEM, 20% fetal calf serum and 4.5% glucose at 37 °C. Differentiation was induced by moving cells to low-serum medium (DMEM, 4% horse serum). Transfection with lipofectamine (Gibco-BRL, UK) followed standard procedures. For immunoprecipitation and kinase assays, day 1 cells from 10-cm dishes were lysed in 200 μl 20 mM HEPES, pH 7.2, 5 mM $MgCl_2$, 1 mM EGTA, 1 mM dithiothreitol (DTT), 1 mM Na_3VO_4, 0.1% Triton X-100 and a cocktail of protease inhibitors (Boehringer).

Antibodies and immunochemistry. An affinity-purified polyclonal rabbit antibody was used against telethonin residues 104–167 (ref. 23). Western blots were performed by standard procedures using the enhanced chemoluminescence (ECL) kit (Amersham). Anti-phosphotyrosine blots were done with the monoclonal antibodies 4G10 (Upstate Biotechnology). Immunoprecipitation was carried out on protein-A beads using the anti-T7-tag (Novagen) or 4G10 monoclonal antibodies, or the anti-titin-kinase polyclonal antibody[9] (see below) using standard protocols.

Phosphorylation assays. Kinase assays from immunoprecipitates were done as described[31] in assay buffer (20 mM HEPES, pH 7.2, 5 mM $MgCl_2$, 0.5 mM $CaCl_2$, 1 mM DTT, 0.2 mM ATP and 1 μCi [γ-^{32}P]ATP, 3,000 Ci mM^{-1}). Assays with purified kinases and telethonin were carried out with 1 ng to 0.4 μg enzyme in 20 μl kinase assay buffer, 20 μg ml^{-1} calmodulin or S100 proteins, and 2 μg substrate protein. Assays were stopped after 20 min at 30 °C by addition of SDS sample buffer[32] and analysed on 15–18% SDS–polyacrylamide gels[32]. Tyrosine kinase assays were done with 100 μg ml^{-1} kin1 and 1 μl cell lysate supernatant in 20 mM MOPS, pH 7, 5 mM $MnCl_2$, 1 mM Na_3VO_4, 0.1 mM ATP and 2 μCi [γ-^{32}P]ATP, 3,000 Ci mM^{-1}, as described[33]. Dried gels were autoradiographed for 2–12 h. Quantification of calmodulin- and S100-induced activation and basal activity was done as described[26]. Phosphoproteins were enriched for microsequencing by chelating chromatography[27] and by SDS–PAGE.

Mass spectrometry. The phosphorylated protein was purified on a self-assembled 100-nl Poros[a] R2 column (Perceptive Biosystems) and analysed on a triple-quadrupole mass spectrometer (API III, PE-Sciex)[34]. The protein was digested in 0.1 M NH_4HCO_3 by adding 0.1 μg trypsin. Half of the resulting peptide mixture was purified on a 100-nl Poros[a] R3 column into a nano electrospray capillary. The sample was analysed on the triple-quadrupole mass spectrometer in negative-ion mode. A phosphorylated peptide was detected and its mass measured using a precursor ion scan of the PO_{3^-} ion (0.079K)[33]. The second half of the peptide mixture was desalted as described[34], but eluted in 60% methanol, 35% water and 5% formic acid and analysed in positive-ion mode. The doubly charged peptide ion $(M + 2H)^{2+}$ was selected for fragmentation in the mass spectrometer. Its sequence and the phosphorylation site could be deduced from the fragment spectrum.

Expression and purification. Human kin1 or the mutant kin1(Y170E), which includes the catalytic domain and the C-terminal regulatory extension (EMBL accession number X90568; aa 24,731–25,054; for simplicity, the residue numbering has been changed by assigning the first residue of the construct as '1'), was purified to homogeneity from sf9 cells infected with recombinant baculoviruses. Purified titin kinase was dialysed into 20 mM Tris-HCl, pH 7.4, 50 mM NaCl, 1 mM EDTA, 1 mM DTT and 1 mM NaN_3 and concentrated in a Centricon-10 to about 4–7 mg ml^{-1} for crystallization. Telethonin fragments were cloned from a primary pGAD10 clone into an N-terminally His_6-tagged pET3d vector. Expression of His_6-fusion proteins followed standard procedures. Recombinant calmodulin, $S100A1_2$, CACY (calcyclin) and CAPL were purified from *Escherichia coli* on phenyl sepharose and anion-exchange chromatography as described[24,35]. For some assays, $S100A1_2$ from bovine brain was used (Sigma) with no differences to the recombinant protein. Purity of proteins was assessed by SDS–PAGE and mass spectrometry.

Crystallization and X-ray structure determination. Crystals were obtained from 1.1 M sodium/potassium tartrate, 2.5% (*v/v*) ethanol, 25 mM sodium acetate, pH 4.9, and 25 mM imidazole, pH 7.5, using the vapour-diffusion method. The crystals grew as thin plates with dimensions of 5 μm × 10 μm × 300 μm. The thickness of these crystals was slightly improved by macroseeding and the use of oil layers over the reservoir as described[36]. Crystals were shock-frozen at 100 K using 12% glycerol as cryo-protectant. A native X-ray data set was collected up to 2.0 Å resolution on the wiggler beamline BW7B (EMBL Hamburg Outstation) using a wavelength of 0.8373 Å. The data were recorded in a high-resolution sweep (132 frames, frame width 0.8 degrees, crystal-to-detector distance 180 mm) and a low-resolution sweep (83 frames, frame width 1.2 degrees, detector distance 340 mm) on a 345-mm Mar imaging plate scanner. We used the DENZO and SCALEPACK software[37] for processing and reduction of data, which consisted of 53,151 unique reflections with a redundancy of 3.7, a completeness of 96% and an R_{sym}(I) of 7.7% overall (21,052 unique reflections in the highest resolution shell, 2.39–2.0 Å, with a multiplicity of 2.6, a completeness of 94.5% and an R_{sym}(I) of 23.9%). The crystals belong to the $P2_12_12_1$ space group with cell dimensions of $a = 78.6$ Å, $b = 89.9$ Å, $c = 113.3$ Å and two molecules per asymmetric unit. The non-crystallographic two-fold axis is parallel to the crystallographic b axis at $x = 0.5$ and $z = 0.09$. Initial phases were obtained from a trimmed model of the catalytic domain of twitchin[10] using the molecular-replacement software package AMoRe[38]. A Sigma-A[39] weighted, $2F_o - F_c$ electron-density map calculated with the initial molecular-replacement phases was used for NCS averaging using the AVE software, with mask creation and manipulation done in MAMA[40]. We used the molecular dynamics slow-cooling protocol of X-PLOR[41] for structure refinement. We applied bulk solvent correction, overall anisotropic B-factor scaling, restrained NCS and restrained individual B-factor refinement. The final model includes 5,206 protein non-hydrogen atoms and 514 solvent atoms for the two copies in the asymmetric unit. The overall R-factor is 20.7% and R_{free} is 24.8% for all observed data between 40 and 2.0 Å resolution (in the outer-resolution shell, 2.28–2.0 Å, these values are 22.4% and 27.4%, respectively). The test data set corresponds to 5% of the total data and was generated with FREERFLAG[39].

Received 27 May; accepted 14 September 1998.

1. Taylor, S. S., Radzio-Andelzelm, E. & Hunter, T. How do protein kinases discriminate between serine/threonine and tyrosine? Structural insights from the insulin receptor tyrosine kinase. *FASEB J.* **9,** 1255–1266 (1995).
2. Johnson, L. N., Noble, M. E. M. & Owen, D. J. Active and inactive protein kinases: structural basis for regulation. *Cell* **85,** 149–158 (1996).
3. Johnson, L. N., Lowe, E. D., Noble, M. E. M. & Owen, D. J. The structural basis for substrate recognition and control by protein kinases. *FEBS Lett.* **430,** 1–11 (1998).
4. Trinick, J. Titin as a scaffold and spring. *Curr. Biol.* **6,** 258–260 (1996).
5. Maruyama, K. Connectin/titin, giant elastic protein of muscle. *FASEB J.* **11,** 341–345 (1997).
6. Heierhorst, J. *et al.* Autophosphorylation of molluscan twitchin and interaction of its kinase domain with calcium/calmodulin. *J. Biol. Chem.* **269,** 21086–21093 (1994).

7. Vibert, P., Edelstein, S. M., Castellani, L. & Elliot, B. W. Mini-titins in striated and smooth molluscan muscles: structure, location and immunological crossreactivity. *J. Muscle Res. Cell Motil.* **14,** 598–607 (1993).
8. Obermann, W. M. J. *et al.* The structure of the sarcomeric M band: localization of defined domains of myomesin, M.-protein and the 250 kD carboxy-terminal region of titin by immunoelectron microscopy. *J. Cell Biol.* **134,** 1441–1453 (1996).
9. Gautel, M. *et al.* A calmodulin-binding sequence in the C-terminus of human cardiac titin kinase. *Eur. J. Biochem.* **230,** 752–759 (1995).
10. Kobe, B. *et al.* Giant protein kinases: domain interactions and structural basis of autoregulation. *EMBO J.* **15,** 6810–6821 (1996).
11. Sebestyén, M. G., Fritz, J. D., Wolff, J. A. & Greaser, M. L. Primary structure of the kinase domain region of rabbit skeletal and cardiac titin. *J. Muscle Res. Cell Motil.* **17,** 343–348 (1996).
12. Heierhorst, J. *et al.* Ca^{2+}/S100 regulation of giant protein kinases. *Nature* **380,** 636–639 (1996).
13. Goldberg, J., Nairn, A. C. & Kuriyan, J. Structural basis for the auto-inhibition of calcium/calmodulin-dependent protein kinase I. *Cell* **84,** 875–887 (1996).
14. Crivici, A. & Ikura, M. Modular and structural basis of target recognition by calmodulin. *Annu. Rev. Biomol. Struct.* **24,** 85–116 (1995).
15. Heierhorst, J. *et al.* Phosphorylation of myosin regulatory light chains by the molluscan twitchin kinase. *Eur. J. Biochem.* **233,** 426–431 (1995).
16. Chin, D., Winkler, K. E. & Means, A. R. Characterisation of substrate phosphorylation and use of calmodulin mutants to address implications from the enzyme crystal structure of calmodulin-dependent protein kinase I. *J. Biol. Chem.* **272,** 31235–31240 (1997).
17. Hubbard, S. R. Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. *EMBO J.* **16,** 5572–5581 (1997).
18. Zhang, F. *et al.* Atomic structure of the MAP kinase ERK2 at 2.3 Å resolution. *Nature* **367,** 704–711 (1994).
19. Songyang, Z. *et al.* A structural basis for substrate specificities of protein Ser/Thr kinases: primary sequence preference of casein kinase I and II, phosphorylatse kinase, calmodulin-dependent kinase II, CDK5 and Erk1. *Mol. Cell. Biol.* **16,** 6486–6493 (1996).
20. Lowe, E. D. *et al.* The crystal structure of a phosphorylase inase peptide substrate complex: kinase substrate recognition. *EMBO J.* **16,** 6646–6658 (1997).
21. Taylor, S. S. *et al.* A template for the protein kinase family. *Trends Biochem. Sci.* **18,** 84–89 (1993).
22. Valle, G. *et al.* Telethonin, a novel sarcomeric protein of heart and skeletal muscle. *FEBS Lett.* **415,** 163–168 (1997).
23. Mues, A. *et al.* Two immunoglobulin-like domains of the Z-disk portion of titin interact in a conformation-dependent way with telethonin. *FEBS Lett.* **428,** 111–114 (1998).
24. Engelkamp, D., Schäfer, B. W., Erne, P. & Heizmann, C. W. S100 alpha, CAPL, and CACY: molecular cloning and expression analysis of three calcium-binding proteins from human heart. *Biochemistry* **31,** 10258–10264 (1992).
25. Hubbard, S. R., Wei, L., Ellis, L. & Hendrickson, W. A. Crystal structure of the tyrosine kinase domain of the human insulin receptor. *Nature* **372,** 746–754 (1994).
26. Heierhorst, J. *et al.* Substrate specificity and inhibitor sensitivity of Ca^{2+}/S100-dependent protein kinases. *Eur. J. Biochem.* **242,** 454–459 (1996).
27. Songyang, Z. *et al.* Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Curr. Biol.* **4,** 973–982 (1994).
28. Konishi, H. *et al.* Activation of protein kinase C by tyrosine phosphorylation in response to H_2O_2. *Proc. Natl Acad. Sci. USA* **94,** 11233–11237 (1997).
29. Andersson, S. *et al.* Cloning, structure, and expression of the mitochondrial cytochrome P-450 sterol 26-hydroxylase, a bile acid biosynthetic enzyme. *J. Biol. Chem.* **264,** 8222–8229 (1989).
30. Young, P., Ferguson, C., Bañuelos, S. & Gautel, M. Molecular structure of the sarcomeric Z-disk: two types of titin interactions lead to an asymmetrical sorting of α-actinin. *EMBO J.* **17,** 1614–1624 (1998).
31. Cohen, O., Feinstein, E. & Kimchi, A. DAP-kinase is a Ca2+/calmodulin-dependent, cytoskeletal-associated protein kinase, with cell-death inducing functions that depend on its catalytic activity. *EMBO J.* **16,** 998–1008 (1997).
32. Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227,** 680–685 (1970).
33. Weijland, A. *et al.* The purification and characterization of the catalytic domain of Src expressed in *Schizosaccharomyces pombe*. *Eur. J. Biochem.* **240,** 756–764 (1996).
34. Wilm, M. & Mann, M. Analytical properties of the nanoelectrospray ion source. *Anal. Chem.* **68,** 1–8 (1996).
35. Dedman, J. R. & Kaetzel, M. A. Calmodulin purification and fluorescence labeling. *Methods Enzymol.* **102,** 1–8 (1983).
36. Chayen, N. E. A novel technique to control the rate of vapour diffusion, giving larger protein crystals. *J. Appl. Crystallogr.* **30,** 198–202 (1997).
37. Otkinowski, Z. in *Proceedings of the CCP4 Study Weekend* (eds Sawyer, L., Isaacs, N. & Bailey, S.) 56–62 (SERC Daresbury Lab., 1993).
38. Navaza, J. AMoRe: an automated package for molecular replacement. *Acta Crystallogr. A* **50,** 577–587 (1994).
39. Collaborative Computaitonal Project 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–767 (1994).
40. Kleywegt, G. T. & Jones, T. A. in *Proceedings of the CCP4 Study Weekend* (eds Sawyer, L., Isaacs, N. & Bailey, S.) 56–62 (SERC, Daresbury Lab., 1994).
41. Brünger, A. T., Kuriyan, J. & Karplus, M. Crystallographic R factor refinement by molecular dynamics. *Science* **235,** 458–466 (1987).
42. Tan, J. L. & Spudich, J. Characterization and bacterial expression of the *Dictyostelium* myosin light chain kinase cDNA. *J. Biol. Chem.* **266,** 16044–16049 (1991).
43. Cohen, G. E. A program to superimpose protein coordinates, accounting for insertions and deletions. *J. Appl. Crystallogr.* **30,** 1160–1161 (1997).
44. Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structure. *J. Appl. Crystallogr.* **24,** 251–259 (1991).
45. Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insight from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11,** 281–296 (1991).
46. Jones, T. A. Diffraction methods for biological macromolecules. Interactive computer graphics: FRODO. *Methods Enzymol.* **115,** 157–171 (1985).
47. Fürst, D. O., Osborn, M., Nave, R. & Weber, K. The organization of titin filaments in the half-sarcomere revealed by monoclonal antibodies in immunoelectron microscopy; a map of ten non-repetitive epitopes starting at the Z line extends close to the M line. *J. Cell Biol.* **106,** 1563–1572 (1988).

Acknowledgements. We thank M. Saraste for support in the early stages of this project and J. Heierhorst for stimulating and fruitful discussions.

Correspondence and requests for materials should be addressed to M.W. (e-mail: Wilmanns@embl-hamburg.de) or M.G. (e-mail: Gautel@embl-heidelberg.de). Coordinates and structure factors have been deposited at the Protein Data Base (accession number 1TKI).

Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*

Richard A. Alm*, Lo-See L. Ling†, Donald T. Moir†, Benjamin L. King*, Eric D. Brown*, Peter C. Doig*, Douglas R. Smith†, Brian Noonan*, Braydon C. Guild†, Boudewijn L. deJonge*, Gilles Carmel†, Peter J. Tummino*, Anthony Caruso†, Maria Uria-Nickelsen*, Debra M. Mills†, Cameron Ives*, Rene Gibson†, David Merberg*, Scott D. Mills*, Qin Jiang‡, Diane E. Taylor‡, Gerald F. Vovis† & Trevor J. Trust*

** Astra Research Center Boston, 128 Sidney Street, Cambridge, Massachusetts 02139-4239, USA*
† Genome Therapeutics Corporation, 100 Beaver Street, Waltham, Massachusetts 02453-8443, USA
‡ Department of Medical Microbiology & Immunology and Canadian Bacterial Diseases Network, University of Alberta, Edmonton T6G 2H7, Canada

***Helicobacter pylori*, one of the most common bacterial pathogens of humans, colonizes the gastric mucosa, where it appears to persist throughout the host's life unless the patient is treated. Colonization induces chronic gastric inflammation which can progress to a variety of diseases, ranging in severity from superficial gastritis and peptic ulcer to gastric cancer and mucosal-associated lymphoma[1]. Strain-specific genetic diversity has been proposed to be involved in the organism's ability to cause different diseases or even be beneficial to the infected host[2,3] and to participate in the lifelong chronicity of infection[4]. Here we compare the complete genomic sequences of two unrelated *H. pylori* isolates. This is, to our knowledge, the first such genomic comparison. *H. pylori* was believed to exhibit a large degree of genomic and allelic diversity, but we find that the overall genomic organization, gene order and predicted proteomes (sets of proteins encoded by the genomes) of the two strains are quite similar. Between 6 to 7% of the genes are specific to each strain, with almost half of these genes being clustered in a single hypervariable region.**

H. pylori strain J99 (*cagA*$^+$ *vacA*$^+$), isolated in the USA in 1994 from a patient with a duodenal ulcer, was subjected to minimal subculturing before being sequenced by us in 1996. We describe this sequence below and compare it with the sequence of strain 26695, which was isolated in the UK before 1987 from a gastritis patient and which had a history of subculturing before being sequenced[5]. The J99 circular chromosome is 1,643,831 base pairs (bp) in size, which is 24,036 bp smaller than the 26695 chromosome. Several features, including the absence of an identifiable origin of relication, the average length of coding sequences and the relative frequency of the different initiation codons, are similar in the two strains (Table 1). We predict that there are 1,495 open reading frames (ORFs) in J99, representing 91% of the genome. Eighty-nine of these ORFs are absent from 26695. Of these J99-specific ORFs, 25 and 8 have sequence similarity to genes of predicted and unknown function, respectively, and 56 share no significant sequence similarity with any genes in public databases. J99 has 95 fewer genes than has been reported for 26695. However, 54 predicted genes of strain 26695 are less than 150 bp in size. In comparison with J99 genes, these 54 small genes either are highly conserved (16) and likely to encode proteins (note that three of these 26695 ORFs are part of larger ORFs in J99), or contain in-frame stop codons or exhibit nucleotide drift (38), as do other intergenic regions, and are therefore unlikely to encode proteins. Thus, we revised the 26695 gene complement to 1,552 genes; 117 of these are unique to 26695 and 26 of these unique genes have a predicted function. Some genes appeared to contain a frameshift in J99 or 26695: 27 J99 genes are the equivalents of 55 predicted genes in 26695, and 7 genes from 26695 are the equivalents of 15 predicted genes in J99. In addition, three single-copy genes in 26695 have complete (gene HP1365; *H. pylori* 26695 genes are numbers preceded by 'HP') or partial (genes HP0818 and HP0928) duplications in J99. There are 1,406 genes in J99 that have counterparts in 26695.

Both genomes contain two 16S and two 23S–5S ribosomal RNA copies in the same relative locations, but strain 26695 contains a further, orphan 5S rRNA. In contrast to most other bacteria, the *H. pylori* rRNA loci are not contiguous, indicating that they may be regulated in a complex way. There are fewer complete insertion-sequence elements and fragments in J99 than in 26695, yet their location in both strains appears to be biased towards one half of the genome (Fig. 1a). Both genomes encode 36 transfer RNA species, each mapping to the same relative location. Neither strain contains Asn or Gln tRNA species; however, we have identified homologues for the *Bacillus subtilis gatABC* genes (gene JHP769/HP0830, JHP603/HP0658 and JHP909/HP0975; *H. pylori* J99 genes are numbers attached to the prefix 'JHP'), which amidate glutamate charged tRNAs to make glutamine-charged tRNAs[6]. Such genes are also likely to be responsible for amidation of appropriate aspartate-charged tRNAs.

Table 1 General comparative features of the *H. pylori* genomes

Genome features	*H. pylori* 26695	*H. pylori* J99
Size (base pairs)	1,667,867	1,643,831
(G + C) content (%)	39	39
Regions of different (G + C) content	8*	9†
AGTGATT repeats at bp = 1	26	2‡
vacA genotype	*slb/ml*	*sla/ml*
Open reading frames		
Per cent of genome (coding)	91.0	90.8
Predicted number	1,590	1,495
Functionally classified	875§	895
Conserved with no fucntion	275§	290
H. pylori specific	345§	367
Number with signal sequence‖	517	502
Average length (base pairs)	954	998
Per cent AUG initiation codons	81.8	82.7
Per cent GUG initiation codons	9.7	6.7
Per cent UUG initiation codons	8.1	10.4
Per cent other initiation codons	0.4¶	0.2#
Insertion elements☆		
Complete IS*605* copies	5	0
Partial IS*605* copies	8	5
Complete IS*606* copies	2	1
Partial IS*606* copies	2	2(4**)
RNA elements		
Per cent of genome (stable RNA)	0.75	0.75
23S–5S rRNA	2††(3‡‡)	2††
16S rRNA	2§§	2§§
tRNAs	36	36

* Includes the five reported previously[5]. Additional regions are HP0051–HP0054 and DNA flanking HP0611–HP0612 and HP0314–HP0316 (translocation 1).
† Four regions match those in 26695 (26695 loci 1 and 3 are joined in J99): JHP43–JHP46, JHP163–JHP165, JHP1422–JHP1423 and JHP414–JHP415 have a lower (G + C) content DNA flanking JHP299–JHP300 (translocation 1) has lower (G + C) content.
‡ Another cluster of these heptamer repeats is present ~2.35 kb upstream; at this position, there are 13 copies of the repeat in J99 and 2 copies in 26695.
§ Total ORFs equal 1,552 as defined during re-analysis of 26695 (see text).
‖ Defined as a *P* value of less than 0.05 using the SPScan algorithm in GCG 9.1.
¶ HP0142 (CUG), HP0655 (AUU), HP0882 and HP0904 (AAA), HP0685 (GGA), and HP0451 (UGC).
# JHP55, JHP402 and JHP600 (AUU).
☆ IS, insertion sequence.
** Two copies are smaller than the other two and are within the 31-bp repeated boundary of *cag* PAI.
†† 23S–5S rRNA is located at nucleotides 1,057,138–1,060,475 and 1,426,976–1,430,313 in J99, and 445,306–448,642 and 1,437,499–1,476,836 in 26695.
‡‡ 26695 orphan 5S rRNA is located at nucleotides 1,045,074–1,045,248.
§§ 16S rRNA is located at nucleotides 1,188,029–1,189,529 and 1,463,047–1,464,547 in J99, and 1,207,583–1,209,081 and 1,511,137–1,512,634 in 26695.

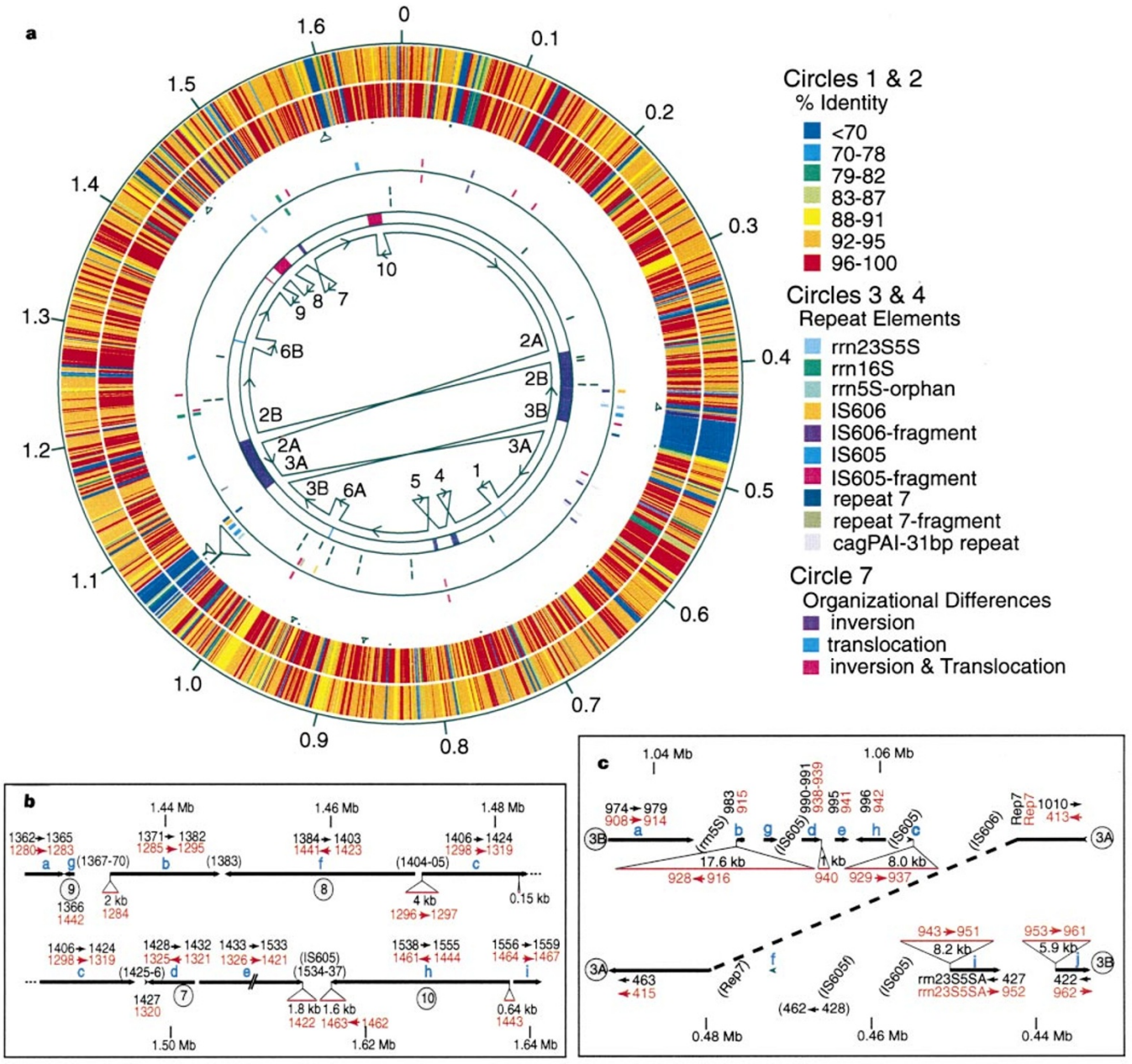

Figure 1 Comparison of the two sequenced *H. pylori* genomes based on the chromosomal organization of strain 26695. **a**, Genome-wide view. Circles are numbered starting from the outermost concentric ring. Circle 1, nucleotide and circle 2, amino-acid similarity between each J99 and 26695 orthologue. The relative location and amount of each J99-specific sequence are shown immediately inside the second circle (the height of each line is proportional to the amount of unique sequence, and for larger regions the size relative to the equivalent 26695 region is indicated by a triangle proportional to the 26695 scale). The largest J99-specific region shown is composed of two segments separated by 150 bp (see **b** for details). Circles 3 and 4, which flank the solid reference circle, show the locations of rRNA, insertion-sequence (IS) and repeat elements, for 26695 (circle 3) and J99 (circle 4). Circles 5 and 6 represent the locations of the *Not*I sites in the 26695 and J99 genomes, respectively. Circle 7 represents the relative transcriptional direction of J99 genes compared to their 26695 orthologues. Regions that are not coloured and translocations are transcribed in the same relative direction in J99 and 26695, whereas inversions result in genes being transcribed in the opposite relative direction in J99. Circle 8 represents the organization of the J99 genome relative to the 26695 genome, incorporating artificial end points needed to allow the alignment. The required inversions and/or translocations are numbered consecutively for J99. **b**, **c**, An expanded view of the complex organizational differences 7–10 (**b**) and 3A/3B (**c**) shown in circle 8 of **a**. The 26695 ORFs are shown in the order and location that they are found (black numbers). The J99 ORFs are shown as red numbers. ORFs and other elements in parentheses are found in 26695 but not J99. The organization of J99 segments that share >90% identity to 26695 are depicted by the solid black lines, with arrows indicating the relative orientation of the J99 segments with respect to 26695 segments. The open triangles represent J99-specific DNA, drawn to scale, with the size and genes shown. The order of these regions in J99 is indicated by lower-case blue letters. The circled numbers correspond to the inversions/translocations referred to in **a**. Sizes of regions in **a** are shown in megabases (Mb).

Severity of *H. pylori* related disease is correlated with the presence of an island of genes (the *cag* pathogenicity island, *cag*PAI) associated with production of the CagA antigen[7] and upregulation of interleukin (IL)-8 in gastric epithelial cells[8]. Both J99 and 26695 contain the complete *cag*PAI flanked by the same chromosomal genes and the previously described 31-bp repeat[7] but lack the insertion-sequence *605* elements that are associated with *cag*PAI in strain NCTC11638 (ref. 7). Comparison of available *cag*PAI gene sequences showed minor differences between the J99 *cag*PAI genes and the other available sequences, such as apparent deletions in the *cag7* gene of J99 and 26695 (JHP476, HP0527) that lead to loss of up to 114 amino acids.

Like 26695, J99 encodes many families of paralogous proteins (337 genes, 22.5% of the total, are members of 113 families). One family contains the *vacA*-encoded vacuolating cytotoxin and three paralogues. Two of the three orthologues differ significantly in size between J99 and 26695: JHP856 encodes a protein that is 130 amino acids shorter than the protein encoded by HP922, and JHP556 represents a fusion between HP0610 and HP0609. The three paralogues in both strains lack the cleavage signal contained within VacA and may not be secreted.

The DNA-sequence differences between orthologues from the two strains are mainly found in the third position of coding triplets, consistent with the variance seen between *H. pylori* strains using methods dependent on the nucleotide sequence or on the sequencing of specific loci in different strains[9–11]. However, this nucleotide variation does not translate into a highly divergent proteome (Fig. 1a). For example, there are only eight genes with ⩾98% nucleotide identity but 310 proteins with ⩾98% amino-acid conservation, including 41 with perfect identity.

To align homologous regions in the two genomes, we needed to artificially invert and/or transpose ten segments, ranging in size from 1 kilobase (kb) to 83 kb, of the J99 sequence. Most of the artificial end points are in intergenic regions and most are associated with insertion elements, repeated sequences or genes, and/or DNA-restriction/modification genes in one or both of the genomes (Fig. 1, Table 2), consistent with a possible role for such elements in generating these organizational differences. Two differences between the genomes are associated with genes encoding members of the large outer membrane protein (Omp) family. Inversion 5 in J99 could have resulted from a simple recombination across the inverted, repeated nucleotide sequence encoding the carboxy-terminal domain of two Omp proteins. Rearrangement 6 in J99 is the result of the equivalent of a reciprocal exchange of the Lewis-antigen-binding adhesin genes *babA* and *babB*[12]; BabA and BabB share similar C-terminal domains. The complex rearrangements 8–10 in J99 consist of both inversions and translocations (Fig. 1b). In both genomes, inversion 3 is associated with a region of (G + C) percentage that is lower (35%) than in the rest of the genome (39%). We named this region a 'plasticity zone' because it contains 46% and 48% of the genes that are unique to 26695 and J99, respectively (Fig. 1c). Although this region is continuous in J99, it is split in 26695 into two domains that are separated by ~600 kb. The presence of *vir* homologues, insertion sequences and a lower percentage of (G + C) DNA indicates that these regions might represent pathogenicity islands. The clustering of DNA with a lower (G + C) percentage is suggestive of horizontal DNA transfer, and the strain-specific sequence differences are consistent with different origins for this DNA. *H. pylori* and *Campylobacter* spp. plasmids have a (G + C) percentage in this lower range[13]. Significantly, two copies of the insertion-sequence *605* element and neighbouring 26695-strain-specific chromosomal DNA from the plasticity zone (genes HP0999–HP1001) are present on the *H. pylori* plasmid pHPM186 (GenBank accession number AF077006). Thus, plasmids may be responsible for the integration of new DNA into the *H. pylori* chromosome and for the transfer of this DNA between strains. Recombination across inverted repeats (repeat 7; ref. 5) in a progenitor strain that resembles strain 26695 would yield an arrangement similar to that of the region of inversion 3A in strain J99, but a similar reciprocal event cannot account for the complexity of the J99 3B locus (Fig. 1c).

To confirm the assembled sequence, we studied the J99 genome by pulsed-field gel electrophoresis (PFGE) and hybridization with specific probes (for J99 genes JHP117, 312, 548, 663, 733 and 1133–1136). Each observed *Not*I fragment was consistent in size with that predicted by the sequence. Hybridization of the 26695 genome *in silico* with these same probes yielded *Not*I fragments that were different in size to those observed with J99 DNA. Differences similar to those which we observed in restriction-fragment sizes and in probe hybridization patterns have been interpreted to mean that *H. pylori* strains are highly diverse in their genomic organization and gene order[14]. The differences in sizes of *Not*I fragments in strains J99 and 26695 are due mainly to silent nucleotide variation within genes. J99 contains twice as many *Not*I sites as 26695 (Fig. 1a), and silent nucleotide changes in 26695 are responsible for the absence of six of the seven *Not*I sites unique to J99; differences at the seventh site result in the alteration of a single amino acid. Similar minor sequence differences account for the variability in the *Nru*I-site content between the strains. Thus, results obtained with lower-resolution techniques such as PFGE and polymerase chain reaction (PCR)–restriction-fragment length polymorphism (RFLP) have probably led to an overestimation of the true extent of genetic

Table 2 Elements associated with the artificial end points required to align the two *H. pylori* chromosomes

Locus	Type*	Size (kb)	Associated elements and genes	Strain
1	TR	1.5†	Lower (G + C)-content DNA (JHP299–JHP300; HP611–HP613); repeat element	Both
2A/B	IN	75	IS*605* in 26695 (HP1095–HP1096); genes of unknown function in J99 (JHP331–JHP332) and 26695 (HP1094)	26695 or J99
3A/B	IN	83	'Orphan' 5S rRNA; inverted copies of repeat 7	26695
4	IN	10	Insertion of DNA-restriction/modification genes (JHP629–JHP630)	J99
5	IN	2.5	Conserved C terminus of *omp* genes (JHP659 and JHP662; HP0722 and HP0725); IS*605* left-end fragment	Both
6A/B	TR	2†	Conserved C terminus of *bab* genes (JHP833 and JHP1164; HP0896 and HP1243) and 5′ repeat element	Both
7	IN	5.5	Repeated, overlapping C terminus of histidine-rich genes (JHP1320–JHP1321; HP1427 and HP1432)	Both
8	IN/TR	24†	DNA-restriction/modification-gene replacement (JHP1296–JHP1297; HP1404/HP1405)	Both
10	IN/TR	21†	IS*605* (HP1534–HP1535)	26695
9	IN/TR	1†	DNA-restriction/modification genes (JHP1442; HP1366); duplication of response regulator (JHP1283 and JHP1442; HP1365)	Both

* TR, translocation; IN, inversion.
† Distance between relative position of translocations in J99 and 26695 are 287 kb (locus 1), 385 kb (locus 6A/B), 146 kb (locus 8), 4 kb (locus 9) and 184 kb (locus 10).

diversity in *H. pylori*[9,10,14]. However, these techniques will continue to be useful for epidemiology and strain discrimination.

To estimate the degree of conservation of gene order between J99 and 26695, we studied the immediate neighbours of each J99 gene and its 26695 orthologue, if present. Of the 1,495 genes in J99, 1,267 (84.7%) have the same neighbour on each side in both genomes; 161 (10.8%) are flanked by one common neighbour and one strain-specific gene; and 40 (2.7%) are flanked by strain-specific genes on both sides. Only 27 (1.8%) have the same neighbour on one side and a common gene that appears in a different position on the other side as the result of an organizational difference. There are 9 conserved gene strings that are more than 50 genes long, representing 46% of the genes common to both strains, with the longest string containing 133 genes. This highly conserved gene order indicates that physical linkage of a few genes (*topA/flaB*[15] and *ftsH/pss/copA* (D.E.T., unpublished observations)) in several strains is the rule rather than the exception. The absence of extensive gene shuffling between J99 and 26695 is consistent with a low level of evolutionary divergence[16].

Of the 1,495 J99 genes and the 1,552 re-annotated 26695 genes, 874 (58.5%) and 895 (57.7%) gene products, respectively, have been assigned putative functions. A total of 275 (18.4%) J99 and 290 (18.7%) 26695 gene products have orthologues of unknown function in other species, and 346 (23.1%) J99 and 367 (23.6%) 26695 genes are *H. pylori* specific (that is, they show no sequence similarity with genes available in public databases). Of these *H. pylori* specific genes, 56 and 69 are specific to strains J99 and 26695, respectively. Excluding the strain-specific ORFs in the plasticity zone, the J99-specific genes are located singly (24 times) or as clusters of two (8 clusters) or three (1 cluster); many of these clusters appear to be organized to permit co-transcription. In one case, six genes (insertion-sequence *606* element and four J99-specific genes) are linked and are flanked by a duplicated region. In 17 corresponding locations, both J99 and 26695 have strain-specific genes. This high proportion of common relative loci for strain-specific ORFs indicates that *H. pylori* may have limited flexibility for containing strain-specific genes. Of the total of 206 strain-specific genes (89 in J99, 117 in 26695), the plasticity zones contain 94 (42 in J99, 52 in 26695); 125 of the 206 strain-specific genes (60.7%) are also specific to *H. pylori*, and 30 (14.6%) share similarity with genes of unknown function. J99- or 26695-specific genes have been assigned to the following categories: DNA restriction or modification (15 and 16, respectively); cell-envelope synthesis (4 and 2); cellular processes, such as DNA transfer and competence proteins (2 and 4); DNA replication (2 and 2); energy metabolism (2 and 1); and phospholipid metabolism (1 in 26695).

The fact that strain-specific DNA-restriction/modification genes have a lower (G + C) content than the remainder of the genome and are associated with regions that are organized differently in the J99 and 26695 genomes indicates that these genes may have been acquired horizontally from other bacterial species or transferred more recently from other *H. pylori* strains by natural transformation. Each *H. pylori* strain contains its own specific complement of these restriction/modification enzymes (R.A.A., unpublished observations). Nine type II methyltransferases are conserved between the two strains but lack identifiable cognate restriction-subunit partners, indicating that *H. pylori* may regulate gene expression by methylation.

The strain-specific genes encoding proteins involved in cell envelope (lipopolysaccharide and outer membrane protein) biosynthesis represent members of four paralogous families. Each strain contains one unique member of the *omp* families (HP0317 and JHP870). J99 and 26695 contain two (JHP820 and JHP1032) and one (HP1578) unique member, respectively, as well as four common members, of the *rfaI/rfaJ*-like family which is involved in lipopolysaccharide biosynthesis. In addition, J99 contains a unique member (JHP562) plus three common members of the *lex2B* lipopolysaccharide-biosynthesis family.

One of the J99-specific genes involved in energy metabolism encodes a second homologue of alcohol dehydrogenase (JHP1429), and the other (JHP585) may be required for amino-acid degradation. The 26695-specific energy-metabolism gene (HP1045) encodes an acetyl-CoA synthase. Strain 26695 has a second, larger acyl-carrier protein (encoded by HP0962) which is involved in phospholipid metabolism. J99 and 26695 have two (JHP919 and JHP931) and one (HP0440) unique genes encoding topoisomerase homologues, respectively, in their plasticity zones, which also contain the strain-specific genes that encode proteins involved in cellular processes. J99 has two adjacent *virB4* homologues (JHP917 and JHP918) which may have once represented a single complete gene, whereas 26695 contains two complete *virB4* (HP0441 and HP0459) and one truncated *virD4* (HP1006) homologues and a protein (encoded by HP0432) with similarity to a human protein kinase C.

The identification of homopolymeric tracts and dinucleotide repeats in *H. pylori* led to the prediction that 'slipped-strand repair' may modulate gene expression[5], which could result in antigenic variation and adaptive evolution. The J99 gene sequences do not support some previously proposed examples of genes which are regulated in this fashion (for example, HP0211 and HP0928)[17]. In other cases, the data obtained from J99 do support this mechanism of control. Repeat lengths in some J99 genes differ from those in 26695 genes, indicating that such genes may be differently expressed in the two strains (Table 3). The same five members of

Table 3 Genes whose expression may be regulated by 'slipped-strand-repair

JHP gene* (repeats:status)	HP gene* (repeats:status)	Repeat	Variation in J99 (#reads@#repeats)
Cell envelope (outer membrane protein)			
7 (6:off)	0009 (11:off)	(CT)	9@6
581 (9:on)	0638 (6:on)	(CT)	7@9
659 (9:off)	0722 (8:off)	(CT)	8@9; 1@8
662 (9:off)	0725 (6:off)	(CT)	7@9
1164 (8:on)	0896 (11:on)	(CT)	†
Cell envelope (lipopolysaccharide biosynthesis)			
86 (13:off)	0093–0094 (14:off)	(C)	18@13
194 (8:off)	0208 (11:off)	(AG)	†
563 (12:off)	0619 (13:off)	(C)	†
596 (5:on‡)	0651 (13:on)	(C)	†
820 (14:on)	Absent§	(C)	†
1002 (13&9:off)	0379 (13&6‡:on)	(C)&(A)	4@13&4@9
Cell envelope (flagella biosynthesis)			
625 (8:on‖)	0684–0685 (9:off)	(C)	†
Regulatory functions			
151 (9:on)	0164–0165 (13:off¶)	(C)	3@8; 10@9; 3@10
Transport and binding proteins			
1129 (9:on‡)	1206 (10:on)	(A)	6@9
DNA restriction/modification			
416 (10:off‡)	0464 (15:on)	(C)	3@9; 3@10
1364 (14:on)	1471 (14:on)	(G)	†
1411 (11:off)	1522 (12:off)	(G)	†
1442 (8:off)	1366 (6:on)	(A)	1@7; 13@8; 5@9
Conserved with no known function			
131 (6:on)	0143 (7:off)	(A)	†
1312 (10:off)	1417 (9:off)	(G)	1@9; 3@10; 1@12
H. pylori specific with no known function			
203 (6&7:off)	0217 (12&6:on)	(G)&(G)	14@6&11@7; 1@6
351 (5:on)	1074 (6:off#)	(A)	12@5
681 (7:off)	0744 (9:off)	(AG)	†
1272 (13&12:off)	1353–1354 (12&15:off)	(C)&(C)	11@13&2@12; 2@13
1326 (11:on)	1433 (5:on‡)	(C)	2@9; 8@10; 3@11; 4@12; 2@13
1392 (7:on)	1499 (6:off#)	(A)	14@7

* Gene number from J99 (JHP) or 26695 (HP). The number of repeats and whether the gene appears in-frame (on) or out-of-frame (off) are shown in parentheses.
† Insufficient sequence coverage was available to be deemed significant.
‡ In addition to the repeats, an extra base pair of different identity to the repeat was found at one end.
§ Not applicable.
‖ The string of C nucleotides in the *fliP* gene (JHP625) has a C-to-A substitution in the middle.
¶ The 26695 gene also has another frameshift upstream of this string of C nucleotides.
# The ORF predicted in 26695 (ref. 5) is truncated by this change in repeat length relative to J99.

the large *omp* paralogous family contain CT dinucleotide repeats in both strains, but the number of repeats differs without affecting the predicted expression status. The comparative data indicate that slipped-strand regulation may operate at two sites in some genes, including the α-(1,3)-fucosyltransferase gene. This regulatory mechanism also operates during laboratory passage of cell cultures: we found changes in the lengths of specific homopolymeric tracts or dinucleotide repeats within different populations of strain J99 (Table 3). We also detected nucleotide substitutions, most of which were found in the third position of coding triplets, at a low frequency.

Several factors could influence the pathophysiology and severity of disease associated with infection by different *cagA*$^+$ *H. pylori* strains. First, strain-specific genes, such as those associated with the plasticity zone, could play a role. Second, differences in gene expression, perhaps mediated by slipped-strand repair, may be important and may affect the ability of the organism to colonize. Third, a human host factor(s) may play a significant, and perhaps unappreciated, part in susceptibility to, and severity of, *H. pylori* infection. In any host–parasite relationship, bacterial, host and environmental factors influence the host's susceptibility to and the clinical outcome of infection. For example, different mice strains exhibit markedly different susceptibilities to *H. pylori* colonization and clinical outcome[18]. Different human populations also show differences in susceptibility to *H. pylori* infection and incidence rates for gastric cancer[19]. Our identification of the minimal genetic diversity between two virulent strains, genes that are conserved between the two strains, and the strain-specific plasticity zone allows a better understanding of the biology of *H. pylori*. Our results suggest the need, and provide a unique opportunity, for a reassessment of the respective roles of bacterial and host factors in diseases associated with *H. pylori*. □

Methods

H. pylori strain J99 was sequenced, assembled and analysed nearly as described[14,20]. The sequences of regions that differ significantly between strains J99 and 26695, including putative frameshifts, were all confirmed by sequencing PCR products of J99 and, where relevant, by diagnostic PCR of 26695. The nucleotide and amino-acid alignments used to determine the identity between orthologues were generated by ALIGN from version 2.0 of the FASTA program package. Paralogues were identified using BLASTP and TBLASTX algorithms. The output was initially grouped such that all members of a family exhibited similarity to at least one other member, using a cut-off of $P < 10^{-10}$, and then checked manually for validity.

Received 4 September; accepted 24 November 1998.

1. Cover, T. L. & Blaser, M. J. *Helicobacter pylori* and gastroduodenal disease. *Annu. Rev. Med.* **42,** 135–145 (1992).
2. Atherton, J. C. Jr, Peek, R. M. Jr, tham, K. T., Cover, T. L. & Blaser, M. J. Clinical and pathological importance of heterogeneity in *vacA*, the vacuolating cytotoxin gene of *Helicobacter pylori*. *Gastroenterology* **12,** 92–99 (1997).
3. Blaser, M. J. Not all *Helicobacter pylori* strains are created equal: should all be eliminated? *Lancet* **349,** 1020–1022 (1997).
4. Logan, R. P. H. & Berg, D. E. Genetic diversity of *Helicobacter pylori*. *Lancet* **348** 1462–1463 (1996).
5. Tomb, J.-F. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388,** 539–547 (1997).
6. Curnow, A. W. *et al.* Glu-tRNAGln amidotransferase: a novel heterotrimeric enzyme required for correct decoding of glutamine codons during translation. *Proc. Natl Acad. Sci. USA* **94,** 11819–11826 (1997).
7. Akopyants, N. S. *et al.* Analyses of the *cag* pathogenicity island of *Helicobacter pylori*. *Mol. Microbiol.* **28,** 37–53 (1998).
8. Censini, S. *et al. cag*, a pathogenecity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. *Proc. Natl Acad. Sci. USA* **93,** 14648–14653 (1996).
9. Akopyanz, N., Bukanov, N. O., Westblom, T. U. & Berg, D. E. PCR-based RFLP analysis of DNA sequence diversity in the gastric pathogen *Helicobacter pylori*. *Nucleic Acids Res.* **20,** 6221–6225 (1992).
10. Han, J., Yu, E., Lee, I. & Lee, Y. Diversity among clinical isolates of *Helicobacter pylori* in Korea. *Mol. Cells* **7,** 544–547 (1997).
11. Kansau, I. *et al.* Genotyping of *Helicobacter pylori* isolates by sequencing of PCR products and comparison with the RAPD technique. *Res. Microbiol.* **147,** 661–669 (1996).
12. Ilver, D. *et al. Helicobacter pylori* adhesin binding fucosylated histo-blood group antigens revealed by retagging. *Science* **279,** 373–377 (1998).
13. Lee, W. K. *et al.* Construction of a *Helicobacter pylori–Escherichia coli* shuttle vector for gene transfer in *Helicobacter pylori*. *Appl. Environ. Microbiol.* **63,** 4866–4871 (1997).
14. Jiang, Q., Hiratsuka, K. & Taylor, D. E. Variability of gene order in different *Helicobacter pylori* strains contributes to genome diversity. *Mol. Microbiol.* **20,** 833–842 (1996).
15. Suerbaum, S., Brauer-Stepkes, T., Labigne, A., Cameron, B. & Drlica, K. Topoisomerase I of *Helicobacter pylori*: juxtaposition with a flagellin gene (*flaB*) and functional requriement of a fourth zinc finger motif. *Gene* **210,** 151–161 (1998).
16. Tatusov, R. L. *et al.* Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.* **6,** 279–291 (1996).
17. Saunders, N. J., Peden, J. F., Hood, D. w. & Moxon, E. R. Simple sequence repeats in the *Helicobacter pylori* genome. *Mol. Microbiol.* **27,** 1091–1098 (1998).
18. Sakagami, T. *et al.* Atrophic gastrities changes in both *Helicobacter felis* and *Helicobacter pylori* infected mice are host dependent and separate from antral gastritis. *Gut* **39,** 639–648 (1996).
19. Fock, K. M. *et al.* Seroprevalence of *Helicobacter pylori* infection. *Gastroenterology* **114,** A596 (1998).
20. Smith, D. R. Complete genome sequence of *Methanobacetrium thermoautotrophicum* deltaH: functional analysis and comparative genomics. *J. Bacteriol.* **179,** 7135–7155 (1997).

Acknowledgements. We thank T. Dorman, M. Rubenfield, C. Butler, B. Andrews and K. MacCormack for assistance in finalizing and editing sequence information. D.E.T. is a member of the Canadian Bacterial Diseases Network and a Scientist with the Alberta Heritage Foundation for Medical Research.

Correspondence and requests for materials should be addressed to R.A.A. (e-mail: richard.alm@arcb.us.astra.com). The annotated genome sequence and further information are available in the *H. pylori* database (www.astra-boston.com/hpylori or www.genomecorp.com/hpylori) and will appear in GenBank under accession number AE001439.