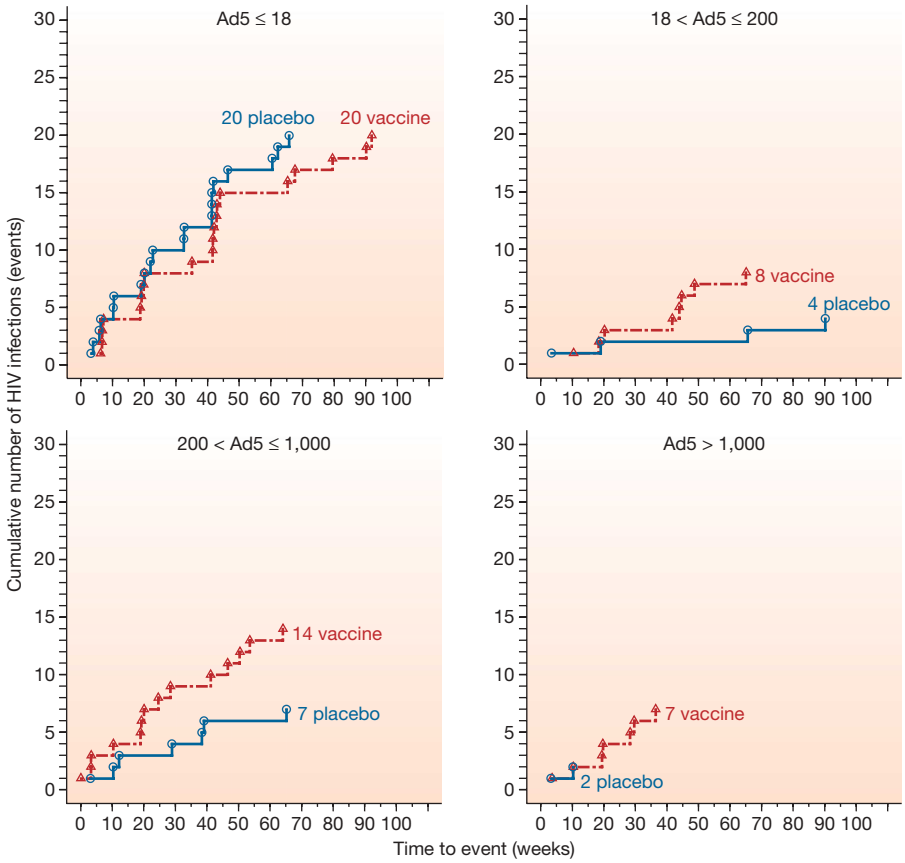

Figure 2 | Cumulative HIV-1 infections in men enrolled in the STEP study stratified by pre-existing Ad5-specific neutralizing antibody titre. Cumulative infections as of 17 October 2007 in men enrolled in the STEP study (HVTN 502) evaluating the Merck rAd5-Gag/Pol/Nef vaccine are depicted. Infections in vaccinees (red) and placebos (blue) are shown in individuals stratified by their pre-existing Ad5-specific neutralizing antibody titres. Data represent the modified intent-to-treat population. Image courtesy of M. Robertson, Merck Research Laboratories.

unpublished observations). An alternative hypothesis is that Ad5-specific neutralizing antibodies may have opsonized rAd5 vectors after immunization, resulting in altered tropism or inflammatory responses. It is also possible that pre-existing Ad5-specific neutralizing antibodies may have been a marker for other confounding variables that have not yet been identified.

A STEP forward?

Despite the disappointing results of the STEP study, several key lessons have already been learned. First, it is clear that the path forward towards an HIV-1 vaccine will be neither simple nor straightforward. Second, the importance of understanding both systemic and mucosal immune responses to vaccine vectors is paramount. Third, the biological determinants of HIV-1 acquisition and the impact that vector-specific and antigen-specific mucosal immune responses may have on this process will require intensive investigation. Fourth, clinical vaccine studies will need to adapt to the safety concerns raised by the STEP study, such as possibly excluding subjects who have pre-existing neutralizing antibodies to the vaccine vector that is used until this phenomenon is more completely understood. Fifth, future T-cell-based vaccine candidates should be prioritized for clinical efficacy studies only if they are convincingly superior to the homologous rAd5-Gag/Pol/Nef regimen that has failed. Sixth, non-human primate challenge models should be recalibrated on the basis of the STEP study to guide future HIV-1 vaccine development.

The protection afforded by the homologous rAd5 regimen against SHIV-89.6P indicates that this model lacks sufficient stringency for the evaluation of T-cell-based vaccine candidates. Although the more stringent SIV challenge model cannot be considered to be validated until there is a successful clinical efficacy study in humans, it seems reasonable to use SIV_{MAC239} or SIV_{MAC251} as challenge viruses for evaluating next-generation vaccine candidates (Box 3). Preclinical challenge studies need to be adequately powered with sufficient follow-up time, and the vaccine schedule and dose should model the proposed clinical regimen. For optimal stringency, studies should exclude rhesus monkeys that express MHC class I alleles that are specifically associated with efficient virologic control, such as Mamu-A*01, Mamu-B*17 and Mamu-B*08. The use of homologous Env antigens that may inappropriately overestimate protective efficacy should also be avoided. Mucosal challenges may offer certain physiological advantages over intravenous challenges, and these challenge models should therefore be developed. Finally, increased emphasis should be placed on assessing the capacity of promising vaccine candidates to protect against highly heterologous SIV challenges, because infecting viruses in humans will almost certainly be heterologous to any vaccine sequence. Because very few heterologous SIV challenge studies have been completed so far, a practical approach may be to determine the protective efficacy of promising vaccine

Box 3 | Recommendations for preclinical challenge studies of T-cell-based vaccines

(1) Use stringent challenge virus (SIV_{MAC239}, SIV_{MAC251}).
(2) Design study with adequate power and follow-up time.
(3) Model clinical regimen with vaccine schedule and dose.
(4) Select rhesus monkeys that lack MHC alleles associated with efficient virologic control (Mamu-A*01, Mamu-B*17, Mamu-B*08).
(5) Avoid the use of a homologous Env antigen.
(6) Assess promising vaccine concepts against both homologous and heterologous viral challenges.

candidates against both homologous and heterologous SIV challenges. It is currently debated whether non-human primate challenge studies should be used as a formal 'gatekeeper' for advancing vaccine candidates into clinical efficacy studies, because the capacity of this model to predict the results of clinical efficacy studies remains unclear. Nevertheless, it would seem reasonable to give a relative priority to the development of vaccine candidates that lead to durable control of setpoint viral loads after SIV_{MAC239} or SIV_{MAC251} challenge.

The STEP study has also had a major impact on other HIV-1 vaccine programmes in the field. HVTN 503 was terminated as it used the same rAd5-based vaccine candidate that was used in HVTN 502. The NIH Vaccine Research Center has developed a DNA prime/rAd5 boost vaccine regimen expressing clade B Gag–Pol and multiclade Env antigens. This vaccine candidate has been shown to be immunogenic in most individuals in phase 1 studies, particularly for the Env antigens[62,68,83]. In preclinical studies, a DNA prime/rAd5 boost vaccine regimen expressing SIV Gag, Pol, Nef and Env antigens afforded a 1.1 log reduction of peak viral loads for 112 days after a homologous SIV_{MAC251} challenge[77]. However, no durable control of setpoint viral loads was observed with this vaccine, although delayed progression to AIDS-related mortality was evident[77]. NIH recently announced that it will not proceed with a large phase 2b efficacy study known as PAVE 100, although a smaller, more focused efficacy study with this vaccine candidate is still under consideration[84]. DNA prime/poxvirus boost regimens are also being evaluated using modified vaccinia Ankara (MVA)[69] and NYVAC[70] vectors, and phase 1 clinical trials have demonstrated immunogenicity in most volunteers. Central to all of these programmes, however, is the hypothesis that DNA priming before vector boosting will improve protective efficacy. This has been observed in some[72] but not all[77] SIV challenge studies, and thus it still remains an open question that requires further investigation and should be considered a high priority.

New rAd vectors derived from Ad serotypes that are rare in human populations are also being explored as a strategy to evade pre-existing Ad5-specific neutralizing antibodies. It is hoped that such vectors may offer immunologic as well as safety advantages as compared with rAd5 vectors by circumventing pre-existing vector-specific neutralizing antibodies. However, these possibilities have not yet been confirmed in clinical trials. Current strategies include the development of rare serotype rAd26, rAd35 and rAd48 vectors[78,79,85]; chimaeric rAd5HVR48 vectors in which dominant Ad5-specific neutralizing antibody epitopes have been exchanged[86]; and non-human rAd vectors[87,88]. Rare serotype rAd vectors are biologically different from rAd5 vectors in terms of their cellular receptors, tropism, intracellular trafficking pathways and innate immune profiles. Moreover, rAd26 and rAd48 vectors have been shown to elicit T lymphocyte responses of a substantially different phenotype as compared with rAd5 vectors[89], and potent heterologous rAd prime-boost regimens can be constructed using serologically distinct rAd vectors. We have recently demonstrated that a heterologous rAd26 prime/Ad5 boost regimen expressing SIV Gag afforded a durable 2.4 log reduction of setpoint viral loads after SIV_{MAC251} challenge of Mamu-A*01-negative rhesus monkeys, whereas a homologous rAd5 regimen provided no protection in this stringent challenge model[90]. These data suggest that vaccine candidates that elicit improved magnitude, breadth and quality of T lymphocyte responses may provide superior protective efficacy as compared with homologous rAd5 regimens.

Perspectives and future directions

To a great extent, HIV-1 vaccine science is still in its infancy. Major unsolved problems remain, and a renewed commitment to basic discovery research in addition to preclinical studies and clinical trials will be required to move the field forward. Clinical trials that are focused on answering specific scientific hypotheses rather than exclusively aimed at product development may be most useful to the field at the present time. Certain vaccine regimens, such as heterologous rAd prime–boost regimens, may offer the possibility of improved magnitude, breadth and quality of T lymphocyte responses as compared with the homologous rAd5 regimen. New antigen concepts, such as centralized consensus[91,92] and mosaic[93] immunogens, may also result in increased breadth of cellular immune responses and improved coverage of viral diversity.

Perhaps the most important research focus should be the development of improved Env immunogens to elicit broadly reactive neutralizing antibodies. Given the scope of this problem, increased basic research regarding the structure, function and immunogenicity of the Env glycoprotein will be required. Innovative and high-risk ideas should be pursued, and promising approaches should be tested as rapidly as possible in preclinical studies and eventually in clinical trials. Ultimately, it is likely that a combination vaccine consisting of separate vaccine components that elicit T lymphocytes and neutralizing antibodies will prove optimal. As a result, development of improved T-cell-based and antibody-based vaccine strategies should be pursued in parallel.

To achieve these goals, it will be critical to attract and to retain talented new investigators to the field. Funding programmes should therefore be expanded to encourage junior investigators to explore innovative ideas that address critical problems in the field. Given the scientific challenges currently facing the HIV-1 field, increased support and encouragement of fellows and junior faculty should be viewed as a top priority by both senior investigators and funding organizations. It will also be important for industry to continue to participate in the HIV-1 vaccine field, as biotechnology and pharmaceutical companies have critical knowledge and capacities that are not available in academia, government and non-profit organizations.

A current debate is whether the HIV-1 vaccine field can 'withstand' another vaccine efficacy study failure. For HIV-1, the scientific challenges are enormous, and thus so are the risks in testing any new vaccine concept. Clearly, the decision to advance a vaccine candidate into efficacy trials should be highly selective and based on a rigorous and transparent analysis of preclinical and clinical data. However, there is no way to determine whether a potentially promising vaccine candidate will afford protection in humans other than by conducting a clinical efficacy study. Multiple efficacy trials may be required, and many concepts will undoubtedly fail. We should therefore be ready to accept multiple failures of efficacy studies as part of the expected pathway towards the ultimate successful development of a safe and effective HIV-1 vaccine.

1. Barre-Sinoussi, F. *et al.* Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220,** 868–871 (1983).
2. Gallo, R. C. *et al.* Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science* **224,** 500–503 (1984).
3. Popovic, M., Sarngadharan, M. G., Read, E. & Gallo, R. C. Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science* **224,** 497–500 (1984).
4. Sarngadharan, M. G., Popovic, M., Bruch, L., Schupbach, J. & Gallo, R. C. Antibodies reactive with human T-lymphotropic retroviruses (HTLV-III) in the serum of patients with AIDS. *Science* **224,** 506–508 (1984).
5. Schupbach, J. *et al.* Serological analysis of a subgroup of human T-lymphotropic retroviruses (HTLV-III) associated with AIDS. *Science* **224,** 503–505 (1984).
6. Fauci, A. S. 25 years of HIV. *Nature* **453,** 289–290 (2008).
7. Quinn, T. C. *et al.* Viral load and heterosexual transmission of human immunodeficiency virus type 1. *N. Engl. J. Med.* **342,** 921–929 (2000).
8. Mascola, J. R. *et al.* Immunization with envelope subunit vaccine products elicits neutralizing antibodies against laboratory-adapted but not primary isolates of human immunodeficiency virus type 1. The National Institute of Allergy and Infectious Diseases AIDS Vaccine Evaluation Group. *J. Infect. Dis.* **173,** 340–348 (1996).
9. Moore, J. P. *et al.* Primary isolates of human immunodeficiency virus type 1 are relatively resistant to neutralization by monoclonal antibodies to gp120, and their neutralization is not predicted by studies with monomeric gp120. *J. Virol.* **69,** 101–109 (1995).
10. Flynn, N. M. *et al.* Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *J. Infect. Dis.* **191,** 654–665 (2005).

11. Pitisuttithum, P. *et al.* Randomized, double-blind, placebo-controlled efficacy trial of a bivalent recombinant glycoprotein 120 HIV-1 vaccine among injection drug users in Bangkok, Thailand. *J. Infect. Dis.* **194,** 1661–1671 (2006).
12. Priddy, F. H. *et al.* Safety and immunogenicity of a replication-incompetent adenovirus type 5 HIV-1 clade B gag/pol/nef vaccine in healthy adults. *Clin. Infect. Dis.* **46,** 1769–1781 (2008).
13. Fauci, A. S. The release of new data from the HVTN 502 (STEP) HIV vaccine study. *NIH News* ⟨http://www3.niaid.nih.gov/about/directors/news/step_11707.htm⟩ (2007).
These data demonstrate that a homologous rAd5-Gag/Pol/Nef vaccine regimen did not protect against HIV-1 in humans and may have increased risk of HIV-1 acquisition in individuals with pre-existing Ad5-specific neutralizing antibodies.
14. Gaschen, B. *et al.* Diversity considerations in HIV-1 vaccine selection. *Science* **296,** 2354–2360 (2002).
15. Walker, B. D. & Korber, B. T. Immune control of HIV: the obstacles of HLA and viral diversity. *Nature Immunol.* **2,** 473–475 (2001).
16. Montefiori, D., Sattentau, Q., Flores, J., Esparza, J. & Mascola, J. Antibody-based HIV-1 vaccines: recent developments and future directions. *PLoS Med.* **4,** e348 (2007).
17. Kwong, P. D. *et al.* Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* **393,** 648–659 (1998).
18. Wyatt, R. *et al.* The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature* **393,** 705–711 (1998).
19. Chen, B. *et al.* Structure of an unliganded simian immunodeficiency virus gp120 core. *Nature* **433,** 834–841 (2005).
20. Wei, X. *et al.* Antibody neutralization and escape by HIV-1. *Nature* **422,** 307–312 (2003).
21. Richman, D. D., Wrin, T., Little, S. J. & Petropoulos, C. J. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. *Proc. Natl Acad. Sci. USA* **100,** 4144–4149 (2003).
22. Li, Y. *et al.* Broad HIV-1 neutralization mediated by CD4-binding site antibodies. *Nature Med.* **13,** 1032–1034 (2007).
23. Zhou, T. *et al.* Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature* **445,** 732–737 (2007).
24. Haynes, B. F. *et al.* Cardiolipin polyspecific autoreactivity in two broadly neutralizing HIV-1 antibodies. *Science* **308,** 1906–1908 (2005).
25. Sun, Z. Y. *et al.* HIV-1 broadly neutralizing antibody extracts its epitope from a kinked gp41 ectodomain region on the viral membrane. *Immunity* **28,** 52–63 (2008).
26. Frey, G. *et al.* A fusion-intermediate state of HIV-1 gp41 targeted by broadly neutralizing antibodies. *Proc. Natl Acad. Sci. USA* **105,** 3739–3744 (2008).
27. Baba, T. W. *et al.* Human neutralizing monoclonal antibodies of the IgG1 subtype protect against mucosal simian-human immunodeficiency virus infection. *Nature Med.* **6,** 200–206 (2000).
28. Mascola, J. R. *et al.* Protection of macaques against vaginal transmission of a pathogenic HIV-1/SIV chimeric virus by passive infusion of neutralizing antibodies. *Nature Med.* **6,** 207–210 (2000).
29. Pantaleo, G. *et al.* Major expansion of $CD8^+$ T cells with a predominant V beta usage during the primary immune response to HIV. *Nature* **370,** 463–467 (1994).
30. Koup, R. A. *et al.* Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. *J. Virol.* **68,** 4650–4655 (1994).
31. Borrow, P., Lewicki, H., Hahn, B. H., Shaw, G. M. & Oldstone, M. B. Virus-specific $CD8^+$ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. *J. Virol.* **68,** 6103–6110 (1994).
32. Musey, L. *et al.* Cytotoxic-T-cell responses, viral load, and disease progression in early human immunodeficiency virus type 1 infection. *N. Engl. J. Med.* **337,** 1267–1274 (1997).
33. Kiepiela, P. *et al.* Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* **432,** 769–775 (2004).
34. Kiepiela, P. *et al.* $CD8^+$ T-cell responses to different HIV proteins have discordant associations with viral load. *Nature Med.* **13,** 46–53 (2007).
35. Schmitz, J. E. *et al.* Control of viremia in simian immunodeficiency virus infection by $CD8^+$ lymphocytes. *Science* **283,** 857–860 (1999).
36. Jin, X. *et al.* Dramatic rise in plasma viremia after $CD8^+$ T cell depletion in simian immunodeficiency virus-infected macaques. *J. Exp. Med.* **189,** 991–998 (1999).
37. Phillips, R. E. *et al.* Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. *Nature* **354,** 453–459 (1991).
38. Allen, T. M. *et al.* Tat-specific cytotoxic T lymphocytes select for SIV escape variants during resolution of primary viraemia. *Nature* **407,** 386–390 (2000).
39. Barouch, D. H. *et al.* Eventual AIDS vaccine failure in a rhesus monkey by viral escape from cytotoxic T lymphocytes. *Nature* **415,** 335–339 (2002).
40. Betts, M. R. *et al.* HIV nonprogressors preferentially maintain highly functional HIV-specific $CD8^+$ T cells. *Blood* **107,** 4781–4789 (2006).
41. Precopio, M. L. *et al.* Immunization with vaccinia virus induces polyfunctional and phenotypically distinctive $CD8^+$ T cell responses. *J. Exp. Med.* **204,** 1405–1416 (2007).
42. Darrah, P. A. *et al.* Multifunctional TH1 cells define a correlate of vaccine-mediated protection against *Leishmania major*. *Nature Med.* **13,** 843–850 (2007).
43. Watkins, D. I., Burton, D. R., Kallas, E. G., Moore, J. P. & Koff, W. C. Nonhuman primate models and the failure of the Merck HIV-1 vaccine in humans. *Nature Med.* **14,** 617–621 (2008).
44. Seder, R. A., Darrah, P. A. & Roederer, M. T-cell quality in memory and protection: implications for vaccine design. *Nature Rev. Immunol.* **8,** 247–258 (2008).
45. Chun, T. W. *et al.* Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387,** 183–188 (1997).
46. Chun, T. W. *et al.* Early establishment of a pool of latently infected, resting $CD4^+$ T cells during primary HIV-1 infection. *Proc. Natl Acad. Sci. USA* **95,** 8869–8873 (1998).
47. Douek, D. C. *et al.* HIV preferentially infects HIV-specific $CD4^+$ T cells. *Nature* **417,** 95–98 (2002).
48. Veazey, R. S. *et al.* Gastrointestinal tract as a major site of $CD4^+$ T cell depletion and viral replication in SIV infection. *Science* **280,** 427–431 (1998).
49. Mattapallil, J. J. *et al.* Massive infection and loss of memory $CD4^+$ T cells in multiple tissues during acute SIV infection. *Nature* **434,** 1093–1097 (2005).
50. Li, Q. *et al.* Peak SIV replication in resting memory $CD4^+$ T cells depletes gut lamina propria $CD4^+$ T cells. *Nature* **434,** 1148–1152 (2005).
51. Brenchley, J. M. *et al.* Microbial translocation is a cause of systemic immune activation in chronic HIV infection. *Nature Med.* **12,** 1365–1371 (2006).
52. Mattapallil, J. J. *et al.* Vaccination preserves CD4 memory T cells during acute simian immunodeficiency virus challenge. *J. Exp. Med.* **203,** 1533–1541 (2006).
53. Daniel, M. D., Kirchhoff, F., Czajak, S. C., Sehgal, P. K. & Desrosiers, R. C. Protective effects of a live attenuated SIV vaccine with a deletion in the *nef* gene. *Science* **258,** 1938–1941 (1992).
54. Wyand, M. S., Manson, K. H., Garcia-Moll, M., Montefiori, D. & Desrosiers, R. C. Vaccine protection by a triple deletion mutant of simian immunodeficiency virus. *J. Virol.* **70,** 3724–3733 (1996).
55. Learmont, J. C. *et al.* Immunologic and virologic status after 14 to 18 years of infection with an attenuated strain of HIV-1. A report from the Sydney Blood Bank Cohort. *N. Engl. J. Med.* **340,** 1715–1722 (1999).
56. Baba, T. W. *et al.* Pathogenicity of live, attenuated SIV after mucosal infection of neonatal macaques. *Science* **267,** 1820–1825 (1995).
57. Baba, T. W. *et al.* Live attenuated, multiply deleted simian immunodeficiency virus causes AIDS in infant and adult macaques. *Nature Med.* **5,** 194–203 (1999).
58. Murphey-Corb, M. *et al.* A formalin-inactivated whole SIV vaccine confers protection in macaques. *Science* **246,** 1293–1297 (1989).
59. Wille-Reece, U. *et al.* HIV Gag protein conjugated to a Toll-like receptor 7/8 agonist improves the magnitude and quality of Th1 and $CD8^+$ T cell responses in nonhuman primates. *Proc. Natl Acad. Sci. USA* **102,** 15190–15194 (2005).
60. Wille-Reece, U. *et al.* Toll-like receptor agonists influence the magnitude and quality of memory T cell responses after prime-boost immunization in nonhuman primates. *J. Exp. Med.* **203,** 1249–1258 (2006).
61. Casimiro, D. R. *et al.* Comparative immunogenicity in rhesus monkeys of DNA plasmid, recombinant vaccinia virus, and replication-defective adenovirus vectors expressing a human immunodeficiency virus type 1 *gag* gene. *J. Virol.* **77,** 6305–6313 (2003).
62. Graham, B. S. *et al.* Phase 1 safety and immunogenicity evaluation of a multiclade HIV-1 DNA candidate vaccine. *J. Infect. Dis.* **194,** 1650–1660 (2006).
63. Barouch, D. H. *et al.* Control of viremia and prevention of clinical AIDS in rhesus monkeys by cytokine-augmented DNA vaccination. *Science* **290,** 486–492 (2000).
64. Chong, S. Y. *et al.* Comparative ability of plasmid IL-12 and IL-15 to enhance cellular and humoral immune responses elicited by a SIVgag plasmid DNA vaccine and alter disease progression following SHIV(89.6P) challenge in rhesus macaques. *Vaccine* **25,** 4967–4982 (2007).
65. Luckay, A. *et al.* Effect of plasmid DNA vaccine design and *in vivo* electroporation on the resulting vaccine-specific immune responses in rhesus macaques. *J. Virol.* **81,** 5257–5269 (2007).
66. Liu, J., Kjeken, R., Mathiesen, I. & Barouch, D. H. Recruitment of antigen-presenting cells to the site of inoculation and augmentation of human immunodeficiency virus type 1 DNA vaccine immunogenicity by *in vivo* electroporation. *J. Virol.* **82,** 5643–5649 (2008).
67. Shiver, J. W. *et al.* Replication-incompetent adenoviral vaccine vector elicits effective anti-immunodeficiency-virus immunity. *Nature* **415,** 331–335 (2002).
68. Catanzaro, A. T. *et al.* Phase 1 safety and immunogenicity evaluation of a multiclade HIV-1 candidate vaccine delivered by a replication-defective recombinant adenovirus vector. *J. Infect. Dis.* **194,** 1638–1649 (2006).
69. Amara, R. R. *et al.* Control of a mucosal challenge and prevention of AIDS by a multiprotein DNA/MVA vaccine. *Science* **292,** 69–74 (2001).
70. Harari, A. *et al.* An HIV-1 clade C DNA prime, NYVAC boost vaccine regimen induces reliable, polyfunctional, and long-lasting T cell responses. *J. Exp. Med.* **205,** 63–77 (2008).
71. Shiver, J. W. & Emini, E. A. Recent advances in the development of HIV-1 vaccines using replication-incompetent adenovirus vectors. *Annu. Rev. Med.* **55,** 355–372 (2004).
72. Casimiro, D. R. *et al.* Attenuation of simian immunodeficiency virus SIVmac239 infection by prophylactic immunization with DNA and recombinant adenoviral vaccine vectors expressing Gag. *J. Virol.* **79,** 15547–15555 (2005).
This manuscript demonstrates that homologous rAd5 vaccine regimens were minimally effective against SIV_{MAC239} challenges in rhesus monkeys.

73. Mothe, B. R. *et al.* Expression of the major histocompatibility complex class I molecule Mamu-A*01 is associated with control of simian immunodeficiency virus SIVmac239 replication. *J. Virol.* **77,** 2736–2740 (2003).
74. Pal, R. *et al.* ALVAC-SIV-gag-pol-env-based vaccination and macaque major histocompatibility complex class I (A*01) delay simian immunodeficiency virus SIVmac-induced immunodeficiency. *J. Virol.* **76,** 292–302 (2002).
75. Zhang, Z. Q. *et al.* Mamu-A*01 allele-mediated attenuation of disease progression in simian-human immunodeficiency virus infection. *J. Virol.* **76,** 12845–12854 (2002).
76. Wilson, N. A. *et al.* Vaccine-induced cellular immune responses reduce plasma viral concentrations after repeated low-dose challenge with pathogenic simian immunodeficiency virus SIVmac239. *J. Virol.* **80,** 5875–5885 (2006).
77. Letvin, N. L. *et al.* Preserved $CD4^+$ central memory T cells and survival in vaccinated SIV-challenged monkeys. *Science* **312,** 1530–1533 (2006).
78. Vogels, R. *et al.* Replication-deficient human adenovirus type 35 vectors for gene transfer and vaccination: efficient human cell infection and bypass of preexisting adenovirus immunity. *J. Virol.* **77,** 8263–8271 (2003).
79. Abbink, P. *et al.* Comparative seroprevalence and immunogenicity of six rare serotype recombinant adenovirus vaccine vectors from subgroups B and D. *J. Virol.* **81,** 4654–4663 (2007).
80. Thorner, A. R. *et al.* Age dependence of adenovirus-specific neutralizing antibody titers in individuals from sub-Saharan Africa. *J. Clin. Microbiol.* **44,** 3781–3783 (2006).
81. Kostense, S. *et al.* Adenovirus types 5 and 35 seroprevalence in AIDS risk groups supports type 35 as a vaccine vector. *AIDS* **18,** 1213–1216 (2004).
82. Fauci, A. S. *et al.* HIV vaccine research: the way forward. *Science* **321,** 530–532 (2008).
This perspective describes revised NIH research priorities for HIV-1 vaccine research.
83. Catanzaro, A. T. *et al.* Phase I clinical evaluation of a six-plasmid multiclade HIV-1 DNA candidate vaccine. *Vaccine* **25,** 4085–4092 (2007).
84. Fauci, A. S. NIAID will not move forward with the PAVE 100 HIV vaccine trial. *NIH News* ⟨http://www3.niaid.nih.gov/news/newsreleases/2008/pave100.htm⟩ (2008).
85. Barouch, D. H. *et al.* Immunogenicity of recombinant adenovirus serotype 35 vaccine in the presence of pre-existing anti-Ad5 immunity. *J. Immunol.* **172,** 6290–6297 (2004).
86. Roberts, D. M. *et al.* Hexon-chimaeric adenovirus serotype 5 vectors circumvent pre-existing anti-vector immunity. *Nature* **441,** 239–243 (2006).
87. Farina, S. F. *et al.* Replication-defective vector based on a chimpanzee adenovirus. *J. Virol.* **75,** 11603–11613 (2001).
88. Fitzgerald, J. C. *et al.* A simian replication-defective adenoviral recombinant vaccine to HIV-1 gag. *J. Immunol.* **170,** 1416–1422 (2003).
89. Liu, J. *et al.* Magnitude and phenotype of cellular immune responses elicited by recombinant adenovirus vectors and heterologous prime-boost regimens in rhesus monkeys. *J. Virol.* **82,** 4844–4852 (2008).
90. Barouch, D. H. Novel adenovirus vector-based vaccines for HIV-1. *Keystone Symposia on HIV Vaccines, Banff, Canada*. Abstract X7 009, page 60 (Keystone Symposia, 27 March–1 April 2008).
91. Liao, H. X. *et al.* A group M consensus envelope glycoprotein induces antibodies that neutralize subsets of subtype B and C HIV-1 primary viruses. *Virology* **353,** 268–282 (2006).
92. Weaver, E. A. *et al.* Cross-subtype T-cell immune responses induced by a human immunodeficiency virus type 1 group m consensus env immunogen. *J. Virol.* **80,** 6745–6756 (2006).
93. Fischer, W. *et al.* Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. *Nature Med.* **13,** 100–106 (2007).
This manuscript proposes the use of polyvalent 'mosaic' antigens to improve immunologic coverage of global HIV-1 diversity.

Acknowledgements The author would like to thank R. Dolin, N. Letvin, J. Mascola and J. McElrath for critically reviewing this manuscript. The author acknowledges support from the National Institutes of Health and the Bill & Melinda Gates Foundation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to D.H.B. (dbarouch@bidmc.harvard.edu).

Photoemission kinks and phonons in cuprates

Arising from: F. Giustino, M. L. Cohen & S. G. Louie *Nature* 452, 975–978 (2008)

One of the possible mechanisms of high transition temperature (T_c) superconductivity is Cooper pairing with the help of bosons, which change the slope of the electronic dispersion as observed by photoemission. Giustino *et al.*[1] calculated that in the high temperature superconductor $La_{1.85}Sr_{0.15}CuO_4$ crystal lattice vibrations (phonons) should have a negligible effect on photoemission spectra and concluded that phonons do not have an important role. Here we show that the calculations used by Giustino *et al.*[1] do not reproduce the huge influence of electron–phonon coupling on important phonons observed in experiments. Thus, we would similarly expect that these calculations do not explain the role of electron–phonon coupling for the electronic dispersion.

Density functional theory (DFT) calculations used by Giustino *et al.*[1] treat electrons and phonons as independent entities, which scatter each other. Because of this scattering, the electronic states acquire finite lifetimes and abrupt changes in dispersions (kinks) at the phonon energies. In addition, the phonons soften and broaden in energy. These effects are calculated from first principles without adjustable parameters. Therefore, if DFT is appropriate for the high T_c cuprates it ought to accurately reproduce the electronic contribution to phonon softening and broadening deduced from neutron or X-ray scattering experiments.

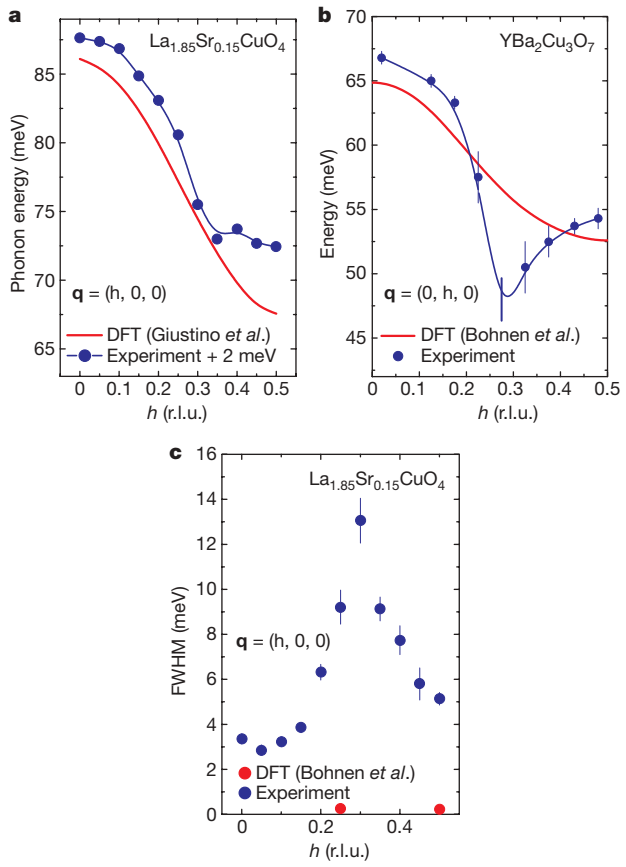

Figure 1 | Comparison of DFT predictions with experimental results for $La_{1.85}Sr_{0.15}CuO_4$ and $YBa_2Cu_3O_7$ at 10 K. a, b, Experimental bond-stretching phonon dispersions[2–4] compared to DFT results[1,12]. The data in **a** are shifted by 2 meV. **c,** Phonon line widths in $La_{1.85}Sr_{0.15}CuO_4$ (refs 2 and 3) compared with DFT results (K.-P. Bohnen, personal communication) on $YBa_2Cu_3O_7$. Giustino *et al.*[1] contains no line-width results for $La_{1.85}Sr_{0.15}CuO_4$ but we expect them to be similar. Error bars represent s.d.; **q** represents reduced wavevector in reciprocal lattice units (r.l.u.). FWHM, full-width at half-maximum.

DFT predicts that the phonon branch, in large part responsible for the calculated electronic dispersion kink, is the optical bond-stretching branch involving the bond-stretching motion of CuO_2 plane oxygen against copper[1]. Several experimental papers have highlighted large anomalous renormalization of these phonons[2–6]. They have huge low temperature dispersion dips and/or line-width maxima around half-way ($h = 0.3$) to the zone boundary in the superconductors $La_{1.85}Sr_{0.15}CuO_4$ (refs 2 and 3) and $YBa_2Cu_3O_7$ (ref. 4). However, DFT predicts a smooth dispersion without any pronounced features in either the dispersion or line width around $h = 0.3$ (Fig. 1). Furthermore, the very small calculated line widths in Fig. 1b illustrate that the calculated electron–phonon coupling is very weak in absolute terms.

Substantial evidence points to an electronic origin of the phonon effect. First, the phonon anomaly weakens at elevated temperatures[2,3], whereas alternatives such as phonon–phonon scattering and structural inhomogeneity should either show the opposite trend or have no temperature dependence. Second, the phonon effect appears at specific wavevectors and is phenomenologically similar to anomalies observed in conventional systems with strong electron–phonon coupling. Third, both phonon renormalization[2,6] and the photoemission kink[7] become bigger when hole concentration decreases from high doping (in which superconductivity is suppressed) towards so-called 'optimal' doping with the maximum superconducting T_c. This simultaneous enhancement of the two features may result from an increase in electron–phonon coupling due to enhanced electronic correlations or reduced screening not included in DFT. The findings of Giustino *et al.*[1] cannot rule out such hypotheses. The same holds for $YBa_2Cu_3O_7$ where there is a similar disagreement between the experimental and DFT results for both the phonon dispersions and the photoemission kink[8].

It is notable that many-body calculations predict a considerable enhancement of the coupling to bond-stretching phonons compared to DFT and describe anomalous doping dependence of the zone boundary phonons[9,10], suggesting that strong correlation effects might be relevant. Recent high resolution photoemission measurements have found an oxygen isotope effect in the dispersion kink at the half-breathing phonon energy, hinting at an important role of oxygen phonons[11]. We conclude that more work is necessary to establish phonon contribution to the photoemission kink.

D. Reznik[1], G. Sangiovanni[2], O. Gunnarsson[3] & T. P. Devereaux[3]

[1]Forschungszentrum Karlsruhe, Institut für Festkörperphysik, PO Box 3640, D-76021 Karlsruhe, Germany.
e-mail: reznik@llb.saclay.cea.fr

[2]Max-Planck-Institut für Festkörperforschung, D-70506 Stuttgart, Germany.

[3]Department of Photon Science, Stanford Linear Accelerator Center, Stanford University, 2575 Sand Hill Road, Menlo Park, California 94025, USA.

Received 1 July; accepted 20 August 2008.

1. Giustino, F., Cohen, M. L. & Louie, S. G. Small phonon contribution to the photoemission kink in the copper oxide superconductors. *Nature* **452**, 975–978 (2008).
2. Reznik, D. *et al.* Electron–phonon coupling reflecting dynamic charge inhomogeneity in copper oxide superconductors. *Nature* **440**, 1170–1173 (2006).
3. Reznik, D. *et al.* Electron–phonon anomaly related to charge stripes: Static stripe phase versus optimally doped superconducting $La_{1.85}Sr_{0.15}CuO_4$. *J. Low Temp. Phys.* **147**, 353–364 (2007).
4. Pintschovius, L. *et al.* Oxygen phonon branches in $YBa_2Cu_3O_7$. *Phys. Rev. B* **69**, 214506 (2004).
5. Uchiyama, H. *et al.* Softening of Cu-O bond stretching phonons in tetragonal $HgBa_2CuO_{4+\delta}$. *Phys. Rev. Lett.* **92**, 197005 (2004).

6. Pintschovius, L., Reznik, D. & Yamada, K. Oxygen phonon branches in overdoped $La_{1.7}Sr_{0.3}Cu_3O_4$. *Phys. Rev. B* **74,** 174514 (2006).
7. Zhou, X. J. *et al.* Universal nodal Fermi velocity. *Nature* **423,** 398 (2003).
8. Heid, R., Bohnen, K.-P., Zeyher, R. & Manske, D. Momentum dependence of the electron-phonon coupling and self-energy effects in superconducting $YBa_2Cu_3O_7$ within the local density approximation. *Phys. Rev. Lett.* **100,** 137001 (2008).
9. Rösch, O. & Gunnarsson, O. Electron–phonon interaction in the three-band model. *Phys. Rev. B* **70,** 224518 (2004).
10. Horsch, P. & Khaliullin, G. Doping dependence of density response and bond-stretching phonons in cuprates. *Physica B* **359–361,** 620–622 (2005).
11. Iwasawa, H. *et al.* An isotopic fingerprint of electron–phonon coupling in high-T_c cuprates. Preprint at <http://arxiv.org/abs/0808.1323> (2008).
12. Bohnen, K.-P., Heid, R. & Krauss, M. Phonon dispersion and electron–phonon interaction for $YBa_2Cu_3O_7$ from first-principles calculations. *Europhys. Lett.* **64,** 104–110 (2003).

doi:10.1038/nature07364

ARTICLES

Speciation through sensory drive in cichlid fish

Ole Seehausen[1,2], Yohey Terai[3], Isabel S. Magalhaes[1,2], Karen L. Carleton[4], Hillary D. J. Mrosso[5], Ryutaro Miyagi[3], Inke van der Sluijs[6]†, Maria V. Schneider[2]†, Martine E. Maan[6]†, Hidenori Tachida[7], Hiroo Imai[8] & Norihiro Okada[3]

Theoretically, divergent selection on sensory systems can cause speciation through sensory drive. However, empirical evidence is rare and incomplete. Here we demonstrate sensory drive speciation within island populations of cichlid fish. We identify the ecological and molecular basis of divergent evolution in the cichlid visual system, demonstrate associated divergence in male colouration and female preferences, and show subsequent differentiation at neutral loci, indicating reproductive isolation. Evidence is replicated in several pairs of sympatric populations and species. Variation in the slope of the environmental gradients explains variation in the progress towards speciation: speciation occurs on all but the steepest gradients. This is the most complete demonstration so far of speciation through sensory drive without geographical isolation. Our results also provide a mechanistic explanation for the collapse of cichlid fish species diversity during the anthropogenic eutrophication of Lake Victoria.

The sensory drive hypothesis for speciation[1,2] predicts that adaptation in sensory and signalling systems to different environments in allopatry may cause premating isolation on secondary contact of populations. Recent theoretical work suggested that sensory drive can lead to the evolution of colour polymorphisms[3,4] and speciation[5], even in the absence of geographical isolation, when the light environment is heterogeneous. However, the only case of sympatric sister species, in which assortative mating has been shown to be facilitated by sensory drive, were sticklebacks in British Columbia[6]. Here we provide ecological, population genetic and molecular evidence for each of the predictions of sensory drive speciation[2] in sympatric cichlid fish inhabiting light gradients in Lake Victoria (East Africa).

Lake Victoria is spatially highly heterogeneous in water clarity and ambient light[7,8], and there is much evidence that the cichlid visual system has been under strong diversifying selection during the adaptive radiation of cichlids into several hundred species in Lake Victoria[9]. Vertebrate visual pigments consist of a light-absorbing component, the chromophore, and a protein moiety, the opsin[10]. Spectral sensitivity is determined by the chromophore (A1 or A2 pigments), and by its interaction with the amino acid residues lining the retinal-binding pocket of the opsin in which the chromophore lies[11]. Eight different visual pigments have been found in all haplochromine cichlids[12–14], but only a subset of these is expressed in any individual species[12,14,15]. Several *Pundamilia* species from Lake Victoria expressed the same complement of four opsin genes: short-wavelength-sensitive opsin gene 2a (*SWS2A*, λ_{max} ~455 nm) in single cones; rhodopsin-like (*RH2*, λ_{max} ~528nm) and long-wavelength-sensitive opsin gene (*LWS*, λ_{max} ~565 nm) in double cones; and rhodopsin (*RH1*, λ_{max} ~505 nm) in rods[16]. Of these, the *LWS* opsin gene is by far the most variable among Lake Victoria cichlids[13,17], with sequence variation being five times greater than in Lake Malawi cichlids despite a tenfold greater age of the latter species flock[18].

Female Lake Victoria cichlids have mating preferences for conspicuously coloured males[19]. Perception of conspicuousness is influenced by ambient and background light, signal transmission, receiver sensitivity and higher level processing[2,20]. Sympatric pairs of closely related cichlid species, one with red and one with blue nuptial colouration (Fig. 1 and Supplementary Fig. 3), are common in Lake Victoria[8]. Visual pigments have been compared for three pairs, and behavioural light detection thresholds measured in two. In each pair, the red species has its LWS λ_{max} at a longer wavelength[16,21], with a lower detection threshold for red but a higher one for blue light[22,23]. These observations are consistent with a role for sensory drive in speciation, whereby interaction between ambient light, natural-selection-driven divergence of visual sensitivities and sexual selection for conspicuous male colours leads to the fixation of different male colours[1,2,16,23].

Examining the role of environmental gradients in speciation requires tests at replicate gradients, as is recognized both in evolutionary ecology[24–26] and in population genomics[27]. A recent model of clinal speciation through sensory drive[5], as well as other models of clinal speciation[28–30], predicts the greatest probability of speciation on gradients of intermediate slope. There, migration rates are sufficiently low to be compensated for by selection, but are sufficiently high to generate significant migration load[31] and intermediate genotypes with a poor fit to the local environment. Migration load and reduced fitness of intermediate genotypes lead to disruptive selection, which may be required for the evolution of assortative mating through reinforcement-like mechanisms[28–30]. Previously we demonstrated adaptive evolution in the *LWS* opsin gene of the Lake Victoria cichlid fish *Neochromis greenwoodi* and *Mbipia mbipi* along very shallow gradients of light colour mediated by variation in turbidity

[1]Institute of Zoology, University of Bern, Baltzerstr. 6, CH-3012 Bern, Switzerland. [2]Eawag, Swiss Federal Institute for Aquatic Science and Technology, Centre of Ecology, Evolution & Biogeochemistry, Department of Fish Ecology & Evolution, 6047 Kastanienbaum, Switzerland. [3]Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama 226-8501, Japan. [4]Department of Biology, University of Maryland, College Park, Maryland 20742, USA. [5]Tanzania Fisheries Research Institute, Mwanza Centre, PO Box 475 Mwanza, Tanzania. [6]Department of Animal Ecology, Institute of Biology, Leiden University, PO Box 9516, 2300 RA Leiden, The Netherlands. [7]Department of Biology, Faculty of Sciences, Kyushu University, Ropponmatsu, Fukuoka 810-8560, Japan. [8]Department of Cellular and Molecular Biology, Primate Research Institute, Kyoto University, 484-8506 Japan. †Present addresses: Department of Biology, McGill University, 1205 Avenue Docteur Penfield, Montréal, Québec H3A 1B1, Canada (I.v.d.S.); The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK (M.V.S.); University of Texas at Austin, Integrative Biology, 1 University Station C0930, Austin, Texas 78712, USA (M.E.M.).

between islands[9]. *LWS* genotype frequencies and male colour morph frequencies formed correlated clines, but, even though populations at opposite ends of one gradient fixed different *LWS* alleles, all populations retained polymorphism for colour, indicating that speciation remained incomplete[9].

Here we investigate populations of cichlid fish living on light gradients primarily mediated by water depth within islands in Lake Victoria. *Pundamilia pundamilia* and *Pundamilia nyererei*[32] (Fig. 1a and Supplementary Fig. 3) are geographically fully sympatric. Within islands, they have narrowly parapatric depth ranges. Where they are phenotypically well differentiated, *P. pundamilia* has blue–grey male nuptial colouration whereas *P. nyererei* nuptial males are yellow with a bright crimson-red dorsum. Females of both are cryptically yellowish and have mating preferences for the nuptial colouration of conspecific males[33,34]. The red *P. nyererei* occurs at greater mean water depths, in more red-shifted ambient light than the blue *P. pundamilia*[23]. *P. nyererei* have a lower threshold for the detection of red light, whereas *P. pundamilia* possess a lower threshold for detection of blue light[23]. Earlier we found that red and blue fish tended to possess different alleles at the *LWS* opsin gene locus[16]. Here we fully develop this system to test predictions of sensory drive speciation.

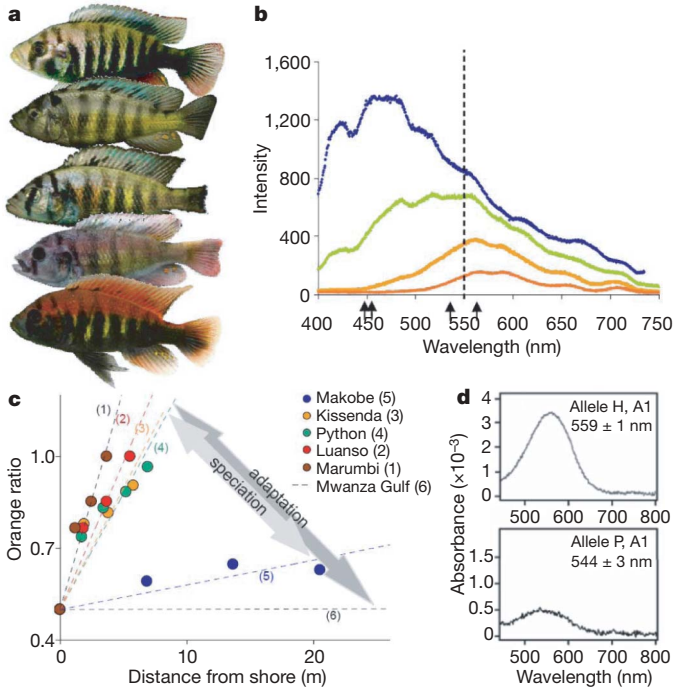

Figure 1 | Male phenotypes, light gradients and *LWS* opsin absorbance. **a**, Variation in male nuptial colouration. Five phenotype classes from 0 ('blue', typical *P. pundamilia*; top) to 4 ('red', typical *P. nyererei*; bottom). **b**, An example of a moderately steep light gradient (Python island): surface light spectrum (blue) and three subsurface light spectra measured at 0.5 m (green), 1.5 m (orange) and 2.5 m (red) water depth. The line through 550 nm indicates the divide used to calculate the transmittance orange ratio. Arrows indicate peak absorbance of two opsin pigments: main allele groups at *LWS* opsin locus (544 nm and 559 nm) and known range of peak absorbance at *SWS2A* locus[16]. **c**, Slopes of seven different light gradients. The lines for two shallow gradients overlay each other and are together labelled 'Mwanza Gulf'. For this line, the *x*-axis represents the distance from clear water (rather than from shore). Significant differentiation in opsin genes was observed on all gradients with slopes equal to or shallower than the Kissenda (orange) line, but speciation was observed only on gradients with slopes between the Kissenda (orange) and the Makobe (blue) lines. The dark grey arrow indicates region with divergent adaption at *LWS* opsin gene, and the light grey arrow indicates region with speciation. **d**, Absorption spectra of *LWS* pigments evaluated by the dark–light difference spectra[9]. The *LWS* pigments were reconstituted from the H allele with A1 retinal (top) and from the P allele with A1 retinal (bottom). λ_{max} values (with standard errors) are indicated.

If sensory drive caused speciation into a red and a blue species, we expected to find: (1) variation in the *LWS* opsin sequence at amino acid positions where they shift λ_{max}; (2) an association of such sequence variation with water depth, such that more red-shifted alleles occur at greater depth; and (3) an association of *LWS* alleles with the predominant male nuptial colouration of a population, such that populations with predominantly red-shifted opsin alleles have predominantly red males. Furthermore, if disruptive selection was required to complete speciation through the evolution of assortative mating, we predicted that the strongest associations between *LWS* alleles, water depth and colour occur on intermediate light slopes (prediction (4)). For testing prediction (4), we compared the data from the depth-mediated gradients of this study with data we had collected earlier on populations occupying the same depth at different islands with different turbidities[9] (see Supplementary Information).

Light, depth and colour

We examined depth-mediated light gradients at five islands. The light climate of Lake Victoria is dominated by effects of particulate (non-phytoplankton) matter, selectively absorbing and scattering light of short wavelengths[35], causing successive shifts of ambient light towards longer wavelengths with increasing water depth (this study), and also with increasing turbidity (earlier study)[7,8]. The rate at which ambient light changes with increasing depth is positively correlated with turbidity[8] (difference between islands in this study). The cichlids we study feed and breed right above and within the rocky substrate. We characterize depth-associated light gradients in their habitat by the change in the 'transmittance orange ratio' that occurs per metre as one moves outwards from the shore into the lake along the lake floor (the 'light slope', see Methods and Fig. 1b). Steeper slopes occur with more turbid water and steeper shores (Table 1). The steepest light slopes occurred at the most turbid sites, Marumbi and Luanso islands (Table 1 and Fig. 1c). Intermediate slopes occurred at Kissenda and Python islands, and the shallowest slope at Makobe island. The latter was still steeper than all the turbidity-mediated light slopes of our earlier work[9]. The size of the light differential between the ends of the gradients was similar between the five depth-mediated gradients, and larger than on the turbidity-mediated gradients (Table 1 and Supplementary Table 1).

Mapping the microdistribution of phenotypes on the five depth-mediated gradients using data from 960 males (Fig. 2a) revealed significant differences between islands. It showed the absence of any association between colour and ambient light (water depth) at Marumbi and Luanso (analysis of variance, ANOVA: $df = 2$, $F = 1.1$, $P = 0.3$, and $df = 2$, $F = 0.3$, $P = 0.7$, respectively), but significant associations at all other sites (ANOVA: $df = 2$ (Kissenda), $df = 1$ (Python, Makobe), $F > 50$, $P < 0.0001$), and increasing strength of association with decreasing light slope (F ratio against slope, logarithmic regression, $df = 4$, $R^2 = 0.87$, $P = 0.021$; Fig. 3). Blue phenotypes are associated with shallow waters (<3 m) in all locations, whereas red phenotypes occur in shallow waters only on the steepest gradients, and become restricted to greater depths with decreasing light slope. Frequency distributions of male nuptial colour phenotypes differ significantly between islands too (Fig. 2b). Distributions are unimodal and skewed towards blue on the two steepest gradients. They are bimodal with few intermediates on gradients of intermediate slope, and consist of two discrete classes, blue and red, on the shallowest within-island gradient.

Table 1 | The five environmental gradients of this study

Island	Water clarity (cm Secchi) (mean ± s.d.)	Shoreline slope (mean ± s.d.)	Light slope	Light differential
Marumbi island	53 ± 8	0.82 ± 0.15	1.4×10^{-1}	0.50
Luanso island	50 ± 10	0.54 ± 0.05	9.6×10^{-2}	0.50
Kissenda island	78 ± 24	0.52 ± 0.12	7.9×10^{-2}	0.50
Python island	96 ± 21	0.58 ± 0.24	7.6×10^{-2}	0.50
Makobe island	225 ± 67	0.15 ± 0.04	8×10^{-3}	0.35

LWS gene variation, light and colour

We observed 13 polymorphic sites (3 synonymous, 10 nonsynonymous) among the *LWS* sequences (Supplementary Table 6). Three nonsynonymous substitutions occurred at high frequencies. From the bovine rhodopsin crystal structure[36] we inferred that two of these variable amino acid positions, 216 (nucleotide site 647) and 275 (823 and 824), are located in or near the retinal-binding pocket. The third one was position 230 (688), one of the tuning sites of human red/

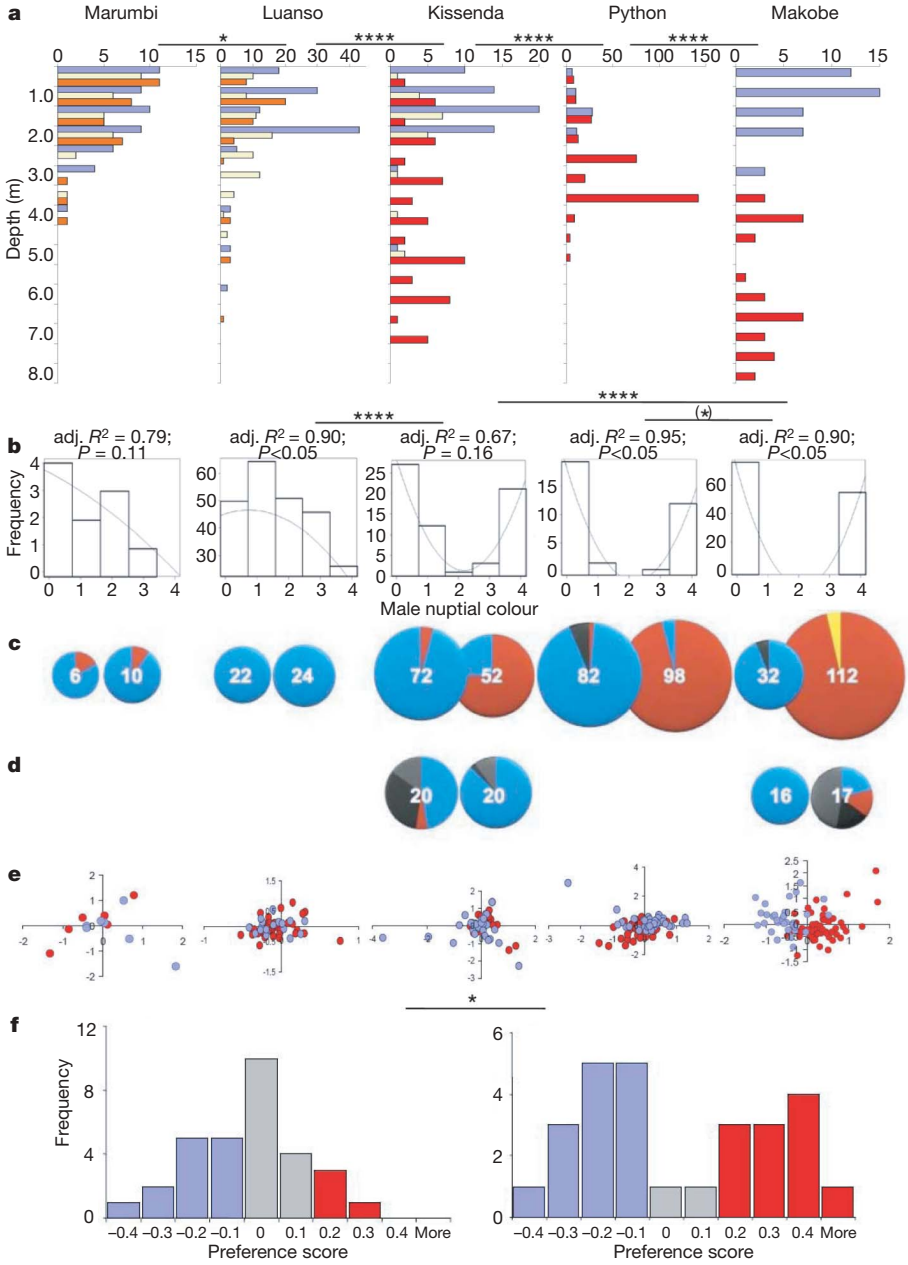

Figure 2 | Ecological, phenotypic, genetic and behavioural differentiation between blue and red *Pundamilia* nuptial phenotypes at five islands. All data for the same island are presented in the same column. Significant differences between islands indicated by asterisks (all tests two-tailed): $*P < 0.05$, $****P < 0.0001$, $(*)P < 0.1$. **a**, Depth distributions of male nuptial colour phenotypes. Blue bars, blue; pale yellow bars, intermediate; and orange or red bars, red (orange if dominated by class 3; red if dominated by class 4). Significance levels of differences between islands in the divergence between red and blue reported as *P* values of *G*-tests. **b**, Frequency distributions of male nuptial colour phenotypes (see Fig. 1a and text). Lines are quadratic fits; R^2 and significance levels indicated. Significance levels of differences between islands reported as *P* values of *G*-tests. **c**, Frequencies of functional allele groups at the *LWS* opsin gene by island and male colour (left, blue; right, red). Numbers report sample sizes of completely sequenced haplotypes. For Marumbi and Luanso islands, only the haplotypes of those individuals are included that could be assigned to 'blueish' and 'reddish' phenotypes (altogether 24 and 54 haplotypes were sequenced from Marumbi and Luanso, respectively). Fish from Marumbi were divided into classes 0 + 1 and classes 2 + 3. Fish from Luanso were divided into classes 0 + 1 and 2–4. At all other islands, only fish of phenotype classes 0 and 4 were included. Alleles of the P group shown in blue, alleles of the H group in red, M3 alleles in yellow, and other alleles in grey. **d**, Allele frequencies at the *SWS2A* opsin gene by island and nuptial colour class. The *SWS2A* P allele shown in blue, the N allele in red, other alleles in black, and alleles not determined in grey. **e**, Individuals plotted on first and second axes of a factorial correspondence analysis of genetic variance calculated from 11 unlinked microsatellite loci. Colours indicate pooled male nuptial colour classes as described in **c**. **f**, Histograms of female mating preferences at Luanso island[41] (left) and Python island[40] (right, includes new data). Blue, preference classes in which most females had statistically significant individual preferences for blue males; red, preference classes in which most females had significant preferences for red males; grey, preference classes in which females had no significant mating preference. Significance level of the difference in the frequency distributions between the two islands reported as *P* value of a *G*-test.

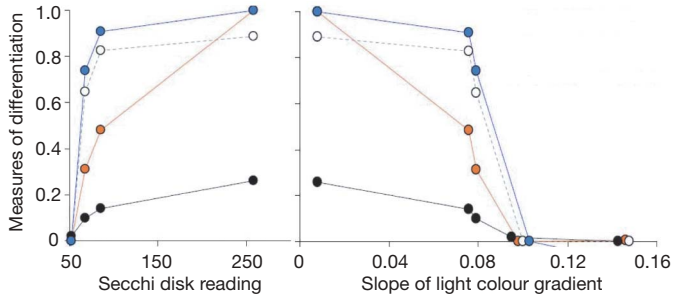

Figure 3 | Measures of differentiation between sympatric *Pundamilia* phenotypes plotted against water transparency (left) and light slope (right). Blue symbols and line: Spearman rank correlations between colour and *LWS* genotype (best fit to water clarity $R^2 = 0.79$, $P = 0.045$, d$f = 4$; best fit to light slope $R^2 = 0.69$, P(one-tailed) $= 0.042$, d$f = 4$.). Open symbols and dashed line: *LWS* F_{ST} between red and blue phenotypes (best fit to water clarity, $R^2 = 0.79$, $P = 0.044$, d$f = 4$; best fit to light slope, $R^2 = 0.65$, P(one-tailed) $= 0.045$, d$f = 4$). Filled orange symbols and orange line: association between colour and water depth (ANOVA F ratios (normalized to range 0–1 for display); best fit to water clarity $R^2 = 0.99$, $P = 0.000$; best fit to light slope $R^2 = 0.87$, $P = 0.021$; both d$f = 4$). Filled black symbols and black line: microsatellite F_{ST} (multiplied by 10 for display) between red and blue phenotypes (best fit to water clarity, $R^2 = 0.99$, $P = 0.000$; best fit to light slope, $R^2 = 0.90$, P(one-tailed) $= 0.013$; both d$f = 4$).

green opsin[37]. Focusing on these three positions, we divided alleles into three groups described previously[9]: the H group (all alleles with 216Y, 230A, 275C), the P group (216F, 230T, 275I) and the M3 group (216Y, 230T, 275I). M3 alleles can be considered recombinants or intermediate between H and P alleles. H and P alleles differed in only the 3 amino acid positions 216, 230 and 275. Substitutions at the other 7 nonsynonymous sites were rare and resulted in other allele variants (Supplementary Table 5).

We reconstituted the LWS pigments from P alleles *in vitro* with A1-derived retinal, and measured their absorption spectra, as previously done for the H alleles[9] (Fig. 1d). The peak spectral sensitivity (λ_{max}) of the A1 pigment of the P allele was blue-shifted by 15 nm relative to the H allele. The λ_{max} values of cone outer segments expressing either P or H pigments were measured previously by microspectrophotometry, reporting too that P pigments were blue-shifted relative to H pigments[16]. Hence, the absorption spectra of P and H alleles seem to be adapted to shallower and deeper water light environments in Lake Victoria, respectively (Fig. 1b, d), supporting prediction (1).

Light gradients with slopes steeper than 0.09 were inhabited by populations with one or two different *LWS* alleles, whereas up to six different alleles were present on less steep gradients (Table 1 and Supplementary Tables 2 and 6). On these gradients of steepness $0.008 \leq x \leq 0.09$, H alleles were strongly associated with red nuptial colouration ($\chi^2 > 66$, d$f = 1$, $P < 0.0001$; Spearman correlation coefficients 0.74, 0.91 and 1, respectively, for slopes 0.079, 0.076 and 0.008; $P < 0.0001$), and were rare in blue phenotypes (Fig. 2c, Supplementary Table 6 and Supplementary Information), supporting prediction (3).

A strong association between *LWS* alleles and water depth emerges from these results, supporting prediction (2): at Marumbi and Luanso islands, most individuals reside in waters less than 3 m deep. P alleles strongly dominate. At all other islands, only the blue phenotype is confined to depths less than 3 m, and P alleles predominate among these fish, even where gene exchange with the red phenotype is frequent (see later). The sweep to high frequency of H alleles in the red phenotype is associated with shifting larger fractions of the population to depths beyond 3 m. At Kissenda island, 75% of the *LWS* alleles of the red population belong to the red-shifted H group. The proportion of H alleles, residing in red individuals, increases to Python island and further to Makobe island, associated with successively increasing fractions of the red population living in deep water (Fig. 2a versus 2c). Red and blue phenotypes were highly significantly differentiated at the *LWS* locus at Kissenda, Python and Makobe islands (F_{ST} (fixation index) 0.65, 0.83, 0.89), but neither at Luanso nor at Marumbi islands (F_{ST} 0.00).

Gene flow at neutral loci

A sensory drive model of speciation predicts that the rate of divergence at the opsin loci should exceed the rate of divergence at neutral loci. Our data are fully consistent with this prediction (Table 2). Pairwise F_{ST} between sympatric blue and red phenotypes estimated from 11 microsatellite loci reveal no differentiation at Marumbi or Luanso islands (Fig. 2e), consistent with the unimodal frequency distributions of male nuptial colour variants and the absence (Marumbi) or rarity of really red males. Pairwise F_{ST} at all other islands suggest significant, albeit weak, differentiation, consistent with the strongly bimodal frequency distributions of male nuptial colour variants and the emergence of the really red phenotype at those islands. Whereas F_{ST} at the *LWS* locus jumps from 0 at Marumbi and Luanso to 0.65 at Kissenda, F_{ST} at microsatellite loci increases gradually and much more slowly (Figs 2c, e and 3). The number of microsatellite loci carrying the signature of differentiation increases steadily from Marumbi and Luanso (0 out of 11) to Makobe island (7 out of 11; Table 2), consistent with the successive disappearance of intermediate phenotypes.

With the exception of Makobe island, all microsatellite F_{ST} among sympatric red and blue phenotypes are smaller than F_{ST} between any two allopatric populations of the blue phenotype, and 7 out of 10 of the red phenotype (Supplementary Fig. 1a and Supplementary Table 3). Even the largest between-phenotype F_{ST} at Makobe is smaller than most within-phenotype F_{ST} between islands. This suggests either more

Table 2 | Pairwise F_{ST} statistics between sympatric phenotypes

Island	Marumbi island	Luanso island	Kissenda island	Python island	Makobe island
Light slope	0.144	0.096	0.079	0.076	0.008
F_{ST} at *LWS* opsin locus	0.000	0.000	**0.648**	**0.826**	**0.890**
F_{ST} at microsatellite loci					
Ppun21	0.000	0.000	0.010	0.006	**0.023**
Ppun7	0.000	0.000	0.003	**0.023**	**0.013**
Ppun5	0.000	0.002	0.002	0.000	0.010
Ppun32	0.041	0.005	0.000	0.016	**0.080**
Ppun17	0.000	0.000	0.000	**0.046**	**0.027**
OSU16d	0.017	0.002	0.011	0.006	**0.020**
OSU20d	0.002	0.000	**0.040**	0.004	**0.008**
OSU19t	0.000	0.022	0.013	0.013	**0.032**
TMO5	0.000	0.013	0.000	0.012	0.010
Pzeb3	0.000	0.000	0.002	0.048	0.107
Pzeb5	0.000	0.000	0.024	0.024	0.049
Multilocus (11μsats)	0.000	0.002	**0.010**	**0.014**	**0.026**

Significant F_{ST} ($P < 0.05$) are shown in bold.

gene flow or more recent divergence between phenotypes within islands than between island populations of the same phenotype. It implies either parallel maintenance of phenotypic differentiation in the face of gene flow, or parallel sympatric speciation. All H alleles as well as the most frequent P allele are shared with several distantly related cichlid species (Supplementary Fig. 4). The two *Pundamilia* H alleles are the most frequent H alleles in those distantly related species too. Either red *Pundamilia* populations acquired these alleles once or multiple times from other species through introgressive hybridization, or the shared ancestor of red and blue *Pundamilia* possessed all the P and H alleles. In either scenario, the H and P allele split must pre-date the origin of the blue and red *Pundamilia* species.

Selection on the *LWS* gene

We analysed sequences up- and down-stream of *LWS* in a population (Python) that exhibits strong divergence in *LWS* but only weak differentiation at microsatellite loci. Sliding-window F_{ST} analysis revealed at least 6 times greater divergence in the *LWS* gene exons and in 2 kilobases (kb) of upstream sequence ($F_{ST} > 0.8$; Supplementary Fig. 2a) than in the downstream sequences ($F_{ST} < 0.15$), and more than 50 times greater divergence than at microsatellite loci (Table 2). Together with results of McDonald tests[38] and HKA tests[39] (Supplementary Table 4 and Supplementary Information), this is consistent with a recent selective sweep in the red species, associated with increased presence in a red-shifted environment.

Divergence in the *SWS2A* opsin gene

We sequenced the *SWS2A* opsin gene at two islands to test for divergence at the short-wavelength end of the light spectrum. Out of 10 variable nucleotide positions, 5 were synonymous and 5 were located in introns (Supplementary Table 7). At Kissenda, the *SWS2A* sequences were variable in both phenotypes, and differentiated between them (F_{ST} 0.1, $P < 0.01$). At Makobe, a single *SWS2A* sequence variant was almost fixed in *P. pundamilia*, and the species were more strongly differentiated, although not as strongly as in *LWS* (F_{ST}: 0.437, $P < 0.001$; Fig. 2d).

Female mating preferences

Experiments and field data suggest that female *Pundamilia* use male colour as an important mate choice cue[19,33,34]. Most wild and laboratory-bred Python island females prefer either blue or red males, but laboratory-bred F_1-hybrid females, most laboratory-bred F_2-hybrid females and most Luanso females have no preference between blue and red males[40,41]. Combining published data[40,41] with previously unpublished data for 11 females from Python island, we find that the frequency distributions of female mating preferences differ between the islands (*G*-test, $P = 0.02$), roughly resembling those of male nuptial colour (compare Fig. 2f with 2b). The distribution at Luanso (38 females) had a single mode on no preference, and a skew towards blue preference. The distribution at Python (27 females) was bimodal.

We analysed Python island non-hybrid and laboratory-bred F_2 hybrid females to ask whether the *LWS* genotype directly determines mating preference. For non-hybrids and hybrids combined, we observed a significant association between individual *LWS* genotype and mating preference ($\chi^2 = 22$, $df = 10$, $P = 0.03$, 10,000 randomizations). However, this relationship was not significant when restricted to F_2 hybrid females ($\chi^2 = 10.2$, $df = 6$, $P = 0.13$, 10,000 randomizations). Hence, variation in the *SWS2A-SWS2B-LWS* chromosomal region alone does not strongly predict visual mating preferences in a laboratory environment: some component of mating preference seems independent of it, consistent with biometric estimates that implied that the difference in mating preferences between *P. pundamilia* and *P. nyererei* was due to more than one factor[40]. Modelling light detection, using solar spectrum, water transmission, *Pundamilia* colour patch reflection and *Pundamilia* visual pigment absorption, suggested that a λ_{max} shift of 4 nm towards longer wavelengths causes a 10% increase in quantum catch for a fish looking at a red patch[16]. It seems probable that, in interaction with ambient light in the natural environment, the opsin genotype more strongly determines mating preference than it does under standard laboratory light conditions.

Discussion

Our data on ambient light colour, male nuptial colour, visual pigment λ_{max} and female mating preference indicate sensory drive speciation, which occurred or is maintained by selection without geographical isolation. However, we only observed this under a restricted range of environmental conditions. At all sites with moderately shallow to moderately steep light gradients, two differentiated populations emerged with strong associations between water depth, *LWS* alleles, colouration and preferences (Fig. 3). Strong bimodalities in the quantitative traits colour and preference, strong heterozygote deficiencies at the *LWS* opsin gene, and differentiation at microsatellite loci clearly indicate speciation initiated by strong selection on *LWS*. Very steep light gradients, in contrast, were inhabited by single panmictic populations that showed little variation in *LWS*, even though they contained some variation in colour and mating preference.

The following sensory drive speciation scenario is fully consistent with our data. First, divergent natural selection between light regimes at different water depths acts on *LWS*. Second, sexual selection for conspicuous colouration is also divergent because perceptual biases differ between light regimes. Third, their interaction generates initial deviation from linkage equilibrium between *LWS* and nuptial colour alleles as observed on all but the steepest gradients. Fourth, subsequent disruptive selection due to reduced fitness of genotypes with a mismatch between *LWS* and colour alleles causes speciation, perhaps involving reinforcement-like selection for mating preferences, whereby male nuptial colour may serve as a marker trait for opsin genotype.

The strong association between *LWS* alleles and male nuptial colouration with few or no mismatch genotypes in sympatric species pairs is not restricted to *P. pundamilia* and *P. nyererei* (Table 3, Supplementary Fig. 3 and Supplementary Information). In contrast with these results, we did not find any such discontinuities in the frequency distribution of opsin genotypes along very shallow (between-island) gradients investigated earlier[9]—that is, intermediate *LWS* genotypes predominated in large sections of each gradient. This suggested the presence of divergent selection but the absence of disruptive selection (or the absence of an evolutionary response to disruptive selection). This is consistent with the low migration load predicted from the very small difference in ambient light that migrants between adjacent islands experience (Supplementary Table 1). Despite positive correlations between frequencies of *LWS* alleles and male nuptial colour morphs, and complete fixation of different *LWS* alleles between some populations, speciation as would be indicated first, by strong association between *LWS* and colour and, second, by genotypic and phenotypic discontinuities was not observed on these gradients. This may be due to a difference between the taxa that we studied, but it may also imply that speciation requires disruptive selection, and hence migration and gene flow between habitats[5,28–30,42]. In contrast, when migration exceeds selection, divergence cannot occur either[43,44]. This explains the absence of speciation on the steepest of our gradients.

Our results are relevant to conservation because they provide a mechanistic explanation for the collapse of cichlid fish species diversity during the anthropogenic eutrophication of Lake Victoria[8]. Eutrophication changes the slope of environmental light gradients, and, by steepening them, potentially moves sites from the region in parameter space that is permissive of species coexistence into the region that is not. We hope these results help focus attention of biodiversity conservation efforts in Lake Victoria and other lakes to issues of water quality.

Table 3 | *LWS* opsin allele-group frequency (%) and male nuptial colouration in species of *Pundamilia*

Species	Population	Male nuptial colour type	P	M3	H	Others	*n*†
P. "Luanso"	Luanso island	Predominantly blue	100	0	0	0	54
P. "Marumbi"	Marumbi island	Predominantly blue	92	0	8	0	24
P. pundamilia	Makobe island	Blue	94	0	0	6	32
P. pundamilia	Igombe island	Blue	83	17	0	0	6
P. pundamilia-like*	Kissenda island	Blue	96	0	4	0	70
P. pundamilia-like*	Python island	Blue	90	1	4	5	82
P. azurea[16]	Ruti island	Blue	100	0	0	0	6
P. nyererei-like*	Kissenda island	Red dorsum	25	0	75	0	52
P. nyererei-like*	Python island	Red dorsum	4	0	96	0	98
P. nyererei	Makobe island	Red dorsum	0	4	96	0	112
P. igneopinnis	Igombe island	Red dorsum	0	0	100	0	6
P. "red head"[16]	Zue island	Red chest	0	0	100	0	6
Total							548

* Hybridizing populations (neither *P. pundamilia* nor *P. nyererei*, but the hybridizing blue (*P. pundamilia*-like) and red (*P. nyererei*-like) populations shown in Fig. 2 (this study)).
† *n* represents *n* haplotypes sequenced.

METHODS SUMMARY

Ambient, absorbance and transmittance light spectra were measured with an Ocean Optics PS 1000 spectrophotometer and a 100 μm optical fibre, in the shade between 8:50 and 9:00 in the morning. We calculated the 'transmittance orange ratio' as the ratio of transmittance in the 550–700 nm range over the total visible range. The 'light slope' was obtained by regressing the transmittance orange ratio against distance (m) from the shore along the lake floor. Male fish in breeding colouration were collected by angling and netting; 480 males were photographed immediately in a photo cuvette. Water depth was measured and recorded to the nearest 0.5 m for each of 960 males. To determine the functional relevance of the observed amino acid substitutions in the *LWS* genes, the sequence of the P allele was reconstructed from the H allele by *in vitro* mutagenesis. The pigments were then expressed, reconstituted and purified as described elsewhere[9]. Absorption spectra of reconstituted pigments were measured before and after irradiation with light (>490 nm). DNA was extracted from fin tissue of 305 individuals and amplified using 11 microsatellite primers. The fragments were analysed on a Beckman Coulter CEQ 8000 Genetic Analysis System. Determination of the opsin genes was as described previously[13]. We sequenced exons 2–5 of *LWS* (872 bp), which encode the trans-membrane region, from 263 individuals (526 haplotypes). We sequenced exons 1–5 (including introns) of the *SWS2A* gene from males of Makobe island and Kissenda island. For detection of selection, the *LWS* gene and its 5-kb upstream and 3.5-kb downstream flanking sequences (total 10.5 kb) were amplified by long PCR[9] and sequenced from 10 red and 9 blue males from Python island. To determine female mating preferences, we conducted laboratory two-way mate choice assays with females from Luanso island and Python island and laboratory-bred F_2 hybrids from Python island[40].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 April; accepted 25 July 2008.

1. Schluter, D. & Price, T. Honesty, perception and population divergence in sexually selected traits. *Proc. R. Soc. Lond. B* **253,** 117–122 (1993).
2. Boughman, J. W. How sensory drive can promote speciation. *Trends Ecol. Evol.* **17,** 571–577 (2002).
3. Gray, S. M. & McKinnon, J. S. Linking color polymorphism maintenance and speciation. *Trends Ecol. Evol.* **22,** 71–79 (2007).
4. Chunco, A. J., McKinnon, J. S. & Servedio, M. R. Microhabitat variation and sexual selection can maintain male color polymorphisms. *Evolution* **61,** 2504–2515 (2007).
5. Kawata, M., Shoji, A., Kawamura, S. & Seehausen, O. A genetically explicit model of speciation by sensory drive within a continuous population in aquatic environments. *BMC Evol. Biol.* **7,** 99 (2007).
6. Boughman, J. W. Divergent sexual selection enhances reproductive isolation in sticklebacks. *Nature* **411,** 944–948 (2001).
7. Levring, T. & Fish, G. R. The penetration of light in some tropical East African waters. *Oikos* **7,** 98–109 (1956).
8. Seehausen, O., van Alphen, J. J. M. & Witte, F. Cichlid fish diversity threatened by eutrophication that curbs sexual selection. *Science* **277,** 1808–1811 (1997).
9. Terai, Y. *et al.* Divergent selection on opsins drives incipient speciation in Lake Victoria cichlids. *PLoS Biol.* **4,** 2244–2251 (2006).
10. Shichida, Y. *The Retinal Basis of Vision: Visual pigment: photochemistry and molecular evolution* (ed. Toyoda, J.-I.) 23–35 (Elsevier Science, 1999).
11. Yokoyama, S., Blow, N. S. & Radlwimmer, F. B. Molecular evolution of color vision of zebra finch. *Gene* **259,** 17–24 (2000).
12. Carleton, K. L. & Kocher, T. D. Cone opsin genes of African cichlid fishes: Tuning spectral sensitivity by differential gene expression. *Mol. Biol. Evol.* **18,** 1540–1550 (2001).
13. Terai, Y., Mayer, W. E., Klein, J., Tichy, H. & Okada, N. The effect of selection on a long wavelength-sensitive (*LWS*) opsin gene of Lake Victoria cichlid fishes. *Proc. Natl Acad. Sci. USA* **99,** 15501–15506 (2002).
14. Parry, J. W. L. *et al.* Mix and match color vision: Tuning spectral sensitivity by differential opsin gene expression in Lake Malawi cichlids. *Curr. Biol.* **15,** 1734–1739 (2005).
15. Carleton, K. *et al.* Visual sensitivities tuned by heterochronic shifts in opsin gene expression. *BMC Biol.* **6,** 22 (2008).
16. Carleton, K. L., Parry, J. W. L., Bowmaker, J. K., Hunt, D. M. & Seehausen, O. Colour vision and speciation in Lake Victoria cichlids of the genus *Pundamilia*. *Mol. Ecol.* **14,** 4341–4353 (2005).
17. Spady, T. C. *et al.* Adaptive molecular evolution in the opsin genes of rapidly speciating cichlid species. *Mol. Biol. Evol.* **22,** 1412–1422 (2005).
18. Genner, M. J. *et al.* Age of cichlids: New dates for ancient lake fish radiations. *Mol. Biol. Evol.* **24,** 1269–1282 (2007).
19. Maan, M. E. *et al.* Intraspecific sexual selection on a speciation trait, male coloration, in the Lake Victoria cichlid *Pundamilia nyererei*. *Proc. R. Soc. Lond. B* **271,** 2445–2452 (2004).
20. Endler, J. A. Some general comments on the evolution and design of animal communication systems. *Phil. Trans. R. Soc. Lond. B* **340,** 215–225 (1993).
21. Vandermeer, H. J., Anker, G. C. & Barel, C. D. N. Ecomorphology of retinal structures in zooplanktivorous haplochromine cichlids (Pisces) from Lake Victoria. *Environ. Biol. Fishes* **44,** 115–132 (1995).
22. Smit, S. A. & Anker, G. C. Photopic sensitivity to red and blue light related to retinal differences in two zooplanktivorous haplochromine species (Teleostei, Cichlidae). *Neth. J. Zool.* **47,** 9–20 (1997).
23. Maan, M. E., Hofker, K. D., van Alphen, J. J. M. & Seehausen, O. Sensory drive in cichlid speciation. *Am. Nat.* **167,** 947–954 (2006).
24. Endler, J. A. Gene flow and population differentiation. *Science* **179,** 243–250 (1973).
25. Schluter, D. & Nagel, L. M. Parallel speciation by natural selection. *Am. Nat.* **146,** 292–301 (1995).
26. Nosil, P., Egan, S. R. & Funk, D. J. Heterogeneous genomic differentiation between walking-stick ecotypes: "Isolation by adaptation" and multiple roles for divergent selection. *Evolution* **62,** 316–336 (2008).
27. Stinchcombe, J. T. & Hoektsra, H. E. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* **100,** 158–170 (2008).
28. Doebeli, M. & Dieckmann, U. Speciation along environmental gradients. *Nature* **421,** 259–264 (2003).
29. Gavrilets, S. *Fitness Landscapes and the Origin of Species* (Princeton Univ. Press (2004).
30. Leimar, O., Doebeli, M. & Dieckmann, U. Evolution of phenotypic clusters through competition and local adaptation along an environmental gradient. *Evolution* **62,** 807–822 (2008).
31. Nosil, P., Vines, T. H. & Funk, D. J. Perspective: Reproductive isolation caused by natural selection against immigrants from divergent habitats. *Evolution* **59,** 705–719 (2005).
32. Seehausen, O. *Lake Victoria Rock Cichlids. Taxonomy, Ecology and Distribution.* (Verduijn Cichlids, 1996).
33. Seehausen, O. & van Alphen, J. J. M. The effect of male coloration on female mate choice in closely related Lake Victoria cichlids (*Haplochromis nyererei* complex). *Behav. Ecol. Sociobiol.* **42,** 1–8 (1998).
34. Stelkens, R. B., Pierotti, M. E. R., Joyce, D. A., Smith, A. M., van der Sluijs, I. & Seehausen, O. Female mating preferences facilitate disruptive sexual selection on male nuptial colouration in hybrid cichlid fish. *Phil. Trans. R. Soc. B* 363 2861–2870 (2008).
35. Okullo, W. *et al.* Parameterization of the inherent optical properties of Murchison Bay, Lake Victoria. *Appl. Opt.* **46,** 8553–8561 (2007).
36. Palczewski, K. *et al.* Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **289,** 739–745 (2000).
37. Asenjo, A. B., Rim, J. & Oprian, D. D. Molecular determinants of human red/green color discrimination. *Neuron* **12,** 1131–1138 (1994).

38. McDonald, J. H. Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. *Mol. Biol. Evol.* **15,** 377–384 (1998).
39. Hudson, R. R., Kreitman, M. & Aguade, M. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116,** 153–159 (1987).
40. Haesler, M. P. & Seehausen, O. Inheritance of female mating preference in a sympatric sibling species pair of Lake Victoria cichlids: implications for speciation. *Proc. R. Soc. B* **272,** 237–245 (2005).
41. van der Sluijs, I., van Alphen, J. J. M. & Seehausen, O. Preference polymorphism for coloration but no speciation in a population of Lake Victoria cichlids. *Behav. Ecol.* **19,** 177–183 (2008).
42. Nosil, P., Crespi, B. J. & Sandoval, C. P. Reproductive isolation driven by the combined effects of ecological adaptation and reinforcement. *Proc. R. Soc. Lond. B* **270,** 1911–1918 (2008).
43. Nosil, P. & Crespi, B. J. Does gene flow constrain adaptive divergence or vice versa? A test using ecomorphology and sexual isolation in *Timema cristinae* walking-sticks. *Evolution* **58,** 102–112 (2004).
44. Rasanen, K. & Hendry, A. Disentangling interactions between adaptive divergence and gene flow when ecology drives diversification. *Ecol. Lett.* **11,** 624–636 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge the Tanzania Commission for Science & Technology for research permissions, the Tanzania Fisheries Research Institute, and its Muranza Centre director E. F. B. Katunzi, for hospitality and logistical support; M. Kayeba, M. Haluna, S. Mwaiko, M. Haesler and E. Burgerhout for help with data and fish collection; H. Araki, L. Excoffier, L. Harmon, B. Ibelings, I. Keller, T. Kocher, P. Nosil, M. Pierotti, D. Schluter, A. Sivasundar and O. Svensson for comments on the manuscript; and M. Kawata, J. J. M. van Alphen, K. Young, R. Stelkens and E. Bezault for discussion. This work was supported by Swiss National Science Foundation project 3100A0-106573 (to O.S.), and by the Ministry of Education, Culture, Sports, Science and Technology of Japan (to N.O.).

Author Contributions O.S. conceived and designed the study, collected, photographed and identified fish, measured light and shore slopes, supervised field work, conducted the hybridization experiments, supervised microsatellite analyses and mate choice experiments, and did the statistical data analyses and the writing. Y.T. designed experiments on opsins, did most of the laboratory work and data analysis on opsins, and contributed to writing. I.S.M. collected depth distribution data and did all microsatellite analyses. K.L.C. determined *LWS* sequences from experimental females and contributed to writing. H.D.J.M. collected depth distribution, light data and fish. R.M. determined *LWS* and *SWS2A* sequences with Y.T. I.v.d.S. collected fish and conducted mate choice experiments. M.V.S. helped with the microsatellite analysis. M.E.M. collected fish and measured light. H.T. performed analysis of selection pressure with Y.T. H.I. measured opsin pigment absorbance with Y.T. N.O. designed and supervised the laboratory work on opsins and contributed to the writing.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to O.S. (ole.seehausen@aqua.unibe.ch) or N.O. (nokada@bio.titech.ac.jp).

METHODS

Ambient light gradients and water clarity. Water transparency was measured using a white Secchi disk. Ambient, absorbance and transmittance light spectra between 400 nm and 750 nm were measured every metre between the surface and 3 m water depth with an Ocean Optics PS 1000 spectrophotometer and an optical fibre (100 μm), using SpectraWin 4.16 software (Avantes). Measurements were taken in the shade, between 8:50 and 9:00 in the morning. We calculated at every depth the 'transmittance orange ratio', which is a property of the water unaffected by variation in solar irradiance, as the ratio of transmittance in the 550–700 nm range (yellow, orange, red) over the total visible range (400–700 nm). The steepness of the light gradient, the 'light slope', was calculated by regressing the transmittance orange ratio against the mean distance (m) from the shore, measured along the lake floor in three transects for every island. The turbidity-mediated between-island light slopes were calculated by regressing the transmittance orange ratio measured at every island at 2 m water depth against the distance (m) from the clear water end of each gradient. The light differential was measured for both types of gradients as the difference between the transmittance orange ratios at the end points of a gradient. The largest possible value is 0.5, which is given when there is no longer any detectable blue light at the deep end of a gradient (transmittance orange ratio = 1 (that is, orange is the only transmitted light); whereas at the surface the full amounts of both blue and orange light are present (that is, transmittance orange ratio = 0.5)).

Frequency and depth distribution of male colouration. Males were collected by angling and gill nets in April and August 2001, February 2003, and January and May 2005. Photos were taken of 11 (Marumbi), 241 (Luanso), 64 (Kissenda), 34 (Python) and 130 (Makobe) males in breeding dress—480 in total—immediately on capture in specially designed photographic cuvettes. Photos were scored on a 5-point (0–4) colour phenotype scale by two to five independent observers, and the mean value was used[41] (Fig. 1). Phenotype scoring of different observers was very similar (Spearman correlations between 0.605 and 0.729, $P < 0.05$). Linear regressions with a quadratic term were fitted to the log-transformed counts of the colour phenotypes from each island separately using R[45]. Frequency distributions were compared between islands by G-tests.

Water depth was measured and recorded to the nearest 0.5 m for each of 960 males. The association between phenotype and water depth was tested for each island separately using ANOVA tests. These males were assigned to colour classes in the field, and only three robust classes were used: blue, intermediate and red (corresponding to classes 0 + 1, 2 and 3 + 4). G-tests were performed to compare depth distributions between islands. The curve-fitting procedure in SPSS (SPSS Inc. 2005) was used to quantify the relationship between strength of association (F-value) and steepness of the light slope.

***LWS* absorption spectra.** *In vitro* mutagenesis of *LWS* for construction of the sequence of P alleles, expression, reconstitution, purification and measurement were performed as described previously[9] with minor modifications. We measured absorption spectra of reconstituted pigments before and after irradiation with light (>490 nm). On the basis of the λ_{max} values determined by 3 independent difference spectra calculated from the measurements using independent preparations, we determined the absorption maximum values for each allele with standard errors.

Population genetics of neutral loci. DNA was extracted from fin tissue of 305 individuals (Marumbi 13, Luanso 61, Kissenda 59, Python 84, Makobe 88) and amplified using 11 microsatellite primers developed for these or other haplochromine species (see Supplementary Methods). We used Arlequin[46] to calculate observed and expected heterozygosities, to test for significance of departure from Hardy–Weinberg equilibrium for each locus in each population (1 million MCMC permutations), and for significant deviations from linkage equilibrium (10.000 permutations). After sequential Bonferroni correction[47], 3 out of 55 tests revealed significant deviations from Hardy–Weinberg equilibrium (1 locus each in *P. pundamilia* and *P. nyererei* from Makobe, 1 in *P. pundamilia* from Kissenda), and 2 tests of linkage equilibrium were significant: 1 in *P. pundamilia* from Python island and 1 in *P. pundamilia* from Kissenda island. Because there was no indication of any consistent linkage disequilibrium across populations between any pair of loci, all loci were retained for subsequent analysis. Molecular variance among individuals within and between phenotype groups was visualized in a factorial correspondence analysis performed over individuals in Genetix 4.05 (ref. 48). F_{ST} estimates and their significance were calculated over 100 permutations, as implemented in Arlequin[46].

Population genetics of opsin genes. Determination of the *LWS* gene was as described previously[13]. We determined the sequences of exons 2–5 of *LWS* (872 bp), which encode the transmembrane region, from 263 individuals (526 haplotypes): Marumbi (12 individuals; 24 haplotypes), Luanso (27; 54), Kissenda (62; 124), Python (90; 180) and Makobe (72; 144). Additionally, we sequenced exons 2–5 of several hundred individuals of other species of Lake Victoria cichlids (Supplementary Fig. 4). Determination of the *SWS2A* gene is described in Supplementary Methods. We sequenced exons 1–5 (including introns) from males of Makobe (16 *P. pundamilia* and 17 *P. nyererei*) and Kissenda (20 blue and 20 red males). F_{ST} values for *LWS* and *SWS2A* sequences were calculated using DnaSP 4.0 (ref. 49). The *SWS2A* sequence (1,930 bp) was split into two putative alleles for the analysis.

Molecular signature of selection on *LWS*. Determination of the *LWS* flanking sequences and the tests for detection of selection were performed as described previously[9] with minor modifications. The *LWS* gene and its 5 kb upstream and 3.5 kb downstream flanking sequences (total 10.5 kb) were amplified by long PCR[9] from 10 red and 9 blue males. To reflect the approximate frequencies of *LWS* alleles in the two phenotype populations, we included one heterozygous (H/P) individual of each nuptial colour. The McDonald test[38] was calculated with the recombination parameter set to 2, 4, 10, 32 and 1,000 replicates.

Female mating preferences. We conducted laboratory two-way mate choice assays as described elsewhere[40]. Each female was tested in at least 5 trials with 5 different male pairs. A G-test was used to compare the frequency distributions of mating preferences between islands.

45. Venables, W. N. & Ripley, B. D. Modern applied statistics with S. (Springer, 2002).
46. Excoffier, L., Laval, G. & Schneider, S. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evol. Bioinform. Online* **1,** 47–50 (2005).
47. Rice, W. R. Analyzing tables of statistical tests. *Evolution* **43,** 223–225 (1989).
48. Belkhir K., Borsa P. & Chikhi L.. Raufaste, N. & Bonhomme, F. Genetix Version 4.05 for Windows Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier (France) (1996–2004); http://www.genetix.univ-montp2.fr/genetix/genetix.htm.
49. Rozas, J., Sanchez-DelBarrio, J. C., Messeguer, X., Rozas, R. & Dna, S. P. DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19,** 2496–2497 (2003).

ARTICLES

In vivo reprogramming of adult pancreatic exocrine cells to β-cells

Qiao Zhou[1], Juliana Brown[2], Andrew Kanarek[1], Jayaraj Rajagopal[1] & Douglas A. Melton[1]

One goal of regenerative medicine is to instructively convert adult cells into other cell types for tissue repair and regeneration. Although isolated examples of adult cell reprogramming are known, there is no general understanding of how to turn one cell type into another in a controlled manner. Here, using a strategy of re-expressing key developmental regulators *in vivo*, we identify a specific combination of three transcription factors (*Ngn3* (also known as *Neurog3*) *Pdx1* and *Mafa*) that reprograms differentiated pancreatic exocrine cells in adult mice into cells that closely resemble β-cells. The induced β-cells are indistinguishable from endogenous islet β-cells in size, shape and ultrastructure. They express genes essential for β-cell function and can ameliorate hyperglycaemia by remodelling local vasculature and secreting insulin. This study provides an example of cellular reprogramming using defined factors in an adult organ and suggests a general paradigm for directing cell reprogramming without reversion to a pluripotent stem cell state.

Cells of adult organisms arise from sequential differentiation steps that are generally thought to be irreversible[1]. Biologists often describe this process of development as proceeding from an undifferentiated (embryonic) cell to a terminally differentiated cell that forms part of an adult tissue or organ. There are rare examples, however, in which cells of one type can be converted to another type in a process called cellular reprogramming or lineage reprogramming[2,3]. Various forms of cellular reprogramming are referred to in the literature as transdifferentiation, dedifferentiation or transdetermination[4]. For example, cellular reprogramming occurs in amphibian limb regeneration and fly imaginal disc identity switches[5,6], and it may be central to certain types of pathological metaplasia[4]. There is long-standing interest and fascination in reprogramming studies, in part because of the promise of harnessing this phenomenon for regenerative medicine whereby abundant adult cells that can be easily collected would be converted to other medically important cell types to repair diseased or damaged tissues.

Somatic cell nuclear transfer (SCNT), developed in the 1960s, demonstrated that nuclei from differentiated adult cells could be reprogrammed to a totipotent state after injection into enucleated eggs[2,7]. More recently, it was shown that a small number of transcription factors can reprogram cultured adult skin cells to induced pluripotent stem (iPS) cells[8–13]. These studies point to the possibility of regenerating mammalian tissues by first reverting skin or other adult cells to pluripotent stem cells and then redifferentiating these into various cell types. Alternatively, it should be possible to convert one cell type into another directly, without the need to first revert the cell to an undifferentiated pluripotent state. Indeed, there are examples in the literature that suggest that this approach is feasible. For example, studies with embryonic cells have shown that dermal fibroblasts and retinal epithelial cells can be converted into muscle-like cells[14], and pancreatic tissue to liver[15]. In adult animals, mature B lymphocytes have been reprogrammed into macrophages[16] or pro-B cells[17]. Today, well documented examples of cellular reprogramming, especially in adult animals, remain rare and have generally been restricted to cases in which a single inducing factor is involved. The recent work on iPS formation suggests that a specific combination of multiple factors, instead of a single one, might be the most effective way to reprogram adult cells[8–13].

We developed a strategy to identify adult cell reprogramming factors by re-expressing multiple embryonic genes in living adult animals. Our focus on embryonic genes is based in part on regeneration studies in newts, frogs and fish, wherein it has been shown that dedifferentiation of adult cells to progenitors, a form of cellular reprogramming, is accompanied by reactivation of embryonic regulators[5,18,19]. These studies suggest that re-expression of appropriate embryonic genes may reprogram differentiated cells.

To search for factors that could reprogram adult cells into β-cells, we focused on transcription factors, a class of genes enriched for factors that regulate cell fates during embryogenesis. An *in situ* hybridization screen of more than 1,100 transcription factors identified groups of transcription factors with cell-type-specific expressions in the embryonic pancreas[20]. There are at least 20 transcription factors expressed in mature β-cells and their immediate precursors, the endocrine progenitors (Supplementary Table 1). Of these, 9 genes exhibited β-cell developmental phenotypes when mutated[21,22], and these were selected for initial reprogramming experiments.

We chose mature exocrine cells of the adult pancreas as target cells for reprogramming. Exocrine cells derive from pancreatic endoderm, as do β-cells[23], and exocrine cells can turn on endocrine programs when dissociated and cultured *in vitro*[24,25]. We carried out our experiments *in vivo* so that any induced β-cells would reside in their native environment, which might promote their survival and/or maturation. In addition, this approach allows for a direct comparison of endogenous and induced β-cells. The transcription factors were delivered into the pancreas in adenoviral vectors. It has been shown that adenovirus preferentially infects pancreatic exocrine cells, but not islet cells[26], and, because most endogenous β-cells reside within islets (Fig. 1b), any newly formed (induced) β-cells could be easily detected as extra-islet insulin$^+$ cells.

Induction of insulin$^+$ cells in adult mice

Adenovirus that co-expresses each transcription factor together with nuclear GFP (nGFP) was purified. All nine viruses were pooled and injected as a mixture (referred to as M9, for mixture of nine) into the pancreata of 2-month-old adult mice (Fig. 1a). The immune-deficient

[1]Department of Stem Cell and Regenerative Biology, Howard Hughes Medical Institute, Harvard Stem Cell Institute, Harvard University, 7 Divinity Avenue, Cambridge, Massachusetts 02138, USA. [2]Department of Pathology, Children's Hospital, Boston, Harvard Medical School, Harvard Stem Cell Institute, 300 Longwood Avenue, Boston, Massachusetts 02115-5724, USA.

Rag1$^{-/-}$ strain was used to avoid complications associated with viral-elicited immune response[27]. One month after viral delivery, immunohistochemistry revealed a modest increase of extra-islet insulin$^+$ cells among viral infected cells (nGFP$^+$) in two out of three animals (Fig. 1d). To determine which of the nine factors are required, individual factors were removed from the pool one at a time. Pools lacking Nkx2.2, Nkx6.1 or Pax4 continued to produce increased extra-islet insulin$^+$ cells (data not shown), suggesting that these genes are dispensable. Results for the other six genes were inconclusive. We conducted another round of factor withdrawal with mixtures of the remaining six genes (M6); three of them, *Ngn3*, *Pdx1* and *Mafa*, proved to be absolutely required (Fig. 1d). The combination of these three factors (referred to as M3) converted >20% of infected cells to insulin$^+$ cells (red cells with green nuclei, Fig. 1c, e). Notably, single factors or combinations of any two factors did not elicit this effect (Fig. 1e). Antibody labelling confirmed that these three inducing factors are co-expressed in the induced insulin$^+$ cells (Supplementary Fig. 1). NeuroD (also known as Neurod1) can functionally replace Ngn3 in M3, but the resulting cocktail has reduced induction efficiency (Fig. 1e).

We noticed that the percentage of insulin$^+$ cells among infected cells increases with progressive removal of factors from the pool such that M3 induces more insulin$^+$ cells than M6, whereas M6 is better than M9 (Fig. 1d, e). This is probably due to the fact that a constant volume of virus was injected into each animal, regardless of the viral combinations. The effective concentration of Ngn3, Pdx1 and Mafa viruses in a cocktail, therefore, increases when fewer factors are included. New insulin$^+$ cells were detected 3 days after injection, but the expression level was low. The intensity of insulin staining increased gradually so that, by day 10, the level was comparable to that of endogenous β-cells (Supplementary Fig. 2). These new insulin$^+$ cells were still present after 3 months, the longest time point that we analysed, and remained as scattered individual cells or small clusters and did not form islets (Fig. 1c). The reprogramming effect of the three factors appeared to be rather specific for pancreatic exocrine cells: infection of skeletal muscle *in vivo* or fibroblasts *in vitro* with M3 did not induce insulin expression, despite extensive co-expression of the three factors in the target cells (Supplementary Fig. 1).

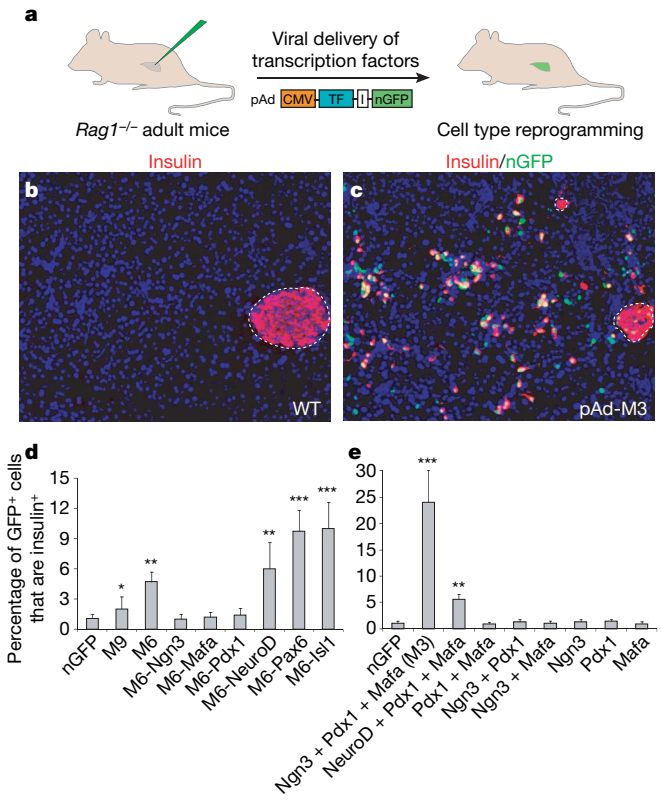

Figure 1 | A combination of three transcription factors induces insulin$^+$ cells in adult mouse pancreas *in vivo*. **a**, Schematic diagram of the experimental strategy. Adenoviruses encoding bicistronic transcription factor (TF) and nGFP linked by an IRES element (I) were injected into the pancreas of an adult mouse (*Rag*$^{-/-}$). CMV, cytomegaloviral promoter. **b**, Wild type (WT) pancreas is predominantly exocrine tissue with insulin$^+$ β-cells in the islet (outlined). Nuclei were stained blue with DAPI. **c**, One month after infection with a combination of Ngn3, Pdx1 and Mafa viruses (pAd-M3), numerous insulin$^+$ cells appear outside of islets. **d**, **e**, Quantification of induction one month after infection. M9, M6: mixture of 9 and 6 different viruses, respectively. Data are presented as mean + s.d.; $n = 3$ animals. ~1,000 nGFP$^+$ cells were counted per animal. Asterisk, $P < 0.05$; two asterisks, $P < 0.01$; three asterisks, $P < 0.001$.

New insulin$^+$ cells come from exocrine cells

Lineage analysis was performed to determine the origin of the new insulin$^+$ cells. The five major cell types in the adult pancreas can be detected with lineage-specific molecular markers: exocrine (amylase), duct (Ck19), endocrine (insulin, glucagon, somatostatin and pancreatic polypeptide), vascular (PECAM) and mesenchymal (nestin and vimentin) cells. On injection with a control nGFP virus, most infected cells (>95%) were found to be mature amylase$^+$ exocrine cells (Fig. 2a, b), consistent with previous reports[26]. Non-exocrine cells together accounted for approximately 5% of the infected population. Because more than 20% of M3-infected cells become insulin$^+$ 10 days after viral delivery, it suggests that non-exocrine cells can contribute, at most, to a minor fraction of these new insulin$^+$ cells. As there is little cell death and no enhanced proliferation during this reprogramming (Supplementary Fig. 3), most insulin$^+$ cells would thus appear to originate from mature exocrine cells. To confirm the exocrine origin of the new insulin$^+$ cells, we genetically labelled mature exocrine cells with a mouse line (*Cpa1CreER*^{T2}) that expresses an inducible form of Cre recombinase (*CreER*^{T2}) specifically in adult exocrine cells[20] (Fig. 2c). When crossed with the *R26R* reporter line, tamoxifen induction in double heterozygous *Cpa1CreER*^{T2};*R26R* adults indelibly labelled 5–10% of mature

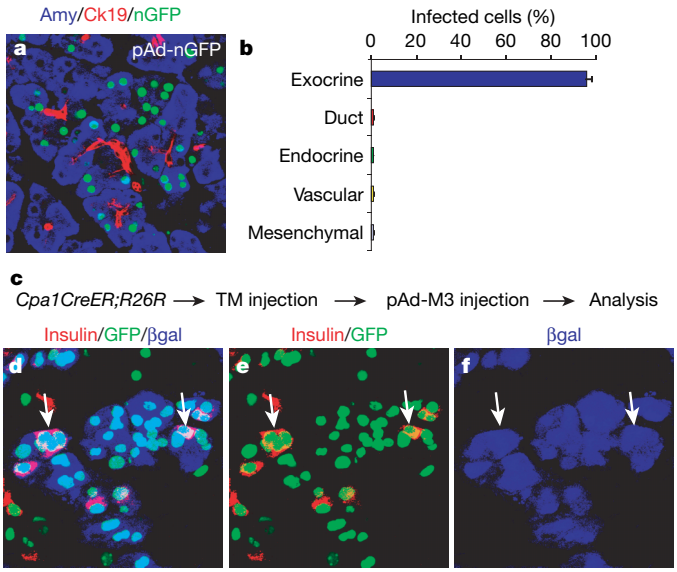

Figure 2 | Induced new β-cells originate from differentiated exocrine cells. **a**, Ten days after nGFP viral infection, most infected cells are amylase$^+$ (Amy$^+$) mature exocrine cells, not duct cells (Ck19$^+$). **b**, Quantification of nGFP-infected cell types. Data are presented as mean + s.d., $n = 3$ animals. ~1,000 nGFP$^+$ cells were counted. **c**, Double heterozygous *Cpa1CreER*^{T2};*R26R* adult mice are injected with tamoxifen (TM), which labels the mature exocrine cells with β-galactosidase (βgal). Reprogramming is subsequently induced by infection with pAd-M3. **d**–**f**, Ten days after infection, many βgal$^+$insulin$^+$ cells (arrows) are present. **e** and **f** are insulin (red)/GFP (green) and βgal (blue) channels of **d**, respectively.

exocrine cells with β-galactosidase (Fig. 2d, f); no label was found in other cell types. After pAd-M3 injection, many β-galactosidase$^+$ cells become insulin$^+$ (Fig. 2d–f, pink cells), providing direct evidence that mature exocrine cells give rise to new insulin$^+$ cells.

Induced β-cells closely resemble islet β-cells

We next examined the new insulin$^+$ cells to determine the extent to which they have been reprogrammed. Morphologically, exocrine cells are large with a cobble stone appearance (Fig. 3a, b) whereas islet β-cells are much smaller and spindle shaped (Fig. 3a). When dissociated into single cells, the diameter of amylase$^+$ exocrine cells range from 25 μm to 17 μm whereas insulin$^+$ β-cells range from 9 μm to 15 μm. The induced cells are indistinguishable from islet β-cells in size and shape (Fig. 3b, c).

At the ultrastructural level, the reprogrammed cells have all the hallmarks of islet β-cells (Fig. 3d, e). They possess the small dense secretory granules characteristic of insulin granules, and lack the large zymogen granules and dense assemblies of endoplasmic reticulum that are characteristic of exocrine cells (Fig. 3d, e). Immunoelectron microscopy further showed that the induced β-cells express both GFP in the nucleus and abundant insulin in the granules (Supplementary Fig. 4). Interestingly, the induced β-cells often appeared on the electron micrograph as intercalated within exocrine acinar rosettes (Fig. 3e). In wild-type pancreatic samples, rare single or small clusters of β-cells reside outside islets, but they often associate with duct but not exocrine cells. The unique position of induced cells probably reflects their exocrine origin.

Molecular marker analysis reveals that most of the insulin$^+$ cells co-express genes essential for β-cell endocrine function including glucose transporter 2 (Glut2, also known as Slc2a2, expressed in 92.8% of the new insulin$^+$ cells), glucokinase (GCK, 96.7%), prohormone convertase (PC1/3, also known as Pcsk1, 86.7%; Fig. 4a–c and Supplementary Fig. 5), and the key β-cell transcription factors NeuroD (88.9%), Nkx2.2 (85.3%) and Nkx6.1 (85.9%; Fig. 4d–f and Supplementary Fig. 5). The induced insulin$^+$ cells express C-peptide (part of proinsulin; Fig. 4h). Expression profile analysis of the reprogrammed cells further indicates a strong overlap of endocrine-enriched genes between reprogrammed cells and islet cells, suggesting a high degree of similarity between their endocrine programs (Supplementary Fig. 6).

The new β-cells do not express exocrine genes such as amylase or Ptf1a, the duct marker Ck19 (also known as Krt19), mesenchymal markers nestin and vimentin, nor the neuronal marker Tuji (β-tubulin III, also known as Tubb3) (Fig. 4g, Supplementary Fig. 5 and data not shown). Nor do the new β-cells express any other pancreatic hormones such as glucagon, somatostatin or pancreatic polypeptide (Fig. 4h, i and Supplementary Fig. 5). Thus, the new β-cells do not exhibit a hybrid or mixed phenotype, indicating silencing of non-β-cell programs.

The primary function of β-cells is to synthesize and release insulin. To facilitate the release of insulin into the circulation, β-cells, unique among pancreatic cell types, synthesize vascular endothelial growth factor (VEGF), which promotes local angiogenic remodelling[28]. Notably, induced β-cells similarly synthesize VEGF and induce angiogenesis so that blood vessels form next to these new cells (Fig. 5a, b). Quantification indicates that, in nGFP controls, 32% of infected cells lie adjacent to blood vessels whereas 61% and 83% of induced β-cells are directly juxtaposed to blood vessels 10 days and 30 days after induction, respectively.

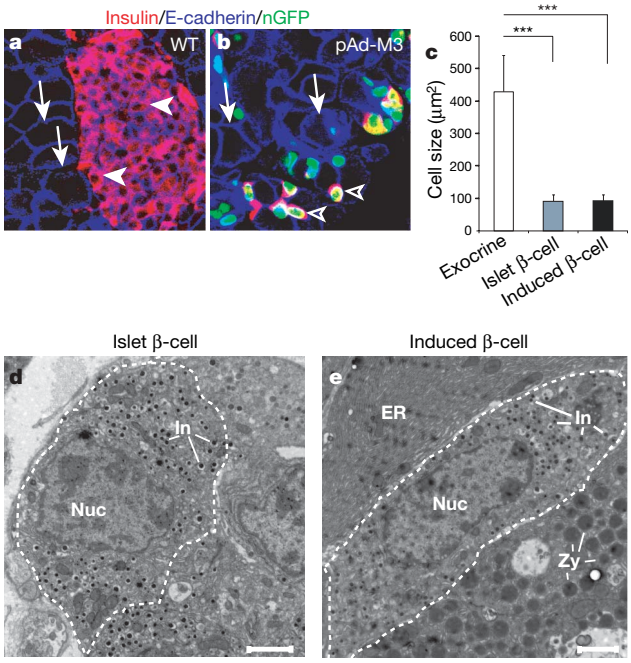

Figure 3 | Endogenous and induced β-cells are indistinguishable in morphology and ultrastructure. **a**, **b**, Islet β-cells (**a**, arrowheads) and induced β-cells (**b**, arrowheads) are similar in size and shape but distinctly different from exocrine cells (**a**, **b**, arrows). E-cadherin staining was used to visualize cell boundaries. **c**, Size comparison of exocrine cells, islet β-cells and induced β-cells. Data are presented as mean + s.d., $n = 3$ animals. >100 cells per animal were used. Three asterisks, $P < 0.001$. **d**, Electron micrograph of a β-cell (outlined) in an islet. **e**, Example of an induced β-cell situated between two exocrine cells. Endogenous and induced β-cells contain small insulin granules (In) and lack zymogen granules (Zy) of exocrine cells and extensive endoplasmic reticulum (ER). Nuc, nucleus. Scale bars, 2 μm.

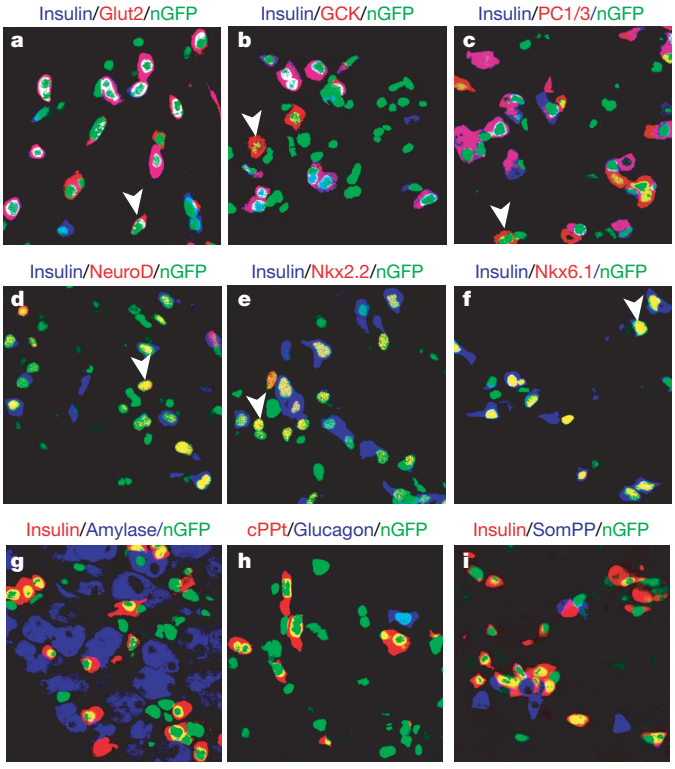

Figure 4 | Molecular marker characterization of induced β-cells. **a**–**f**, One month after infection with pAd-M3, most insulin$^+$ induced β-cells co-express endocrine genes: glucose transporter 2 (Glut2, **a**), glucokinase (GCK, **b**), prohormone convertase 1/3 (PC1/3, **c**) and β-cell transcription factors NeuroD (**d**), Nkx2.2 (**e**) and Nkx6.1 (**f**). Arrowheads indicate examples of cells that express marker genes but not insulin. **g**–**i**, Induced β-cells do not express amylase (**g**), glucagon (**h**) or somatostatin/pancreatic polypeptide (**i**, SomPP). cPPt, c-peptide.

To test whether induced β-cells release insulin, mice were rendered diabetic by streptozotocin (STZ) injection, which specifically ablates islet β-cells. When subsequently injected with pAd-M3, fasting blood glucose levels of hyperglycaemic animals showed a significant and long-lasting improvement compared to animals injected with control (nGFP) virus (Fig. 5d). In addition, the pAd-M3 animals showed increased glucose tolerance (Supplementary Fig. 7), had increased insulin levels in the serum (non-fasting, $P < 0.01$, Fig. 5e) and possessed large numbers of induced β-cells (Fig. 5f). Polymerase chain reaction with reverse transcription (RT–PCR) analysis and direct observation revealed that virus injected into the pancreas does not spread to other internal organs such as liver and intestine that, theoretically, could modulate insulin secretion and/or response (Supplementary Fig. 8). In addition, we found no evidence that STZ-treated animals show spontaneous conversion of exocrine cells to β-cells (Supplementary Fig. 8). As the data in Fig. 5 show, the total number of induced β-cells is rather small compared to the number of β-cells in normal animals and this may account for the limitation to the effectiveness in restoring glucose homeostasis. Alternatively, because the new β-cells are not reorganized into islet structures, this may limit their effectiveness. Together, these data show that induced β-cells can produce and secrete insulin *in vivo*.

Inducing factors are required only transiently

Our results thus far support the contention that a combination of three transcription factors fully reprograms exocrine cells to β-cells *in vivo*. To determine whether continued presence of these factors is required to maintain the phenotype of reprogrammed cells, we used RT–PCR and primers specific to viral transgenes to detect their presence. Transgene expression from all three viruses was substantially diminished after 1 month and was undetectable after 2 months (Supplementary Fig. 9). Ngn3 protein was undetectable by antibody staining 1 month after infection (Supplementary Fig. 9). Pdx1 and Mafa protein expression in the induced β-cells, however, remains consistently strong even after 2 months, indicating the activation of endogenous genes (Supplementary Fig. 9). These results are consistent with the fact that endogenous islet β-cells do not express Ngn3, but do express Pdx1 and Mafa[21,22]. Thus, a transient expression of the inducing factors is sufficient to convert exocrine cells to a stable new β-cell state.

β-cell reprogramming does not involve dedifferentiation

In principle, the conversion of exocrine cells to β-cells could be direct or involve dedifferentiation to common progenitors that then redifferentiate into β-cells. Indeed, exocrine and β-cells share a common progenitor during embryogenesis that is characterized by rapid division and expression of genes including *Sox9* and *Hnf6* (also known as *Onecut1*; ref. 20). Continuous 5-bromodeoxyuridine (BrdU) labelling over the first 10 days of reprogramming, however, shows that few induced β-cells (3.2%) have divided (Supplementary Fig. 3). In comparison, 12.9% of endogenous islet β-cells in the same animals incorporated BrdU (Supplementary Fig. 3). In addition, we detected no induction of Sox9 or Hnf6 (data not shown). These results suggest that *in vivo* reprogramming of exocrine to β-cells is a direct conversion of cell types and does not involve dedifferentiation. We can not formally exclude the possibility that a very transient or partial dedifferentiation may occur, but our results indicate that extensive replication and reversion to a dedifferentiated cell for an appreciable time does not occur.

Discussion

Our results provide evidence that fully differentiated exocrine cells can be directly reprogrammed into cells that closely resemble β-cells

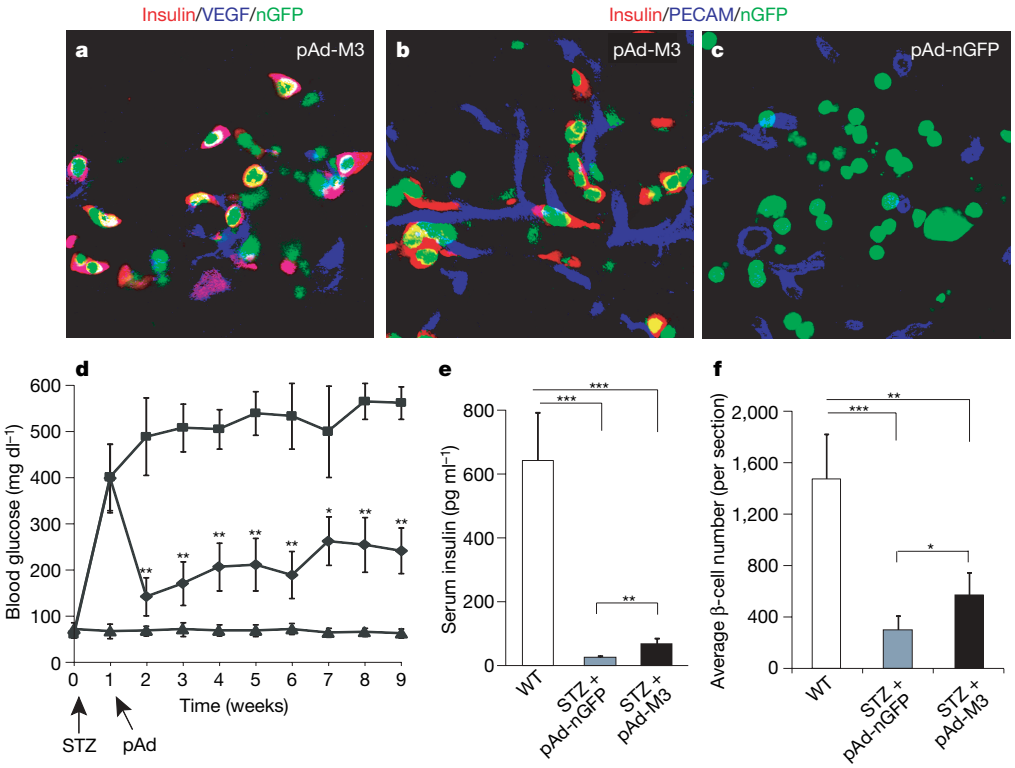

Figure 5 | Induced new β-cells remodel vasculature and ameliorate hyperglycaemia. **a–c**, New β-cells synthesize VEGF (**a**) and induce local angiogenic remodelling (**b**). Note the proximity of blood vessels (PECAM$^+$) to the induced β-cells (**b**) versus control infected cells (**c**). **d**, Improvement of fasting blood glucose level in diabetic mice after injection with pAd-M3 (diamond) compared to controls with nGFP virus (square). Triangle, non-diabetic controls. STZ, streptozotocin. Arrows indicate timing of injection. $n = 6–8$ animals. **e**, Non-fasting serum insulin levels 6 weeks after injection. $n = 6–8$ animals. **f**, Average insulin$^+$ β-cell number per section 8 weeks after injection. $n = 3$ animals. Both islet and induced β-cells were counted for the pAd-M3 samples. One asterisk, $P < 0.05$; two asterisks, $P < 0.01$; three asterisks, $P < 0.001$. Data are presented as mean + s.d.

in adult animals by a combination of just three transcription factors. The three reprogramming factors, *Ngn3*, *Pdx1* and *Mafa*, are known to be important in the embryonic development of pancreas and β-cells[21,22]. In contrast, many additional factors are also required for β-cell development[21,22]. Further studies will be necessary to understand why this particular combination is sufficient for adult β-cell reprogramming.

The induced β-cells do not organize into islet structures and remain as single cells or small clusters. Signalling between β-cells inhibits basal insulin secretion and enhances glucose-stimulated insulin secretion[29]. The lack of organization of induced β-cells undoubtedly impairs their function. Strategies that promote aggregation of the induced β-cells in adult should help to restore full glucose responsiveness.

There have been previous attempts to convert adult liver cells to β-cells *in vivo* by expressing pancreatic transcription factors[27,30–32]. These factors were able to induce expression of some pancreatic genes, but not phenotypic or morphological conversion into functional β-cells[27,30–32]. Mature exocrine cells can turn on endocrine programs when dissociated and cultured *in vitro*[24,25,33,34]. Interestingly, dissociation itself is apparently sufficient to initiate endocrine programs whereas the addition of growth factors is necessary for cell survival[24,25,33,34]. However, the molecular mechanisms of this process remain largely unknown. Other studies have shown that pancreatic duct cells and liver cells could be induced to express certain β-cell gene products in culture[35–37]. Most of these studies, however, did not address whether these cells possess a hybrid phenotype. In addition, RT–PCR on populations of cells, instead of single-cell resolution immunohistochemistry, was routinely used to evaluate the expression of β-cells markers. It is unclear how many cells actually expressed these markers or at what level. Finally, β-cells exhibit highly unstable phenotypes when cultured and appear to transform into fibroblast-like cells[38,39]. *In vitro* generation of β-cells will probably require suitable culture conditions that have yet to be discovered.

It is surprising that the reprogramming of exocrine cells to β-cells does not involve multiple rounds of cell proliferation. It is generally thought that epigenetic changes that underlie reprogramming events are most easily made during cell division[2]. It may be the case that many reprogramming events do indeed involve obligatory proliferation steps[4]. In contrast, reprogramming of B lymphocytes to macrophages seems to be cell-cycle-independent[16]. Early SCNT experiments also provided evidence for reprogramming without DNA replication[40].

Reprogramming of exocrine cells to β-cells occurs at a relatively fast speed, with the first insulin$^+$ cells appearing at day 3, and with efficiency of up to 20%. This is in contrast with recent reports of reprogramming fibroblasts to embryonic stem cells[8–13], where it takes a considerably longer time (7–30 days) and the efficiency is much reduced (typically less than 0.1%). This may be due to the fact that pancreatic exocrine and β-cells are closely related cell types and share much of their epigenomes whereas the epigenomes of fibroblasts and embryonic stem cells are largely dissimilar. Conversion between exocrine and β-cells may therefore require fewer epigenetic changes.

Recent advances in mammalian cellular reprogramming with defined genes collectively point to the possibility that a limited number of factors could reprogram any given adult cell to a different type of cell such as a stem cell, a committed progenitor or another mature cell type. All these studies relied on knowledge of the normal development of these cell types, which enabled the manipulation of key developmental regulators in adult cells. This approach may prove to be a general strategy for directing adult cell reprogramming. The recent reprogramming of human skin cells to iPS cells raises the possibility of generating patient-specific human embryonic stem lines for therapies[9,10,13]. This would be the first step in a process that will then require directed differentiation of the iPS cells to produce therapeutically important cell types such as neurons, cardiomyocytes or pancreatic β-cells. In principle, patient-specific cell therapies could be achieved more directly by reprogramming abundant and easily accessible patient-specific human cells such as fibroblasts, blood cells or adipocytes.

METHODS SUMMARY

Adenovirus construction and purification. Genes of interest were first cloned into a shuttle vector containing an internal ribosome entry site linked to nuclear GFP (*IRES-nGFP*), and then into the pAd/CMV/V5-DEST adenoviral vector (Invitrogen). High titre virus ($>1 \times 10^{10}$ plaque-forming units (p.f.u.) per ml) was obtained by purification with the AdEasy Kit (Stratagene).

Animals, surgery and physiological studies. *Rag1*$^{-/-}$ and *Rag1*$^{-/-}$*;NOD* animals were obtained from Jackson Laboratories. Adult animals (>2 months old) were injected with 100 μl ($>1 \times 10^9$ p.f.u.) of purified adenovirus directly into the splenic lobe of the dorsal pancreas. Blood glucose was measured with Ascensia Elite blood glucose meter. Insulin levels were determined with an Ultrasensitive insulin ELISA kit (Alpco).

Immunohistochemistry, BrdU labelling and TUNEL analysis. This was performed as previously described[41]. BrdU (1 mg ml^{-1}) was provided in drinking water for BrdU labelling after surgery. Apoptotic cells were recognized by TUNEL (terminal dUTP nick-end labelling) with a TMR red cell death kit (Roche).

Electron microscopy. Dissected pancreas was fixed in 4% paraformaldehyde and 0.1% glutaraldehyde for 2 h at room temperature (24 °C). For conventional transmission electron microscopy, samples were further fixed by osmium tetroxide, embedded in Epon resin and sectioned at 60–80 nm. For immunogold labelling, ultrathin sections were cut at −120 °C and stained with gold-conjugated antibodies. Images were obtained with a Tecnai G^2 Spirit BioTWIN transmission electron microscope.

FACS analysis and gene profiling. Pancreas was digested with liberase and elastase (Roche) to single cells. GFP$^+$ cells were isolated by FACS with FACSaria (BD Bioscience). Biotin-labelled complementary RNA probes were synthesized with the Illumina TotalPrep RNA Amplification kit (Ambion). Gene profiling was performed with Sentrix BeadChip Array MouseRef-8 v1.1 (Illumina). Data were analysed with the BeadStudio software.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 26 June; accepted 6 August 2008.
Published online 27 August 2008.

1. Weissman, I. L. Stem cells: units of development, units of regeneration, and units in evolution. *Cell* **100,** 157–168 (2000).
2. Hochedlinger, K. & Jaenisch, R. Nuclear reprogramming and pluripotency. *Nature* **441,** 1061–1067 (2006).
3. Orkin, S. H. & Zon, L. I. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132,** 631–644 (2008).
4. Slack, J. M. Metaplasia and transdifferentiation: from pure biology to the clinic. *Nature Rev. Mol. Cell Biol.* **8,** 369–378 (2007).
5. Brockes, J. P. & Kumar, A. Plasticity and reprogramming of differentiated cells in amphibian regeneration. *Nature Rev. Mol. Cell Biol.* **3,** 566–574 (2002).
6. Hadorn, E. Transdetermination in cells. *Sci. Am.* **219,** 110–114 (1968).
7. Gurdon, J. B. From nuclear transfer to nuclear reprogramming: the reversal of cell differentiation. *Annu. Rev. Cell Dev. Biol.* **22,** 1–22 (2006).
8. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (2006).
9. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131,** 861–872 (2007).
10. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318,** 1917–1920 (2007).
11. Meissner, A., Wernig, M. & Jaenisch, R. Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. *Nature Biotechnol.* **25,** 1177–1181 (2007).
12. Wernig, M. *et al. In vitro* reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* **448,** 318–324 (2007).
13. Park, I. H. *et al.* Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451,** 141–146 (2008).
14. Choi, J. *et al.* MyoD converts primary dermal fibroblasts, chondroblasts, smooth muscle, and retinal pigmented epithelial cells into striated mononucleated myoblasts and multinucleated myotubes. *Proc. Natl Acad. Sci. USA* **87,** 7988–7992 (1990).
15. Shen, C. N., Slack, J. M. & Tosh, D. Molecular basis of transdifferentiation of pancreas to liver. *Nature Cell Biol.* **2,** 879–887 (2000).
16. Xie, H., Ye, M., Feng, R. & Graf, T. Stepwise reprogramming of B cells into macrophages. *Cell* **117,** 663–676 (2004).
17. Cobaleda, C., Jochum, W. & Busslinger, M. Conversion of mature B cells into T cells by dedifferentiation to uncommitted progenitors. *Nature* **449,** 473–477 (2007).

18. Whitehead, G. G., Makino, S., Lien, C. L. & Keating, M. T. *fgf20* is essential for initiating zebrafish fin regeneration. *Science* **310,** 1957–1960 (2005).
19. Tanaka, E. M. Cell differentiation and cell fate during urodele tail and limb regeneration. *Curr. Opin. Genet. Dev.* **13,** 497–501 (2003).
20. Zhou, Q. *et al.* A multipotent progenitor domain guides pancreatic organogenesis. *Dev. Cell* **13,** 103–114 (2007).
21. Murtaugh, L. C. & Melton, D. A. Genes, signals, and lineages in pancreas development. *Annu. Rev. Cell Dev. Biol.* **19,** 71–89 (2003).
22. Jensen, J. Gene regulatory factors in pancreatic development. *Dev. Dyn.* **229,** 176–200 (2004).
23. Gu, G., Dubauskaite, J. & Melton, D. A. Direct evidence for the pancreatic lineage: $NGN3^{+}$ cells are islet progenitors and are distinct from duct progenitors. *Development* **129,** 2447–2457 (2002).
24. Baeyens, L. *et al. In vitro* generation of insulin-producing beta cells from adult exocrine pancreatic cells. *Diabetologia* **48,** 49–57 (2005).
25. Minami, K. *et al.* Lineage tracing and characterization of insulin-secreting cells generated from adult pancreatic acinar cells. *Proc. Natl Acad. Sci. USA* **102,** 15116–15121 (2005).
26. Wang, A. Y., Peng, P. D., Ehrhardt, A., Storm, T. A. & Kay, M. A. Comparison of adenoviral and adeno-associated viral vectors for pancreatic gene delivery *in vivo*. *Hum. Gene Ther.* **15,** 405–413 (2004).
27. Wang, A. Y., Ehrhardt, A., Xu, H. & Kay, M. A. Adenovirus transduction is required for the correction of diabetes using Pdx-1 or Neurogenin-3 in the liver. *Mol. Ther.* **15,** 255–263 (2007).
28. Lammert, E. *et al.* Role of VEGF-A in vascularization of pancreatic islets. *Curr. Biol.* **13,** 1070–1074 (2003).
29. Konstantinova, I. *et al.* EphA–Ephrin-A-mediated beta cell communication regulates insulin secretion from pancreatic islets. *Cell* **129,** 359–370 (2007).
30. Ferber, S. *et al.* Pancreatic and duodenal homeobox gene 1 induces expression of insulin genes in liver and ameliorates streptozotocin-induced hyperglycemia. *Nature Med.* **6,** 568–572 (2000).
31. Kaneto, H. *et al.* PDX-1/VP16 fusion protein, together with NeuroD or Ngn3, markedly induces insulin gene transcription and ameliorates glucose tolerance. *Diabetes* **54,** 1009–1022 (2005).
32. Miyatsuka, T. *et al.* Ectopically expressed PDX-1 in liver initiates endocrine and exocrine pancreas differentiation but causes dysmorphogenesis. *Biochem. Biophys. Res. Commun.* **310,** 1017–1025 (2003).
33. Minami, K. & Seino, S. Pancreatic acinar-to-beta cell transdifferentiation *in vitro*. *Front. Biosci.* **13,** 5824–5837 (2008).
34. Okuno, M. *et al.* Generation of insulin-secreting cells from pancreatic acinar cells of animal models of type 1 diabetes. *Am. J. Physiol. Endocrinol. Metab.* **292,** E158–E165 (2007).
35. Sapir, T. *et al.* Cell-replacement therapy for diabetes: generating functional insulin-producing tissue from adult human liver cells. *Proc. Natl Acad. Sci. USA* **102,** 7964–7969 (2005).
36. Heremans, Y. *et al.* Recapitulation of embryonic neuroendocrine differentiation in adult human pancreatic duct cells expressing neurogenin 3. *J. Cell Biol.* **159,** 303–312 (2002).
37. Gasa, R. *et al.* Proendocrine genes coordinate the pancreatic islet differentiation program *in vitro*. *Proc. Natl Acad. Sci. USA* **101,** 13245–13250 (2004).
38. Morton, R. A., Geras-Raaka, E., Wilson, L. M., Raaka, B. M. & Gershengorn, M. C. Endocrine precursor cells from mouse islets are not generated by epithelial-to-mesenchymal transition of mature beta cells. *Mol. Cell. Endocrinol.* **270,** 87–93 (2007).
39. Gershengorn, M. C. *et al.* Epithelial-to-mesenchymal transition generates proliferative human islet precursor cells. *Science* **306,** 2261–2264 (2004).
40. De Robertis, E. M. & Gurdon, J. B. Gene activation in somatic nuclei after injection into amphibian oocytes. *Proc. Natl Acad. Sci. USA* **74,** 2470–2474 (1977).
41. Dor, Y., Brown, J., Martinez, O. I. & Melton, D. A. Adult pancreatic beta-cells are formed by self-duplication rather than stem-cell differentiation. *Nature* **429,** 41–46 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to M. Ericsson for expert assistance on electron microscopy, R. Hellmiss-Peralta for advice on graphics, and B. Tilton and P. Rogers for FACS. We thank R. Martinez and G. Kenty for technical assistance; H. Edlund for the gift of Ptf1a antiserum; A. Kweudjeu for microarray analysis; members of the Melton laboratory for advice and feedback; and J. Sneddon, J. Annes and W. Anderson for critical reading of the manuscript. Q.Z. was supported by a Damon-Runyon Cancer Research Foundation Postdoctoral Fellowship and a Pathway to Independence (PI) Award from the National Institute of Health. D.A.M. is an HHMI investigator and this work was supported in part by the Harvard Stem Cell Institute and the NIH.

Author Information The microarray data were deposited in the Gene Expression Omnibus (GEO) under accession number GSE12025. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to D.A.M. (dmelton@harvard.edu).

METHODS

Viral injection and tissue collection. For adult pancreas, ~100 µl purified virus was injected directly into 2–3 foci of the dorsal splenic lobe with a 3/10 cc Insulin Syringe (Becton Dickinson). For skeletal muscle, ~20 µl virus was injected into the upper thigh. At the time of tissue collection, the infected portion of the tissue was visualized by GFP fluorescence and dissected out. For adult pancreas, typically ~50% of the dorsal pancreas was taken.

Immunohistochemistry. Adult mouse pancreata were fixed by immersion in 4% paraformaldehyde for 2 h at 4 °C. Samples were subsequently incubated in 30% sucrose solution overnight (6–12 h) and embedded with optimal cutting temperature compound (Tissue-Tek).

The following primary antibodies were used: rat anti-E-cadherin (Zymed), rat anti-Pecam1 (Pharmingen), goat anti-Ngn3 (Santa Cruz), guinea-pig anti-insulin (Dako), guinea-pig anti-glucagon (Linco), guinea-pig anti-Pancreatic polypeptide (Linco), rabbit anti-somatostatin (Dako), rabbit anti-pancreatic polypeptide (Dako), goat anti-somatostatin (Santa Cruz), goat anti-Pdx1 (Santa Cruz), guinea-pig anti-Pdx1 (gift from C. Wright), goat anti-β-galactosidase (Biogenesis), goat anti-amylase (Santa Cruz), mouse anti-BrdU (Amersham), rabbit anti-mafA (Bethyl), chick anti-nestin (Ames), chick anti-vimentin (Chemicon), goat anti-Glut2 (Santa Cruz), goat anti-VEGF (R&D), rabbit anti-PC1/3 (Chemicon), goat anti-glucokinase (Santa Cruz), rabbit anti-Ck19 (Melton laboratory stock), rabbit anti-chromogranin A/B (RDI), rabbit anti-Ptf1a (gift from H. Edlund), goat anti-NeuroD (Santa Cruz), rabbit anti-Nkx6.1 (BCBC), rabbit anti-Sox9 (Santa Cruz), goat anti-Nkx2.2 (Santa Cruz) and rabbit anti-c-peptide (Linco).

Rodamin-Red-X-, FITC-, Cy5- and Alexa-dye-conjugated donkey secondary antibodies were obtained from the Jackson Immunoresearch Laboratories and Molecular Probes Inc. Tyramide amplification system (PerkinElmer) was used for PC1/3 and glucokinase staining. Immunofluorescence pictures were taken with a Zeiss LSM 510 META confocal microscope.

***Cpa1CreER^{T2}* labelling of mature exocrine cells.** *Cpa1CreER^{T2};R26R* double heterozygous animals were generated by mating homozygous *Cpa1CreER^{T2}* males with *R26R* homozygous females (Jackson laboratory). Two-month-old *Cpa1CreER^{T2};R26R* adults were injected with tamoxifen at 6 mg per animal every third day four times to label mature exocrine cells.

Physiological studies. Diabetic animals were produced with intraperineal injection of streptozotocin (120 µg per g body weight) after overnight fasting with 2-month-old adult animals of the Rag1 strain (Jackson laboratory). Hyperglycaemic animals that displayed $>250\ mg\ dl^{-1}$ fasting blood glucose levels for at least two consecutive days were used for experiments.

Fasting blood glucose was measured on tail-vein blood with an Ascensia Elite glucometer (Bayer) after 6–8 h fasting. The non-fasting insulin level was determined from tail-vein blood collected around 9 to 10 am with an Ultrasensitive Insulin ELISA kit (Alpco).

The average β-cell number per section was determined by sectioning through the entire pancreas at 15 µm and collecting every third section. Twenty randomly selected sections were immunostained for insulin and 4,6-diamidino-2-phenylindole (DAPI) to visualize individual β-cells. The total number of β-cells was counted and averaged from three animals.

The glucose tolerance test was performed by fasting animals overnight (12 h), followed by intraperineal injection of glucose (3 g per kg body weight).

Electron microscopy. Small pieces of pancreatic samples (1–2 mm) were fixed with 4% paraformaldehyde and 0.1% glutaraldehyde for 2 h at room temperature.

For conventional electron microscopy, samples were further refixed with a mixture of 1% osmiumtetroxide (OsO_4) plus 1.5% potassium ferrocyanide ($KFeCN_6$) for 2 h, were washed in water and stained in 1% aqueous uranyl acetate for 1 h followed by dehydration in grades of alcohol (50%, 70%, 95%, 2 × 100%) and propyleneoxide (1 h), and then infiltrated in propyleneoxide:Epon 1:1 overnight and embedded in TAAB Epon (Marivac Canada Inc.). Ultrathin sections (about 60–80 nm) were cut on a Reichert Ultracut-S microtome, picked up on to copper grids, stained with 0.2% lead citrate and examined in a Tecnai G^2 Spirit BioTWIN transmission electron microscope. Images were taken with an AMT CCD camera.

For immunoelectron microscopy, fixed samples were infiltrated with 2.3 M sucrose in PBS for 30 min then frozen in liquid nitrogen. Frozen samples were sectioned at −120 °C, the sections transferred to formvar–carbon-coated copper grids and floated on PBS until the immunogold labelling was carried out.

The gold labelling was carried out at room temperature on a piece of parafilm. All antibodies and protein-A gold were diluted in 1% BSA. The diluted antibody solution was centrifuged for 1 min at >10,000*g* before labelling to avoid possible aggregates. Grids were floated on drops of 1% BSA for 10 min to block unspecific labelling, transferred to 5-µl drops of primary antibody and incubated for 30 min. The grids were then washed in four drops of PBS for a total of 15 min, transferred to 5-µl drops of protein-A gold (G. Posthuma) for 20 min, and washed in four drops of PBS for 15 min and six drops of double-distilled water.

For double labelling, after the first protein-A gold incubation, grids were washed in four drops of PBS for a total of 15 min and then transferred to a drop of 0.2% glutaraldehyde in PBS for 5 min, and washed in four drops of PBS/0.15 M glycine (to quench free aldehyde groups). Following this, the second primary antibody was applied, followed by PBS wash and different size protein-A gold as described previosuly. The antibodies used were rabbit anti-GFP (Invitrogen) and guinea-pig anti-insulin (Dako).

Contrasting/embedding of the labelled grids was carried out on ice in 0.3% uranyl acetete (Electron Microcopy Sciences) in 2% methyl cellulose (Sigma) for 10 min. Grids were picked up with metal loops (diameter slightly larger than the grid) and the excess liquid was removed by streaking on a filter paper (Whatman, number 1), leaving a thin coat of methyl cellulose (bluish interference colour when dry).

The grids were examined in a Tecnai G^2 Spirit BioTWIN transmission electron microscope and images were recorded with a 2k AMT CCD camera.

FACS analysis, islet isolation and gene profiling. For FACS sorting of GFP$^+$ cells, pancreata infected by the M3 inducing factors for one month were perfused through the common bile duct, digested with liberase and elastase (Roche), and further dissociated into single cells with EDTA incubation. GFP$^+$ cells were isolated by FACS with FACSaria (BD Bioscience). Staining of sorted cells indicates that ~70% of total sorted cells are GFP$^+$ and ~22% are insulin$^+$.

Islets were isolated by liberase digestion of the pancreas of Pdx1–GFP animals. Islets were picked manually under a fluorescent dissecting scope. Pancreatic cells devoid of GFP$^+$ islets were collected as the non-islet sample.

RNA was extracted with Trizol reagent (Invitrogen). Biotin-labelled cRNA probes were synthesized with the Illumina TotalPrep RNA amplification kit (Ambion). Gene profiling was performed with Sentrix BeadChip Array MouseRef-8 v1.1 (Illumina) that contains probes for ~19,000 genes. Data were analysed with BeadStudio software. For identifying differentially enriched genes, the following parameters suggested by Illumina were used: P value < 0.05, Diff score > 30, average signal > 100.

RT–PCR. Pancreatic tissues were harvested and immediately frozen in liquid nitrogen. Total RNA was extracted with the RNeasy kit (Qiagen). First-strand cDNA was synthesized with Superscript III kit (Invitrogen). Thirty cycles of semiquantitative RT–PCR were performed using the standard protocol. The following primer pairs were used: *Ngn3* viral transgene: ~350 bp, *Ngn3*.F: CAGACGCTGCGCATAGCGGACCAC, IRES2.R: GCGGCTTCGGCCAGTAACGTTAG. *Pdx1* viral transgene: ~1.2 kb, *Pdx1*.F: GGAGCAAGATTGTGCGGTGACCTC, IRES2.R: GCGGCTTCGGCCAGTAACGTTAG. *Mafa* viral transgene: ~300 bp, *Mafa*.F: ACATTCTGGAGAGCGAGAAGTGCC, IRES2.R: GCGGCTTCGGCCAGTAACGTTAG. *GADPH*: ~400 bp, F: ACCACAGTCCATGCCATCAC, R: TCCACCACCCTGTTGCTGTA.

ARTICLES

Structure of the *Tribolium castaneum* telomerase catalytic subunit TERT

Andrew J. Gillis[1], Anthony P. Schuller[1] & Emmanuel Skordalakes[1]

A common hallmark of human cancers is the overexpression of telomerase, a ribonucleoprotein complex that is responsible for maintaining the length and integrity of chromosome ends. Telomere length deregulation and telomerase activation is an early, and perhaps necessary, step in cancer cell evolution. Here we present the high-resolution structure of the *Tribolium castaneum* catalytic subunit of telomerase, TERT. The protein consists of three highly conserved domains, organized into a ring-like structure that shares common features with retroviral reverse transcriptases, viral RNA polymerases and B-family DNA polymerases. Domain organization places motifs implicated in substrate binding and catalysis in the interior of the ring, which can accommodate seven to eight bases of double-stranded nucleic acid. Modelling of an RNA–DNA heteroduplex in the interior of this ring demonstrates a perfect fit between the protein and the nucleic acid substrate, and positions the 3′-end of the DNA primer at the active site of the enzyme, providing evidence for the formation of an active telomerase elongation complex.

Telomerase is active in the early stages of life to maintain telomere length and therefore the chromosomal integrity of frequently dividing cells, and it becomes dormant in most somatic cells during adulthood[1,2]. The ability of telomeres to provide genomic stability is diminished over time owing to both the natural loss of telomeric structure with every cell division, and the loss of telomerase activity—a process which leads to ageing[3,4]. In cancer cells, however, telomerase becomes reactivated and works tirelessly to maintain the short length of telomeres of rapidly dividing cells, leading to their immortality[5,6]. The essential role of telomerase in cancer and ageing makes it an important target for the development of therapies to treat cancer and other age-associated disorders.

Telomerase functions as both a monomer and a dimer[7–10], and consists of a protein subunit (TERT) and an integral RNA component (TER) which contains the template that TERT uses to add several DNA repeats to the 3′-end of linear chromosomes[11,12]. TERT, the catalytic subunit of telomerase, is highly conserved among phylogenetic groups and shares common motifs with conventional reverse transcriptases, suggesting an overall conservation of the basic catalytic mechanism between these two classes of enzymes[13,14]. Although TER varies considerably in size, sequence and structure between species, core structural elements are conserved, suggesting that there is a common mechanism of telomere replication among organisms[15,16].

A functional telomerase holoenzyme requires the stable association of the ribonucleoprotein complex, a process mostly carried out by the RNA-binding domain (TRBD)[17,18]. Weak interactions have been reported between TER and both the far amino-terminal domain (a low conservation region of TERT) and the polymerase domain (reverse transcriptase)[18,19]. Current evidence suggests that TRBD binds to the template boundary element of TER, usually a stem loop or a pseudoknot flanked by regions of single-stranded RNA[20–23]. The TRBD–TER association also promotes repeat addition processivity, which is a unique feature of telomerase[19–22,24]. Telomerase repeat addition processivity is also attributed to the IFD (insertion in fingers domain) motif of reverse transcriptase and the carboxy-terminal extension (CTE) proposed to constitute the putative 'thumb' domain of telomerase[25–27].

Initiation of telomere synthesis requires the loading of telomerase onto the end of the chromosomes and the pairing of the 3′-end of the linear DNA substrate with the templating region (usually one and a half repeats of the telomeric repeat)[28–30]. Pairing of the DNA with the RNA template places the 3′-end of the DNA substrate at the active site of the enzyme for nucleotide addition, whereas the RNA template provides the platform for the successive rounds of nucleotide addition and selectivity. RNA–DNA pairing alone is not sufficient for a stable and active telomerase elongation complex and requires extensive contacts of the DNA substrate with both the reverse transcriptase and the putative thumb domain of TERT[25,31]. In some organisms, contacts between the far N-terminal domain and a DNA site upstream of the RNA–DNA hybridization region allow the enzyme to remain attached to the end of the chromosomes during translocation[32,33].

Here we present, to our knowledge, the first high-resolution structure of the catalytic subunit of telomerase. This structure, together with previous biochemical data, provides insights into TERT–TER–DNA assembly and elongation complex formation.

Architecture of the TERT structure

We have solved the structure of the full-length catalytic subunit of the *T. castaneum* active telomerase[34,35], TERT, to 2.71 Å resolution. There is a dimer in the asymmetric unit; however, the protein alone is clearly monomeric in solution as indicated by gel filtration and dynamic light scattering (results not shown) suggesting that the dimer we observe in the crystal is the result of crystal packing. This notion is further supported by the fact that a different crystal form (Supplementary Table 1) of the same protein also contains a dimer in the asymmetric unit of a different configuration than the one presented here. It is worth noting that the TERT from this organism does not contain an N-terminal domain, a low conservation region of telomerase (Fig. 1a, b).

The TERT structure is composed of three distinct domains: an RNA-binding domain (TRBD), the reverse transcriptase domain and the CTE thought to represent the putative thumb domain of TERT (Fig. 1a, c). The TRBD is mostly helical and contains an indentation on its surface formed by two conserved motifs (CP

[1]Gene Expression and Regulation Program, The Wistar Institute, 3601 Spruce Street, Philadelphia, Pennsylvania 19104, USA.

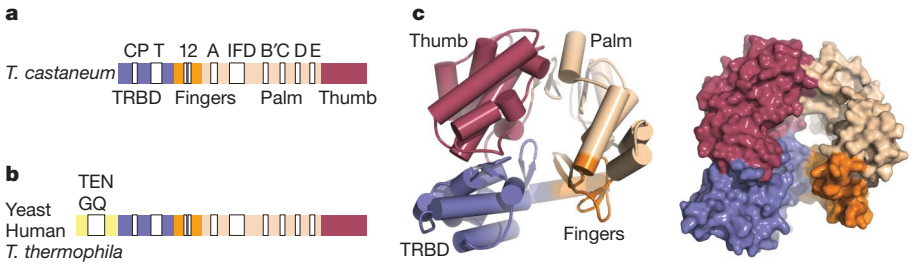

Figure 1 | The structure of TERT. **a**, The primary structure and conserved motifs of the *T. castaneum* TERT are shown. **b**, The primary structure of human, yeast and *T. thermophila* is shown for comparison. **c**, TERT domain organization (cartoon and surface representation); the RNA-binding domain (TRBD) is shown in violet, the fingers domain is in orange, the palm domain is in tan and the thumb domain is in red.

and T) known to bind the double- and single-stranded RNA regions of the template boundary element, respectively[24] (Fig. 2a). Structural comparison of the TRBD from *T. castaneum* with that of the previously determined structure from *Tetrahymena thermophila*[24] shows similarity between the two structures (root mean squared deviation (r.m.s.d.) 2.7 Å), suggesting that a high degree of structural conservation occurs between these domains across organisms of diverse phylogenetic groups.

The reverse transcriptase domain is a mixture of α-helices and β-strands organized into two subdomains that are most similar to the 'fingers' and 'palm' subdomains of retroviral reverse transcriptases[36], viral RNA polymerases[37] and B-family DNA polymerases[38] (Supplementary Fig. 1a–d), and contains important signature motifs that are hallmarks of these families of proteins[14] (Fig. 2b). Structural comparisons of TERT with the HIV reverse transcriptases show that the fingers subdomain of TERT is arranged in the open configuration with respect to the palm subdomain, which is in good agreement with the conformation adopted by HIV reverse transcriptases in the absence of bound nucleotide and nucleic acid substrates[39]. One notable difference between the putative palm domain of TERT and the HIV reverse transcriptases is a long insertion between motifs A and B' of TERT; this is referred to as the IFD motif and is required for telomerase processivity[27]. In the TERT structure, the IFD insertion consists of two antiparallel α-helices (α13 and α14) located on the outside periphery of the ring and at the interface of the fingers and the palm subdomains (Fig. 2b). These two helices are almost in a parallel position with the central axis of the plane of the ring, make extensive contacts with helices α10 and α15, and have an important role in the structural organization of this part of the reverse transcriptase domain. A similar structural arrangement is also present in viral polymerases, and the equivalent of helix α10 in these structures is involved in direct contacts with the nucleic acid substrate[40] (Supplementary Fig. 1c).

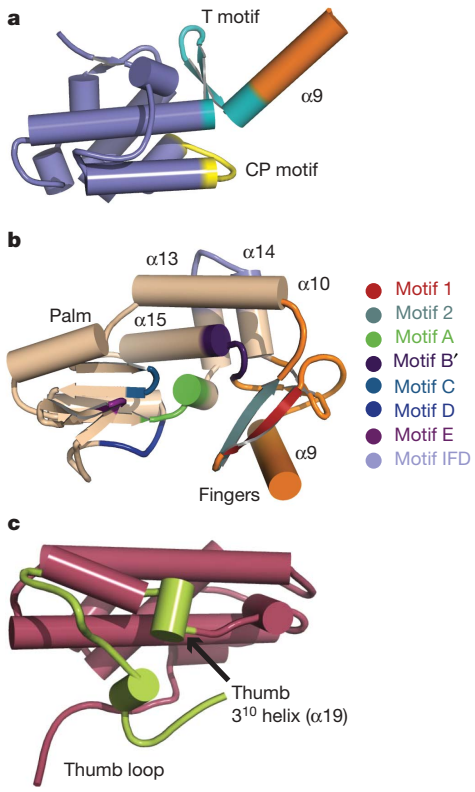

Figure 2 | The TERT domain fold and main signature motifs. **a**, The almost-all helical TRBD binds RNA using two conserved motifs: T (shown in cyan) and CP (shown in yellow). **b**, The reverse transcriptase domain consists of the palm and fingers subdomains. The fingers subdomain contains motifs 1 (red) and 2 (grey) and is implicated in nucleotide and RNA binding. The palm subdomain contains motifs A (green), B' (dark purple), C (blue), D (navy blue), E (magenta) and IFD (light blue) and is involved in nucleotide and nucleic acid binding and DNA synthesis. **c**, The helical thumb domain of TERT contains a loop (thumb loop; light green) which is involved in domain organization and DNA binding.

In contrast to the reverse transcriptase domain, the CTE is an elongated helical bundle that contains several surface-exposed long loops (Fig. 2c). A search in the protein structure database using the secondary-structure matching software (http://www.ebi.ac.uk/msd-srv/ssm)[41] produced no structural homologues, suggesting that the CTE domain of telomerase adopts a new fold. Structural comparison of TERT with the HIV reverse transcriptase, with the viral RNA polymerases and with the B-family DNA polymerases places the thumb domain of these enzymes and the CTE domain of TERT in the same spatial position with respect to the fingers and palm subdomains. This suggests that the CTE domain of telomerase is the thumb domain of the enzyme, a finding that is in good agreement with previous biochemical studies[25] (Supplementary Fig. 2).

TERT domain organization brings the TRBD and thumb domain—which constitute the terminal domains of the molecule—together, an arrangement that leads to the formation of a ring-like structure that is reminiscent of the shape of a doughnut (Fig. 1a, b). Several lines of evidence suggest that the domain organization of the TERT structure presented here is biologically relevant. First, the domains of four TERT monomers observed in two different crystal forms (two in each asymmetric unit) all have the same organization (average r.m.s.d. = 0.76 Å between all four monomers). Second, contacts between the N- and the C-terminal domains of TERT are extensive (1,677 Å^2) and largely hydrophobic in nature, an observation that is consistent with previous biochemical studies[42] (Supplementary Fig. 3). Third, TERT domain organization is similar to that of the polymerase domain (p66 minus the RNase H domain) of its closest homologue, HIV reverse transcriptase[36]. It is also similar to the domain organization of the viral RNA polymerases[37] and that of the B-family DNA polymerases, particularly RB69 (ref. 38; Supplementary Fig. 1a–d). The arrangement of the TERT domains creates a hole in the interior of the particle that is ~26 Å wide and

~21 Å deep, sufficient to accommodate double-stranded nucleic acid approximately seven to eight bases long and in good agreement with existing biochemical data[43,44].

The TERT ring binds double-stranded nucleic acid

To understand better how the TERT ring associates with RNA–DNA to form a functional elongation complex, we modelled double-stranded nucleic acid into its interior using the complex of HIV reverse transcriptase with DNA[36], the closest structural homologue of TERT (Fig. 3a). The TERT–RNA–DNA model immediately shows some notable features that support our model of TERT–nucleic-acid associations. The hole of the TERT ring, and where the nucleic acid heteroduplex is projected to bind, is lined with several key signature motifs that are hallmarks of this family of polymerases and have been implicated in nucleic acid association, nucleotide binding and DNA synthesis (Fig. 3a). Moreover, the organization of these motifs results in the formation of a spiral in the interior of the ring that resembles the geometry of the backbone of double-stranded nucleic acid (Fig. 3b). Several of the motifs, identified as contact points with the DNA substrate, are formed partly by positively charged residues, the side chains of which extend towards the centre of the ring and are poised for direct contact with the backbone of the DNA substrate. For example, the side chain of the highly conserved K210 (Supplementary Fig. 4) that forms part of helix α10, is within coordinating distance of the backbone of the modelled DNA, thus providing the stability required for a functional telomerase enzyme. Helix α10 lies in the upper segment of the reverse transcriptase domain and faces the interior of the ring. The location and stabilization of this helix is heavily influenced by its extensive contacts with the IFD motif implicated in telomerase processivity[27]. Disruption of the IFD contacts with helix α10—by deletion or mutation of this motif—would lead to displacement of helix α10 from its current location, which would in turn effect DNA-binding and telomerase function.

Structural elements of the thumb domain that localize to the interior of the ring also make several contacts with the modelled DNA substrate (Fig. 3a). In particular, the loop (thumb loop) that connects the palm to the thumb domain and constitutes an extension of motif E, also known as the 'primer grip' region of telomerase, preserves the geometry of the backbone of double-stranded nucleic acid to a notable degree (Fig. 3b). The side chains of several lysines and asparagines that form part of this loop extend towards the centre of the TERT molecule and are in coordinating distance of the backbone of modelled double-stranded nucleic acid. Of particular interest is K406, located in proximity of motif E. The side chain of this lysine extends towards the nucleic acid heteroduplex and it is poised for direct contacts with the backbone of the nucleotides located at the 3′-end of the incoming DNA primer. It is therefore possible that the side chain of this lysine, together with motif E, help facilitate placement of the 3′-end of the incoming DNA substrate at the active site of the enzyme during telomere elongation. Sequence alignments of the thumb domain of TERTs from a wide spectrum of phylogenetic groups show that the residues predicted to contact the DNA substrate are always polar (Supplementary Fig. 4). Another interesting feature of the thumb domain, which supports double-stranded nucleic acid binding, is helix α19 (Fig. 2c). This is a 3^{10} helix (thumb 3^{10} helix) that extends into the interior of the ring and seems to dock itself into the minor groove of the modelled double-stranded nucleic acid, thus facilitating RNA–DNA hybrid binding and stabilization (Fig. 3b). Deletion or mutation of the corresponding residues in both yeast and human TERT results in severe loss of TERT processivity, clearly indicating the important role of this motif in TERT function[25,26,45].

The active site of TERT and nucleotide binding

The TERT structure presented here was crystallized in the absence of nucleotide substrates and magnesium; however, the location and organization of TERT's active site and nucleotide-binding pocket can be predicted on the basis of existing biochemical data[14] and structural comparison with the polymerase domain of its closest homologue, the HIV reverse transcriptase[46]. The TERT active site consists of three invariant aspartic acids (D251, D343 and D344) that form part of motifs A and C, which are two short loops located on the palm subdomain and adjacent to the fingers of TERT (Fig. 4a). Structural comparisons of TERT with HIV reverse transcriptases, as well as with RNA and DNA polymerases, show a high degree of similarity between the active sites of these families of proteins (Fig. 4b), suggesting that telomerase also uses a two-metal mechanism for catalysis. Alanine mutants of these TERT aspartic acids resulted in complete loss of TERT activity, indicating that the role of these residues in telomerase function is essential[14].

The telomerase nucleotide-binding pocket is located at the interface of the fingers and palm subdomains of TERT (Fig. 4a) and consists of conserved residues that form the motifs 1, 2, A, C, B' and D which are implicated in template and nucleotide binding[47,48] (Supplementary Fig. 5). Structural comparisons of TERT with viral HIV reverse transcriptases bound to ATP[46] support the presence of a nucleotide substrate in this location. Two highly conserved surface-exposed residues Y256 and V342 of motifs A and C, respectively, form a hydrophobic pocket adjacent to and above the three catalytic

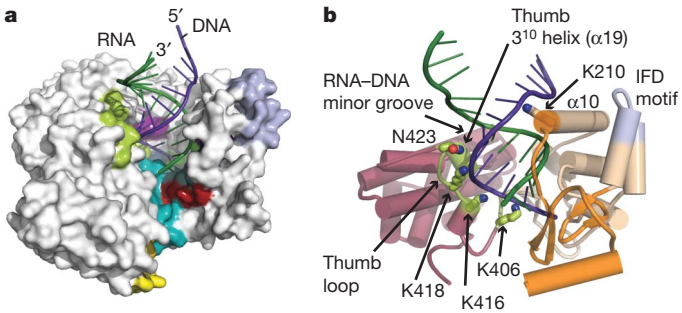

Figure 3 | Model of the TERT–RNA–DNA complex. **a**, Surface representation of TERT with the modelled RNA (dark green) and DNA (dark purple) shown in cartoon. The main signature motifs are coloured as in Fig. 2. The RNA-binding pocket of TERT is located in the deep cavity on the side of the ring. **b**, Contacts between the TERT ring and a modelled RNA–DNA heteroduplex are shown. The path of the incoming DNA primer within the TERT ring is lined up with several conserved positively charged residues poised for direct contacts with its backbone. The thumb loop and the thumb 3^{10} helix are shown in light green; the IFD motif is shown in light blue.

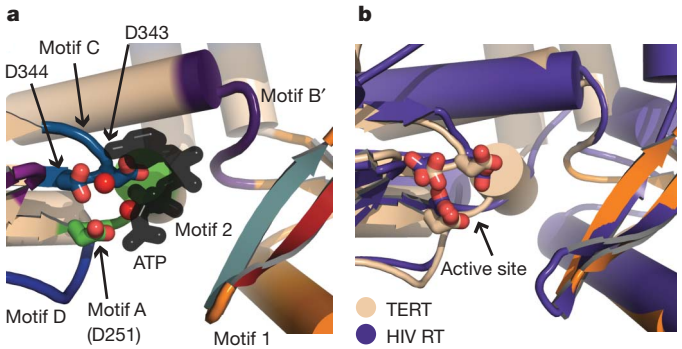

Figure 4 | The active site and nucleotide-binding pocket of telomerase. **a**, The active site of telomerase is formed by three invariant aspartic acids that form part of motifs A (D251) and C (D343 and D344). The nucleotide-binding pocket is located at the interface of the fingers and palm subdomains of reverse transcriptase; a modelled nucleotide using the HIV reverse transcriptase (PDB code 2IAJ) is shown as a black stick. **b**, The overlay of the active site residues of TERT (tan) and HIV reverse transcriptase (RT; purple; PDB code 1N5Y) shows that there is a high degree of similarity between the pockets of the two enzymes.

aspartates and this could accommodate the base of the nucleotide substrate. Binding of the nucleotide in this oily pocket places the triphosphate moiety in proximity of the enzyme's active site for coordination with one of the Mg^{2+} ions. In contrast, it positions the ribose group within coordinating distance of an invariant glutamine (Q308) that forms part of motif B', which is thought to be an important determinant of substrate specificity[49]. Protein contacts with the triphosphate moiety of the nucleotide are mediated by motif D, a long loop located beneath the active site of the enzyme. In particular, the side chain of the invariant K372 is within coordinating distance of the γ-phosphate of the nucleotide, an interaction that probably helps position and stabilize the triphosphate group during catalysis. The side chains of the highly conserved K189 and R192 of motifs 1 and 2, which together form a long β-hairpin that forms part of the fingers subdomain, are also within coordinating distance of both the sugar and triphosphate moieties of the modelled nucleotide. Contacts with either or both the sugar moiety and the triphosphate moiety of the nucleotide substrate would facilitate nucleotide binding and positioning for coordination to the 3′-end of the incoming DNA primer.

TRBD facilitates template positioning at the active site

As with most DNA and RNA polymerases, nucleic acid synthesis by telomerase requires pairing of the templating region (usually seven to eight bases or more) of TER with the incoming DNA primer[28]. TRBD–reverse-transcriptase domain organization forms a deep cavity on the surface of the protein that spans the entire width of the wall of the molecule, forming a gap that allows entry into the hole of the ring from its side (Fig. 3a). The arrangement of this cavity with respect to the central hole of the ring provides an elegant mechanism upon TERT–TER assembly for the placement of the RNA template in the interior of the ring and where the enzyme's active site is located. Of particular significance is the arrangement of the β-hairpin that forms part of the T motif. This hairpin extends from the RNA-binding pocket and makes extensive contacts with the thumb loop and motifs 1 and 2 (Fig. 3a). Contacts between this hairpin and both the fingers and the thumb domains place the opening of the TRBD pocket that faces the interior of the ring in proximity to the active site of the enzyme (Fig. 5). It is therefore possible that this β-hairpin acts as an allosteric effector switch that couples RNA binding in the interior of the ring and placement of the RNA template at the active site of the enzyme. Placement of the template into the interior of the molecule would facilitate its pairing with the incoming DNA substrate, which together would form the RNA–DNA hybrid required for telomere elongation. RNA–DNA pairing is a prerequisite of telomere synthesis in that it brings the 3′-end of the incoming DNA primer in proximity to the active site of the enzyme for nucleotide addition, and the RNA component of the heteroduplex provides the template for the addition of identical repeats of DNA at the ends of chromosomes. Notably, modelling of the RNA–DNA heteroduplex in the interior of the TERT ring places the 5′-end of the RNA substrate at the entry of the RNA-binding pocket and where TERT is expected to associate with TER, whereas it places the 3′-end of the incoming DNA primer at the active site of TERT providing a snapshot of the organization of a functional telomerase elongation complex (Fig. 5).

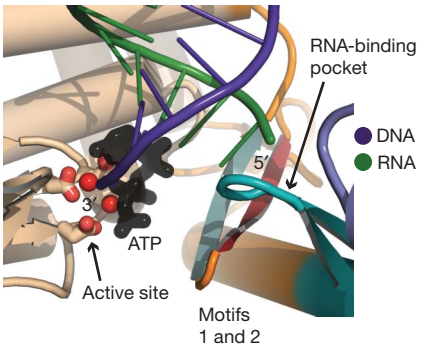

Figure 5 | Localization of the RNA–DNA ends in the interior of the ring. TERT domain organization places the RNA-binding pocket (cyan) and the active site of the enzyme in close proximity to each other. Modelling of the RNA–DNA heteroduplex in the TERT ring places the 3′-end of the DNA substrate (dark purple) at the active site of the enzyme, and the 5′-end of the RNA substrate (dark green) at the RNA-binding pocket of telomerase.

Conclusions

The structure presented here provides a view of the full-length catalytic subunit of telomerase. The structure shows that TERT is organized into an unexpected ring configuration that resembles—both structurally and functionally—the HIV reverse transcriptases, the viral RNA polymerases and the B-family DNA polymerases, suggesting that there is an evolutionary link between these families of enzymes. It also provides insights into the mechanism of TERT and RNA–DNA association, which in turn explains how TERT may assemble with RNA–DNA and offers a snapshot of a functional telomerase elongation complex required for telomere synthesis. Moreover, because telomerase has a critical role in both cancer and ageing, these findings could potentially assist our efforts to identify and develop inhibitors and/or activators of this enzyme for the treatment of cancer and ageing, respectively.

METHODS SUMMARY

The full-length TERT of *T. castaneum* was overexpressed in bacteria and purified by nickel, ion-exchange and gel-filtration chromatography. Co-crystallization of the protein–telomeric-DNA ($(TCAGG)_3$) produced two crystal forms (orthorhombic and hexagonal), which were grown by the vapour diffusion, sitting-drop method. Data were collected at the National Synchrotron Light Source (NSLS) at beamline X6A and were processed with MOSFILM (Supplementary Table 1). Phases for the orthorhombic crystal were obtained by the method of single isomorphous replacement with anomalous signal using a mercury derivative (CH_3HgCl; Supplementary Table 1). The model from the orthorhombic crystal was subsequently used to solve the hexagonal crystal form by molecular replacement. Both models were refined to good stereochemistry (Supplementary Table 1).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 3 June 2008; accepted 23 July 2008.
Published online 31 August 2008.

1. Blackburn, E. H. Telomeres: no end in sight. *Cell* **77,** 621–623 (1994).
2. Greider, C. W. & Blackburn, E. H. Identification of a specific telomere terminal transferase activity in *Tetrahymena* extracts. *Cell* **43,** 405–413 (1985).
3. Wright, W. E. & Hayflick, L. Nuclear control of cellular aging demonstrated by hybridization of anucleate and whole cultured normal human fibroblasts. *Exp. Cell Res.* **96,** 113–121 (1975).
4. Wright, W. E. & Shay, J. W. Cellular senescence as a tumor-protection mechanism: the essential role of counting. *Curr. Opin. Genet. Dev.* **11,** 98–103 (2001).
5. Bodnar, A. G. *et al.* Extension of life-span by introduction of telomerase into normal human cells. *Science* **279,** 349–352 (1998).
6. Campisi, J., Kim, S. H., Lim, C. S. & Rubio, M. Cellular senescence, cancer and aging: the telomere connection. *Exp. Gerontol.* **36,** 1619–1637 (2001).
7. Beattie, T. L., Zhou, W., Robinson, M. O. & Harrington, L. Functional multimerization of the human telomerase reverse transcriptase. *Mol. Cell. Biol.* **21,** 6151–6160 (2001).
8. Bryan, T. M., Goodrich, K. J. & Cech, T. R. *Tetrahymena* telomerase is active as a monomer. *Mol. Biol. Cell* **14,** 4794–4804 (2003).
9. Moriarty, T. J., Huard, S., Dupuis, S. & Autexier, C. Functional multimerization of human telomerase requires an RNA interaction domain in the N terminus of the catalytic subunit. *Mol. Cell. Biol.* **22,** 1253–1265 (2002).
10. Prescott, J. & Blackburn, E. H. Functionally interacting telomerase RNAs in the yeast telomerase complex. *Genes Dev.* **11,** 2790–2800 (1997).
11. Feng, J. *et al.* The RNA component of human telomerase. *Science* **269,** 1236–1241 (1995).
12. Nakamura, T. M. *et al.* Telomerase catalytic subunit homologs from fission yeast and human. *Science* **277,** 955–959 (1997).
13. Counter, C. M., Meyerson, M., Eaton, E. N. & Weinberg, R. A. The catalytic subunit of yeast telomerase. *Proc. Natl Acad. Sci. USA* **94,** 9202–9207 (1997).

14. Lingner, J. *et al.* Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science* **276,** 561–567 (1997).
15. Chen, J. L. & Greider, C. W. An emerging consensus for telomerase RNA structure. *Proc. Natl Acad. Sci. USA* **101,** 14683–14684 (2004).
16. Ly, H., Blackburn, E. H. & Parslow, T. G. Comprehensive structure–function analysis of the core domain of human telomerase RNA. *Mol. Cell. Biol.* **23,** 6849–6856 (2003).
17. Lai, C. K., Mitchell, J. R. & Collins, K. RNA binding domain of telomerase reverse transcriptase. *Mol. Cell. Biol.* **21,** 990–1000 (2001).
18. O'Connor, C. M., Lai, C. K. & Collins, K. Two purified domains of telomerase reverse transcriptase reconstitute sequence-specific interactions with RNA. *J. Biol. Chem.* **280,** 17533–17539 (2005).
19. Friedman, K. L. & Cech, T. R. Essential functions of amino-terminal domains in the yeast telomerase catalytic subunit revealed by selection for viable mutants. *Genes Dev.* **13,** 2863–2874 (1999).
20. Chen, J. L. & Greider, C. W. Template boundary definition in mammalian telomerase. *Genes Dev.* **17,** 2747–2752 (2003).
21. Lai, C. K., Miller, M. C. & Collins, K. Template boundary definition in *Tetrahymena* telomerase. *Genes Dev.* **16,** 415–420 (2002).
22. Seto, A. G. *et al.* A template-proximal RNA paired element contributes to *Saccharomyces cerevisiae* telomerase activity. *RNA* **9,** 1323–1332 (2003).
23. Tzfati, Y., Fulton, T. B., Roy, J. & Blackburn, E. H. Template boundary in a yeast telomerase specified by RNA structure. *Science* **288,** 863–867 (2000).
24. Rouda, S. & Skordalakes, E. Structure of the RNA-binding domain of telomerase: implications for RNA recognition and binding. *Structure* **15,** 1403–1412 (2007).
25. Hossain, S., Singh, S. & Lue, N. F. Functional analysis of the C-terminal extension of telomerase reverse transcriptase. A putative "thumb" domain. *J. Biol. Chem.* **277,** 36174–36180 (2002).
26. Huard, S., Moriarty, T. J. & Autexier, C. The C terminus of the human telomerase reverse transcriptase is a determinant of enzyme processivity. *Nucleic Acids Res.* **31,** 4059–4070 (2003).
27. Lue, N. F., Lin, Y. C. & Mian, I. S. A conserved telomerase motif within the catalytic domain of telomerase reverse transcriptase is specifically required for repeat addition processivity. *Mol. Cell. Biol.* **23,** 8440–8449 (2003).
28. Lee, M. S. & Blackburn, E. H. Sequence-specific DNA primer effects on telomerase polymerization activity. *Mol. Cell. Biol.* **13,** 6586–6599 (1993).
29. Lingner, J., Hendrick, L. L. & Cech, T. R. Telomerase RNAs of different ciliates have a common secondary structure and a permuted template. *Genes Dev.* **8,** 1984–1998 (1994).
30. Shippen-Lentz, D. & Blackburn, E. H. Functional evidence for an RNA template in telomerase. *Science* **247,** 546–552 (1990).
31. Finger, S. N. & Bryan, T. M. Multiple DNA-binding sites in *Tetrahymena* telomerase. *Nucleic Acids Res.* **36,** 1260–1272 (2008).
32. Jacobs, S. A., Podell, E. R. & Cech, T. R. Crystal structure of the essential N-terminal domain of telomerase reverse transcriptase. *Nature Struct. Mol. Biol.* **13,** 218–225 (2006).
33. Lue, N. F. A physical and functional constituent of telomerase anchor site. *J. Biol. Chem.* **280,** 26586–26591 (2005).
34. Osanai, M., Kojima, K. K., Futahashi, R., Yaguchi, S. & Fujiwara, H. Identification and characterization of the telomerase reverse transcriptase of *Bombyx mori* (silkworm) and *Tribolium castaneum* (flour beetle). *Gene* **376,** 281–289 (2006).
35. Richards, S. *et al.* The genome of the model beetle and pest *Tribolium castaneum*. *Nature* **452,** 949–955 (2008).
36. Sarafianos, S. G. *et al.* Structures of HIV-1 reverse transcriptase with pre- and post-translocation AZTMP-terminated DNA. *EMBO J.* **21,** 6614–6624 (2002).
37. Di Marco, S. *et al.* Interdomain communication in hepatitis C virus polymerase abolished by small molecule inhibitors bound to a novel allosteric site. *J. Biol. Chem.* **280,** 29765–29770 (2005).
38. Wang, J. *et al.* Crystal structure of a pol α family replication DNA polymerase from bacteriophage RB69. *Cell* **89,** 1087–1099 (1997).
39. Ding, J. *et al.* Structure and functional implications of the polymerase active site region in a complex of HIV-1 RT with a double-stranded DNA template-primer and an antibody Fab fragment at 2.8 Å resolution. *J. Mol. Biol.* **284,** 1095–1111 (1998).
40. Ferrer-Orta, C. *et al.* Structure of foot-and-mouth disease virus RNA-dependent RNA polymerase and its complex with a template-primer RNA. *J. Biol. Chem.* **279,** 47212–47221 (2004).
41. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions, *Acta Cryst. D* **60,** 2256–2268 (2004).
42. Arai, K. *et al.* Two independent regions of human telomerase reverse transcriptase are important for its oligomerization and telomerase activity. *J. Biol. Chem.* **277,** 8538–8544 (2002).
43. Forstemann, K. & Lingner, J. Telomerase limits the extent of base pairing between template RNA and telomeric DNA. *EMBO Rep.* **6,** 361–366 (2005).
44. Hammond, P. W. & Cech, T. R. Euplotes telomerase: evidence for limited base-pairing during primer elongation and dGTP as an effector of translocation. *Biochemistry* **37,** 5162–5172 (1998).
45. Banik, S. S. *et al.* C-terminal regions of the human telomerase catalytic subunit essential for *in vivo* enzyme activity. *Mol. Cell. Biol.* **22,** 6234–6246 (2002).
46. Das, K. *et al.* Crystal structures of clinically relevant Lys103Asn/Tyr181Cys double mutant HIV-1 reverse transcriptase in complexes with ATP and non-nucleoside inhibitor HBY 097. *J. Mol. Biol.* **365,** 77–89 (2007).
47. Bosoy, D. & Lue, N. F. Functional analysis of conserved residues in the putative "finger" domain of telomerase reverse transcriptase. *J. Biol. Chem.* **276,** 46305–46312 (2001).
48. Haering, C. H., Nakamura, T. M., Baumann, P. & Cech, T. R. Analysis of telomerase catalytic subunit mutants *in vivo* and *in vitro* in *Schizosaccharomyces pombe*. *Proc. Natl Acad. Sci. USA* **97,** 6367–6372 (2000).
49. Smith, R. A., Anderson, D. J. & Preston, B. D. Hypersusceptibility to substrate analogs conferred by mutations in human immunodeficiency virus type 1 reverse transcriptase. *J. Virol.* **80,** 7169–7178 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Marmorstein for critical reading of this manuscript. This project was funded by the Pennsylvania Department of Health and The Ellison Medical Foundation.

Author Contributions E.S. designed the experimental plan, carried out the research and wrote the paper. A.J.G. and A.P.S. assisted with the experimental work.

Author Information The atomic coordinates and structure factors have been deposited in the Protein Data Bank under accession numbers 3DU5 and 3DU6. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.S. (skorda@wistar.org).

METHODS

Protein expression and purification. The synthetic gene of *T. castaneum* full-length TERT was cloned into a modified version of the pET28b vector containing a cleavable hexahistidine tag at its N terminus. The protein was overexpressed in *Escherichia coli* BL21 (pLysS) at 30 °C for 4 h. The cells were lysed by sonication in 50 mM Tris-HCl, pH 7.5, 10% glycerol, 0.5 M KCl, 5 mM β-mercaptoethanol and 1 mM phenylmethyl sulphonyl fluoride, on ice. The protein was purified over a Ni-NTA column followed by TEV cleavage of the hexahistidine tag overnight at 4 °C. The TERT–TEV mixture was dialysed to remove the excess imidazole and the protein was further purified over a second Ni-NTA column that was used to remove all His-tagged products. The Ni-NTA flow through was then passed over a POROS-HS column (Perseptive Biosystems) to remove any trace amounts of protein contaminants. At this stage the protein was more than 99% pure. The protein was finally purified over a sephedex-S200 sizing column pre-equilibrated with 50 mM Tris-HCl, pH 7.5, 10% glycerol, 0.5 M KCl and 1 mM Tris(2-carboxyethyl)phosphine (TCEP) to remove any TERT aggregates, and the protein was concentrated to 10 mg ml^{-1} using an Amicon 30K cutoff (Millipore) and stored at 4 °C for subsequent studies. Stock protein was dialysed in 10 mM Tris-HCl, pH 7.5, 200 mM KCl and 1 mM TCEP before crystallization trials.

Protein crystallization and data collection. Initial crystal trials of the protein alone did not produce crystals. Co-crystallization of the protein with single-stranded telomeric DNA ($(TCAGG)_3$) produced two rod-like crystal forms, one of which belongs to the orthorhombic space group $P2_12_12_1$ and diffracted to 2.71 Å, and the other belongs to the hexagonal space group $P6_1$ and diffracted to 3.25 Å resolution. The protein–nucleic-acid mix was prepared before setting crystal trials by mixing one volume of dialysed protein with a 1.2-fold excess of the DNA substrate. Both crystal forms were grown by the vapour diffusion, sitting-drop method by mixing one volume of the protein–DNA mix with one volume of reservoir solution. Orthorhombic crystals were grown in the presence of 50 mM HEPES, pH 7.0, and 1.5 M $NaNO_3$, whereas hexagonal crystals were grown in 100 mM Tris, pH 8.0, and 2 M $(NH_4)_2SO_4$, and both crystal forms were grown at room temperature. Orthorhombic crystals were collected into cryoprotectant solution that contained 50 mM HEPES, pH 7.0, 25% glycerol, 1.7 M $NaNO_3$, 0.2 M KCl and 1 mM TCEP and were flash frozen in liquid nitrogen. Hexagonal crystals were collected into cryoprotectant solution that contained 100 mM Tris, pH 8.0, 25% glycerol, 2 M $(NH_4)_2SO_4$, 0.2 M KCl and 1 mM TCEP and were also flash frozen in liquid nitrogen. Data were collected at the NSLS at beamline X6A and processed with HKL-2000 (ref. 50; Supplementary Table 1). Both crystal forms contain a dimer in the asymmetric unit.

Structure determination and refinement. Initial phases for the orthorhombic crystals were obtained using the method of single isomorphous replacement with anomalous signal using two data sets collected from two different mercury-derivatized crystals at two different wavelengths (Hg1, 1.00850 Å; Hg2, 1.00800 Å; Supplementary Table 1). The derivatives were prepared by soaking the crystals with 5 mM methyl mercury chloride (CH_3HgCl) for 15 min. Initially, twelve heavy atom sites were located using SOLVE[51], and they were refined and new phases were calculated with MLPHARE[52]. The MLPHARE-improved phases were used to identify the remaining heavy atom sites (22 in total) by calculating an anomalous difference map to a resolution of 3.5 Å. The MLPHARE phases obtained using all the heavy atom sites were then used in DM (density modification package) with two-fold non-crystallographic symmetry and phase extension, using the high-resolution (2.71 Å) data set collected at 1.00800 Å wavelength to calculate starting experimental maps. These maps were of sufficient quality for model building which was carried out in COOT[53]. The electron density map shows clear density for all 596 residues of the protein. Notably, we did not observe density for the nucleic acid substrate in the structure. The model was refined using both CNS-SOLVE[54] and REFMAC5 (ref. 55). The last cycles of refinement were carried out with TLS restraints as implemented in REFMAC5 (Supplementary Table 1). The $P2_12_12_1$ refined model was used to solve the structure of the TERT crystallized in the $P6_1$ crystal form (data collected at 0.97980 Å wavelength) by molecular replacement with PHASER[56].

50. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
51. Terwilliger, T. C. SOLVE and RESOLVE: automated structure solution and density modification. *Methods Enzymol.* **374,** 22–37 (2003).
52. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
53. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
54. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
55. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53,** 240–255 (1997).
56. Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. A graphical user interface to the CCP4 program suite. *Acta Crystallogr. D* **59,** 1131–1137 (2003).

LETTERS

An 84-μG magnetic field in a galaxy at redshift $z = 0.692$

Arthur M. Wolfe[1], Regina A. Jorgenson[1], Timothy Robishaw[2], Carl Heiles[2] & Jason X. Prochaska[3]

The magnetic field pervading our Galaxy is a crucial constituent of the interstellar medium: it mediates the dynamics of interstellar clouds, the energy density of cosmic rays, and the formation of stars[1]. The field associated with ionized interstellar gas has been determined through observations of pulsars in our Galaxy. Radio-frequency measurements of pulse dispersion and the rotation of the plane of linear polarization, that is, Faraday rotation, yield an average value for the magnetic field of $B \approx 3\,\mu$G (ref. 2). The possible detection of Faraday rotation of linearly polarized photons emitted by high-redshift quasars[3] suggests similar magnetic fields are present in foreground galaxies with redshifts $z > 1$. As Faraday rotation alone, however, determines neither the magnitude nor the redshift of the magnetic field, the strength of galactic magnetic fields at redshifts $z > 0$ remains uncertain. Here we report a measurement of a magnetic field of $B \approx 84\,\mu$G in a galaxy at $z = 0.692$, using the same Zeeman-splitting technique that revealed an average value of $B = 6\,\mu$G in the neutral interstellar gas of our Galaxy[4]. This is unexpected, as the leading theory of magnetic field generation, the mean-field dynamo model, predicts large-scale magnetic fields to be weaker in the past rather than stronger[5].

We detected Zeeman splitting of the $z = 0.692$, 21-cm absorption line in the direction of the quasar 3C 286 (refs 6, 7) using the 100-m Robert C. Byrd Green Bank Telescope (GBT) of the National Radio Astronomy Observatory. The absorption arises in a damped Lyman α (Lyα) system (henceforth denoted DLA-3C286) that is drawn from a population of neutral gas layers widely thought to be the progenitors of modern galaxies[8]. The radio data for DLA-3C286 are summarized in Fig. 1, which shows the line-depth spectra constructed from the measurable quantities used to describe polarized radiation, that is, the Stokes parameters. We show line-depth spectra constructed from the $I(v)$ and $V(v)$ Stokes parameters (where v denotes frequency) near the 839.4-MHz frequency centroid of the redshifted 21-cm absorption line. Figure 1a shows the line-depth spectrum constructed from $I(v)$. A Gaussian fit to the absorption line in Fig. 1a yields a redshift of $z = 0.6921526 \pm 0.0000008$, a central optical depth of $\tau_0 = 0.095 \pm 0.006$, and a velocity dispersion of $\sigma_v = 3.75 \pm 0.20\,\mathrm{km\,s^{-1}}$, which are in good agreement with previous results[6,7].

In Fig. 1b, we plot the line-depth spectrum constructed from $V(v)$, which shows the classic 'S curve' pattern expected for Zeeman splitting. From our least-squares fit to the data, we find that $B_{\rm los} = 83.9 \pm 8.8\,\mu$G, where $B_{\rm los}$ is the magnetic field component projected along the line of sight (we note that the direction of $B_{\rm los}$ is unknown because the instrumental sense of circular polarization was not calibrated). This magnetic field differs in two respects from the magnetic fields obtained from Zeeman splitting arising in interstellar clouds in the Galaxy. First, the field strength corresponds to the line-of-sight component of the mean field $\langle B_{\rm los}\rangle$ averaged over transverse dimensions exceeding 200 pc, as very-long-baseline interferometry observations of the 21-cm absorption line show that the gas layer must extend across more than 0.03″ to explain the difference between the velocity centroids of the fringe amplitude and phase-shift spectra[9] (although the data are consistent with a magnetic field coherence length of less than 200 pc, the resulting gradient in magnetic pressure would produce velocity differences exceeding the shift of $\sim 3\,\mathrm{km\,s^{-1}}$ across 200 pc detected by very-long-baseline interferometry). By contrast, the transverse dimensions of radio beams subtended at neutral interstellar clouds in the Galaxy are typically less than 1 pc. Second, this field is at least an order of magnitude stronger than the 6-μG average of magnetic fields inferred from Zeeman splitting for such clouds[4].

We obtained further information about conditions in the absorbing gas in DLA-3C286 from accurate optical spectra acquired with the HIRES echelle spectrograph on the Keck I 10-m telescope. Figure 2 shows velocity profiles for several resonance absorption lines arising from dominant low-ionization states of abundant elements. The results of our least-squares fit of Voigt profiles to the data are shown in Table 1, where the optical redshift is displaced $+3.8 \pm 0.2\,\mathrm{km\,s^{-1}}$ from the 21-cm redshift. This solution also yields ionic column densities from which we derived the logarithmic metal abundances with respect to solar abundances, [M/H], and dust-to-gas ratios with respect to the Galactic interstellar medium, [D/G]. These are among the lowest values of [M/H] and [D/G] deduced for damped Lyα systems at $z = 0.7$ (refs 10, 11). The low metallicity indicates a history of low star formation rates. Because the intensity of far-ultraviolet radiation emitted by young massive stars is proportional to the concurrent star formation rate per unit area, $\Sigma_{\rm SFR}$, low values of $\Sigma_{\rm SFR}$ should result in low grain photoelectric heating rates per hydrogen atom, $\Gamma_{\rm pe}$ (ref. 11). This is consistent with the low upper limit, $\Gamma_{\rm pe} < 10^{-27.4}\,\mathrm{erg\,s^{-1}}$ per hydrogen atom, obtained by combining the assumption of thermal balance with the absence of C II* absorption (that is, absorption from C II in the excited $^2P_{3/2}$ fine-structure state) at a wavelength of 1,335.7 Å in the previous low-resolution Hubble Space Telescope spectra of quasar 3C 286 (ref. 12), and indicates that $\Sigma_{\rm SFR} < 10^{-2.9} M_\odot\,\mathrm{yr^{-1}\,kpc^{-2}}$ (95% confidence level), which is less than the solar-neighbourhood value of $10^{-2.4} M_\odot\,\mathrm{yr^{-1}\,kpc^{-2}}$ (ref. 13).

As a result, we have detected an unusually strong magnetic field at $z = 0.692$ with a coherence length that probably exceeds 200 pc in neutral gas that is quiescent, metal poor, nearly dust free, and presents little evidence of star formation. To model this configuration, we first consider the magnetostatic equilibrium of a plane-parallel sheet with in-plane magnetic field $B_{\rm plane}$ orthogonal to the vertical gravitational field exerted by gas with perpendicular mass surface density Σ. In magnetostatic equilibrium, the total mid-plane pressure, $B^2_{\rm plane}/8\pi + \rho\sigma_v^2$, equals the 'weight' of the gas, $\pi G\Sigma^2/2$, where ρ is the mass volume density of the gas and G is the gravitational constant. However, because the pressure-to-weight ratio exceeds

[1]Department of Physics and Center for Astrophysics and Space Sciences, University of California, San Diego, La Jolla, California 92093-0424, USA. [2]Astronomy Department, University of California, Berkeley, California 94720-3411, USA. [3]UCO-Lick Observatory; University of California, Santa Cruz, Santa Cruz, California 95464, USA.

715 in DLA-3C286, the magnetized gas cannot be confined by its self-gravity. Therefore, self-consistent magnetostatic configurations are ruled out unless the contribution of stars to Σ exceeds $\sim 350 M_{\odot}\,\mathrm{pc}^{-2}$. Although this is larger than the $50 M_{\odot}\,\mathrm{pc}^{-2}$ surface density perpendicular to the solar neighbourhood, such surface densities are common in the central regions of galaxies. In fact, high surface densities of stars probably confine the highly magnetized gas in the nuclear rings of barred spirals. These exhibit total field strengths of ~100 μG, inferred by assuming equipartition of magnetic and cosmic-ray energy densities[1]. However, because the rings are associated with regions of active star formation, high molecular content and high dust content, they are unlikely sites for the magnetic field detected in DLA-3C286.

On the other hand, the absorption site might consist of highly magnetized gas confined by the gravity exerted by a disk of old stars. The H I disks found at the centres of early-type S0 and elliptical galaxies[14] are possible prototypes. Support for this idea stems from a high-resolution image obtained with the Hubble Space Telescope: a Wide Field and Planetary Camera 2 (WFPC2) I-band image, from which the quasar has been subtracted, reveals residual emission spread over angular scales of ~1″ (ref. 15). The asymmetry of the light distribution with respect to the point-source quasar suggests

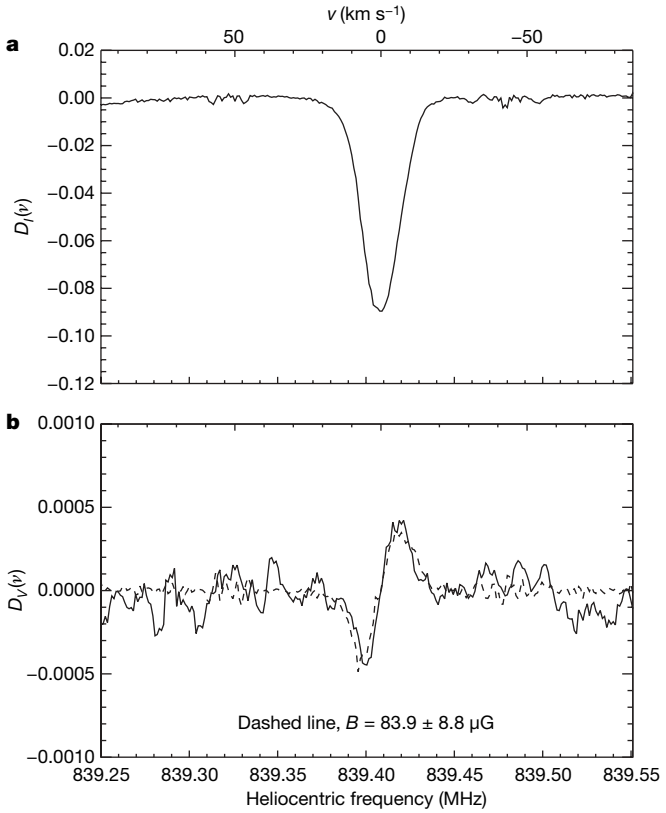

Figure 1 | Line-depth spectra of Stokes parameters. Data acquired in 12.6 hours of on-source integration with the GBT radio antenna. Because the GBT feeds detect only orthogonal, linearly polarized signals, whereas Zeeman splitting requires measuring circular polarization to construct $V(v)$, we generated $V(v)$ by cross-correlation techniques[23]. The velocity $v = 0\,\mathrm{km\,s^{-1}}$ corresponds to $z = 0.6921526$. **a**, Line-depth function $D_I(v) \equiv (I(v) - I_c(v))/I_c(v)$. Here $I(v) \equiv s_0 + s_{90}$, with s_θ the power measured in linear-polarization position angle θ, corresponds to the total intensity spectrum, and $I_c(v)$ is a model fit to the $I(v)$ continuum. $D_I(v) = \exp(-\tau(v)) - 1$, where $\tau(v) \equiv (\tau(v)_0 + \tau(v)_{90})/2$ is the average optical depth in the two orthogonal states of linear polarization[4]. **b**, Line-depth function $D_V(v) \equiv V(v)/I_c(v)$, where $V(v) \equiv s_{RCP} - s_{LCP}$ is the difference in power between the right-hand and left-hand circularly polarized (respectively RCP and LCP) signals. Here $D_V(v) = -(\tau_V(v)/2)\exp(-\tau(v))$, where $\tau_V(v) \equiv \tau_{RCP}(v) - \tau_{LCP}(v) \ll 1$ (ref. 4) is the difference between the optical depths of RCP and LCP photons. For Zeeman splitting of the 21-cm line, the degeneracy of the $F = 0$ to $F = 1$ hyperfine transition is removed because the $m_F = -1, 0, +1$ states differ in energy. This results in a small frequency difference between absorbed LCP photons ($m_F = -1$) and RCP photons ($m_F = +1$). $V(v)$ is crucial for detecting Zeeman splitting because the orthogonal, circularly polarized states of the photon are eigenstates of the spin angular momentum operator with eigenvalues $\pm\hbar$, that is, angular momenta directed along or opposite to the direction of photon propagation[24]. When $B_{105} = B$, transitions between the hyperfine $F = 0$ and $F = 1$ states occur exclusively through absorption of LCP or RCP photons through excitation of the $m_F = -1$ and $m_F = +1$ hyperfine states, respectively. Because $V(v)$ is the difference in the RCP and LCP intensities, the resulting $V(v)$ line profile is the difference between two Gaussian absorption profiles with frequency centroids shifted by $\Delta\nu_B = 2.8 B_{los}(1+z)^{-1}$ Hz (where B_{los} is measured in microgauss). The 'S curve' pattern is due to the sign flip in RCP-minus-LCP intensity difference as v passes through the line centre.

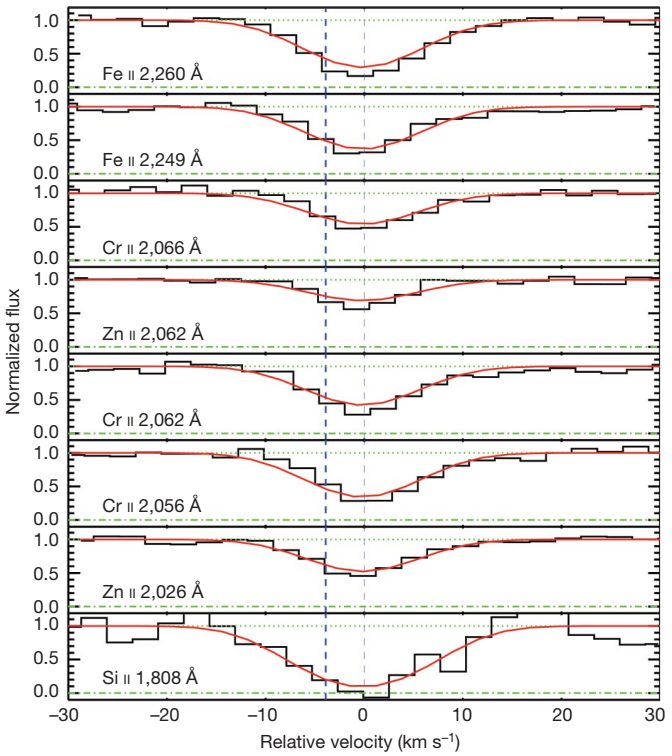

Figure 2 | HIRES velocity profiles for dominant low-ionization states of abundant elements in the 21-cm absorber in the direction of quasar 3C 286. Spectral resolution is $\Delta v = 7.0\,\mathrm{km\,s^{-1}}$ and the average signal-to-noise ratio per 2.1-km-s^{-1} pixel is about 30:1. The bold dashed vertical line denotes the velocity centroid of the single-dish 21-cm absorption feature and the faint dashed vertical lines denotes the velocity centroid of the resonance line shown in the figure. Our least-squares fit of Voigt profiles (red) to the data (black) yields ionic column densities as well as the redshift centroid and velocity dispersion shown in Table 1 (lower and upper green horizontal lines refer to zero and unit normalized fluxes, respectively). Because refractory elements such as Fe and Cr can be depleted onto dust grains[25], we used the volatile elements Si and Zn to derive a logarithmic metal abundance with respect to solar abundances of [M/H] = −1.30. The depletion ratios [Fe/Si] and [Cr/Zn] were then used to derive a conservative upper limit on the logarithmic dust-to-gas ratio relative to Galactic values of [D/G] < −1.8.

Table 1 | Physical parameters of DLA-3C286 inferred from optical absorption

Ion, X	$\log_{10}[N(X)\,(\mathrm{cm}^{-2})]$	[X/H]
H I	21.25 ± 0.02	—
Fe II	15.09 ± 0.01	−1.66 ± 0.02
Cr II	13.44 ± 0.01	−1.48 ± 0.02
Zn II	12.53 ± 0.03	−1.39 ± 0.03
Si II	>15.48	>−1.31

Redshift, $z = 0.69217485 \pm 0.00000058$; velocity dispersion, $\sigma_v = 3.08 \pm 0.13\,\mathrm{km\,s^{-1}}$. $N(X)$, column density of ion X.

that some of the light is emitted by a foreground galaxy with a brightness centroid displaced less than 0.5″ from the quasar. The location of diffuse emission in the direction of an amorphous object detected 2.5″ from the quasar in ground-based imaging[16] further suggests that the diffuse emission comes from central regions of the amorphous object. A recent reanalysis of the WFPC2 image shows the amorphous object to be a filament resembling a spiral arm or tidal tail (H.-W. Chen, personal communication), that is, the outer appendage of a galaxy centred within a few kiloparsecs of the quasar sightline.

However, the magnetic field detected in DLA-3C286 may not be confined by gravity in an equilibrium configuration. Rather, the detected field may be enhanced by a shock (F. H. Shu, personal communication). Assuming a typical value of $B_{\rm plane} \approx 5\,\mu$G for the equilibrium field of the pre-shock gas, we find that a shock-front velocity of $\sim$250 km s^{-1} will result in a post-shock field strength of $\sim$100 μG in the limit of flux freezing in a radiative shock with post-shock density of $\sim$10 cm^{-3}. This scenario seems plausible because 250 km s^{-1} is a reasonable value for the impact velocity generated by the merger between the gaseous disks of two late-type galaxies, and the WFPC2 image is consistent with the presence of two foreground galaxies. But the second disk would create another set of absorption lines displaced $\geq$250 km s^{-1} from the redshift of DLA-3C286, which is the only redshift observed. By contrast, the merger between a gaseous disk and an elliptical galaxy could result in only one damped Lyα system redshift, as a significant fraction of elliptical galaxies do not contain H I disks[12]. In this case, a shock front moving in the plane of the disk galaxy would be generated by the gravitational impulse induced by the elliptical galaxy moving normal to the plane. Preliminary estimates indicate that an elliptical galaxy with a modest mass, $M = 2 \times 10^{11} M_{\odot}$, and impact velocity of $\sim$300 km s^{-1} would produce a cylindrical shock of sufficient strength to boost an initial field with $B_{\rm plane} \approx 10\,\mu$G to a final field of $\sim$100 μG.

Let us examine these scenarios more closely. The quiescent velocity field of the gas fits in naturally with the 'magnetostatic equilibrium' scenario, because the low value of $\Sigma_{\rm SFR}$ suggests a low rate of energy injection into the gas by supernovae[17], which could result in a velocity dispersion of $\sigma_v \approx 4$ km s^{-1}. Moreover, the weak radio jets associated with early-type galaxies containing central H I disks are natural sources of magnetic fields for these disks. However, 21-cm absorption measurements of such disks in nearby galaxies reveal the presence of absorption line widths far broader than the narrow line width of DLA-3C286 (ref. 18). Also, it is unclear whether or not the high surface density of old stars required to confine the magnetic fields are present in these disks, and whether or not the build-up of $B_{\rm plane}$ to 100 μG is possible in the 4–5-Gyr age of the disk. In the 'merger scenario', the dynamo need only build up to $\sim$10 μG in the same time interval, but it is then necessary to explain why the post-shock velocity field averaged over length scales of 200 pc is so quiescent. Furthermore, the probability, p, of detecting $\sim$100-μG magnetic fields in a random sample of 21-cm absorbers is small. Our estimates, based on the merger fraction of galaxies with $z \approx 1$ (ref. 19) and on the duration time for magnetic field enhancement, suggest that $p \approx 0.005$–0.03: either we were lucky, or some characteristic of DLA-3C286, such as narrow line width, is a signature of strong magnetic fields.

Therefore, it is premature to decide among these and other possible models to explain the presence of the 84-μG magnetic field in DLA-3C286. However, our data support the inference from recent tentative evidence for Faraday rotation in high-z quasars[20] that magnetic fields are generic features of galaxies at high redshifts, which potentially have a more important role in galaxy formation and evolution[21] than hitherto realized. Specifically, the highly magnetized gas that we have detected could suppress gravitational collapse and, hence, may be a reason for the low *in situ* star formation rates of high-z damped Lyα systems[22]. We plan to test this hypothesis by using the GBT to search for Zeeman splitting in high-redshift damped Lyα systems exhibiting 21-cm absorption.

Received 4 April; accepted 15 July 2008.

1. Beck, R. in *Cosmic Magnetic Fields* 41–68 (Lect. Notes Phys. 664, Springer, 2005).
2. Han, J. L., Manchester, R. N., Lyne, A. G., Qiao, G. J. & van Straten, W. Pulsar rotation measures and the large-scale structure of galactic magnetic fields. *Astrophys. J.* **642,** 868–881 (2006).
3. Kronberg, P. P. *et al.* A global probe of cosmic magnetic fields to high redshifts. *Astrophys. J.* **676,** 70–79 (2008).
4. Heiles, C. & Troland, T. H. The millennium Arecibo 21 centimeter absorption-line survey. III. Techniques for spectral polarization and results for Stokes *V*. *Astrophys. J. Suppl. Ser.* **151,** 271–297 (2004).
5. Parker, E. The origin of magnetic fields. *Astrophys. J.* **160,** 383–404 (1970).
6. Brown, R. L. & Roberts, M. S. 21-centimeter absorption at z=0.692 in the quasar 3C 286. *Astrophys. J.* **184,** L7–L10 (1973).
7. Davis, M. M. & May, L. S. New observations of the radio absorption line in 3C 286, with potential application to the direct measurement of cosmological deceleration. *Astrophys. J.* **219,** 1–4 (1978).
8. Wolfe, A. M., Gawiser, E. & Prochaska, J. X. Damped Lyα systems. *Annu. Rev. Astron. Astrophys.* **43,** 861–918 (2005).
9. Wolfe, A. M., Broderick, J. J., Condon, J. J. & Johnston, K. J. 3C 286: A cosmological QSO? *Astrophys. J.* **208,** L47–L50 (1976).
10. Meiring, J. D. *et al.* Elemental abundance measurements in low-redshift damped Lyα absorbers. *Mon. Not. R. Astron. Soc.* **370,** 43–62 (2006).
11. Wolfe, A. M., Prochaska, J. X. & Gawiser, E. CII* absorption in damped Lyα systems. I. Star formation rates in a two-phase medium. *Astrophys. J.* **593,** 215–234 (2003).
12. Boisse, P., Le Brun, V., Bergeron, J. & Deharveng, J.-M. A HST spectroscopic study of QSOs with intermediate redshift damped Lyα systems. *Astron. Astrophys.* **333,** 841–863 (1998).
13. Kennicutt, R. C. Jr. Star formation in galaxies along the Hubble sequence. *Annu. Rev. Astron. Astrophys.* **36,** 189–231 (1998).
14. Morganti, R. *et al.* Neutral hydrogen in nearby elliptical and lenticular galaxies: the continuing formation of early-type galaxies. *Mon. Not. R. Astron. Soc.* **371,** 157–169 (2006).
15. Le Brun, V., Bergeron, J. & Deharveng, J. M. The nature of intermediate-redshift damped Lyα absorbers. *Astron. Astrophys.* **321,** 733–748 (1997).
16. Steidel, C. C., Pettini, M., Dickinson, M. & Persson, S. E. Imaging of two damped Lyman-alpha absorbers at intermediate redshifts. *Astron. J.* **108,** 2046–2053 (1994).
17. McKee, C. F. & Ostriker, J. P. A theory of the interstellar medium: three components regulated by supernova explosions in an inhomogeneous substrate. *Astrophys. J.* **218,** 148–169 (1977).
18. Morganti, R., Greenhill, L. J., Peck, A. B., Jones, D. L. & Henkel, C. Disks, tori, and cocoons: emission and absorption diagnostics of AGN environments. *N. Astron. Rev.* **48,** 1195–1209 (2004).
19. Lotz, J. M. *et al.* The evolution of galaxy mergers and morphology at *z* 1.2 in the extended Groth strip. *Astrophys. J.* **672,** 177–197 (2008).
20. Bernet, M. L. Miniati, F. Lilly, S. J., Kronberg, P. P. & Dessauges–Zavadsky, M. Strong magnetic fields in normal galaxies at high redshift. *Nature* **454,** 302–304 (2008).
21. Rees, M. J. Origin of cosmic magnetic fields. *Astron. Nachr.* **327,** 395–398 (2006).
22. Wolfe, A. M. & Chen, H.-W. Searching for low surface brightness galaxies in the Hubble ultra deep field: implications for the star formation efficiency in neutral gas at $z \sim 3$. *Astrophys. J.* **652,** 981–993 (2006).
23. Heiles, C. Cross-correlation spectropolarimetry in single-dish radio astronomy. *Publ. Astron. Soc. Pacif.* **113,** 1243–1246 (2001).
24. Baym, G. *Lectures on Quantum Mechanics* Ch. 1 (Benjamin, 1981).
25. Savage, B. D. & Sembach, K. R. Interstellar abundances from absorption- line observations with the Hubble-Space Telescope. *Annu. Rev. Astron. Astrophys.* **34,** 279–329 (1996).

Acknowledgements We wish to thank F. H. Shu for suggesting the merger model and H.-W. Chen for providing us with her reanalysed images of 3C 286. We also thank F. H. Shu, E. Gawiser and A. Lazarian for comments and the US National Science Foundation for financial support. The GBT is one of the facilities of the National Radio Astronomy Observatory, which is a center of the National Science Foundation operated under cooperative agreement by Associated Observatories, Inc. A.M.W., R.A.J. and J.X.P. are Visiting Astronomers at the W. M. Keck Telescope. The Keck Observatory is a joint facility of the University of California, the California Institute of Technology and the National Aeronautics and Space Administration.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.M.W. (awolfe@ucsd.edu).

LETTERS

Clustered star formation as a natural explanation for the Hα cut-off in disk galaxies

Jan Pflamm-Altenburg[1] & Pavel Kroupa[1]

The rate of star formation in a galaxy is often determined by the observation of emission in the Hα line, which is related to the presence of short-lived massive stars. Disk galaxies show a strong cut-off in Hα radiation at a certain galactocentric distance, which has led to the conclusion that star formation is suppressed in the outer regions of disk galaxies. This is seemingly in contradiction to recent observations[1] in the ultraviolet which imply that disk galaxies have star formation beyond the Hα cut-off, and that the star-formation-rate surface density is linearly related to the underlying gas surface density, which is a shallower relationship than that derived from Hα luminosities[2]. In a galaxy-wide formulation, the clustered nature of star formation has recently led to the insight that the total galactic Hα luminosity is nonlinearly related to the galaxy-wide star formation rate[3]. Here we show that a local formulation of the concept of clustered star formation naturally leads to a steeper radial decrease in the Hα surface luminosity than in the star-formation-rate surface density, in quantitative agreement with the observations, and that the observed Hα cut-off arises naturally.

The integrated galactic initial mass function (IGIMF) describes the mass spectrum of all newly formed stars in a galaxy. The IGIMF is calculated by adding all stars of all newly formed star clusters[4,5], and falls off more steeply with increasing stellar masses for massive stars[5] than the canonical initial mass function (IMF) in each star cluster, owing to the combination of two effects: the masses of the young star clusters are distributed according to the embedded cluster mass function (ECMF), for which the upper mass limit is a function of the total star formation rate[6] (SFR), and the stellar upper mass limit of the IMF is a function of the total star cluster mass[7]. Consequently, the total fraction of massive stars and, therefore, the total Hα luminosity decreases faster than linearly with decreasing SFR (ref. 3). The IGIMF theory has already been shown to lead naturally to the observed mass–metallicity relation of galaxies[8] and has received recent empirical verification in a study of IMF variations among galaxies[9].

To construct a quantitative local IGIMF theory, we introduce the local embedded cluster mass function (LECMF)

$$\xi_{\mathrm{LECMF}}(M_{\mathrm{ecl}},x,y)=\frac{\mathrm{d}N_{\mathrm{ecl}}}{\mathrm{d}M_{\mathrm{ecl}}\,\mathrm{d}x\,\mathrm{d}y}$$

which defines the number of newly formed star clusters, $\mathrm{d}N_{\mathrm{ecl}}$, in the mass interval $[M_{\mathrm{ecl}}, M_{\mathrm{ecl}}+\mathrm{d}M_{\mathrm{ecl}}]$ per unit area at the location (x, y) in a disk galaxy. Observations[10] of Galactic star-forming regions show that this function is a single-part power law, $\xi_{\mathrm{LECMF}} \propto M_{\mathrm{ecl}}^{-\beta}$, with an index of $\beta = 2$. The least massive cluster, with mass[5] $M_{\mathrm{ecl,min}} = 5M_{\odot}$, should form at any place in the galaxy, whereas the mass, $M_{\mathrm{ecl,max,loc}}(x, y)$, of the most massive star cluster that can form locally is expected to depend on the local gas density, that is, on how much material is locally available for star cluster formation. Observations[5,6] show that the mass, $M_{\mathrm{ecl,max}}$, of the most massive star cluster in the whole galaxy is a function of the total galactic star formation rate. To express the dependence of the upper limit of the LECMF on the local gas surface density, we write

$$M_{\mathrm{ecl,max,loc}}(x,y)=M_{\mathrm{ecl,max}}\left(\frac{\Sigma_{\mathrm{gas}}(x,y)}{\Sigma_{\mathrm{gas,0}}}\right)^{\gamma} \quad (1)$$

where $\Sigma_{\mathrm{gas}}(x, y)$ and $\Sigma_{\mathrm{gas,0}}$ are the gas densities at the location (x, y) and at the origin, respectively.

The local mass of all star clusters between the two mass limits is determined by the local star formation rate, $\Sigma_{\mathrm{SFR}}(x, y)$, which is described by the Kennicutt–Schmidt law[2,11]: $\Sigma_{\mathrm{SFR}}(x, y) = A\ \Sigma_{\mathrm{gas}}^{N}(x, y)$. Here A is a proportionality constant and N is widely accepted to have the value[2] 1.4, although $N = 0.99$ follows from recent ultraviolet observations[1]. As ultraviolet emission is a star formation tracer that is much less sensitive to the presence of OB stars than Hα emission, the true exponent N must be much closer to the value derived from ultraviolet observations. Thus, we chose $N = 1$. We note that the high-gas-density part of the Kennicutt–Schmidt plot[2] based on far-infrared observations also has a flatter slope, of $N = 1.08$, and that galaxy evolution models suggest N must not exceed unity if observed radial density profiles of disk galaxies are to be reproduced[12], confirming our choice.

We calculate the mass spectrum of all newly formed stars per unit area—the local integrated galactic initial mass function (LIGIMF)—by adding all newly formed stars of all young star clusters, and the Hα surface density follows by adding the Hα flux contributions of all newly formed stars. The newly formed stars in each young star cluster are distributed according to the invariant canonical IMF (refs 13, 14) with a fixed lower mass limit but an upper mass limit depending on the total cluster mass[7]. In terms of Hα emission, young star clusters above $\sim 3{,}000M_{\odot}$ have constant light-to-mass ratios, whereas smaller clusters are increasingly underluminous[3]. With decreasing star-formation-rate surface density, the upper mass limit of the LECMF decreases and, consequently, the fraction of underluminous star clusters increases. Thus, ultraviolet and Hα emissions scale differently with the star-formation-rate surface density, gas surface density or galactocentric radius. A detailed explanation of how the Hα surface luminosity is calculated is given in the Supplementary Information.

Next, we apply the LIGIMF theory to a sample of disk galaxies[15] with measured gas surface densities and Hα surface luminosities of H II regions averaged over annuli at different galactocentric radii. It is known that ionising photons emitted by massive stars can escape from well-defined H II regions and lead to recombinations and, thus, Hα radiation in the surrounding diffuse ionised gas[16]. Using Hα emission as a star formation tracer, this kind of photon leakage has to be taken into account to get an estimate of the true star formation rate. A previous study[16] of the galaxies NGC 247 and NGC 7793 allows us to construct a correction procedure to obtain the total

[1]Argelander-Institut für Astronomie, Universität Bonn, 53121 Bonn, Germany.

Hα surface luminosities from the surface luminosities of H II regions only (see Supplementary Information).

For a linear star formation law ($N = 1$) as derived from ultraviolet observations[1], the LIGIMF theory predicts a Hα surface luminosity as a function of the gas surface density that is in full agreement with the observations (Fig. 1). Additionally, the radial Hα profile derived in the LIGIMF theory matches the observations perfectly (Fig. 2). The concept of clustered star formation resolves the discrepancies between Hα and ultraviolet observations completely.

At first sight it might be objected that the LIGIMF theory contradicts observations of the ultraviolet sources in the outer disks of galaxies: 5–10% of all clusters in the outer disks of galaxies detected in the ultraviolet have associated Hα emission[17]. The age estimates of the ultraviolet knots range up to 400 Myr. Clusters with Hα emission have ages $\leq$20 Myr, as they are powered by short-lived massive stars; therefore, 5% of all observed ultraviolet knots are expected to have associated Hα emission, in agreement with observations. The LIGIMF theory predicts an overabundance of Hα-underluminous star clusters beyond the Hα cut-off and a smaller number ratio of Hα-emitting to non-Hα-emitting ultraviolet knots is expected. Underluminosity does not mean that there is no Hα emission, however. In the LIGIMF theory, each young ultraviolet cluster is a Hα source, too, but ultraviolet and Hα luminosities scale differently with the cluster mass. Thus, this finding[17] is entirely consistent with the LIGIMF theory. The observed ultraviolet knots in the outer disk of the galaxy M83 are always systematically smaller than their counterparts in the inner disk[18], in agreement with the fundamental basics of the LIGIMF theory. There is one outstanding massive young star cluster in the outer region of M83, but this does not contradict the theory and is instead expected from a statistical point of view (see Supplementary Information). Furthermore, the M83 far-ultraviolet luminosity function of outer-disk stellar complexes is steeper than that of the inner-disk population[18].

A similar trend is reported in the galaxy NGC 628 for the Hα luminosity function of H II regions[19]. In the LIGIMF theory, inner-disk LECMFs have higher upper mass limits than outer-disk LECMFs. Integration of the LECMFs over the outer and inner regions leads to an outer-disk ECMF steeper than the inner-disk ECMF, indicating that outer-disk star formation complexes are systematically less massive than those in the inner disk. This integration effect is fundamentally equivalent to the IGIMF being steeper for dwarf galaxies with low global star formation rates than for disk galaxies with high star formation rates[3].

Previously, the Hα cut-off has been explained[15] by a drop of the local gas density below a critical value determined by the stability condition of a thin isothermal disk[20,21] where no star formation can occur. In contradiction to this explanation, recent ultraviolet observations[1] reveal star formation outside the Hα cut-off, and dwarf galaxies[22] show star formation although their average gas density is lower than the critical value. Indeed, it has been shown that in regions with densities lower than the critical value, star formation can be driven by instabilities other than thermal[23]. It has been argued that H II regions powered by the same massive stars are larger in a thin environment—that is, at large galactocentric radii—than in a dense one, and identical H II regions thus become fainter in the outer galaxy. Therefore, it has been concluded[23] that the Hα surface luminosity should decrease faster than the star-formation-rate surface density. Indeed, the surface brightness of individual H II regions should

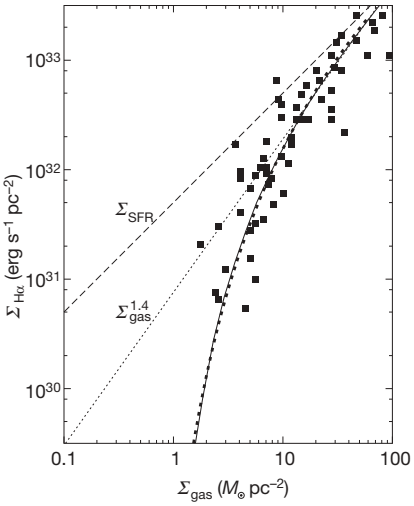

Figure 1 | Hα luminosity surface density ($\Sigma_{H\alpha}$) versus total gas surface density (Σ_{gas}). We plot data (black squares) observed for seven disk galaxies[15], averaged over annuli at different galactocentric radii, after correcting for photon leakage from H II regions (see Supplementary Information). These galaxies have a mean star formation rate of $6.9 M_\odot$ yr^{-1} ($3.2 M_\odot$–$16.4 M_\odot$ yr^{-1})[2,15], a mean total gas mass of $2.1 \times 10^{10} M_\odot$ ($0.6 \times 10^{10} M_\odot$–$3.6 \times 10^{10} M_\odot$)[2,15] and a mean scale length of 4.4 kpc (3.9–5.2 kpc)[25–28]. These mean values define our model standard disk galaxy. For a choice of $\gamma = 3/2$ (see equation (1)) the LIGIMF theory predicts a $\Sigma_{H\alpha}$–Σ_{gas} relation that matches the observations excellently (solid line). We note that the underlying true star-formation-rate surface density as derived from ultraviolet observations[1] is directly proportional to the gas surface density ($N = 1$); it is displayed after being converted into Hα surface luminosity using the wrong linear Kennicutt Hα–SFR relation[2,29] (dashed line) and shows the expected $\Sigma_{H\alpha}$–Σ_{gas} relation based on the classical picture, which is in disagreement with the observations. Furthermore, the Hα luminosity surface density in the high-luminosity part ($\Sigma_{H\alpha} \geq 10^{32.5}$ erg s^{-1} pc^{-2}) depends, for the correct LIGIMF theory, on the gas surface density raised to the power of 1.4 (dotted line, extrapolated to low Hα surface luminosity), in agreement with the classical Kennicutt–Schmidt slope of $N = 1.4$. The LIGIMF theory puts the hitherto inconsistent Hα and ultraviolet observations in perfect agreement with each other. The steeper high-luminosity slope of $N = 1.4$ and the Hα cut-off at low gas density are simultaneous outcomes of the LIGIMF theory. The thick dotted line which almost coincides with the thick solid curve shows a fitting function for the LIGIMF model (see Supplementary Information).

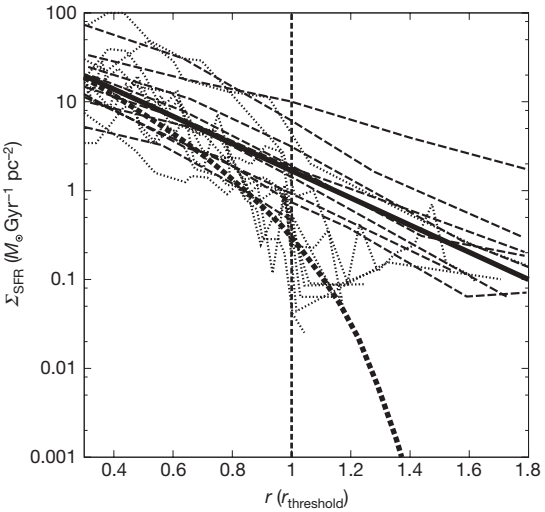

Figure 2 | Star-formation-rate surface density (Σ_{SFR}) versus galactocentric radius (r). Radial distribution of the star-formation-rate surface density of nine disk galaxies based on ultraviolet[1] (thin dash–dot lines) and Hα (ref. 30; thin dotted lines) observations that rely on a wrong linear conversion[29] between the corresponding Hα luminosity surface density and star-formation-rate surface density after correction for photon leakage (see Supplementary Information). The galactocentric radius, r, is expressed in units of $r_{threshold}$, the Hα threshold radius[30]. Also plotted are the true underlying star-formation-rate surface density of our standard disk galaxy (thick solid line) as defined in Fig. 1 and the model Hα surface luminosity (thick dotted line) converted into a star-formation-rate surface density using the same linear conversion[29,30]. The LIGIMF theory thus naturally accounts for the discrepant Σ_{SFR} value at a particular radius.

be fainter in the outer galaxy. However, the Hα surface density considered in star formation laws refers to the total Hα luminosity per unit area of the galaxy and not to the cross-section of the H II region. Identically powered H II regions contribute equally to the Hα surface luminosity independently of their location in a thin or a dense gas environment. Thus, the proposed solution[23] explains neither the Hα cut-off nor the different slopes of the ultraviolet-based and Hα-based star formation laws. It has been shown recently that a required minimum column density for massive star formation might exist[24], implying star formation with no massive stars in low-density environments. However, this model predicts a top-heavy IMF for cloud column densities much larger than this threshold, for which no observational evidence exists[14], and allows no quantitative linkage of Hα luminosity and the star formation rate.

Contrary to this previously existing work, the LIGIMF theory developed here is in excellent agreement with the observed radial Hα and ultraviolet luminosity profiles (Fig. 2) and the Kennicutt–Schmidt star formation law (Fig. 1), and also allows the determination of star formation rates even in Hα-faint galaxy regions.

Received 6 March; accepted 14 July 2008.

1. Boissier, S. *et al.* Radial variation of attenuation and star formation in the largest late-type disks observed with GALEX. *Astrophys. J. Suppl. Ser.* **173,** 524–537 (2007).
2. Kennicutt, R. C. Jr. The global Schmidt law in star-forming galaxies. *Astrophys. J.* **498,** 541–552 (1998).
3. Pflamm-Altenburg, J., Weidner, C. & Kroupa, P. Converting Hα luminosities into star formation rates. *Astrophys. J.* **671,** 1550–1558 (2007).
4. Vanbeveren, D. Theoretical evolution of massive stellar aggregates. *Astron. Astrophys.* **124,** 71–76 (1983).
5. Weidner, C. & Kroupa, P. The variation of integrated star initial mass functions among galaxies. *Astrophys. J.* **625,** 754–762 (2005).
6. Weidner, C., Kroupa, P. & Larsen, S. S. Implications for the formation of star clusters from extragalactic star formation rates. *Mon. Not. R. Astron. Soc.* **350,** 1503–1510 (2004).
7. Weidner, C. & Kroupa, P. Evidence for a fundamental stellar upper mass limit from clustered star formation. *Mon. Not. R. Astron. Soc.* **348,** 187–191 (2004).
8. Köppen, J., Weidner, C. & Kroupa, P. A possible origin of the mass-metallicity relation of galaxies. *Mon. Not. R. Astron. Soc.* **375,** 673–684 (2007).
9. Hoversten, E. A. & Glazebrook, K. Evidence for a nonuniversal stellar initial mass function from the integrated properties of SDSS galaxies. *Astrophys. J.* **675,** 163–187 (2008).
10. Lada, C. J. & Lada, E. A. Embedded clusters in molecular clouds. *Annu. Rev. Astron. Astrophys.* **41,** 57–115 (2003).
11. Kennicutt, R. C. Jr *et al.* Star formation in NGC 5194 (M51a). II. The spatially resolved star formation law. *Astrophys. J.* **671,** 333–348 (2007).
12. Zasov, A. V. & Abramova, O. V. The star-formation efficiency and density of the disks of spiral galaxies. *Astron. Rep.* **50,** 874–886 (2006).
13. Kroupa, P. On the variation of the initial mass function. *Mon. Not. R. Astron. Soc.* **322,** 231–246 (2001).
14. Kroupa, P. The initial mass function of stars: Evidence for uniformity in variable systems. *Science* **295,** 82–91 (2002).
15. Kennicutt, R. C. Jr. The star formation law in galactic disks. *Astrophys. J.* **344,** 685–703 (1989).
16. Ferguson, A. M. N., Wyse, R. F. G., Gallagher, J. S. III & Hunter, D. A. Diffuse ionized gas in spiral galaxies: Probing Lyman continuum photon leakage from H II regions? *Astron. J.* **111,** 2265–2279 (1996).
17. Zaritsky, D. & Christlein, D. On the extended knotted disks of galaxies. *Astron. J.* **134,** 135–141 (2007).
18. Thilker, D. A. *et al.* Recent star formation in the extreme outer disk of M83. *Astrophys. J.* **619,** L79–L82 (2005).
19. Lelièvre, M. & Roy, J.-R. The H II regions of the extreme outer disk of NGC 628. *Astron. J.* **120,** 1306–1315 (2000).
20. Toomre, A. On the gravitational stability of a disk of stars. *Astrophys. J.* **139,** 1217–1238 (1964).
21. Cowie, L. L. Cloud fluid compression and softening in spiral arms and the formation of giant molecular cloud complexes. *Astrophys. J.* **245,** 66–71 (1981).
22. Hunter, D. A., Elmegreen, B. G. & van Woerden, H. Neutral hydrogen and star formation in the irregular galaxy NGC 2366. *Astrophys. J.* **556,** 773–800 (2001).
23. Elmegreen, B. G. & Hunter, D. A. Radial profiles of star formation in the far outer regions of galaxy disks. *Astrophys. J.* **636,** 712–720 (2006).
24. Krumholz, M. R. & McKee, C. F. A minimum column density of 1 g cm^{-2} for massive star formation. *Nature* **451,** 1082–1084 (2008).
25. Kenney, J. D. P. & Young, J. S. The effects of environment on the molecular and atomic gas properties of large Virgo cluster spirals. *Astrophys. J.* **344,** 171–199 (1989).
26. Wong, T. & Blitz, L. The relationship between gas content and star formation in molecule-rich spiral galaxies. *Astrophys. J.* **569,** 157–183 (2002).
27. Schuster, K. F., Kramer, C., Hitschfeld, M., Garcia-Burillo, S. & Mookerjea, B. A complete ^{12}CO 2–1 map of M 51 with HERA. I. Radial averages of CO, HI, and radio continuum. *Astron. Astrophys.* **461,** 143–151 (2007).
28. Crosthwaite, L. P. & Turner, J. L. CO(1–0), CO(2–1), and neutral gas in NGC 6946: Molecular gas in a late-type, gas-rich, spiral galaxy. *Astron. J.* **134,** 1827–1842 (2007).
29. Kennicutt, R. C. Jr, Tamblyn, P. & Congdon, C. E. Past and future star formation in disk galaxies. *Astrophys. J.* **435,** 22–36 (1994).
30. Martin, C. L. & Kennicutt, R. C. Jr. Star formation thresholds in galactic disks. *Astrophys. J.* **555,** 301–321 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. S. de Boer for discussions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.P.-A. (jpflamm@astro.uni-bonn.de).

LETTERS

Nanoscale magnetic sensing with an individual electronic spin in diamond

J. R. Maze[1], P. L. Stanwix[2], J. S. Hodges[1,3], S. Hong[1], J. M. Taylor[4], P. Cappellaro[1,2], L. Jiang[1], M. V. Gurudev Dutt[5], E. Togan[1], A. S. Zibrov[1], A. Yacoby[1], R. L. Walsworth[1,2] & M. D. Lukin[1]

Detection of weak magnetic fields with nanoscale spatial resolution is an outstanding problem in the biological and physical sciences[1–5]. For example, at a distance of 10 nm, the spin of a single electron produces a magnetic field of about 1 μT, and the corresponding field from a single proton is a few nanoteslas. A sensor able to detect such magnetic fields with nanometre spatial resolution would enable powerful applications, ranging from the detection of magnetic resonance signals from individual electron or nuclear spins in complex biological molecules[5,6] to readout of classical or quantum bits of information encoded in an electron or nuclear spin memory[7]. Here we experimentally demonstrate an approach to such nanoscale magnetic sensing, using coherent manipulation of an individual electronic spin qubit associated with a nitrogen-vacancy impurity in diamond at room temperature[8]. Using an ultra-pure diamond sample, we achieve detection of 3 nT magnetic fields at kilohertz frequencies after 100 s of averaging. In addition, we demonstrate a sensitivity of 0.5 μT Hz$^{-1/2}$ for a diamond nanocrystal with a diameter of 30 nm.

Sensitive solid-state magnetometers typically use phenomena such as superconducting quantum interference in SQUIDs[2,3] or the Hall effect in semiconductors[4]. Intriguing avenues such as magnetic resonance force microscopy are also currently being explored[5,6]. Our approach to magnetic sensing[8] uses the coherent manipulation of a single quantum system, an electronic spin qubit. As illustrated in Fig. 1, the electronic spin of an individual nitrogen-vacancy impurity in diamond can be polarized by optical pumping and measured through state-selective fluorescence. Conventional electron spin resonance (ESR) techniques are used to coherently manipulate its orientation. To achieve magnetic sensing, we monitor the electronic spin precession, which depends on external magnetic fields through the Zeeman effect. This method is directly analogous to precision measurement techniques in atomic and molecular systems[9], which are widely used to implement ultra-stable atomic clocks[10–12] and sensitive magnetometers[13].

The principal challenge for achieving high sensitivity using solid-state spins is their strong coupling to the local environment, which limits the free precession time and thus the magnetometer's sensitivity. Recently, there has been great progress in understanding the local environment of nitrogen-vacancy spin qubits, including ^{13}C nuclear spins[7,14–17] and electronic spin impurities[18–20]. Here we use coherent control over a coupled electron–nuclear system[8,16], similar to techniques used in magnetic resonance, to decouple the magnetometer spin from its environment. As illustrated in Fig. 1d, a spin-echo sequence refocuses the unwanted evolution of the magnetometer spin due to environmental fields fluctuating randomly on timescales much longer than the length of the sequence. However, oscillating external magnetic fields matching the echo period will affect the spin dynamics constructively, allowing sensitive detection of its amplitude.

The ideal preparation, manipulation and detection of an electronic spin would yield a so-called quantum-projection-noise-limited minimum detectable magnetic field[12]

$$\delta B_{\min} \approx \frac{\hbar}{g\mu_B\sqrt{T_2 T}} \quad (1)$$

where T_2 is the electronic spin coherence time, T is the measurement time, μ_B is the Bohr magneton, $\hbar$ is Planck's constant divided by 2π, and $g \approx 2$ is the electronic Landé g-factor. In principle, for typical values of $T_2 \approx 0.1$–1 ms, sensitivity of the order of a few nT Hz$^{-1/2}$ can be achieved with a single nitrogen-vacancy centre. Although this is less sensitive than for state-of-the-art macroscopic magnetometers[1,3], a key feature of our sensor is that it can be localized within a region of about 10 nm, either in direct proximity to a diamond surface or within a nano-sized diamond crystal (Fig. 1a). Sensitive magnetic detection on a nanometre scale can then be performed with such a system under ambient conditions. Supplementary Fig. 1 provides a comparison between magnetic field sensitivity and detector volume for several state-of-the-art magnetometers and the nitrogen-vacancy diamond systems demonstrated here.

To establish the sensitivity limits of a single electronic spin magnetometer, we carried out a series of proof-of-principle experiments involving single nitrogen-vacancy centres in bulk ultra-pure single-crystal diamond and in commercially available diamond nanocrystals. Our experimental methodology is outlined schematically in Fig. 1; further details about our experimental set-up and diamond samples are given in Methods. We first focus on the single-crystal diamond bulk sample. Figure 2a shows a typical spin-echo signal observed from an individual nitrogen-vacancy centre. The periodic modulation of the echo is caused by a bath of spin-1/2 ^{13}C nuclei (1.1% natural abundance), which create an effective precessing magnetic field at the nitrogen-vacancy centre of a few microteslas. In the presence of an applied static magnetic field B_{DC}, the periodic Larmor precession of the nuclear field causes the nitrogen-vacancy spin-echo signal to collapse and revive[16] at half the rate of the Larmor frequency of ^{13}C, $\omega_L = \gamma_{13C} B_{DC}$, where γ_{13C} is the carbon gyromagnetic ratio. Note that substantial spin-echo revivals exist even after a free evolution of 0.6 ms. To detect an external AC magnetic field with the highest sensitivity, we must eliminate the contribution from the ^{13}C nuclear field. To this end, the revival rate of the spin-echo signal is adjusted by varying the strength of B_{DC}, such that the frequency of the echo revival peaks coincides with multiples of the AC field frequency (ν) to be detected.

[1]Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA. [2]Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA. [3]Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [4]Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02138, USA. [5]Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA.

As shown in Fig. 2b, the observed peak of the spin-echo signal varies periodically as the amplitude of the external AC field (B_{AC}) is increased. This signal variation results from phase accumulated by the nitrogen-vacancy spin due to the external AC magnetic field and the resultant time-varying Zeeman shift during the spin's precession; converting this phase into a spin population difference gives rise to variations in the detected fluorescence, which serves as the magnetometer signal. Note that the period of this signal oscillation depends on the spin-echo interval, $\tau = 1/\nu$. For a given value of B_{AC}, the phase accumulated by the electronic spin over one period will increase as the frequency of the external AC field decreases. At the conclusion of a single run of the magnetometry pulse sequence, the measurable spin-echo signal S_B is proportional to the probability of the nitrogen-vacancy spin being in the $m_s = 0$ state: $S_B \propto P_0(B_{AC}) = [1 + F(\tau)\cos(\delta\phi)]/2$, where $\delta\phi = 4g\mu_B B_{AC}/2\pi\nu$ and $F(\tau)$ is the amplitude of the spin-echo signal envelope in the absence of the external AC magnetic field (Fig. 2a).

The sensitivity of the nitrogen-vacancy magnetometer to small variations in B_{AC}, as depicted in the measurements shown in Fig. 2b, is given by $\delta B_{min} = \sigma_S^N / dS_B$, where σ_S^N is the standard deviation of the spin-echo measurement after N averages and dS_B is the slope of the spin-echo signal variation with B_{AC}. Since maximum sensitivity (that is, smallest δB_{min}) occurs at maximum slope, all magnetometer sensitivity measurements were conducted at this point. This maximum slope is proportional to the spin-echo amplitude divided by the frequency of the oscillating field, $dS_B \propto F(1/\nu)/\nu$. For a shot-noise-limited signal with uncertainty σ_S in a single measurement, $\sigma_S^N = \sigma_S/\sqrt{N}$, where $N = T/\tau$. Hence the magnetometer sensitivity is expected to scale as $\delta B_{min} \propto \sqrt{\nu}/F(1/\nu)$.

Figure 3a shows example measurements of the sensitivity δB_{min} after 1 s of averaging as a function of the AC magnetic field frequency, $\nu = 1/\tau$. As this frequency decreases, the accumulated Zeeman phase

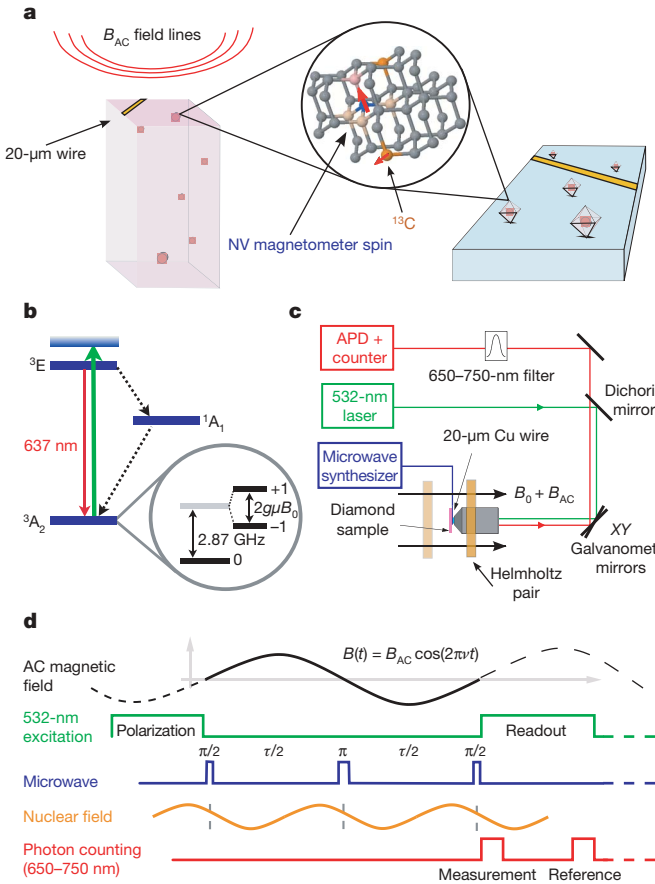

Figure 1 | Principles of the magnetic sensor, which is based on individual nitrogen-vacancy electronic spins in diamond. **a**, A single nitrogen-vacancy impurity (NV) proximal to the surface of an ultra-pure bulk single-crystal diamond sample (left) or localized within a diamond nanocrystal (right) is used to sense an externally applied AC magnetic field (B_{AC}, top left). A 20-μm-diameter wire (yellow) generates microwave pulses to manipulate the electronic spin states. **b**, Level structure of the nitrogen-vacancy centre; see Methods for details. **c**, Diagram of the experimental approach. Single nitrogen-vacancy centres are imaged and localized with ~170 nm resolution using confocal microscopy. The position of the focal point is moved near the sample surface, using a galvanometer mounted mirror to change the beam path and a piezo-driven objective mount. A pair of Helmholtz coils is used to provide both AC and DC magnetic fields. Experiments are then performed on single nitrogen-vacancy centres, as verified by photon correlation measurements. **d**, Optical and microwave spin-echo pulse sequence used for sensing an AC magnetic field, $B_{AC}(\tau)$. An individual centre is first polarized into the $m_s = 0$ sublevel. A coherent superposition between the states $m_s = 0$ and $m_s = 1$ is created by applying a microwave π/2 pulse tuned to this transition. The system freely evolves for a period of time $\tau/2$, followed by a π refocusing pulse. After a second $\tau/2$ evolution period, the electronic spin state is projected onto the $m_s = 0,1$ basis by a final π/2 pulse, at which point the ground state population is detected optically via spin-dependent fluorescence. The DC magnetic field is adjusted to eliminate the contribution of the randomly phased field produced by ^{13}C nuclear spins (gold curve) by choosing $\tau = 2n/\omega_L$, for integer n.

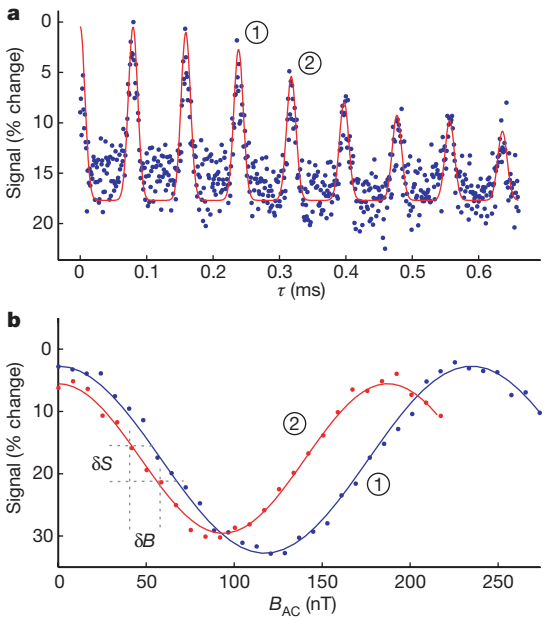

Figure 2 | Demonstration of spin-echo-based magnetometry with an individual nitrogen-vacancy electronic spin in a bulk diamond sample. **a**, Example of electronic spin-echo measurement. We plot the normalized echo signal corresponding to a fractional change of nitrogen-vacancy centre fluorescence. Maximal signal corresponds to an average number of photons $\langle n \rangle = 0.03$ detected during the 324-ns photon counting window of a single experimental run. Collapses and revivals are due to interactions with a ^{13}C nuclear spin bath. The revivals occur at half the rate of the Larmor frequency of ^{13}C (here set by $B_{DC} = 22$ G). The spin-echo signal envelope was fitted with an exponential decay function modulated by a strongly interacting pair of nearby ^{13}C (see Methods). Magnetometer sensitivity experiments are performed at spin-echo revival peaks to maximize signal. Revivals 1 and 2, treated in **b**, are indicated. **b**, Examples of measured spin-echo signal as a function of B_{AC} for two operating frequencies, $\nu_1 = 3.15$ kHz (red) and $\nu_2 = 4.21$ kHz (blue), corresponding to revivals 1 and 2 indicated in **a**. Each displayed point is a result of $N = 7 \times 10^5$ averages of spin-echo sequences. The magnetometer is most sensitive to variations in the AC magnetic field amplitude (δB) at the point of maximum slope, with the sensitivity being limited by the uncertainty in the spin-echo signal measurement (δS). We note that the cosine behaviour of the signal with respect to AC magnetic field amplitude can be changed to a sine by adjusting the phase of the third microwave pulse by 90°. This change moves the point of maximum magnetometer sensitivity to near zero AC field amplitude.

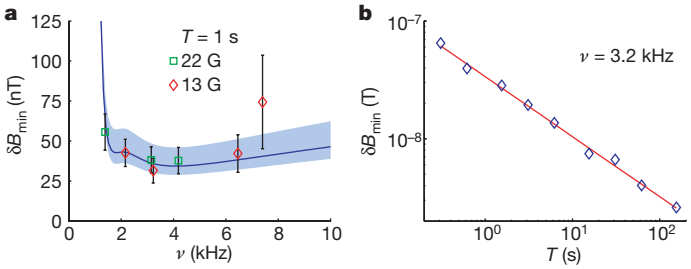

Figure 3 | Characterization of magnetometer sensitivity and minimum measurable AC magnetic field. a, Measured sensitivity of a single nitrogen-vacancy spin magnetometer in a bulk diamond sample over a range of frequencies for the external AC magnetic field after averaging for one second ($T = 1$ s). Error bars, standard deviation (s.d.) for a sample size of 30. Also shown is the theoretically predicted sensitivity (solid blue line), with the shaded region representing uncertainty due to variations in photon collection efficiency (see Methods). Measurements were carried out at two different DC fields, $B_{DC} = 13$ G (in red) and 22 G (in green). **b**, The minimum measurable AC magnetic field as a function of averaging time, for AC field frequency $\nu = 3.2$ kHz and $B_{DC} = 13$ G. Fit to this data (red curve) shows that the sensitivity improves as the square root of the averaging time, and is consistent with theoretical estimates based on photon-shot-noise limited detection.

shift of the nitrogen-vacancy spin during one period increases. This makes the nitrogen-vacancy spin more sensitive to variations of B_{AC} as the frequency is reduced, until the point at which the nitrogen-vacancy spin decoheres during a single period of the external AC magnetic field's oscillation. This decoherence decreases the magnetometer's sensitivity by decreasing the contrast of the spin-echo signal ($F(1/\nu) \rightarrow 0$) and therefore the slope dS_B. At high frequencies or short times, $F(1/\nu) \rightarrow 1$, and the sensitivity scales as $\sqrt{\nu}$. Hence, the magnetometer sensitivity is optimized for frequencies comparable with the longest time for which substantial echo signal is still observable. We note that it is possible to measure at higher frequencies without further loss of sensitivity by using multiple spin-echo pulses in a given measurement period[8]. Figure 3b shows examples of measured nitrogen-vacancy magnetometer sensitivity for a fixed ν as a function of T. The solid line is a fit to $\delta B_{min} \propto T^{-\alpha}$, where $\alpha = 0.5 \pm 0.01$, indicating that magnetic fields as small as few nanoteslas are resolvable after 100 s of averaging.

As noted above, a key feature of our technique is that at specific times, determined by echo revivals, the nitrogen-vacancy electronic spin can be essentially decoupled from ^{13}C nuclear spins. In practice, the decoupling is not perfect, owing to internal dynamics of the electronic environment other than simple spin precession. In fact, the overall decay of the echo signal shown in Fig. 2a does not follow the simple exponential decay associated with typical ESR on bulk samples. This can be understood by noting that the echo dynamics of a single nitrogen-vacancy centre near its revivals is probably determined by a few nearby ^{13}C atoms, which interact strongly with the electronic spin[7,14–16,21], yielding multiple characteristic timescales for echo decay (see Methods).

The absolute sensitivity of the nitrogen-vacancy magnetometer depends on the signal-to-noise ratio in the readout of the nitrogen-vacancy electronic spin state. In the present demonstration, this is limited by photon collection efficiency, which is ~0.1%. The resulting photon shot noise[1,8] is about an order of magnitude larger than the ideal quantum projection noise limit given by equation (1), resulting in a corresponding degradation of magnetometer sensitivity. Our theoretical prediction of magnetometer sensitivity (solid curve in Fig. 3a) combines the nitrogen-vacancy coherence properties shown in Fig. 2a with the noise due to photon counting statistics and imperfect collection efficiency (see Methods). This prediction is in excellent agreement with our experimental results, indicating that our magnetometer is photon-shot-noise limited.

To demonstrate magnetic sensing within a nanoscale detection volume, we also performed similar experiments with single nitrogen-vacancy centres in diamond nanocrystals. We used commercially available nanocrystals that contain a large number of impurities, which shorten the electronic spin coherence time[22] to values ranging from 4 to 10 μs. Sensitive detection of AC magnetic fields is still possible, as demonstrated experimentally in Fig. 4. Here, the echo signal from a single nitrogen-vacancy centre in a 30-nm-size nanocrystal decays on a timescale of ~4 μs. The absence of characteristic collapses and revivals, associated with couplings to ^{13}C nuclear spins, indicates that the echo decay is probably due to other spin impurities, such as paramagnetic substitutional nitrogen atoms containing unpaired electron spins. Magnetic sensing with such a nanocrystal at $\nu = 380$ kHz is demonstrated in Fig. 4b. From these measurements, we estimate a magnetometer sensitivity of $\delta B_{min} \approx 0.5 \pm 0.1\ \mu T\ Hz^{-1/2}$ for this nanocrystal.

Improved magnetometer sensitivity for bulk and nanocrystal diamond may be achieved in several ways. By using isotopically pure diamond with low concentrations of both ^{13}C and nitrogen electron spin impurities, much longer coherence and interrogation times should be possible. For diamond nanocrystals, however, the ultimate sensitivity will eventually be limited by surface effects[19,23]. Increases to the signal-to-noise ratio may also be possible by improving the measurement readout efficiency. Near single-shot readout of an electronic spin in diamond has been achieved with cryogenic cooling using resonant excitation[24]. Photon collection efficiency at room temperature can also be substantially improved using either conventional far-field optics or evanescent, near-field coupling to optical waveguides[25]. Finally, further improvements can probably be obtained by using magnetic sensing with multiple nitrogen-vacancy centres and by using more complex pulse sequences[8].

Our results demonstrate that electronic spins in diamond can be used for precision measurements of nanoscale magnetic fields. This approach opens a new regime of magnetic sensing, enabling detection of single-electron and even nuclear spins separated from

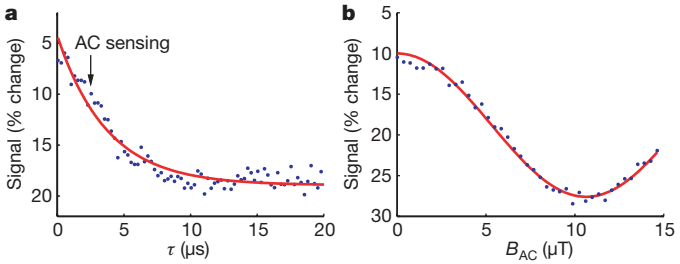

Figure 4 | Demonstration of magnetic sensing with a single nitrogen-vacancy electronic spin in a diamond nanocrystal. a, Example of electronic spin-echo signal from a single nitrogen-vacancy centre contained in a diamond nanocrystal with diameter of 34 ± 12 nm as determined by atomic force microscopy. Maximum signal corresponds to an average number of photons $\langle n \rangle = 0.02$ counted during a 324-ns photon counting window. The arrow indicates the time at which magnetic sensing is performed in **b**. **b**, Example of electronic spin-echo signal as a function of B_{AC} at a frequency of $\nu = 380$ kHz. For these data, $N = 2 \times 10^6$ averages of spin-echo sequences were used. The resulting standard deviation yields a magnetometer sensitivity of $0.5 \pm 0.1\ \mu T\ Hz^{-1/2}$.

nitrogen-vacancy centres by a few tens of nanometres (see Supplementary Information for details). For example, by combining our spin-echo based method with the recently demonstrated[26] transport and manipulation of nanocrystals using an atomic force microscope, a new kind of nanoscale scanning magnetic sensor may be created. Such a sensor could have a wide range of applications, ranging from biological and materials science to quantum information processing and fundamental tests of quantum mechanics. With the aid of field gradients, used for example in approaches based on magnetic resonance force microscopy[5,6], nitrogen-vacancy diamond magnetometers may allow sensing and resolving of individual nuclear spins, with applications in structural biology[8,27]. Our sensing technique also provides an efficient method for measuring single electronic spins in various quantum computing architectures. Furthermore, this technique may allow non-destructive mapping of quantum states into nitrogen-vacancy centres, operating as a quantum magnetic 'head'[28], with possibilities for mechanical transport of quantum information. Finally, we note that our technique could be used for detecting the quantum motion of magnetic mechanical resonators[29,30], with new possibilities for creating non-classical states of mechanical motion and for testing quantum mechanics on a macroscopic scale.

METHODS SUMMARY

AC magnetometry was performed at room temperature on nitrogen-vacancy centres found in both a bulk single-crystal diamond sample and in synthetic diamond nanocrystals (30 nm mean diameter). Single nitrogen-vacancy centres were isolated and probed by confocal microscopy. Phonon-mediated fluorescent emission (630–750 nm) was detected under coherent optical excitation ($\lambda = 532$ nm) using a single photon counting module (APD). As single spots in the confocal image may constitute many nitrogen-vacancy centres, single centres were identified by observing photon antibunching in the measurement of the second-order correlation function.

Green excitation of a nitrogen-vacancy centre also polarized the electronic spin by optical pumping to the $m_s = 0$ sublevel of the $^3A_2$ ground state. The mechanism responsible for optical pumping also provided a means for spin-sensitive detection, as the rate of fluorescence differs for the $m_s = 0$ and $m_s = \pm 1$ states. Coherent manipulation of the spin states was achieved by applying microwave radiation resonant with the $|0\rangle \rightarrow |1\rangle$ transition through a 20 μm wire. A pair of Helmholtz coils provided a static magnetic field to split the degenerate $|\pm 1\rangle$ levels; these coils also produced the external AC magnetic fields sensed with the nitrogen-vacancy magnetometer.

In performing magnetometry, pulsed laser and microwave excitations were defined with an acousto-optic modulator and microwave switch, respectively. As described in Fig. 1d, magnetometer measurements were made for an external AC magnetic field with amplitude B_{AC} and frequency ν, properly phased with respect to the microwave pulses. When the length of the spin-echo sequence (τ) equalled $1/\nu$, the accumulated phase of the electronic spin was proportional to B_{AC}. The fluorescence rate was directly related to this phase. A counting window of 324 ns provided optimal contrast of the fluorescent readout. Many spin-echo cycles were typically averaged to reduce the uncertainty of the photon statistics associated with the low count rate (<1 photon per readout). This technique was sensitive to the projection of the AC magnetic field onto the quantization axis of the electronic spin, corresponding to a vector magnetometer.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 April; accepted 18 July 2008.

1. Budker, D. & Romalis, M. Optical magnetometry. *Nature Phys.* **3,** 227–234 (2007).
2. Bending, S. J. Local magnetic probes of superconductors. *Adv. Phys.* **48,** 449–535 (1999).
3. Kleiner, R., Koelle, D., Ludwig, F. & Clarke, J. Superconducting quantum interference devices: State of the art and applications. *Proc. IEEE* **92,** 1534–1548 (2004).
4. Owston, C. N. A Hall effect magnetometer for small magnetic fields. *J. Sci. Instrum.* **44,** 798–800 (1967).
5. Rugar, D., Budakian, R., Mamin, H. J. & Chui, B. W. Single spin detection by magnetic resonance force microscopy. *Nature* **430,** 329–332 (2004).
6. Mamin, H. J., Poggio, M., Degen, C. L. & Rugar, D. Nuclear magnetic resonance imaging with 90-nm resolution. *Nature Nanotechnol.* **2,** 301–306 (2007).
7. Dutt, M. V. G. *et al.* Quantum register based on individual electronic and nuclear spin qubits in diamond. *Science* **316,** 1312–1316 (2007).
8. Taylor, J. *et al.* High-sensitivity diamond magnetometer with nanoscale resolution. *Nature Phys.* (in the press); preprint at ⟨http://arXiv.org/abs/0805.1367v1⟩ (2008).
9. Budker, D. F., Kimball, D. F. & DeMille, D. P. *Atomic Physics: An Exploration Through Problems and Solutions* (Oxford Univ. Press, 2004).
10. Ludlow, A. D. *et al.* Sr lattice clock at 1×10^{-16} fractional uncertainty by remote optical evaluation with a Ca clock. *Science* **319,** 1805–1808 (2008).
11. Rosenband, T. *et al.* Frequency ratio of Al^+ and Hg^+ single-ion optical clocks; metrology at the 17th decimal place. *Science* **319,** 1808–1812 (2008).
12. Wineland, D. J., Bollinger, J. J., Itano, W. M., Moore, F. L. & Heinzen, D. J. Spin squeezing and reduced quantum noise in spectroscopy. *Phys. Rev. A* **46,** R6797 (1992).
13. Kominis, I. K., Kornack, T. W., Allred, J. C. & Romalis, M. V. A subfemtotesla multichannel atomic magnetometer. *Nature* **422,** 596–599 (2003).
14. Jelezko, F., Gaebel, T., Popa, I., Gruber, A. & Wrachtrup, J. Observation of coherent oscillations in a single electron spin. *Phys. Rev. Lett.* **92,** 076401 (2004).
15. Jelezko, F. *et al.* Observation of coherent oscillation of a single nuclear spin and realization of a two-qubit conditional quantum gate. *Phys. Rev. Lett.* **93,** 130501 (2004).
16. Childress, L. *et al.* Coherent dynamics of coupled electron and nuclear spin qubits in diamond. *Science* **314,** 281–285 (2006).
17. Jiang, L. *et al.* Coherence of an optically illuminated single nuclear spin qubit. *Phys. Rev. Lett.* **100,** 073001 (2008).
18. Hanson, R., Mendoza, F. M., Epstein, R. J. & Awschalom, D. D. Polarization and readout of coupled single spins in diamond. *Phys. Rev. Lett.* **97,** 087601 (2006).
19. Gaebel, T. *et al.* Room-temperature coherent coupling of single spins in diamond. *Nature Phys.* **2,** 408–413 (2006).
20. Hanson, R., Dobrovitski, V. V., Feiguin, A. E., Gywat, O. & Awschalom, D. D. Coherent dynamics of a single spin interacting with an adjustable spin bath. *Science* **320,** 352–355 (2008).
21. Maze, J. R., Taylor, J. M. & Lukin, M. D. Electron spin decoherence of single nitrogen-vacancy defects in diamond. Preprint at ⟨http://arXiv.org/abs/0805.0327⟩ (2008).
22. Rabeau, J. R. *et al.* Single nitrogen vacancy centers in chemical vapor deposited diamond nanocrystals. *Nano Lett.* **7,** 3433–3437 (2007).
23. Rabeau, J. R. *et al.* Implantation of labelled single nitrogen vacancy centers in diamond using ^{15}N. *Appl. Phys. Lett.* **88,** 023113 (2006).
24. Wrachtrup, J. & Jelezko, F. Processing quantum information in diamond. *J. Phys. Condens. Matter* **18,** S807–S824 (2006).
25. Chang, D. E., Sorensen, A. S., Hemmer, P. R. & Lukin, M. D. Quantum optics with surface plasmons. *Phys. Rev. Lett.* **97,** 053002 (2006).
26. Balasubramanian, G. *et al.* Nanoscale imaging magnetometry with diamond spins under ambient conditions. *Nature* doi:10.1038/nature07278 (this issue).
27. Degen, C. L. Scanning magnetic field microscope with a diamond single-spin sensor. Preprint at ⟨http://arXiv.org/abs/0805.1215v2⟩ (2008).
28. Cirac, J. I. & Zoller, P. A scalable quantum computer with ions in an array of microtraps. *Nature* **404,** 579–581 (2000).
29. Treutlein, P., Hunger, D., Camerer, S., Hansch, T. W. & Reichel, J. Bose-Einstein condensate coupled to a nanomechanical resonator on an atom chip. *Phys. Rev. Lett.* **99,** 140403 (2007).
30. Rabl, P. *et al.* Strong magnetic coupling between an electronic spin qubit and a mechanical resonator. Preprint at ⟨http://arXiv.org/abs/0806.3606⟩ (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge A. Akimov, D. Budker, F. Jelezko, F. Koppens, A. Trifonov, P. Hemmer and J. Wratchtrup for many discussions and experimental assistance. This work was supported by the NSF, DARPA, the Packard Foundation and Harvard CNS.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.D.L. (lukin@fas.harvard.edu).

METHODS

Samples. AC magnetometry was performed at room temperature on nitrogen-vacancy centres in both a bulk single-crystal diamond sample (1 mm × 1 mm × 0.5 mm, natural diamond with an atypically low nitrogen concentration) and in diamond nanocrystals (monocrystalline, synthetic diamonds, 30 nm mean diameter, purchased from Microdiamant) deposited on a quartz coverslip.

Confocal set-up. Single nitrogen-vacancy centres were isolated and probed via confocal microscopy. Phonon-mediated fluorescent emission (630–750 nm) was detected under coherent optical excitation ($\lambda = 532$ nm) using a single photon counting module (Perkin-Elmer SPCM-AQRH-13). The density of nitrogen-vacancy centres in both the bulk single-crystal and nanocrystal samples were sufficiently low that single bright spots (within the approximate confocal volume of 200 nm × 200 nm × 500 nm) were resolvable from the background fluorescence. As single spots in the confocal image may constitute many nitrogen-vacancy centres, single centres were identified by observing photon antibunching in the measurement of the second-order correlation function. This emission was separated from the excitation path using a dichroic mirror, and also notch and longpass filters. Samples were imaged with an oil immersion objective lens (Nikon CFI Plan Fluor Series, NA = 1.3, 100× magnification) over a 50 μm × 50 μm area in the plane normal to the optical path. Two galvanometer controlled mirrors steered the beam path for rapid imaging of this area. Experimental drifting of the focal plane due to thermal effects was compensated by using closed-loop feedback of the galvanometer and objective piezo voltages.

Single-centre electron spin resonance. The nitrogen-vacancy centre $^3A_2$ ground state consists of two unpaired electrons in a triplet configuration leading to a zero-field splitting ($\Delta = 2.87$ GHz) between the $m_s = 0$ and $m_s = \pm 1$ sublevels. Coherent optical excitation at $\lambda = 532$ nm optically pumped the ground state into its $m_s = 0$ sublevel. In addition, an external static magnetic field produced by a pair of Helmholtz coils split the degeneracy between the $m_s = \pm 1$ states. It was then possible to selectively address transitions between the $m_s = 0$ and $m_s = 1$ (or $m_s = -1$) states with microwave radiation (Fig. 1b) and manipulate a two-level subspace of the spin triplet (for example with spin-echo pulse sequences). Microwave radiation was applied by using the magnetic field emanating from a 20 μm wire placed on the surface of the samples.

The excited (^3E) state decay rates, also responsible for optical pumping, provided a means for spin-sensitive detection, as the rate of fluorescence was reduced for the $m_s = \pm 1$ states compared to the $m_s = 0$ states, with >35% contrast. The spin state in the ground electronic state was measured by pulsing on green excitation and monitoring the total number of photons collected within the optimal measurement interval, 324 ns. A 300 MHz PulseBlaster ESR pulse generator was employed for timing the triggering of counters, microwave pulses, the AC magnetic field, and the excitation laser. Microwave pulses were provided by gating the output of a frequency synthesizer with a microwave switch, while green laser pulses were generated using an acousto-optic modulator. The π and π/2 pulses used for the spin-echo sequence were calibrated from the Rabi nutation curves between the two spin states.

AC magnetometry. As described in Fig. 1d and in ref. 8, demonstration magnetometer measurements were performed for an externally applied AC magnetic field with amplitude B_{AC}, frequency ν, and phase ϕ_{AC} during a cycle of a spin-echo sequence with a period τ. The accumulated phase of the spin superposition state

$$\delta\phi = \frac{4g\mu_B B_{AC}}{2\pi\nu}\sin^2\left(\frac{\pi\nu\tau}{2}\right)\cos(\pi\nu\tau + \phi_{AC}) \quad (2)$$

contained information about the projection of the AC magnetic field amplitude onto the quantization axis of the electronic spin, corresponding to a vector magnetometer. Oscillatory magnetic fields from 1–10 kHz were generated by modulating the current through a Helmholtz pair also used to apply a bias DC magnetic field. For application of higher frequency AC fields (100–300 kHz), a single coil (60 turns) was resonantly driven and placed near the sample.

The measured signal intensity S_B was a function of the accumulated phase $\delta\phi$, as given by the probability of being in the $m_s = 0$ state after the spin-echo pulse sequence: $S_B \propto P_0(B_{AC}) = [1 + F(\tau)\cos(\delta\phi)]/2$. Ideally, for a single-shot measurement of B_{AC} the sensitivity was maximized for a particular ν by setting $\tau = 1/\nu$. In practice, many spin-echo cycles were averaged to reduce the uncertainty in photon statistics given the low single-shot count rate. To this end, the period of the entire measurement sequence (including polarization and readout, Fig. 1d) was matched to $1/\nu$ in order to avoid multiple offset phases ϕ_{AC} when the periods were incommensurate. The dependence on ϕ_{AC} was removed entirely by appropriately shifting the time origin of the measurement pulse train. As the polarization ($\tau_p = 1$ μs) and readout ($\tau_r = 3$ μs) periods were short compared to the oscillation periods for typical 1–10 kHz AC magnetic fields, this choice introduced a slight deviation ε from the optimal $\delta\phi$, as $\tau_p + \tau_r + \tau = 1/\nu \rightarrow \tau\nu = 1 - \varepsilon$. The overall sensitivity was thus slightly reduced from its optimal value as $1 - \mathcal{O}(\varepsilon^2)$. For all experiments presented here, $\nu\tau = 0.88$ was used. The envelope of the spin-echo signal, $F(\tau)$ (see, for example, Fig. 2a) was modelled with an exponential decay modulated by the effect of a pair of nearby strongly interacting ^{13}C nuclear spins. In this model[21], $F(\tau) = \exp(-(\tau/T_2)^4)(1 - [(a^2 - b^2)/a^2]\sin^2(a\tau)\sin^2(b\tau))$; where for the data in Fig. 2a we found $T_2 = 676$ μs, $b = 478$ Hz (corresponding to the dipolar interaction between the two nuclei) and $a = 626$ Hz (corresponding to the interactions between the nuclei and the nitrogen-vacancy spin). Using these experimentally determined parameters, the above model provided a prediction for the magnetometer sensitivity[8] $\eta_{AC} = \pi\hbar/(g\mu_B C\sqrt{\nu}F(1/\nu))$ as a function of frequency (solid curve in Fig. 3a), where $g \approx 2$ is the electron g-factor, μ_B is the Bohr magneton, and $C^{-2} = 1 + 2(a_0 + a_1 + a_0a_1)/(a_0 - a_1)^2$ is a factor that estimates[8] the photon shot noise when the average photon number during the readout window of 324 ns is much less than 1. The values $a_0 = 0.03 \pm 0.006$ and $a_1 = 0.018 \pm 0.004$ were the average numbers of detected photons for the electronic spin states $m_s = 0$ and $m_s = \pm 1$, respectively.

LETTERS

Nanoscale imaging magnetometry with diamond spins under ambient conditions

Gopalakrishnan Balasubramanian[1], I. Y. Chan[2]†, Roman Kolesov[1], Mohannad Al-Hmoud[1], Julia Tisler[1], Chang Shin[3], Changdong Kim[3], Aleksander Wojcik[3], Philip R. Hemmer[3], Anke Krueger[4], Tobias Hanke[5], Alfred Leitenstorfer[5], Rudolf Bratschitsch[5], Fedor Jelezko[1] & Jörg Wrachtrup[1]

Magnetic resonance imaging and optical microscopy are key technologies in the life sciences. For microbiological studies, especially of the inner workings of single cells, optical microscopy is normally used because it easily achieves resolution close to the optical wavelength. But in conventional microscopy, diffraction limits the resolution to about half the wavelength. Recently, it was shown that this limit can be partly overcome by nonlinear imaging techniques[1,2], but there is still a barrier to reaching the molecular scale. In contrast, in magnetic resonance imaging the spatial resolution is not determined by diffraction; rather, it is limited by magnetic field sensitivity, and so can in principle go well below the optical wavelength. The sensitivity of magnetic resonance imaging has recently been improved enough to image single cells[3,4], and magnetic resonance force microscopy[5] has succeeded in detecting single electrons[6] and small nuclear spin ensembles[7]. However, this technique currently requires cryogenic temperatures, which limit most potential biological applications[8]. Alternatively, single-electron spin states can be detected optically[9,10], even at room temperature in some systems[11–14]. Here we show how magneto-optical spin detection can be used to determine the location of a spin associated with a single nitrogen-vacancy centre in diamond with nanometre resolution under ambient conditions. By placing these nitrogen-vacancy spins in functionalized diamond nanocrystals, biologically specific magnetofluorescent spin markers can be produced. Significantly, we show that this nanometre-scale resolution can be achieved without any probes located closer than typical cell dimensions. Furthermore, we demonstrate the use of a single diamond spin as a scanning probe magnetometer to map nanoscale magnetic field variations. The potential impact of single-spin imaging at room temperature is far-reaching. It could lead to the capability to probe biologically relevant spins in living cells.

The nitrogen-vacancy centre in diamond is a unique solid state system that allows ultrasensitive and rapid detection of single electronic spin states under ambient conditions[12]. The nitrogen-vacancy defect is a naturally occurring impurity that is responsible for the pink colouration of diamond crystals when present in high concentration. It was demonstrated that this colour centre can be produced in diamond nanocrystals by electron irradiation. Fluorescing nitrogen-vacancy diamond nanocrystals can be used as markers for bioimaging applications[15]. Such markers have attracted widespread interest because of their unprecedented photostability and non-toxicity[16,17]. It was recognized recently that the magnetic properties of such fluorescent labels can in principle be used for novel microscopy[18,19]. Here we demonstrate the realization of a magneto-optic microscope using nitrogen-vacancy diamond as the magnetic fluorescent label that moreover does not bleach or blink.

Figure 1c and d show the fluorescence and atomic force microscope image of nanocrystals containing nitrogen-vacancy defects. By careful choice of irradiation doses, we were able to control the number of nitrogen-vacancy centres per nanocrystal. The particular sample presented in Fig. 1 has on average a single nitrogen-vacancy defect per 40 nm nanocrystal (confirmed by fluorescence correlation measurements, Fig. 1e).

The energy level scheme and structure of the nitrogen-vacancy defect is shown in Fig. 1a and b. Two out of six electrons of the centre are unpaired, forming an electron spin triplet in the electronic ground and first excited state. Broadband optical excitation of the centre polarizes it by optical pumping into the $m_s = 0$ spin sublevel. Laser-assisted detection of the spin state of a single nitrogen-vacancy centre makes use of differences in the absorption and emission properties of the spin sublevels. Specifically, the spin sublevel with magnetic quantum number $m_s = 0$ (bright state) scatters ~30% more photons than $m_s = \pm 1$ states. Hence, when a resonant microwave field induces magnetic dipole transitions between these electronic spin sublevels, it destroys the optically pumped spin polarization, resulting in a significant decrease of the nitrogen-vacancy centre fluorescence. An example of such an optically detected electron spin resonance (ESR) spectrum of a single nitrogen-vacancy electronic spin is shown in Fig. 1f.

The spin Hamiltonian of the nitrogen-vacancy defect (neglecting electron–nuclear spin coupling) can be written as the sum of zero-field and Zeeman terms, $H = D\left(S_z^2 - (1/3)[S(S+1)]\right) + E\left(S_x^2 - S_y^2\right) + g\mu_B \mathbf{B}\cdot\mathbf{S}$, where D and E are zero-field splitting parameters, $S = 1$, μ_B is the Bohr magneton and g is the electron g-factor ($g = 2.0$). Owing to the magnetic dipole interaction between the two unpaired electrons even at zero external magnetic field, the sublevels $m_s = 0$ and $m_s = \pm 1$ are separated ($D = 2{,}870$ MHz). Owing to symmetry, the $m_s = \pm 1$ sublevels of the nitrogen-vacancy defect are degenerate at zero magnetic field ($E = 0$), resulting in a single resonance line appearing in the ESR spectrum (Fig. 1f). An external magnetic field lifts the degeneracy of $m_s = \pm 1$, leading to the appearance of two lines. By measuring the positions of the ESR resonances ω_1 and ω_2, it is possible to calculate the magnitude of the external field B according to $(g\mu B)^2 = (1/3)\left(\omega_1^2 + \omega_2^2 - \omega_1\omega_2 - D^2\right) - E^2$ (see Methods for details).

From the above-mentioned relations, it can be seen that, when combined with nano-positioning instrumentation, the single spin associated with a nitrogen-vacancy defect can be used as an atom-sized scanning probe vector magnetometer. Similarly, when placed in

[1]3 Physikalisches Institut, Universität Stuttgart, 70550 Stuttgart, Germany. [2]Department of Chemistry, Brandeis University, Waltham, Massachusetts 02454, USA. [3]Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas 77843, USA. [4]Otto-Diels-Institut für Organische Chemie, Christian-Albrechts-Universität zu Kiel, 24098 Kiel, Germany. [5]University of Konstanz and Center for Applied Photonics, 78457 Konstanz, Germany. †Present address: 3 Physikalisches Institut, Universität Stuttgart, 70550 Stuttgart, Germany.

an inhomogeneous magnetic field with a known field gradient, the defect can be used as a magneto-optical spin marker for suboptical-wavelength tagged imaging. As a demonstration, two-dimensional spin imaging experiments were performed using a single nitrogen-vacancy centre and the highly inhomogeneous magnetic field produced by the magnetic tip of an atomic force microscope (AFM). The experimental set-up is shown in Fig. 2a. A commercial AFM was combined with a confocal microscope. The magnetic probe, commonly used in magnetic force microscopy, consists of a sharp silicon tip coated with 30 nm of magnetic material: the exact magnetic field profile of the cantilever is not known a priori, and must be determined. For this, we have used our single-spin nitrogen-vacancy magnetometer. The magnetic cantilever was first placed at a known distance from the diamond nanocrystal, and the magnetic field experienced by the single nitrogen-vacancy centre was recorded in steps (corresponding to several hundred nanometre displacements of the cantilever) by acquiring ESR spectra such as those in Fig. 1f at each location. The experimentally obtained data points were then fitted using a Lorentzian function, inferred from numerical simulation of the field created by the cantilever (Fig. 2b). This gives the magnetic field profile of the cantilever in one dimension. Similarly, the profile along an orthogonal axis is recorded to give the two-dimensional profile as well as the exact position of the nitrogen-vacancy centre.

To visualize the resolving power of our gradient imaging technique, the magnetic cantilever was scanned in the vicinity of a nanocrystal containing a single nitrogen-vacancy defect while simultaneously exciting with a fixed-frequency microwave field. When a confocal image is acquired, each point of the optical image corresponds to a well-defined magnetic field value (as measured in

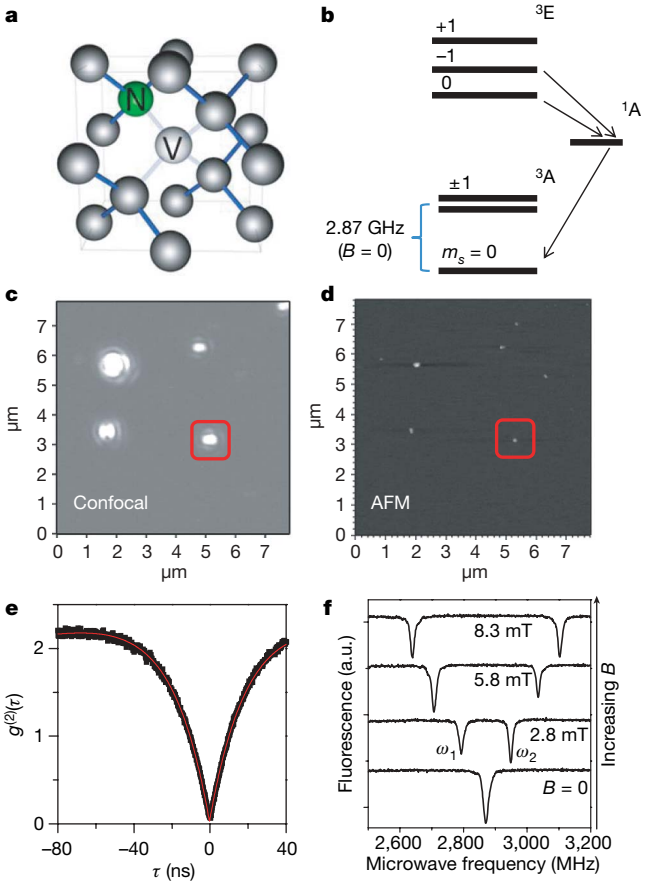

Figure 1 | Nitrogen-vacancy defect in diamond. **a**, Structure and energy level scheme of the nitrogen-vacancy (NV) defect in diamond. Optical pumping initializes the centre into the $m_s = 0$ spin state via spin selective shelving into the metastable singlet state, ^{1}A. This state decays preferentially into the $m_s = 0$ sublevel of the ground state, leading to optically induced spin polarization (more than >90% at room temperature). **c**, **d**, Simultaneously acquired optical (**c**) and AFM (**d**) image of diamond nanocrystals containing single nitrogen-vacancy defects. **e**, Fluorescence autocorrelation function, confirming that the nanocrystal contains a single nitrogen-vacancy defect. The contrast of $g^2(\tau)$ at zero delay time scales as $1/N$, where N is the number of emitters. **f**, Optically detected magnetic resonance spectra for a single nitrogen-vacancy defect at increasing magnetic field (from bottom to top).

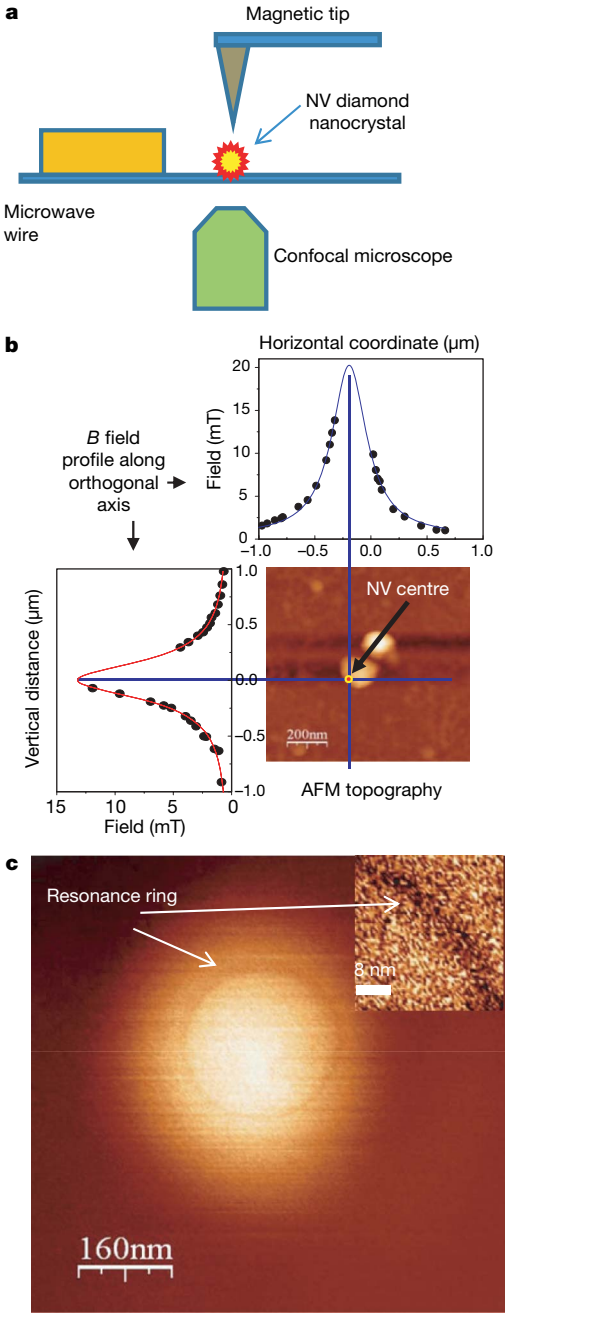

Figure 2 | Gradient imaging with single spins. **a**, Two-dimensional imaging is achieved using a field gradient created by a magnetic cantilever. **b**, The experimental profile of the cantilever's magnetic field for two orthogonal axes. The magnetic tip was placed at several points parallel to the blue lines, and the ESR spectra were measured. The calculated (fitted) magnetic field profile allows estimation of the location of the single nitrogen-vacancy centre (shown in the AFM topography). **c**, Two-dimensional magnetic resonance image of a single nitrogen-vacancy centre, showing resonance rings corresponding to a magnetic field of 3 mT (resonance frequency of 2,780 MHz). Inset, an enlarged section of a ring with a width of approximately 5 nm.

the previous experiment). At particular positions (pixels) when the microwave frequency is resonant with the corresponding spin sublevel splitting, the fluorescence intensity is reduced. This results in a dark ring (Fig. 2), which marks a two-dimensional cut through the B field corresponding to a constant magnetic field projection along the nitrogen-vacancy quantization axis. The width of the rings is given by the magnetic resonance linewidth divided by the field gradient ($80\,\mu T\,nm^{-1}$). This ring width also defines the ultimate resolution limit of this technique for spin imaging. Rings with 5 nm width are observed, and shown in Fig. 2c inset. Note that this width is smaller than the sizes of both the magnetic tip and the diamond nanocrystal, and is only possible because a single nitrogen-vacancy centre is localized to a fraction of a nanometre in the diamond lattice. The dark ring shown in Fig. 2c is only seen if simultaneously the nitrogen-vacancy axis is oriented vertically, and the magnetic field is radially symmetric. This special case was selected to simplify understanding of the technique. It is interesting to note that a vibrating cantilever (a.c. mode AFM was used in all the experiments) induces significant line broadening when the magnetic cantilever comes very close to the spin (see Supplementary Information).

Being an atomic-sized non-perturbing magnetic field sensor, the single nitrogen-vacancy centre can be incorporated into the cantilever instead of a magnetic coating, and used as a scanning probe magnetometer to achieve subwavelength imaging resolution. To demonstrate the feasibility of this approach, we attached a nanocrystal containing a single nitrogen-vacancy centre to the tip of a cantilever, and used it to profile the magnetic field produced by a nanometre-sized magnetic structure. Details of the set-up are shown in Fig. 3a. Microwaves are tuned into resonance with the nitrogen-vacancy spin at the tip of the cantilever (see Fig. 3b) for a particular magnetic field projection. Hence the resonance conditions in the vicinity of the magnetic nanostructures are satisfied along well-defined lines of constant B_z, where z is along the nitrogen-vacancy quantization axis. Figure 3c shows a magneto-optical image of a triangular magnetic structure obtained with a single nitrogen-vacancy defect as light source (as expected, the structure appears as a shadow in our detection geometry). The narrow dark line close to the corner represents spatial regions where the conditions for magnetic resonance of the nitrogen-vacancy centre on the tip are fulfilled ($B_z = 5$ mT). Note that the image represents raw data acquired in just 4 minutes. The magnetic field resolution is given by the width of the dark lines, which are about 20 nm, multiplied by the magnetic field gradient of $25\,\mu T\,nm^{-1}$ (measured by recording several resonance lines at different microwave frequencies, data not shown). It corresponds to a measurement resolution of 0.5 mT. The limiting factor here is oscillatory motion of the nanodiamond attached to the AFM tip (see Methods for details).

The resolution could be significantly improved by phase locking of the detection system to the oscillatory motion of the cantilever, and using echo-based techniques with an echo period matched to a single oscillation period of the cantilever. This essentially corresponds to measuring a.c. instead of d.c. magnetic fields. The advantage of using echoes is that the effective ESR linewidth is narrower than the inhomogeneous linewidth[20], and for a long spin phase memory time T_2 is effectively given by the AFM vibration frequency, which is 100 kHz for standard AFM cantilevers. Hence we expect an improvement of field measurement accuracy by a factor of 150 (3 μT) using this technique. For the magnetic field gradient caused by the structure imaged in Fig. 3c, this would correspond to subnanometre spatial resolution.

Ultrasensitive magnetometry with single spins in diamond not only allows high spatial resolution imaging, but also might be applied to image single external spins under ambient conditions. Here the magnetic sublevels of the nitrogen-vacancy centre are shifted by the magnetic fields produced by (for example) other single electron or nuclear magnetic dipoles in nearby molecules. To show the feasibility of single electron and nuclear spin detection, we estimate the ultimate sensitivity limit of a scanning spin microscope based on nitrogen-vacancy centres in diamond. The magnetic field created by a single electron spin located at distance r from the nitrogen-vacancy spin is $B^{dip} = (\mu_0\mu/4\pi)\sqrt{3\cos^2\theta + 1}/r^3$, where μ the single spin magnetic moment, and θ is the angle between the vector connecting the two spins and the vector of the external magnetic field. When we substitute $\mu = -(1/2)g_e\mu_B \approx 10^{-23}\,JT^{-1}$, and $\mu_0/4\pi \approx 10^{-7}\,NA^{-2}$, a field of 10^{-5} T can be obtained for a distance between the electron and nitrogen-vacancy spins of 5 nm. For the nitrogen-vacancy centre this gives up to 0.3 MHz of ESR frequency shift, which is within the projected detection limit for the single nitrogen-vacancy nanocrystals used in this demonstration. Imaging single nuclear spins is more challenging, as the lower nuclear magnetic moment results in fields three orders of magnitude lower (10^{-8} T). This value corresponds to a kilohertz shift of the nitrogen-vacancy resonance.

In general, the sensitivity of our nitrogen-vacancy magnetometer is determined by the linewidth of the ESR transition. Experiments presented here were carried out using continuous wave (c.w.) ESR

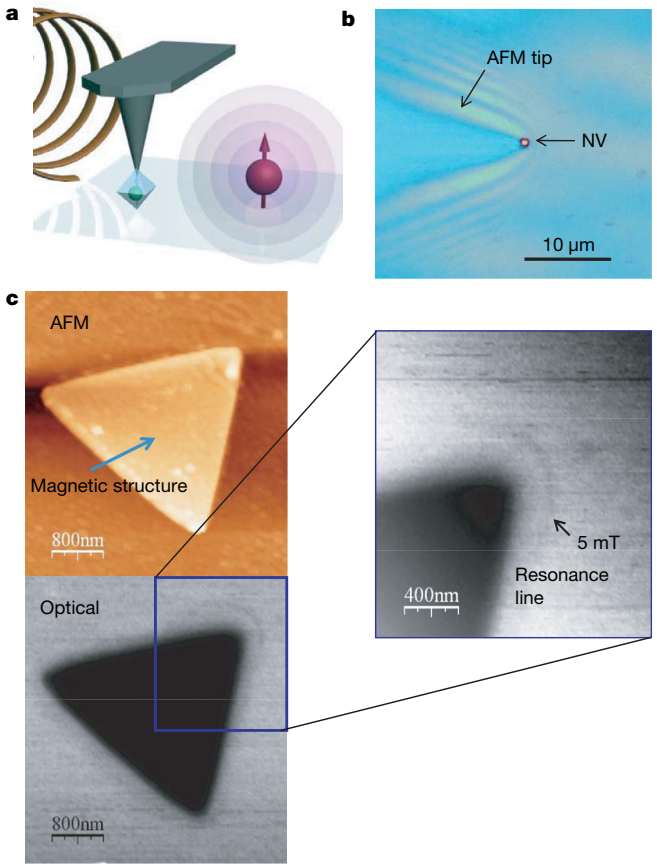

Figure 3 | Scanning probe magnetometry. a, Diagram of the magnetic field imaging experiment. A nanoscale magnetic particle (red) is imaged with a single nitrogen-vacancy defect (green, within the blue nanocrystal) fixed at the scanning probe tip (black). **b**, Optical image of a diamond nanocrystal attached to an AFM tip (view from the bottom). The scattered light image of the tip is overlapped with the fluorescence image of the nanocrystal. The bright spot (arrowed) represents fluorescence of a single nitrogen-vacancy defect. Fluorescence autocorrelation function (data not shown) shows a pronounced antibunching dip, indicating a single nitrogen-vacancy defect in the nanocrystal on the AFM tip. **c**, Field reconstruction using the scanning probe single spin magnetometer. Top left, an AFM image of a nickel magnetic nanostructure prepared by electron beam lithography; bottom left, a magneto-optical image of the same structure, recorded using a single nitrogen-vacancy centre on the AFM tip as light source and magnetometer. Inset (right), the fluorescence signal from the scanned nitrogen-vacancy centre attached to the apex of the AFM tip when resonant microwaves at 2,750 MHz are applied (the arrowed point corresponds to 5 mT resonance line with the magnetic field tilted by 45° relative to the nitrogen-vacancy axis).

and technical grade diamonds. Hence the linewidth of the spin resonance line (a few MHz) was limited by fast (μs) decoherence and microwave-field-induced broadening. However, it was recently shown that the phase memory time for ultrapure diamond reaches one millisecond when echo-based techniques are used for detection[21]. This corresponds to a linewidth of the order of 0.3 kHz. Taking this value, single nuclear spins can be detected at 5 nm distance under ambient conditions[22]. As these spin linewidths were obtained under ambient conditions, this approach will potentially enable the use of nitrogen-vacancy defects in diamond nanocrystals as a probe for intracellular electron (and possibly nuclear) spin imaging.

METHODS SUMMARY

Magnetic imaging and magnetometry experiments were performed using a home-built scanning confocal microscope combined with an AFM (MFP-3D Asylum Research). Nitrogen-vacancy defects were excited with a frequency doubled c.w. Nd:YAG laser (Coherent Compass) focused on to the sample with a high NA objective (Olympus PlanAPO, NA = 1.35). Luminescence light was collected by the same objective and filtered from the excitation light using a dichroic beamsplitter (640 DCXR, Chroma) and long-pass filter (647 LP, Chroma). Photon counting of the filtered light was performed using two avalanche photodiodes (SPQR-14, Perkin-Elmer). Fluorescence autocorrelation histograms were recorded using a fast multichannel analyser (Fastcomtec, P7889). Optically detected magnetic resonance measurements were performed using a commercial microwave source (Rhode & Schwarz GmbH, SMIQ 03) amplified by a travelling wave tube amplifier (Hughes 8020H). Commercially available magnetic cantilevers (Team Nanotec) were used for generation of high magnetic field gradients. UV curing glue (Thorlabs) was used to attach diamond nanocrystals to the AFM tip for the scanning probe magnetometry.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 April; accepted 18 July 2008.

1. Willig, K. I., Rizzoli, S. O., Westphal, V., Jahn, R. & Hell, S. W. STED microscopy reveals that synaptotagmin remains clustered after synaptic vesicle exocytosis. *Nature* **440,** 935–939 (2006).
2. Betzig, E. *et al.* Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313,** 1642–1645 (2006).
3. Aguayo, J. B., Blackband, S. J., Schoeniger, J., Mattingly, M. A. & Hintermann, M. Nuclear-magnetic-resonance imaging of a single cell. *Nature* **322,** 190–191 (1986).
4. Ciobanu, L., Seeber, D. A. & Pennington, C. H. 3D MR microscopy with resolution 3.7 μm by 3.3 μm by 3.3 μm. *J. Magn. Reson.* **158,** 178–182 (2002).
5. Rugar, D., Yannoni, C. S. & Sidles, J. A. Mechanical detection of magnetic resonance. *Nature* **360,** 563–566 (1992).
6. Rugar, D., Budakian, R., Mamin, H. J. & Chui, B. W. Single spin detection by magnetic resonance force microscopy. *Nature* **430,** 329–332 (2004).
7. Mamin, H. J., Poggio, M., Degen, C. L. & Rugar, D. Nuclear magnetic resonance imaging with 90-nm resolution. *Nature Nanotechnol.* **2,** 301–306 (2007).
8. Glover, P. & Mansfield, P. Limits to magnetic resonance microscopy. *Rep. Prog. Phys.* **65,** 1489–1511 (2002).
9. Kohler, J. *et al.* Magnetic resonance of a single molecular spin. *Nature* **363,** 242–244 (1993).
10. Wrachtrup, J., von Borczyskowski, C., Bernard, J., Orrit, M. & Brown, R. Optical detection of magnetic resonance in a single molecule. *Nature* **363,** 244–245 (1993).
11. Hanson, R., Dobrovitski, V. V., Feiguin, A. E., Gywat, O. & Awschalom, D. D. Coherent dynamics of a single spin interacting with an adjustable spin bath. *Science* **320,** 352–355 (2008).
12. Gruber, A. *et al.* Scanning confocal optical microscopy and magnetic resonance on single defect centers. *Science* **276,** 2012–2014 (1997).
13. Epstein, R. J., Mendoza, F. M., Kato, Y. K. & Awschalom, D. D. Anisotropic interactions of a single spin and dark-spin spectroscopy in diamond. *Nature Phys.* **1,** 94–98 (2005).
14. Childress, L. *et al.* Coherent dynamics of coupled electron and nuclear spin qubits in diamond. *Science* **314,** 281–285 (2006).
15. Fu, C. C. *et al.* Characterization and application of single fluorescent nanodiamonds as cellular biomarkers. *Proc. Natl Acad. Sci. USA* **104,** 727–732 (2007).
16. Liu, K. K., Cheng, C. L., Chang, C. C. & Chao, J. I. Biocompatible and detectable carboxylated nanodiamond on human cell. *Nanotechnology* **18,** 325102 (2007).
17. Neugart, F. *et al.* Dynamics of diamond nanoparticles in solution and cells. *Nano Lett.* **7,** 3588–3591 (2007).
18. Chernobrod, B. M. & Berman, G. P. Spin microscope based on optically detected magnetic resonance. *J. Appl. Phys.* **97,** 014903 (2005).
19. Kuhn, S., Hettich, C., Schmitt, C., Poizat, J. P. H. & Sandoghdar, V. Diamond colour centres as a nanoscopic light source for scanning near-field optical microscopy. *J. Microsc.* **202,** 2–6 (2001).
20. Taylor, J. M. *et al.* High-sensitivity diamond magnetometer with nanoscale resolution. *Nature Phys.* (in the press); preprint at ⟨http://arXiv.org/abs/0805.1367v1⟩ (2008).
21. Gaebel, T. *et al.* Room-temperature coherent coupling of single spins in diamond. *Nature Phys.* **2,** 408–413 (2006).
22. Maze, J. R. *et al.* Nanoscale magnetic sensing with an individual electronic spin in diamond. *Nature* doi:10.1038/nature07279 (this issue).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. D. Lukin for drawing our attention to advanced echo-based techniques, and R. Kamella for technical assistance. This work was supported by the EU (QAP, EQUIND, NANO4DRUGS, NEDQIT), DFG (SFB/TR21 and FOR730) and Landesstiftung BW.

Author Contributions G.B., I.Y.C, R.K., M.A.-H., J.T., C.S., C.K., A.W., J.W. and F.J. performed the experiments; A.K. prepared diamond nanocrystals; T.H., A.L. and R.B. prepared magnetic nanostructures; P.R.H., J.W. and F.J. designed and coordinated the experiments; and F.J. wrote the paper. All authors discussed the results, analysed the data and commented on the manuscript.

Author Information Correspondence and requests for materials should be addressed to F.J. (f.jelezko@physik.uni-stuttgart.de).

METHODS

Vector magnetometry using a single nitrogen-vacancy defect electron spin. The resonance frequencies of the $m_s = 0 \leftrightarrow \pm 1$ spin transitions of a spin-1 system allow one to extract the magnitude of the local magnetic field and, under some approximations, the angle between the magnetic field and the symmetry axis of the system. The spin Hamiltonian of an $S = 1$ system having a distorted C_{3v} symmetry is given by the following expression:

$$H = \mu_B g \mathbf{B}\cdot\mathbf{S} + D\left(S_z^2 - S(S+1)/3\right) + E\left(S_x^2 - S_y^2\right)$$

where D and E are the zero-field splitting parameters, $S = 1$, μ_B is the Bohr magneton, $g = 2$ is the g-factor, and $\mathbf{B}$ is the external magnetic field. Imperfect axial symmetry is reflected by the asymmetry parameter E. These parameters unambiguously determine the natural local coordinate system, with the z axis being along the axis of the nitrogen-vacancy centre and x and y axes being along the principal axes of the distortion ellipsoid. In such a coordinate system, it is convenient to describe the magnetic field by its magnitude B and the two angles, θ (polar) and φ (azimuthal). All parameters B, θ and φ can be obtained from analysis of the ESR spectrum. The positions of spin levels are given by the solutions of the characteristic equation:

$$x^3 - \left(\frac{D^2}{3} + E^2 + \beta^2\right)x - \frac{\beta^2}{2}\left(D\cos 2\theta + 2E\cos 2\varphi \sin^2\theta\right) - \frac{D}{6}\left(4E^2 + \beta^2\right) + \frac{2D^3}{27} = 0$$

where $\beta = \mu_B g B$. Denoting the frequency of the $S_z = 0$ state as x_0, one finds that the positions of the $S_z = \pm 1$ states are given by $x_\pm = x_0 + \nu_{0\pm}$, where $\nu_{0\pm}$ are the experimentally measured frequencies of $0 \leftrightarrow \pm 1$ spin transitions. Since $x_\pm$ must satisfy the above-mentioned equation, it is possible to obtain three equations for the four unknowns (x_0, B, θ and φ), two of which, θ and φ, form a unique combination $\Delta = D\cos 2\theta + 2E\cos 2\varphi \sin^2\theta$. This set of equations gives the following solutions for β and Δ:

$$\beta^2 = \frac{1}{3}\left(\nu_1^2 + \nu_2^2 - \nu_1\nu_2 - D^2\right) - E^2,$$

$$\Delta = \frac{7D^3 + 2(\nu_1+\nu_2)\left(2\left(\nu_1^2+\nu_2^2\right) - 5\nu_1\nu_2 - 9E^2\right) - 3D\left(\nu_1^2 + \nu_2^2 - \nu_1\nu_2 + 9E^2\right)}{9\left(\nu_1^2 + \nu_2^2 - \nu_1\nu_2 - D^2 - 3E^2\right)}$$

Since for nitrogen-vacancy centres $D \gg E$, $\Delta \approx D\cos 2\theta$. Thus, knowing the zero-field splitting parameters and the frequencies of the $0 \leftrightarrow \pm 1$ ESR resonances, one can find B and θ. The solution of the inverse problem of finding the two ESR frequencies given the known B and θ is presented in Supplementary Information.

Modelling the magnetic field of the cantilever. The magnetic field of a tip having a ferromagnetic coating was simulated in the following manner. The surface of the tip is assumed to be an axially symmetric cone with a round apex. It can be simulated by the following simple analytical formula:

$$h = r\tanh\frac{r}{2r_0}$$

where h is the height of the surface point above the apex, r is the radial coordinate, and r_0 is the curvature radius of the rounded apex of the tip. It is assumed that the tip surface is covered with a thin layer of ferromagnetic material. The magnetic field produced by an infinitely small element of the surface was approximated as that of a magnetic dipole. The magnetization of the surface is assumed to be uniform and the direction of the magnetization of each surface element assumed the same. The contributions of all surface elements were then integrated over the surface of the tip. We are interested in the magnetic field in a plane somewhat below the tip apex and perpendicular to the tip axis. In the simplest case of the magnetization pointing along the tip axis, the magnetic field has only radial and axial components. The typical result of a simulation is shown in Supplementary Information. It justifies the Lorentzian field distribution model used to find the two-dimensional position of the nitrogen-vacancy centre.

LETTERS

Thresholds for Cenozoic bipolar glaciation

Robert M. DeConto[1], David Pollard[2], Paul A. Wilson[3], Heiko Pälike[3], Caroline H. Lear[4] & Mark Pagani[5]

The long-standing view of Earth's Cenozoic glacial history calls for the first continental-scale glaciation of Antarctica in the earliest Oligocene epoch (~33.6 million years ago[1]), followed by the onset of northern-hemispheric glacial cycles in the late Pliocene epoch, about 31 million years later[2]. The pivotal early Oligocene event is characterized by a rapid shift of 1.5 parts per thousand in deep-sea benthic oxygen-isotope values[3] (Oi-1) within a few hundred thousand years[4], reflecting a combination of terrestrial ice growth and deep-sea cooling. The apparent absence of contemporaneous cooling in deep-sea Mg/Ca records[4–6], however, has been argued to reflect the growth of more ice than can be accommodated on Antarctica; this, combined with new evidence of continental cooling[7] and ice-rafted debris[8,9] in the Northern Hemisphere during this period, raises the possibility that Oi-1 represents a precursory bipolar glaciation. Here we test this hypothesis using an isotope-capable global climate/ice-sheet model that accommodates both the long-term decline of Cenozoic atmospheric CO_2 levels[10,11] and the effects of orbital forcing[12]. We show that the CO_2 threshold below which glaciation occurs in the Northern Hemisphere (~280 p.p.m.v.) is much lower than that for Antarctica (~750 p.p.m.v.). Therefore, the growth of ice sheets in the Northern Hemisphere immediately following Antarctic glaciation would have required rapid CO_2 drawdown within the Oi-1 timeframe, to levels lower than those estimated by geochemical proxies[10,11] and carbon-cycle models[13,14]. Instead of bipolar glaciation, we find that Oi-1 is best explained by Antarctic glaciation alone, combined with deep-sea cooling of up to 4 °C and Antarctic ice that is less isotopically depleted (−30 to −35‰) than previously suggested[15,16]. Proxy CO_2 estimates remain above our model's northern-hemispheric glaciation threshold of ~280 p.p.m.v. until ~25 Myr ago, but have been near or below that level ever since[10,11]. This implies that episodic northern-hemispheric ice sheets have been possible some 20 million years earlier than currently assumed (although still much later than Oi-1) and could explain some of the variability in Miocene sea-level records[17,18].

Evidence for the onset of Antarctic glaciation comes from a combination of marine geochemical and sea-level records[3,19,20], and more direct records of ice-rafted debris and glaciomarine sediments from around the Antarctic margin[1]. Proposed mechanisms include the opening of Southern Ocean gateways[21], mountain uplift and orbital forcing[22]; however, recent modelling studies implicate low atmospheric CO_2 as the most important factor[22]. The growth of ice sheets in the Northern Hemisphere is thought to have begun much later than in Antarctica, beginning on southern Greenland in the late Miocene or early Pliocene[23] and culminating with the onset of major glacial cycles around 2.7–3.0 Myr ago[2]. Atmospheric CO_2 is also considered a critical factor for Northern Hemispheric glaciation, perhaps with additional influence from ocean gateways and mountain uplift[24]. The recent discovery of much older Eocene, Oligocene and Miocene ice-rafted debris in the Greenland Sea[8,9] and Arctic Ocean[25] has called this long-standing view of Earth history into question, although the amount of Northern Hemispheric ice responsible for these sediments remains controversial.

Our current understanding of cryospheric evolution comes largely from marine oxygen isotope[26] and Mg/Ca records[5,6] from foraminiferal calcite, and stratigraphic sea-level records from passive continental margins[19,20]. Benthic isotope records reflect the combined effects of ice volume and ocean temperature, and in principle Mg/Ca ratios provide an independent record of temperature. These data can be combined to deconvolve the ice-volume component of the isotope records, which show a number of sudden and large (>1‰) shifts and excursions throughout the Cenozoic (for example, the Oi and Mi events)[3]. These shifts are thought to represent Antarctic glaciation events because direct evidence for significant northern-hemispheric ice is lacking before ~3.0 Myr ago. The stepwise shift in benthic $\delta^{18}O$ in the earliest Oligocene (Oi-1) is the largest in the Cenozoic. As shown in a highly resolved record from Ocean Drilling Program site 1218 in the eastern tropical Pacific[4], the event began around 34 Myr ago, during a minimum in the 1.2-Myr obliquity cycle producing a long interval of low seasonality (Fig. 1). This observation is in good agreement with time-continuous simulations from a combined general circulation model (GCM)/ice-sheet model accounting for gradual Cenozoic CO_2 decline and idealized orbital forcing[22]. In these simulations an Antarctic ice sheet grows suddenly once CO_2 reaches a critical threshold value around 2.8 to 2.6 times the 'pre-industrial' atmospheric level (PAL, taken to be 280 p.p.m.v., attaining a volume (21×10^6 km^3) comparable to the modern East Antarctic ice sheet. The simulated ice sheet expands in two rapid jumps in response to height/mass-balance and albedo feedbacks initiated when snowlines intersect high plateaux during orbital periods producing cold austral summers[22]. The simulation bears a striking resemblance to the form of the site 1218 record[4], which shows a two-step shift in oxygen isotopes separated by ~200 kyr, with each step occurring within one 41-kyr obliquity cycle (Fig. 1).

The problem lies in the magnitude of the isotope shift, which, if taken as the total change in benthic $\delta^{18}O$ from the latest Eocene to the peak of Oi-1, is ~1.5‰. Assuming that the isotopic composition of Palaeogene Antarctic ice was similar to today (−45 to −57‰ for West and East Antarctica, respectively[27]), and ignoring changes in deep-sea temperature, the implied increase in ice volume is $\sim 40 \times 10^6$ km^3 or ~100 m of equivalent sea level. This is 135% of modern Antarctic ice volume and nearly twice the volume of the simulated Oi-1 ice sheet[22]. If the isotopic composition of precipitation falling on a warmer and thinner Oligocene ice sheet were less depleted than today as suggested previously[4,22], the missing ice-volume problem would be greatly exacerbated.

To test the hypothesis that ancient ice on a warmer Antarctica was isotopically less depleted than today, we ran a set of oxygen-isotope simulations spanning the Eocene/Oligocene climate transition, using an isotope tracer-capable version[28] of the same GCM used in our prior

[1]Department of Geosciences, University of Massachusetts, Amherst, Massachusetts 01003, USA. [2]Earth and Environmental Systems Institute, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [3]National Oceanography Centre, University of Southampton, Southampton SO14 3ZH, UK. [4]School of Earth and Ocean Sciences, Cardiff University, Cardiff CF10 3YE, UK. [5]Department of Geology and Geophysics, Yale University, New Haven, Connecticut 06520, USA.

simulations of Oi-1 (ref. 22; see Methods). Our results indicate that the first ice to accumulate in the earliest Oligocene would have had an isotopic composition of −20‰ to −25‰ (SMOW), becoming more negative as the ice sheet grew, but not exceeding −42‰, even on the highest ice elevations (Fig. 2). The average isotopic composition of precipitation in the accumulation zone of the fully developed earliest Oligocene ice sheet is −30 to −35‰. In the absence of deep-sea cooling, Oi-1 would require the growth of $\sim 65 \times 10^6$ km^3 of ice (160 m equivalent sea level), far in excess of the holding capacity of a warmer Antarctic continent (see Methods and Supplementary Information).

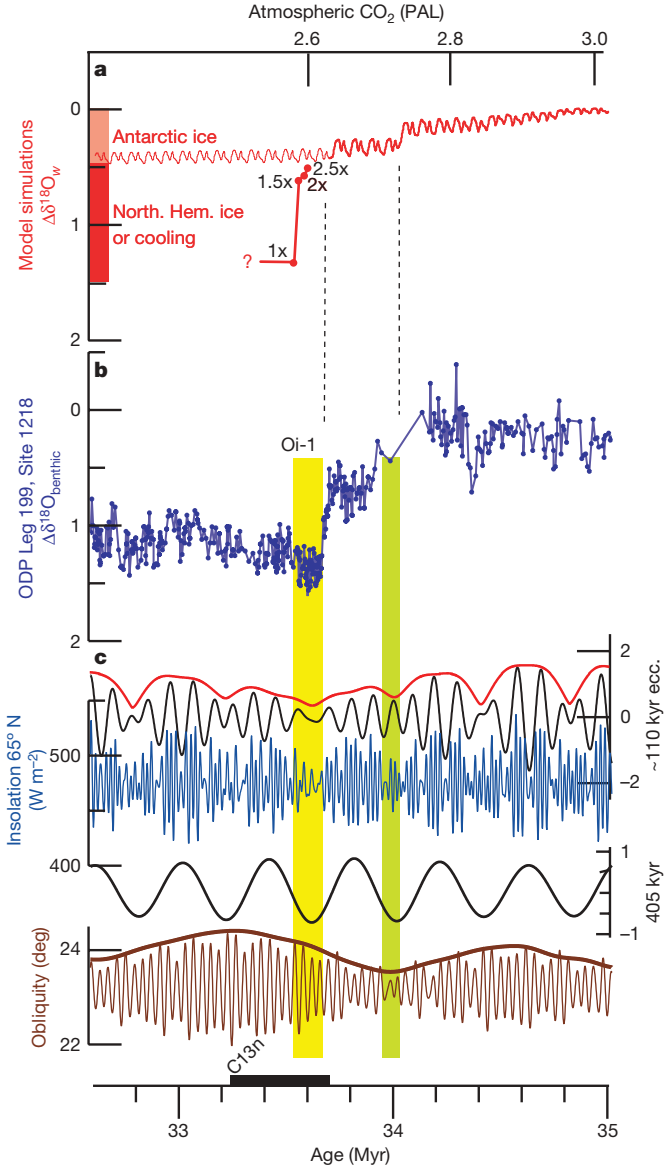

Figure 1 | Changes in the isotopic composition of the ocean across the Eocene/Oligocene transition. Isotopic, orbital and model time series are shown on the same astronomically tuned timescale[4], with the simulated and observed stepwise timing of glaciation aligned (dashed lines) for comparison. **a**, The simulated change in mean ocean $\delta^{18}O_w$ ($\delta^{18}O = [(^{18}O/^{16}O)_{sample}/(^{18}O/^{16}O)_{standard}] - 1$, where standard is SMOW) from coupled GCM/ice-sheet simulations assuming the isotopic composition of ice is −35‰. Red bars at left show the relative contributions from Antarctic ice (light red) and from deep-sea cooling and/or Northern Hemisphere ice (dark red). The thin red line shows a prior simulation[22] ignoring Northern Hemisphere ice and assuming a continuous decline in CO_2. Carbon dioxide levels (top *x* axis) are the average of two simulations assuming open and closed Southern Ocean gateways[22]. A second scheme accounting for Northern Hemisphere ice sheets and rapidly decreasing CO_2 beginning after Antarctic glaciation is shown by the thick red line, with red dots corresponding to the added effect of the ice shown in Fig. 3. **b**, High-resolution benthic $\delta^{18}O$ data from Ocean Drilling Program site 1218 (ref. 4). **c**, Orbital parameters[12] include filtered and normalized eccentricity values for 110-kyr and 405-kyr periodicities (black lines) and the long-term envelope for the 110 kyr (red line). The initial step in Antarctic ice growth corresponds to low obliquity variance (green bar) and the second main step occurs during an interval of reduced summer insolation at high latitudes (yellow bar) as assumed in our Northern Hemisphere simulations.

To test the possibility that Oi-1 includes a contribution from northern-hemispheric ice sheets, we ran a series of 30-kyr simulations using a GCM/ice-sheet model[18,22], developed specifically for simulating the time-continuous evolution of climate and ice sheets over long timescales. The simulations used an earliest Oligocene palaeogeography, with an Antarctic ice sheet already in place[22]. Orbital values were initialized with a favourable but reasonable configuration for northern-hemispheric glaciation, with some experiments updated every 10 kyr to account for subsequent orbital variations (see Methods). The initial cold boreal summer orbit (well within the range that occur on Cenozoic timescales[12]) was chosen to constrain the highest level of CO_2 that would allow Northern Hemisphere glaciation, given the position and topography of Oligocene continents. Atmospheric CO_2 was lowered by 0.5 × PAL in each successive simulation beginning with a starting value of 2.5 × PAL, just below the model's Antarctic CO_2-glaciation threshold. To substantiate our results in light of poorly constrained geographical boundary conditions, the entire sequence was repeated with most Northern Hemisphere continental elevations reduced by 50% (see Supplementary Information).

Simulated ice sheets, total ice volumes and associated changes in the mean isotopic composition of the ocean are shown in Figs 1 and 3, and in Supplementary Information. At relatively high levels of CO_2 near the Antarctic glaciation threshold, small ice caps form on the highest elevations of western North America, northeast Asia, East Greenland and other locations where prevailing storm tracks intersect coastlines with steep relief. Despite the initial prescription of a favourable cool summer orbit, the onset of major glaciation does not occur until CO_2 reaches pre-industrial levels (280 p.p.m.v.), with the largest continental ice sheet forming on Greenland because of its broad plateau and moist maritime climate. In simulations using lower topography (see Supplementary Information), major glaciation is delayed until CO_2 drops below 180 p.p.m.v. Assuming that simulated northern-hemispheric ice sheets had an average isotopic composition similar to modern Greenland ice (−35‰)[27], they would have enriched mean ocean $\delta^{18}O_w$ by nearly 0.8‰. This, in addition to Antarctica's model-derived contribution of 0.5‰ (Fig. 1), could produce most of the observed shift at Oi-1 without invoking deep-sea cooling; but this would only be possible if atmospheric CO_2 fell to pre-industrial levels or below (Fig. 4).

In a situation of decreasing Cenozoic CO_2, our model first produces small isolated ice caps in the Antarctic interior during cold austral summer orbits at CO_2 levels as high as 6 × PAL (ref. 22), but major Antarctic glaciation does not occur until CO_2 reaches ~2.7 × PAL. In the Northern Hemisphere, small ice caps appear on eastern Greenland and the highest elevations of the surrounding continents over a broad range of CO_2 (Fig. 3), but major glaciation only occurs when CO_2 falls near or below 1 × PAL. The lower CO_2 threshold for the large Northern Hemisphere continents is due to their greater seasonality, lower latitudes and consequently warmer summers. Most proxy-based estimates for late Eocene to middle Oligocene CO_2 range between 500 and 1,200 p.p.m.v. (ref. 10), well above our simulated Northern Hemisphere glaciation threshold (Fig. 4). Our results are consistent with the recent discovery of Eocene and Oligocene ice-rafted debris in the Greenland Sea[8] (Fig. 3a, b), but at these relatively high levels of CO_2 they support a picture of small isolated ice caps and alpine outlet glaciers as the source rather than continental-scale ice sheets as recently suggested[9].

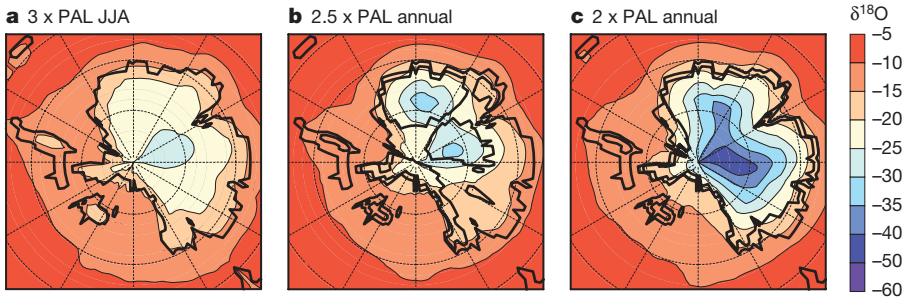

Figure 2 | Simulated isotopic composition of snowfall on a glaciating Antarctic continent. **a**, Austral winter (June, July, August) isotopic averages (‰) on an ice-free Antarctica at 3 × PAL, just above the glaciation threshold. The austral winter average is shown for the 3 × PAL case because most summer precipitation in the ice-free case is rain. **b**, Annual mean isotopic composition of precipitation at 2.5 × PAL and with an intermediate ice sheet (black outlines in continental interior). **c**, Same as **b** except with 2 × PAL and a fully developed early Oligocene East Antarctic ice sheet. Intermediate and fully glaciated ice-sheet geometries in **b** and **c** (extent and surface elevations) are taken from simulations in ref. 22.

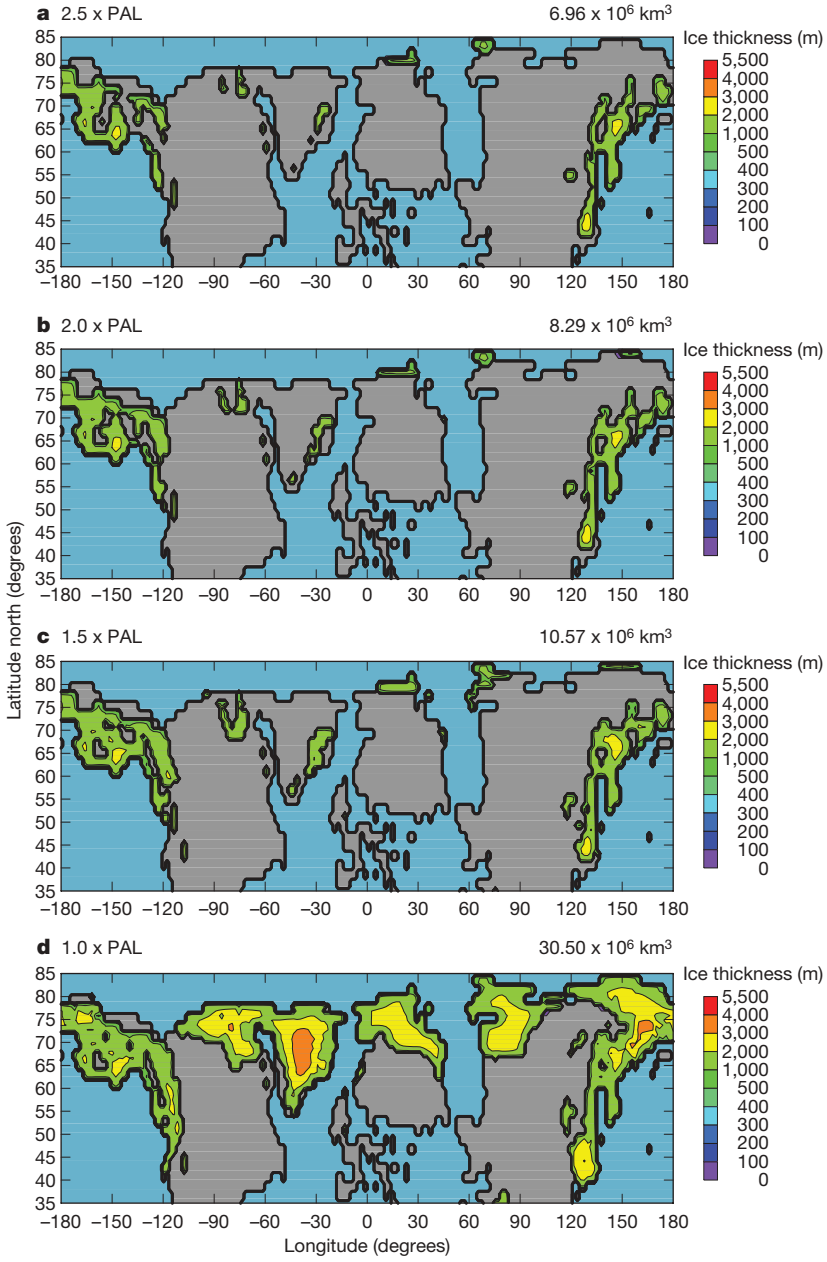

Figure 3 | Simulations of Northern Hemisphere ice sheets for progressively lower values of CO_2. Simulations are for earliest Oligocene palaeogeography and favourable cold boreal summer orbit, and are made with a coupled GCM/ice-sheet model. Ice thicknesses are shown in metres, with corresponding total Northern Hemisphere ice volumes (10^6 km^3) above the right corner of each panel. Palaeogeographical boundary conditions and an alternative set of simulations with continental elevations reduced by 50% are shown in the Supplementary Information.

For major bipolar glaciation to have occurred at Oi-1, CO_2 would first have to cross the Antarctic glaciation threshold (~750 p.p.m.v.) and then fall more than 400 p.p.m.v. within ~200 kyr to reach the Northern Hemisphere threshold (Fig. 4). Increased sea ice and upwelling in the Southern Ocean[13,29] and falling sea level[14] could have acted as feedbacks accelerating CO_2 drawdown at the time of Oi-1. This is supported by CO_2 proxy records and carbon-cycle model results showing a drop in CO_2 across the Eocene/Oligocene transition[10,13,14], but none of these reconstructions reach the low levels required for Northern Hemisphere glaciation. We therefore conclude that major bipolar glaciation at the Eocene/Oligocene transition is unlikely, and Mg/Ca-based estimates of deep-sea temperatures across the boundary[5] are unreliable. Our findings lend support to the hypothesis that the 1-km deepening of the carbonate compensation depth and the associated carbonate ion effect on deep-water calcite mask a cooling signal in the Mg/Ca records[4,5]. Therefore, the observed isotope shift at Oi-1 is best explained by Antarctic glaciation[22] accompanied by 4.0 °C of cooling in the deep sea or slightly less (~3.3 °C) if there was additional ice growth on West Antarctica (see Methods and Supplementary Information). This explanation is in better agreement with sequence stratigraphic estimates of sea-level fall at Oi-1 (70 ± 20 m)[19,20] equivalent to 70–120% of modern Antarctic ice volume, and coupled GCM/ice-sheet simulations showing 2–5 °C cooling and expanding sea ice in the Southern Ocean in response to Antarctic glaciation[29]. Additional support for ocean cooling is provided by new records from Tanzania[16] and the Gulf of Mexico[15], where Mg/Ca temperature estimates show ~2.5 °C cooling in shallow, continental shelf settings during the first step of the Eocene/Oligocene transition.

In summary, our model results show that the Northern Hemisphere contained glaciers and small, isolated ice caps in high elevations through much of the Cenozoic, especially during favourable orbital periods (Fig. 3a–c). However, major continental-scale Northern Hemisphere glaciation at or before the Oi-1 event (33.6 Myr) is unlikely, in keeping with recently published high-resolution Eocene isotope records[30]. Proxy reconstructions of Cenozoic carbon dioxide[10,11] remain well above our model's threshold for Northern Hemisphere glaciation until around the Oligocene/Miocene boundary. Since that time, transient Northern Hemisphere ice sheets could have grown during favourable orbital periods and may help to account for the magnitude of Neogene isotope and sea-level variability[17] despite pronounced hysteresis in Antarctic ice-sheet dynamics[18]. The first major event to be considered in this context is Mi-1 (~23.1 Myr ago)[3], an ephemeral 200-kyr isotopic excursion similar in magnitude to Oi-1 and coeval with a prolonged interval of low obliquity variance[26] favourable for ice-sheet development. Although no definitive evidence of widespread northern-hemispheric glaciation exists before ~2.7 Myr ago, pre-Pliocene records from subsequently glaciated high northern latitudes are generally lacking. More highly resolved CO_2 records focusing on specific events, along with additional geological information from high northern latitudes, will help to unravel the Cenozoic evolution of the cryosphere. According to these results, this evolution may have included an episodic northern-hemispheric ice component for the past 23 million years.

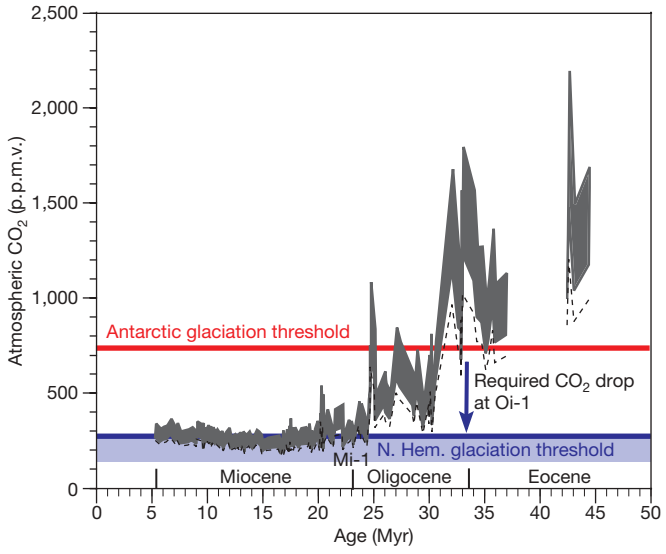

Figure 4 | Model-generated CO_2 thresholds for Antarctic and Northern Hemisphere glaciation superposed on a Cenozoic record of atmospheric CO_2. The CO_2 record is taken from stable carbon isotopic values of di-unsaturated alkenones[10]. The dashed line represents a lowermost limit, assuming $\delta^{18}O$-derived temperatures used in the calculation of CO_2 partial pressure are accurate, and the lower and upper bounds of the shaded (grey) area assume temperatures are 3 °C and 6 °C warmer, respectively. The blue arrow shows the drop in CO_2 required for Northern Hemisphere glaciation at Oi-1. The blue shading shows the range of uncertainty based on alternate Northern Hemisphere simulations with lower continental elevations (see text and Supplementary Information).

METHODS SUMMARY

The GCM and thermomechanical ice-sheet models are interactively coupled, whereby net annual surface mass balance on the ice sheet is calculated from monthly mean GCM meteorological fields of temperature and precipitation horizontally interpolated to the higher-resolution ice-sheet grid. Simulations in Fig. 3 were run to equilibrium (30 kyr) using a cold boreal summer orbit with high eccentricity (0.05), low obliquity (22.5°) and precession placing aphelion in July. The simulations producing large ice sheets (Fig. 3d, h) were repeated in asynchronous coupled mode[22] accounting for climate–ice feedbacks and time-continuous orbital forcing to confirm that the fixed-orbit results in Fig. 3 are representative of those with orbital variations.

A modified version of the ice-sheet model accounting for floating ice shelves and migrating grounding lines was used to determine the potential for additional ice growth over West Antarctica at Oi-1. In this model version, the buttressing effect of an expanding proto-Ross ice shelf assists the growth of some additional ice, but only if ocean temperatures (and sub-ice melt rates) are assumed to be similar to modern (see Supplementary Fig. 1).

Ice volumes simulated by the ice-sheet model are converted to eustatic sea level according to the global ocean-area fraction in our 34 Myr palaeogeography (0.731). Equivalent $\Delta\delta^{18}O_w$ (change in the average isotopic composition of the ocean) reflects either assumed isotopic ice compositions mentioned in the text, or those derived from the simulated isotopic composition of precipitation falling on the ice sheets using the stable isotope physics described previously[28]. In these calculations, the isotopic composition of the ocean was given a uniform global value of –1.2‰, consistent with ice-free conditions at the beginning of the experiment. With Antarctic ice at –35‰, $\Delta\delta^{18}O_w$ is 0.0246 per 10^6 km^3 of grounded ice.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 3 April; accepted 12 August 2008.

1. Barrett, P. J. Antarctic paleoenvironment through Cenozoic times: A review. *Terra Antarct.* **3**, 103–119 (1996).
2. Shackleton, N. J. *et al.* Oxygen isotope calibration of the onset of ice-rafting and history of glaciation in the North Atlantic region. *Nature* **307**, 620–623 (1984).
3. Miller, K. G., Fairbanks, R. G. & Mountain, G. S. Tertiary oxygen isotope synthesis, sea level history, and continental margin erosion. *Paleoceanography* **1**, 1–20 (1987).
4. Coxall, H. K., Wilson, P. A., Pälicke, H., Lear, C. & Backman, J. Rapid stepwise onset of Antarctic glaciation and deeper calcite compensation in the Pacific Ocean. *Nature* **433**, 53–57 (2005).
5. Lear, C. H., Rosenthal, Y., Coxall, H. K. & Wilson, P. A. Late Eocene to early Miocene ice-sheet dynamics and the global carbon cycle. *Paleoceanography* **19**, PA4015, doi: 10.1029/2004PA001039 (2004).
6. Billups, K. & Schrag, D. P. Application of benthic foraminiferal Mg/Ca ratios to questions of Cenozoic climate change. *Earth Planet. Sci. Lett.* **209**, 181–195 (2003).
7. Zanazzi, A., Kohn, M. J., MacFadden, B. J. & Terry, D. O. Jr. Large temperature drop across the Eocene–Oligocene transition in central North America. *Nature* **445**, 639–642 (2007).
8. Eldrett, J. S., Harding, I. C., Wilson, P. A., Butler, E. & Roberts, A. P. Continental ice in Greenland during the Eocene and Oligocene. *Nature* **466**, 176–179 (2007).
9. Tripati, A. *et al.* Evidence for Northern Hemisphere glaciation back to 44 Ma from ice-rafted debris in the Greenland Sea. *Earth Planet. Sci. Lett.* **265**, 112–122 (2008).

10. Pagani, M., Zachos, J. C., Freeman, K. H., Tipple, B. & Bohaty, S. M. Marked decline in atmospheric carbon dioxide concentrations during the Paleogene. *Science* **309,** 600–603 (2005).
11. Pearson, P. N. & Palmer, M. R. Atmospheric carbon dioxide over the past 60 million years. *Nature* **406,** 695–699 (2000).
12. Laskar, J. *et al.* A long-term numerical solution for the insolation quantities of the Earth. *Astron. Astrophys.* **428,** 261–285 (2004).
13. Zachos, J. & Kump, L. Carbon cycle feedbacks and the initiation of Antarctic glaciation in the earliest Oligocene. *Global Planet. Change* **47,** 51–66 (2005).
14. Merico, A., Tyrrell, T. & Wilson, P. A. Eocene/Oligocene ocean de-acidification linked to Antarctic glaciation by sea level fall. *Nature* **452,** 979–982 (2008).
15. Katz, M. E. *et al.* Stepwise transition from the Eocene greenhouse to the Oligocene icehouse. *Nature Geosci.* **1,** 329–334 (2008).
16. Lear, C., Bailey, T. R., Pearson, P. N., Coxhall, H. K. & Rosenthal, Y. Cooling and ice growth across the Eocene–Oligocene transition. *Geology* **36,** 251–354, doi:10.1130/G1124 (2008).
17. Pekar, S. & DeConto, R. M. High-resolution ice-volume estimates for the early Miocene: Evidence for a dynamic ice sheet in Antarctica. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **231,** 101–109 (2006).
18. Pollard, D. & DeConto, R. M. Hysteresis in Cenozoic Antarctic ice sheet variations. *Global Planet. Change* **45,** 9–21 (2005).
19. Pekar, S. F. & Christie-Blick, N. Resolving apparent conflicts between oceanographic and Antarctic climate records and evidence for a decrease in pCO_2 during the Oligocene through early Miocene (34–16 Ma). *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **260,** 41–49 (2008).
20. Kominz, M. A. & Pekar, S. F. Oligocene eustasy from two-dimensional sequence stratigraphic backstripping. *Geol. Soc. Am. Bull.* **113,** 291–304 (2001).
21. Kennett, J. P. Cenozoic evolution of Antarctic glaciation, the circum-Antarctic oceans and their impact on global paleoceanography. *J. Geophys. Res.* **82,** 3843–3859 (1977).
22. DeConto, R. M. & Pollard, D. Rapid Cenozoic glaciation of Antarctica induced by declining atmospheric CO_2. *Nature* **421,** 245–249 (2003).
23. Larsen, H. C. *et al.* Seven million years of glaciation in Greenland. *Science* **264,** 952–955 (1994).
24. Raymo, M. E. & Ruddiman, W. F. Tectonic forcing of late Cenozoic climate. *Nature* **359,** 117–122 (1992).
25. St John, K. Cenozoic ice-rafting history of the central Arctic Ocean: terrigenous sands on the Lomonosov Ridge. *Paleoceanography* **23,** PA1S05 (2008).
26. Zachos, J., Pagani, M., Sloan, L. & Thomas, E. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292,** 686–693 (2001).
27. Lhomme, N., Clarke, G. K. C. & Ritz, C. Global budget of water isotopes inferred from polar ice sheets. *Geophys. Res. Lett.* **32,** L20502, doi:10.1029/2005GL023774 (2005).
28. Mathieu, R. *et al.* Simulation of stable water isotope variations by the GENESIS GCM for modern conditions. *J. Geophys. Res.* **107,** doi:10.1029/2001JD900255 (2002).
29. DeConto, R. M., Pollard, D. & Harwood, D. Sea ice feedback and Cenozoic evolution of Antarctic climate and ice sheets. *Paleoceanography* **22,** PA3214, doi:10.1029/2006PA001350 (2007).
30. Edgar, K. M., Wilson, P. A., Sexton, P. F. & Suganuma, Y. No extreme bipolar glaciation during the Eocene calcite compensation shift. *Nature* **488,** 908–911 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This material is based on work supported by the National Science Foundation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.M.D. (deconto@geo.umass.edu).

METHODS

Global climate/ice-sheet model. The GCM and ice-sheet components of our model are the same as those used in prior simulations of Antarctic glaciation[18,22], allowing interhemispheric comparisons of glaciation potential at different levels of atmospheric CO_2. The horizontal resolution of the atmospheric component of the GCM is T31 ($\sim$3.75° by $\sim$3.75°) with 18 vertical layers. Surface models including a 50-m slab ocean, dynamic–thermodynamic sea ice, and multi-layer models of snow, soil and vegetation are on a finer $2^\circ \times 2^\circ$ grid. The GCM is coupled to a thermomechanical ice-sheet model. The ice model grid over Antarctica is polar stereographic with a resolution of 40 km by 40 km, and 1° latitude by 1° longitude over the Northern Hemisphere continents. Ice-sheet evolution is driven by surface mass balance forcing from the GCM. Terrestrial ice flow is modelled using the shallow ice approximation, while accounting for internal ice temperatures, basal sliding, bedrock isostatic relaxation and lithospheric flexure.

Monthly mean meteorological fields (temperature and precipitation) used in the calculation of net-annual surface mass balance are horizontally interpolated from the GCM to the higher-resolution ice-sheet grids, using a lapse-rate adjustment to account for local topographic offsets between model components. A positive degree–day parameterization with an imposed diurnal cycle is used to calculate ablation while accounting for refreezing of meltwater. Mass balance is re-calculated every 200 model years to account for evolving surface elevations.

All simulations in Fig. 3 were run to equilibrium using a cold boreal summer orbit with high eccentricity (0.05) and low obliquity (22.5°), with the longitude of precession placing aphelion in July. Simulations producing large ice sheets were repeated in asynchronous coupled mode, with the GCM rerun every 10,000 ice-model years to account for changing albedo, topography and evolving orbital parameters (see Supplementary Information Fig. 4). This was done to confirm that summer warming during unfavourable phases of the precession cycle was insufficient to stop the onset of glaciation initiated during a cold boreal summer orbit, and that the results shown in Fig. 3 are representative of those with orbital variations. Although the CO_2-glaciation thresholds shown here are model-dependent, the GCM's sensitivity to $2 \times$ PAL (2.5 °C) is average among models used in future and palaeoclimate modelling studies, and the use of a different atmospheric component with similar CO_2 sensitivity is expected to produce similar results.

The Oi-1 Antarctic glaciation experiment[22] was repeated using a modified version of the ice-sheet model with extensions accounting for floating ice shelves and migrating grounding lines[31] to determine the potential for additional ice growth over West Antarctica. In this model, a combined set of scaled equations for sheet and shelf flow accounts for both horizontal shear ($\partial u/\partial z$) and stretching ($\partial u/\partial x$) dominant in grounded and floating ice, respectively. In this simulation, the buttressing effect of an expanding proto-Ross ice shelf aids the growth of additional West Antarctic ice. If ocean temperatures and sub-ice melt rates are assumed to be similar to modern, our model builds only an additional $7.5 \times 10^6\,km^3$ of ice on West Antarctica (Supplementary Information Fig. 1), not enough to ameliorate the Oi-1 amplitude problem significantly. Further expansion of grounding lines to the continental shelf as occurred at the Last Glacial Maximum[32] is unlikely, owing to warmer conditions in the early Oligocene. Equilibrium ice-free bedrock elevations for the model are obtained from modern observed bathymetry[33], isostatically rebounded with modern ice removed. If West Antarctic bedrock elevations were higher in the early Oligocene[34] with more land area above sea level and shallower continental shelves, this could have allowed somewhat greater amounts of early West Antarctic ice than in Supplementary Information Fig. 1, which in turn would require less ocean cooling to explain the magnitude of Oi-1.

Water isotopes. The water isotope tracer model[28] passively tracks the hydrological cycle in the GCM atmosphere and applies relevant fractionation physics during phase transitions. The model considers $^1H_2{}^{18}O$ and $^1HD^{16}O$ and accounts for evaporative, condensational and post-condensational processes. Reservoir effects are differentiated over ocean, land (vegetation) and ice sheets. In the case of our Palaeogene Antarctic glaciation experiments, the isotopic composition of the ocean was given a uniform global value of –1.2‰. In modern control simulations[28], the seasonal and spatial distribution of $\delta^{18}O$ of precipitation generated by the model is close to observed values, except in the highest elevations of the Antarctic interior, where, as in most GCMs, model surface temperatures are warmer than observed and the $\delta^{18}O$ of precipitation is 5–10‰ too heavy. This bias is not relevant for the initially ice-free conditions and smaller ice sheets in the Palaeogene simulations, because of their lower topography and warmer temperatures than modern.

31. Pollard, D. & DeConto, R. M. in *Glacial Sedimentary Processes and Products* (eds Hambrey, M. *et al.*), 37–52 (Internat. Assoc. Sedimentologists Spec. Publ. 39, Blackwell, 2007).
32. Huybrechts, P. Sea-level changes at the LGM from ice-dynamic reconstructions of the Greenland and Antarctic ice sheets during the glacial cycles. *Quat. Sci. Rev.* **21**, 203–231 (2002).
33. Bamber, J. A. & Bindschadler, R. A. An improved elevation dataset for climate and ice-sheet modelling: validation with satellite imagery. *Ann. Glaciol.* **25**, 439–444 (1997).
34. Sorlien, C. C. *et al.* Oligocene development of the West Antarctic ice sheet recorded in eastern Ross Sea strata. *Geology* **35**, 467–470 (2007).

LETTERS

Crystallographic preferred orientation of akimotoite and seismic anisotropy of Tonga slab

Rei Shiraishi[1], Eiji Ohtani[1], Kyuichi Kanagawa[3], Akira Shimojuku[1,4] & Dapeng Zhao[2]

The mineral akimotoite, ilmenite-structured $MgSiO_3$, exists at the bottom of the Earth's mantle transition zone and within the uppermost lower mantle, especially under low-temperature conditions[1]. Akimotoite is thought to be a major constituent of the harzburgite layer of subducting slabs, and the most anisotropic mineral in the mantle transition zone[2–4]. It has been predicted that if akimotoite crystals are preferentially oriented by plastic deformation, a cold subducted slab would be extremely anisotropic[5]. However, there have been no studies of crystallographic preferred orientations and very few reports of plastic deformation experiments for $MgSiO_3$ ilmenite. Here we present plastic deformation experiments on polycrystalline akimotoite, which were conducted at confining pressures of 20–22 GPa and temperatures of 1,000–1,300 °C. We found a change in crystallographic preferred orientation pattern of akimotoite with temperature, where the *c*-axis maximum parallel to the compression direction develops at high temperature, whereas the *c* axes are preferentially oriented parallel to the shear direction or perpendicular to the compression direction at lower temperature. The previously reported difference in compressional-wave seismic anisotropy between the northern and southern segments of the Tonga slab at depths of the mantle transition zone[6] can conceivably be attributed to the difference in the crystallographic preferred orientation pattern of akimotoite at varying temperature within the slab.

A polycrystalline akimotoite used for the present deformation experiments was synthesized at high pressure and temperature with a Kawai-type multi-anvil apparatus at Tohoku University. The synthesized polycrystalline akimotoite was confirmed to have no crystallographic preferred orientation (CPO).

The akimotoite specimen was then deformed by either uniaxial compression or simple shear geometry at high pressures and temperatures with the Kawai-type multi-anvil apparatus. The experimental conditions and results are given in Table 1.

We performed two types of experiment (types I and II; Table 1). In both type I and type II experiments, pressure was increased at room temperature, and temperature was then increased at the desired pressure. In type I experiments the sample was annealed for either 10 or 60 min, and then quenched. In type II experiments the sample was further deformed by slightly increasing the pressure, and then quenched. Two blank experiments were also conducted: a cold-compression experiment (DI01) and a non-annealing experiment (DI08). In run DI01, pressure was increased and then decompressed immediately. In run DI08, temperature was increased quickly (over about 10 min) to 1,200 °C after increasing the pressure, and the sample was subsequently quenched without annealing.

Recovered samples were cut in half parallel to the compression direction. Thin sections were then prepared and were polished with a colloidal silica suspension. A thin (~10 nm) coating of carbon was applied to decrease specimen charging. In each sample, the crystallographic orientations of 166–271 akimotoite grains were determined by the electron backscatter diffraction (EBSD) technique[7] (Table 1).

The change in sample thickness revealed the total compressional strains of samples deformed by uniaxial compression. The shear strain of specimen DI07 was calculated from the axial displacement of the pistons. The total compressional strains were 0.1–0.3. DI07 was deformed to a shear strain of 0.6. The microstructures and grain sizes (~10 μm) of samples, except specimen DI07, were similar. The grain size of DI07 was about 5 μm, slightly smaller than the other samples.

Equal-area lower-hemisphere projections of akimotoite $<11\bar{2}0>$, $<10\bar{1}0>$ and [0001] directions in the samples, deformed at 1,000–1,300 °C, are given in Fig. 1, where the compression and shear directions are shown by black arrows. Deformed polycrystalline $MgSiO_3$ akimotoite at 1,200–1,300 °C (Fig. 1a–c) has a CPO characterized by a strong *c*-axis maximum subparallel to the compression direction, and $<11\bar{2}0>$ and $<10\bar{1}0>$ axis girdles normal to the compression direction. This CPO suggests a dominant slip system with glide on (0001), which is in good agreement with the observation[8] that the dominant slip system in experimentally deformed akimotoite is 1/3 $<11\bar{2}0>$ (0001). In addition, the basal glide on (0001) was reported in some analogue of akimotoite, such as the trigonal structure of ilmenite[9]. In specimens deformed at 1,000 °C (Fig. 1d, e), the *c* axes

Table 1 | Summary of experimental conditions and results

Run no.	Pressure (GPa)	*P–T* path*	Assembly	Temperature (°C)	Heating duration (min)	Compressional strain	Shear strain	Grain size (μm)	CPO
DI01	21.1	–	Compression	–	–	0.16	–	7.4	Random
DI02	21.1	Type I	Compression	1,200	60	0.09	–	8.0	(0001) $\perp\sigma_1$
DI03	20.0→20.5	Type II	Compression	1,300	140	0.09	–	7.7	(0001) $\perp\sigma_1$
DI04	21.1	Type I	Compression	1,200	10	0.14	–	10.0	(0001) $\perp\sigma_1$
DI06	22.2→22.5	Type II	Compression	1,000	160	0.29	–	8.2	[0001] $\perp\sigma_1$
DI07	22.2	Type I	Shear	1,000	60	–	0.61	5.8	[0001] ||SD
DI08	21.1	–	Compression	1,200	0	0.14	–	8.0	Random

* We examined two types of pressure – temperature (*P–T*) path. In type I experiments, samples were annealed under the target conditions and then quenched. In type II experiments, the sample was further deformed by increasing the pressure slightly. $\perp\sigma_1$, perpendicular to the compression direction; \|\|SD, parallel to the shear direction.

[1]Institute of Mineralogy, Petrology, and Economic Geology. [2]Department of Geophysics, Tohoku University, Sendai 980-8578, Japan. [3]Department of Earth Sciences, Chiba University, Chiba 263-8522, Japan. [4]Department of Earth and Planetary Sciences, Faculty of Sciences, Kyushu University, Fukuoka 812-8581, Japan.

are preferentially orientated parallel to the shear direction or perpendicular to the compression direction. The CPO of the sample deformed in uniaxial compression is not axially symmetric, suggesting a deviation of its deformation from uniaxial geometry. The texture of the sample deformed in simple shear (DI07) is axially symmetric about the shear direction. Both $<11\bar{2}0>$ and $<10\bar{1}0>$ axes spread along a girdle around the *c*-axis maximum. This CPO therefore suggests the dominance of slip in the [0001] direction on multiple planes. In contrast, no clear CPO was developed in the sample deformed at room temperature (DI01) or in the non-annealing sample (DI08). From these blank tests, the CPOs observed in samples DI02–DI07 are considered to have developed during the plastic deformation at the target pressures and temperatures.

In addition, there is no difference in CPO pattern resulting from the difference between the *P*–*T* paths in the type I and type II experiments, judging from the fact that the CPO pattern of DI02 is the same as that of DI03.

Thus, observed akimotoite CPOs and inferred dominant slip systems differ in their deformation temperatures. Slip in the [0001] direction is probably dominant at a lower temperature (1,000 °C), whereas the basal glide on (0001) becomes dominant at higher temperatures (1,200–1,300 °C). The fabric transition occurs at about 1,100 °C. Fabric transitions with increasing temperature are also reported in other minerals such as quartz[10–12] and olivine[13–15]. For wet quartz it has been reported[16] that there is a possible transition from $<11\bar{2}0>$ (0001) to [0001]{$11\bar{2}0$} with increasing temperature. This observation is similar to that of akimotoite observed in this study.

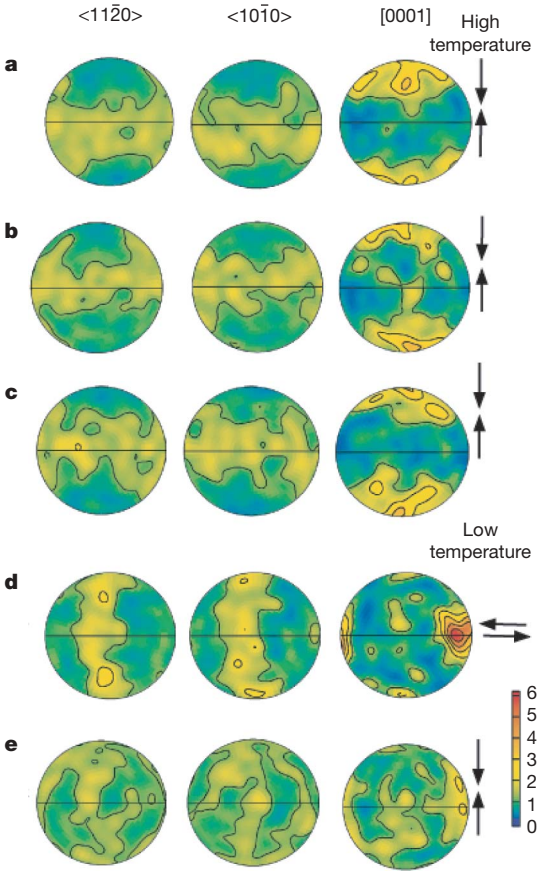

Figure 1 | Equal-area projections of pole figures for $<11\bar{2}0>$, $<10\bar{1}0>$ and [0001] directions of akimotoite in all samples. **a**, DI02 ($T = 1{,}200$ °C; $n = 271$); **b**, DI03 ($T = 1{,}300$ °C; $n = 197$); **c**, DI04 ($T = 1{,}200$ °C; $n = 220$); **d**, DI07 ($T = 1{,}000$ °C; $n = 166$); **e**, DI06 ($T = 1{,}000$ °C; $n = 216$); *n* is the number of grains measured. The projections are coloured according to the density of data points and are contoured at multiples of uniform distribution as shown in the scale at the bottom right. The north–south direction corresponds to the compression direction in **a**–**c** and **e**, and to the shear-plane normal direction in **d**. Pairs of bold arrows represent the compression direction or the shear direction.

We calculated seismic wave velocities from the akimotoite CPO data to examine the relationship between akimotoite CPO and seismic anisotropy. We used the elastic constants and density of akimotoite under the mantle transition zone conditions determined previously[4] with the molecular dynamic approach. The Voigt–Reuss–Hill average was used to calculate the seismic anisotropy. We used the program Anis2k (ref. 17) to calculate bulk elastic constants C_{ij} from the CPO data, as well as P-wave velocities. For the CPO data of DI03 and DI06 deformed in uniaxial compression, we randomly rotated the orientation data about the compression axis five times and used all rotated data for the following calculation, to decrease the deviations of those CPO data from uniaxial symmetry. Single-crystal akimotoite has a V_P anisotropy of 14.4%, which is shown in Fig. 2a. The results calculated from the CPO data for three representative samples (DI03, DI06 and DI07) are shown in Fig. 2b–d.

For the sample deformed at a higher temperature (DI03), the V_P anisotropy is 3.0%. In the sample deformed at a lower temperature, the V_P anisotropy of the compression experiment (DI06) and the simple shear experiment (DI07) are 1.0% and 4.3%, respectively. As regards other mantle transition-zone minerals, it has been reported[18] that the V_P anisotropies of 60% wadsleyite and 40% garnet deformed to shear strains of 1.0 and 0.5 are 2% and 1%, respectively. Although there are no CPO data for the other mantle transition-zone minerals, the anisotropy of a rock composed of 100% akimotoite is at least fourfold to fivefold that of a rock composed of 60% wadsleyite and 40% garnet. Akimotoite therefore has a much greater effect on the seismic anisotropy of subducting slabs at transition-zone depths.

Because of the difference in CPO pattern between the sample deformed at 1,300 °C (DI03) and that deformed at 1,000 °C (DI06 and DI07), the anisotropy pattern also depends on temperature. The P wave propagates most slowly in the shear direction or in the direction perpendicular to the compression direction at 1,000 °C, but in the compression direction at 1,300 °C. This is because the velocity of the P wave through an akimotoite single crystal is slowest in the *c*-axis direction (Fig. 2a).

The spatial variation of seismic anisotropy in the Tonga subducting slab was shown recently[6]. The slab is divided into two segments: the northern segment at latitudes 17–19° S and the southern segment at latitudes 19.5–27° S (Fig. 3). The magnitude of the anisotropy is 5–7% for P waves and 9–12% for S waves, and the direction of maximum velocity is different in each of the two slab segments. In the northern segment, P waves propagate more slowly in the slab normal direction. In contrast, P waves propagate more slowly in

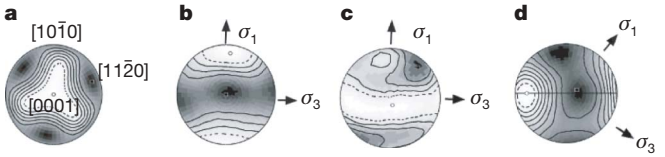

Figure 2 | P-wave anisotropies calculated using Anis2k. V_P contours are shown; black squares, maximum velocities (V_{max}); white circles, minimum velocities (V_{min}). **a**, Akimotoite single crystal. $V_{max} = 12.44$ km s^{-1}; $V_{min} = 10.77$ km s^{-1}. **b**, Akimotoite aggregate experimentally deformed at a relatively high temperature (1,300 °C) by uniaxial compression (DI03). $V_{max} = 11.65$ km s^{-1}; $V_{min} = 11.31$ km s^{-1}. **c**, **d**, Experimental deformation at a relatively low temperature (1,000 °C) by uniaxial compression (**c**; DI06; $V_{max} = 11.59$ km s^{-1}; $V_{min} = 11.47$ km s^{-1}) and by simple shear (**d**; DI07; $V_{max} = 11.67$ km s^{-1}; $V_{min} = 11.18$ km s^{-1}). The north–south direction corresponds to the compression direction in **b** and **c**, and to the shear-plane normal direction in **d**. The horizontal line and the east–west direction in **d** correspond to the shear plane and the shear direction, respectively. The σ_1 direction for the deformed samples is the direction of the advancing pistons.

the slab sinking direction in the southern segment. As regards S waves, the directions of maximum and minimum velocities do not coincide with the slab normal and sinking directions. We compared P-wave anisotropy in experimentally deformed akimotoite aggregates with that observed in the Tonga subducting slab as mentioned above (Figs 2 and 3). We assumed that the maximum compressive direction (σ_1) and the minimum compressive direction (σ_3) in experimentally deformed samples corresponded to the orientation of the principal stress axes estimated from focal mechanisms of deep earthquakes in the Tonga slab[6]. In the southern Tonga slab segment, the P-wave velocity is slower and faster in the σ_1 and σ_3 directions, respectively (Fig. 3c), which is similar to that in the akimotoite aggregate deformed at 1,300 °C (Fig. 2b). In contrast, in the northern Tonga slab segment, the P-wave velocity is faster in the σ_2 direction and slowest in the direction of the bisector between the σ_1 and σ_3 directions (Fig. 3b), which correlates well with that in the akimotoite aggregate deformed at 1,000 °C (Fig. 2d). Thus, the difference in seismic anisotropy between the northern and southern Tonga slab segments is attributable to the difference in CPO patterns of akimotoite resulting from differences in temperature. However, the transition temperature in the Tonga slab between the southern and northern segments could not be determined quantitatively because CPO patterns are also dependent on strain rate, and the geological strain rates are much smaller than the experimental strain rates. The geometry of the Tonga slab is complicated, specifically at greater depths[19,20] (Fig. 3). In addition, it has been reported that there is a difference in the distribution of the low-velocity zones in the mantle wedge between the northern and southern parts of the Tonga back-arc[21]. The low-velocity zone in the deep mantle wedge above the southern part of the Tonga slab may be caused by partial melting or by the existence of fluids from dehydration of the slab. These observations suggest that the temperature of the southern part of the Tonga slab is higher than that of the northern part. There are probably lateral variations in temperature in the Tonga slab that give rise to the difference in the seismic anisotropy pattern in the Tonga slab. It has been suggested that there are strong lateral variations in V_P and V_S, and that these variations are caused by a petrological anomaly, such as compositional or mineralogical variation[22]. In addition to compositional variation, the change in CPO pattern with temperature may have contributed to the differences between the northern and southern segments of the Tonga slab. Because ringwoodite and majorite, which may also be major constituents in the lower part of the mantle transition zone, are almost isotropic, the CPO of akimotoite must control the seismic anisotropy of the slab in the transition zone.

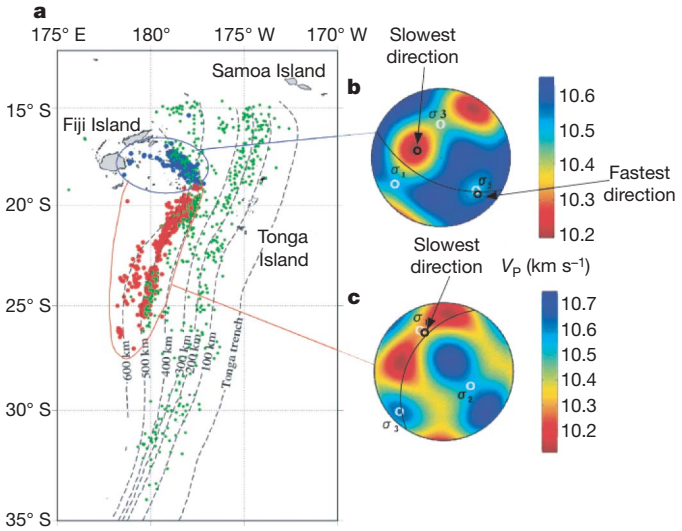

Figure 3 | P-wave anisotropy of the Tonga slab and the deformed akimotoite. a, Map showing the geometry of the subducting Tonga slab and epicentres of earthquakes. **b**, **c**, P-wave anisotropies of the northern (**b**) and southern (**c**) Tonga slab segments[6]. Green dots in **a** represent earthquake foci at depths between 100 and 500 km; blue and red dots show earthquakes at depths greater than 500 km in the northern and southern Tonga segments. The dashed lines are equal-depth contours of the Tonga slab[26]. The V_P contour diagrams (**b**, **c**) are equal-area lower-hemisphere projections, in which the white circles show the directions of the stress axes (σ_1, σ_2 and σ_3) in the Tonga slab deduced from the focal mechanism solutions, and the black arc lines show the intersection of the Tonga slab with the hemisphere at depths greater than 500 km. Black circles show the fastest and the slowest V_P directions expected from the deformed akimotoite.

METHODS SUMMARY

Experimental procedure. The furnace assembly was composed of a sintered ZrO_2 pressure medium (an octahedron with an edge length of 10 mm), Ta electrodes and a $LaCrO_3$ heater. Temperature was measured with W3%Re–W25%Re thermocouples. The starting material, which was placed in a platinum capsule, was an $MgSiO_3$ glass fabricated from oxides. Synthesis experiments were conducted at 20–22 GPa and 1,250–1,550 °C for 60 min.

We measured the water content of the starting material by Fourier-transform infrared spectroscopy with a JASCO MFT-2000 instrument. The water content was determined by integrating the infrared absorption spectrum from 3,200 to 3,750 cm^{-1} using a previous calibration of the extinction coefficient[23]. The water content in akimotoite is 24 p.p.m. by weight, which is extremely low compared with hydrous akimotoite[24].

The sample was sandwiched between hard alumina pistons inserted in the furnace assembly to produce differential stresses during compression. To minimize the deformation during cold compression, crushable alumina was placed at the outer ends of the pistons. Crushable alumina is initially very porous and soft. However, it becomes dense and works as good piston material on compression. These ideas and the cell assemblies were based on ref. 25.

EBSD measurement. EBSD patterns were obtained using a Nordlys II EBSD detector mounted on a Jeol JSM-6460 scanning electron microscope at Chiba University, operating with an accelerating voltage of 20 kV and a beam current of 1.5–2.4 nA, and indexed manually with Channel 5 software (Flamenco) from HKL Technology.

Received 23 October 2007; accepted 1 August 2008.

1. Akaogi, M., Tanaka, A. & Ito, E. Garnet–ilmenite–perovskite transitions in the system $Mg_4Si_4O_{12}$–$Mg_3Al_2Si_3O_{12}$ at high-pressures and high-temperatures: phase equilibria, calorimetry and implications for mantle structure. *Phys. Earth Planet. Inter.* **132,** 303–324 (2002).
2. Weidner, D. J. & Ito, E. Elasticity of $MgSiO_3$ in the ilmenite phase. *Phys. Earth Planet. Inter.* **40,** 65–70 (1985).
3. Da Silva, C. R. S., Karki, B. B., Stixrude, L. & Wentzcovitch, R. M. *Ab initio* study of the elastic behavior of $MgSiO_3$ ilmenite at high-pressure. *Geophys. Res. Lett.* **26,** 943–946 (1999).
4. Zhang, Y., Zhao, D. & Matsui, M. Anisotropy of akimotoite: A molecular dynamics study. *Phys. Earth Planet. Inter.* **151,** 309–319 (2005).
5. Anderson, D. L. *Theory of the Earth* (Blackwell, 1989).
6. Vavryčuk, V. Spatially dependent seismic anisotropy in the Tonga subduction zone: A possible contributor to the complexity of deep earthquakes. *Phys. Earth Planet. Inter.* **155,** 63–72 (2006).
7. Randle, V. *Microtexture Determination and its Applications* 2nd edn (Maney, 2003).
8. Cordier, P. in *Plastic Deformation of Minerals and Rocks* (eds Karato, S. I. & Wenk, H. R.) 137–179 (American Mineralogical Society, 2002).
9. Bascou, J., Raposo, M. I. B., Vauchez, A. & Egydio-Silva, M. Titanohematite lattice-preferred orientation and magnetic anisotropy in high-temperature mylonites. *Earth Planet. Sci. Lett.* **198,** 77–92 (2002).
10. Lister, G. S. Fabric transitions in plastically deformed quartzites: competition between basal, prism and rhomb systems. *Bull. Mineral.* **102,** 232–241 (1979).
11. Schmid, S. M. & Casey, M. Complete fabric analysis of some commonly observed quartz *c*-axis patterns. *Am. Geophys. Un. Geophys. Monogr.* **36,** 263–286 (1986).
12. Mainprice, D., Bouchez, J.-L., Blumenfeld, P. & Tubia, J. M. Dominant *c* slip in naturally deformed quartz; implications for dramatic plastic softening at high temperature. *Geology* **14,** 2181–2202 (1986).
13. Katayama, I. & Karato, S. Effect of temperature on the B- to C-type olivine fabric transition and implication for flow pattern in subduction zones. *Phys. Earth Planet. Inter.* **157,** 33–45 (2006).
14. Carter, N. L. & Ave'Lallemant, H. G. High temperature flow of dunite and peridotite. *Geol. Soc. Am. Bull.* **81,** 2181–2202 (1970).
15. Jung, H. & Karato, S.-I. Water-induced fabric transitions in olivine. *Science* **293,** 1460–1463 (2001).
16. Blacic, J. D. Plastic deformation mechanisms in quartz: The effect of water. *Tectonophysics* **27,** 271–294 (1975).
17. Mainprice, D. A. FORTRAN program to calculate seismic anisotropy from the lattice preferred orientation of minerals. *Comput. Geosci.* **16,** 385–393 (1990).

18. Tommasi, A., Mainprice, D., Cordier, P., Thoraval, C. & Couvy, H. Strain-induced seismic anisotropy of wadsleyite polycrystals and flow patterns in the mantle transition zone. *J. Geophys. Res.* **109,** B12406, doi:10.1029/2005JB004168 (2004).
19. Chen, W.-P. & Brudzinski, M. R. Evidence for a large-scale remnant of subducted lithosphere beneath Fiji. *Science* **292,** 2475–2479 (2001).
20. Chen, W.-P. & Brudzinski, M. R. Seismic anisotropy in the mantle transition zone beneath Fiji–Tonga. *Geophys. Res. Lett.* **30,** 1682, doi:10.1029/2002GL016330 (2003).
21. Zhao, D. *et al.* Depth extent of the Lau back-arc spreading center and its relation to subduction processes. *Science* **278,** 254–257 (1997).
22. Brudzinski, M. R. & Chen, W.-P. A petrologic anomaly accompanying outboard earthquakes beneath Fiji–Tonga: Corresponding evidence from broadband P and S waveforms. *J. Geophys. Res.* **108,** B62299, doi:10.1029/2002JB002012 (2003).
23. Paterson, M. S. The determination of hydroxyl by infrared absorption in quartz, silicate glasses and similar materials. *Bull. Mineral.* **105,** 20–29 (1982).
24. Bolfan-Casanova, N., Keppler, H. & Rubie, D. C. Water partitioning between nominally anhydrous minerals in the $MgO–SiO_2–H_2O$ system up to 24 GPa: implications for the distribution of water in the Earth's mantle. *Earth Planet. Sci. Lett.* **182,** 209–221 (2000).
25. Karato, S. & Rubie, D. C. Toward an experimental study of deep mantle rheology: A new multianvil sample assembly for deformation studies under high pressures and temperatures. *J. Geophys. Res.* **102,** 20111–20122 (1997).
26. Gudmundsson, O. & Sambridge, M. A regionalized upper mantle (RUM) seismic model. *J. Geophys. Res.* **103,** 7121–7136 (1998).

Acknowledgements This work was supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Science, Sport, and Technology of the Japanese Government.

Author Contributions R.S. performed experiments and took the lead in writing the manuscript. E.O. and A.S. designed the study. K.K. performed EBSD analyses. D.Z. worked on the seismological aspects of this study. All co-authors took part in the discussion and interpretation of the results and improving the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.S. (siraisir@ganko.tohoku.ac.jp).

LETTERS

Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960

Michael Worobey[1], Marlea Gemmel[1], Dirk E. Teuwen[2,3], Tamara Haselkorn[1], Kevin Kunstman[4], Michael Bunce[5], Jean-Jacques Muyembe[6,7], Jean-Marie M. Kabongo[6], Raphaël M. Kalengayi[6], Eric Van Marck[8], M. Thomas P. Gilbert[1]† & Steven M. Wolinsky[4]

Human immunodeficiency virus type 1 (HIV-1) sequences that pre-date the recognition of AIDS are critical to defining the time of origin and the timescale of virus evolution[1,2]. A viral sequence from 1959 (ZR59) is the oldest known HIV-1 infection[1]. Other historically documented sequences, important calibration points to convert evolutionary distance into time, are lacking, however; ZR59 is the only one sampled before 1976. Here we report the amplification and characterization of viral sequences from a Bouin's-fixed paraffin-embedded lymph node biopsy specimen obtained in 1960 from an adult female in Léopoldville, Belgian Congo (now Kinshasa, Democratic Republic of the Congo (DRC)), and we use them to conduct the first comparative evolutionary genetic study of early pre-AIDS epidemic HIV-1 group M viruses. Phylogenetic analyses position this viral sequence (DRC60) closest to the ancestral node of subtype A (excluding A2). Relaxed molecular clock analyses incorporating DRC60 and ZR59 date the most recent common ancestor of the M group to near the beginning of the twentieth century. The sizeable genetic distance between DRC60 and ZR59 directly demonstrates that diversification of HIV-1 in west-central Africa occurred long before the recognized AIDS pandemic. The recovery of viral gene sequences from decades-old paraffin-embedded tissues opens the door to a detailed palaeovirological investigation of the evolutionary history of HIV-1 that is not accessible by other methods.

We screened 27 tissue blocks (8 lymph node, 9 liver and 10 placenta) obtained from Kinshasa between 1958 and 1960 by polymerase chain reaction with reverse transcription (RT–PCR); one lymph node biopsy specimen contained HIV-1 RNA. Viral nucleic acids were extracted from this specimen using protocols optimized for the recovery of nucleic acids from ancient or degraded samples[3,4]. After reverse transcription, 12 out of the 14 short HIV-1 complementary DNA fragments in the study (Fig. 1a) were amplified by PCR using a panel of conserved primer pairs from different regions of the viral genome (Supplementary Table 1). Each PCR product was cloned and sequenced. Sequences were reproducible after repeated extractions and were not the result of PCR contamination (see Fig. 1a and Supplementary Table 1 for fragment designations). The results were confirmed independently in two laboratories (Fig. 1b and Supplementary Fig. 1), with the second laboratory successfully identifying the positive 1960 specimen in a blinded assay. The short fragments of the 1960 sample were found to be of subtype A and not to be a mosaic of contemporary sequences (see Supplementary Information for a detailed discussion of the authenticity of the 1960

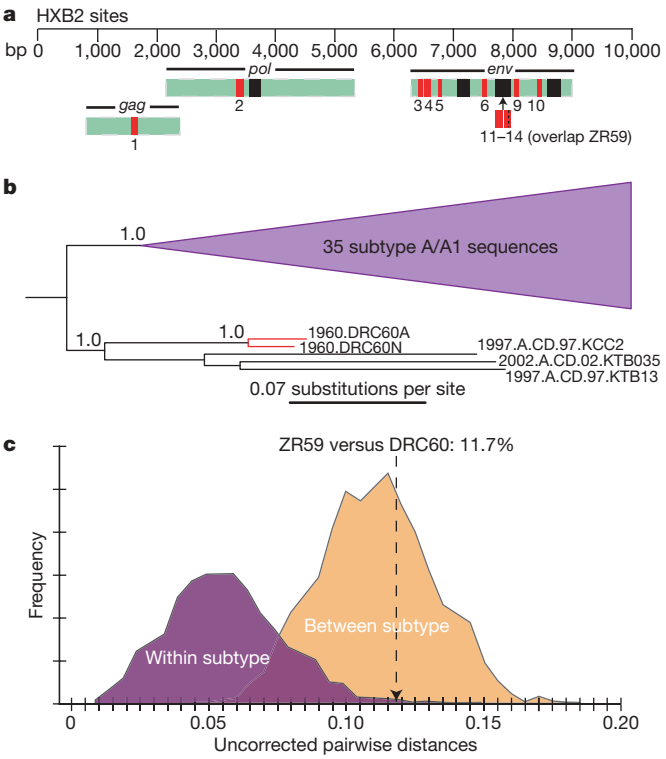

Figure 1 | Fragments amplified from DRC60, and the results of the phylogenetic and sequence analyses. a, The HIV-1 genome fragments that were successfully amplified from DRC60 (red) and are available for ZR59 (black). The numbering for the HIV-1 sequences corresponds to the HXB2 reference sequence (Supplementary Table 1). **b**, The A/A1 subtree from the unconstrained (in which a molecular clock is not enforced) BMCMC phylogenetic analysis. Supplementary Fig. 1 depicts the complete phylogenetic tree (50% majority rule consensus tree of the posterior sample, with branch lengths averaged across the sample). Posterior probabilities are shown on nodes with support >0.95. 1960.DRC60A is the University of Arizona consensus sequence, and 1960.DRC60N is the Northwestern University consensus sequence (that is, the sequences independently recovered in each of the two laboratories). The DRC60 sequences form a strongly supported clade with three modern sequences also sampled in the DRC. **c**, Smoothed histograms of within-subtype (A2, A/A1, B, C, D, F1, F2, H, J, K) and between-subtype distances.

[1]Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA. [2]Sanofi Pasteur, F-69367 Lyon Cedex 07, France. [3]UCB SA Pharma, Braine l'Alleud, BE-1420, Belgium. [4]The Feinberg School of Medicine, Northwestern University, Chicago, Illinois 60611, USA. [5]Ancient DNA Laboratory, School of Biological Sciences and Biotechnology, Murdoch University, Perth, Western Australia 6150, Australia. [6]Department of Anatomy and Pathology, University of Kinshasa, Kinshasa B.P. 864, Democratic Republic of the Congo. [7]National Institute for Biomedical Research, National Laboratory of Public Health, Kinshasa B.P. 1197, Democratic Republic of the Congo. [8]Department of Pathology, University Hospital, University of Antwerp, Antwerp B-2610, Belgium. †Present address: Centre for Ancient Genetics, Biological Institute, University of Copenhagen, Copenhagen DK-2100, Denmark.

sequences). Consensus nucleotide sequences from these short HIV-1 fragments were concatenated for study. The analyses included reference sequences from the Los Alamos National Laboratory HIV sequence database and sequences recovered as part of this study from three paraffin-embedded tissue specimens collected from AIDS patients in Belgium and Canada between 1981 and 1997.

HIV-1 sequences were analysed in MrBayes v3.1.2 (ref. 5) using an unconstrained (in which a molecular clock was not enforced) Bayesian Markov chain Monte Carlo (BMCMC) method. The phylogenetic analyses confirmed that the DRC60 consensus sequences from the two laboratories were derived from a single patient (uncorrected pairwise distance of 1.4%). The sequences were positioned close to the ancestral node of the subtype A lineage (excluding sub-subtype A2), forming a monophyletic clade with three modern sequences from the DRC (Fig. 1b and Supplementary Fig. 1). Assuming a similar rate of evolution along all branches on a tree, the divergence between two sequences reflects the time elapsed since their shared ancestor. As predicted, the DRC60 sequences had a shorter branch length to the A/A1 ancestral node than the contemporary subtype A viruses sampled from the same geographic region ($P = 1.0$).

We validated the time of origin of the 1960 sequence by comparisons of the predicted date to the documented date. With the DRC60 date treated as an unknown, we calculated an evolutionary rate on the basis of the distribution of branch lengths on the unconstrained phylogenetic trees sampled by MrBayes. To limit the effects of evolutionary rate differences between clades and uncertainties in rooting the HIV-1 M group phylogeny, we focused on the subtype A/A1 subtree (Supplementary Fig. 1) and analysed root-to-tip branch lengths relative to the sampling year. The mean estimates for the year of origin of the DRC60 consensus sequences from the University of Arizona and Northwestern University laboratories were 1959 (95% highest probability distribution (HPD) 1902–1984) and 1959 (95% HPD 1915–1985), respectively, corroborating the authenticity of the DRC60 sequences and the existence of a clock-like signal in our data set (see later). Despite initial indications that recombination might seriously confound phylogenetic dating estimates[6], subsequent work has suggested that recombination is not likely to systematically bias HIV-1 dates in one direction or the other, although it is expected to increase variance[7]. The close match between the predicted and the actual dates of both ZR59 (ref. 2) and DRC60 provides support for this view and gives an unambiguous indication that HIV-1 evolves in a fairly reliable clock-like fashion.

The uncorrected pairwise distance between DCR60 and ZR59 in their overlapping *env* region was 11.7% (Fig. 1c). This genetic distance is greater than 99.2% of within-subtype comparisons (within-subtype difference, range 0.01–0.15; between-subtype difference, range 0.05–0.18). Because each subtype represents several decades of independent evolution in the human population[2,8], the extensive divergence between DRC60 and ZR59 indicates that the HIV-1 M group founder virus began to diversify in the human population (and that HIV-1 probably entered Kinshasa) decades before 1960.

We applied a relaxed clock BMCMC coalescent framework as implemented in BEAST v1.4.7 (ref. 9) to estimate the time to the most recent common ancestor (TMRCA) of the HIV-1 M group. This approach robustly incorporates phylogenetic uncertainty and accounts for the possibility of variable substitution rates among lineages and differences in the demographic history of the virus, sampling phylogenies and parameter estimates in proportion to their posterior probability[10]. As with other studies of HIV-1 (ref. 11), comparisons of the marginal likelihoods of strict versus relaxed clock models (both of which are implemented in BEAST) indicated overwhelming support for relaxed clocks (data available on request). Hence, the use of strict clock models with these data would be inappropriate and would probably yield misleadingly small error estimates with regard to both timing and substitution rates.

Using substitution rates calibrated with sequences sampled at different time points, we obtained a posterior distribution of rooted tree topologies with branch lengths in unit time (Fig. 2 and Supplementary Fig. 2). The median estimated substitution rate for the concatenated subregions of the *gag-pol-env* genes was 2.47×10^{-3} substitutions per site per year (95% HPD $1.90–2.95 \times 10^{-3}$). The inclusion of the 1959 and 1960 sequences seemed to improve estimation of the TMRCA of the M group (Table 1), limiting the influence of the coalescent tree prior on the posterior TMRCA distributions compared with the data set that excluded these earliest cases of HIV-1. With DRC60 and ZR59 included, the different demographic/coalescent models gave highly consistent results, with tighter and more similar date ranges compared with the analyses that excluded them and 95% HPDs that extend no later than 1933. The best-fit model incorporated a constant population size demographic model (TMRCA 1921, 95% HPD 1908–1933). The model with a general, non-parametric prior (the Bayesian skyline plot tree prior)[12,13] that indicated a more complex (and biologically plausible) demographic history (Supplementary Fig. 3) had a statistically indistinguishable degree of support (TMRCA 1908, 95% HPD 1884–1924). Moreover, the population expansion demographic model[9], which was a slightly worse fit to the data compared with the constant population and Bayesian skyline plot models, could not be rejected given the Bayes factor comparison of models (Table 1). The inability to strongly reject the model with a constant population size prior is counterintuitive because it is clear that the HIV-1 population size has increased notably. We speculate that this finding might be due to the simplest model providing a good fit to a relatively short,

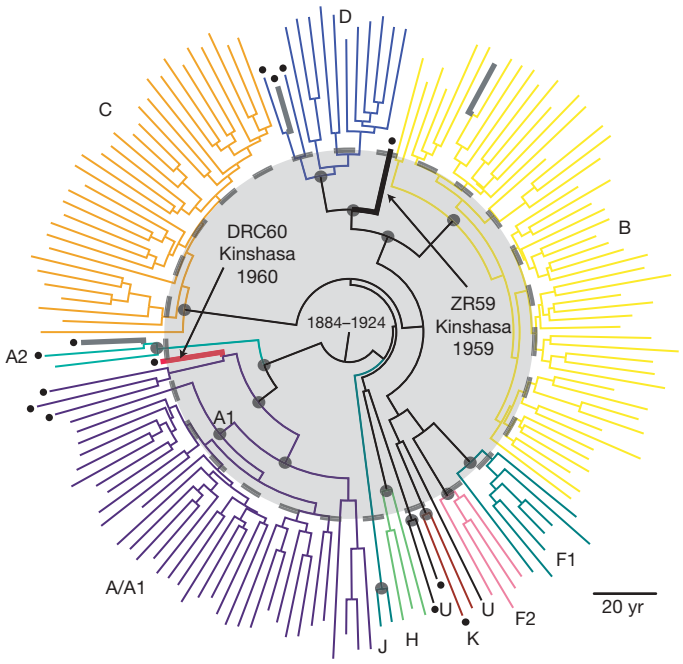

Figure 2 | Maximum clade credibility topology inferred using BEAST v1.4.7 under a Bayesian skyline plot tree prior. Branch lengths are depicted in unit time (years) and represent the median of those nodes that were present in at least 50% of the sampled trees. DRC60 (red), ZR59 (black) and the three control sequences from paraffin-embedded specimens from known AIDS patients (grey) are depicted in bold. The 95% HPD of the TMRCA is indicated at the root of the tree. Nodes (sub-subtype and deeper) with posterior probability of 1.0 are marked with grey circles. Unclassifiable strains are labelled 'U'. Sequences sampled in the DRC are highlighted with a bullet at the tip. DRC60 and the two control sequences from the DRC each form monophyletic clades with previously published sequences from the DRC, whereas the Canadian control sequence clusters, as expected, with subtype B sequences. The dashed circle and shaded area show the extensive HIV-1 diversity in Kinshasa in the 1950s. Supplementary Fig. 2 shows the tree in rectangular form with taxon labels.

information-poor alignment, in comparison with more parameterized models.

Acid-containing fixatives such as Bouin's solution can cause base modifications of nucleic acids, leading to the generation of erroneous bases in sequences derived from such samples[3]. However, the replication of all sequences from independent PCR amplifications and the uncorrected pairwise distance between the consensus sequences from the two laboratories (1.4%) suggest that few of the mutations on the DRC60 lineage are damaged-induced. Moreover, our relaxed clock methods are likely to be fairly robust to the presence of such mutations in one lineage[9]. Nevertheless, additional old sequence data would be helpful for resolving what impact, if any, this possible source of error had on the slightly earlier dates we calculated compared with previous estimates that did not include early calibration points[2,8,14,15]. Interestingly, the best-fit model for the data set that excluded ZR59 and DRC60 (Table 1) gave a TMRCA estimate of 1933 (1919–1945), which is very similar to that of ref. 2. This suggests that the inclusion of the old sequences, rather than the vagaries associated with a much shorter alignment than that analysed by ref. 2, might explain the discrepancy. Also, one earlier study, using sequences from the DRC only[16], produced dating and demography estimates very similar to ours. Overall, there is broad agreement between all of these studies in spite of differences in data and methods.

Our estimation of divergence times, with an evolutionary timescale spanning several decades, together with the extensive genetic distance between DRC60 and ZR59 indicate that these viruses evolved from a common ancestor circulating in the African population near the beginning of the twentieth century; TMRCA dates later than the 1930s are strongly rejected by our statistical analyses. The topology of the HIV-1 group M phylogeny provides further support for this conclusion. Unlike ZR59, which is basal to subtype D[1], DRC60 branches off from the ancestral node of subtype A/A1 (Fig. 2 and Supplementary Figs 1 and 2). Thus, it is clear that phylogenetically distinct subtypes (and/or their progenitors) were already present in the DRC by this early time point (Fig. 2). Notably, DRC60 and ZR59 cluster with other strains from the same geographical region and basal to other members of their respective subtypes, a pattern consistent with the hypothesis that the subtypes spread through lineage founder effects worldwide, whereas a more diverse array of forms remained at the site of origin in Africa[17,18].

The reservoir of the ancestral virus still exists among wild chimpanzee communities in the same area on the African continent[19]. Humans acquired a common ancestor of the HIV-1 M group by cross-species transmission under natural circumstances[20], probably predation[21]. The Bayesian skyline plot (Supplementary Fig. 2), which tracks effective population size through time, suggests that HIV-1 group M experienced an extensive period of relatively slow growth in the first half of the twentieth century. A similar pattern has been inferred using sequences sampled only in the DRC[16]. This pattern, and the short duration between the first presence of urban agglomerations in this area and the timing of the most recent common ancestor of HIV-1 group M (Fig. 3), suggests that the rise of cities may have facilitated the initial establishment and the early spread of HIV-1. Hence, the founding and growth of colonial administrative and trading centres such as Kinshasa[22] may have enabled the region to become the epicentre of the HIV/AIDS pandemic[23].

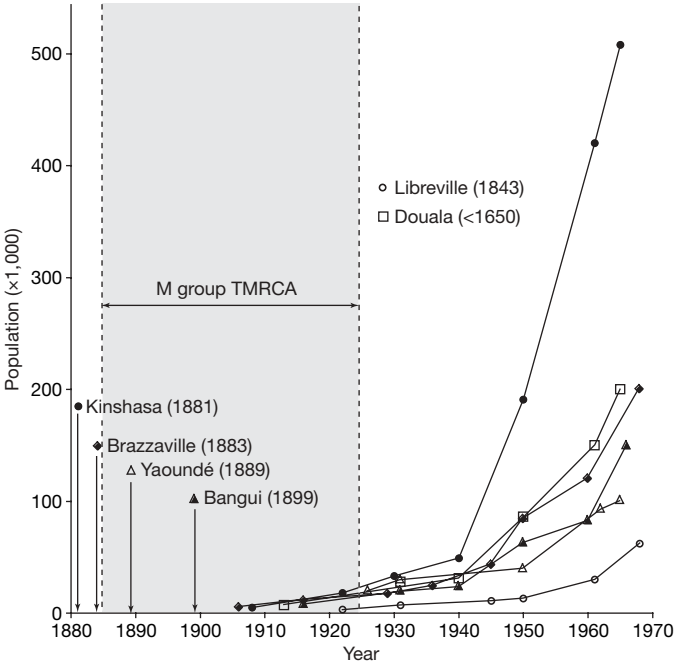

Figure 3 | The origin and growth of the major settlements near the epicentre of the HIV-1 group M epidemic. In the countries surrounding the putative zone of cross-species transmission[19] (current-day Cameroon, Central African Republic, DRC, Republic of Congo, Gabon and Equatorial Guinea) there was not a single site with a population exceeding 10,000 until after 1910. The founding date of each major city in the region is listed beside its name. Most were founded only shortly before the estimated TMRCA of group M. The demographic data are from ref. 23.

The archival banks of Bouin's-fixed paraffin-embedded tissue specimens accumulated by many hospitals in west-central Africa provide a vast source of clinical material for viral genetic analysis. As with the 1918 Spanish influenza pandemic virus[24,25], a deep perspective on the evolutionary history of HIV-1 using sequences resurrected from the earliest cases in Africa could yield important insights into the pathogenesis, virulence and evolution of pandemic AIDS viruses.

METHODS SUMMARY

A total of 813 Bouin's-fixed paraffin-embedded histopathological blocks were recovered from the 1958–1962 archives of the Department of Anatomy and Pathology at the University of Kinshasa. The boxes were stored until transfer to the University of Arizona, where 8 lymph node, 9 liver and 10 placenta samples from 1958–1960 were selected for RNA preservation analysis and HIV-1 RNA screening. We used a human β-2-microglobulin (*B2M*) quantitative RT–PCR assay to assess RNA quality as described[3]. Digestion and extraction of these samples, and of three modern positive-control samples, were performed using QIAamp DNA micro kits (Qiagen) using the protocol described in ref. 3. We used 14 primer sets designed to anneal to highly conserved regions of the *gag*, *pol* and *env* genes of HIV-1 group M and to amplify very short fragments likely to be present even in ancient and/or degraded specimens (Supplementary Table 1). Reverse transcription was performed using the SuperScript III System for RT–PCR (Invitrogen). The cDNA was amplified by PCR using Platinum Taq HiFi enzyme (Invitrogen) and cloned using the TOPO TA Cloning Kit (Invitrogen). We constructed an alignment including 156 published reference sequences plus

Table 1 | HIV-1 M group TMRCA estimates from BEAST analyses under different coalescent tree priors

Coalescent tree prior	DRC60 and ZR59 excluded*	DRC60 and ZR59 included
Constant	**1933 (1919–1945)**†, 0.0	**1921 (1908–1933)**†, 0.0
Exponential	1907 (1874–1932), −3.5 ± 0.8	1914 (1891–1930), −2.1 ± 1.5
Expansion	1882 (1834–1917), −2.7 ± 0.8	**1902 (1873–1922)**†, −1.6 ± 1.5
Logistic	1913 (1880–1937), −2.3 ± 0.8	1913 (1891–1930), −3.2 ± 1.5
Bayesian skyline plot	1882 (1831–1916), −2.7 ± 0.8	**1908 (1884–1924)**†, −0.4 ± 1.5

Shown for each coalescent tree prior is the median, with the 95% highest probability distribution of TMRCA in parentheses. Also shown is the $\log_{10}$ Bayes factor difference in estimated marginal likelihood (± estimated standard error) compared with the coalescent model with strongest support.
*Concatenated *gag-pol-env* fragments available for either or both of ZR59 and DRC60 (994 nucleotides total, 507 from DRC60).
†TMRCAs for the best-fit model and models not significantly worse than it are written in bold.

the sequences recovered in this study, concatenating the 12 (out of 14) fragments successfully amplified from the 1960 sample and the 4 fragments already available from the 1959 sample (994 bases total). We performed an unconstrained (not enforced by a molecular clock) BMCMC analysis in MrBayes v3.1.2 (ref. 5) and used the resulting MCMC sample to test whether the 1960 sequence exhibited properties consistent with its provenance (both age and geography). We used a relaxed molecular clock model, as implemented in BEAST v1.4.7 (ref. 9), to estimate the TMRCA of HIV-1 group M using the 1960 and 1959 samples and to investigate the demographic history of the virus. We also performed pairwise comparisons within and between subtypes for the 163 bases available for both DRC60 and ZR59.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 21 May; accepted 8 September 2008.

1. Zhu, T. F. *et al.* An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* **391,** 594–597 (1998).
2. Korber, B. *et al.* Timing the ancestor of the HIV-1 pandemic strains. *Science* **288,** 1789–1796 (2000).
3. Gilbert, M. T. P. *et al.* The isolation of nucleic acids from fixed, paraffin-embedded tissues — which methods are useful when? *PLoS ONE* **2,** e537 (2007).
4. Worobey, M. Phylogenetic evidence against evolutionary stasis and natural abiotic reservoirs of influenza A virus. *J. Virol.* **82,** 3769–3774 (2008).
5. Huelsenbeck, J. P. & Ronquist, F. MrBayes: Bayesian inference of phylogeny. *Bioinformatics* **17,** 754–755 (2001).
6. Worobey, M. A novel approach to detecting and measuring recombination: insights into evolution in viruses, bacteria, and mitochondria. *Mol. Biol. Evol.* **18,** 1425–1434 (2001).
7. Lemey, P. *et al.* The molecular population genetics of HIV-1 group O. *Genetics* **167,** 1059–1068 (2004).
8. Gilbert, M. T. P. *et al.* The emergence of HIV-1 in the Americas and beyond. *Proc. Natl Acad. Sci. USA* **104,** 18566–18570 (2007).
9. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7,** 214 (2007).
10. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4,** e88 (2006).
11. Salemi, M., de Oliveira, T., Ciccozzi, M., Rezza, G. & Goodenow, M. M. High-resolution molecular epidemiology and evolutionary history of HIV-1 subtypes in Albania. *PLoS ONE* **3,** e1390 (2008).
12. Suchard, M. A., Weiss, R. E. & Sinsheimer, J. S. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* **18,** 1001–1013 (2001).
13. Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22,** 1185–1192 (2005).
14. Sharp, P. M. *et al.* The origins of acquired immune deficiency syndrome viruses: where and when? *Phil. Trans. R. Soc. Lond. B* **356,** 867–876 (2001).
15. Salemi, M. *et al.* Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. *FASEB J.* **15,** 276–278 (2001).
16. Yusim, K. *et al.* Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution. *Phil. Trans. R. Soc. Lond. B* **356,** 855–866 (2001).
17. Vidal, N. *et al.* Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. *J. Virol.* **74,** 10498–10507 (2000).
18. Rambaut, A., Robertson, D. L., Pybus, O. G., Peeters, M. & Holmes, E. C. Human immunodeficiency virus phylogeny and the origin of HIV-1. *Nature* **410,** 1047–1048 (2001).
19. Keele, B. F. *et al.* Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* **313,** 523–526 (2006).
20. Worobey, M. in *Global HIV/AIDS Medicine* (eds Volberding, P. A., Sande, M. A., Lange, J. & Greene, W. C.) 13–21 (Saunders Elsevier, 2008).
21. Hahn, B. H., Shaw, G. M., De Cock, K. M. & Sharp, P. M. AIDS as a zoonosis: scientific and public health implications. *Science* **287,** 607–614 (2000).
22. Hance, W. A. *Population, Migration, and Urbanization in Africa* 209–297 (Columbia Univ. Press, 1970).
23. Chitnis, A., Rawls, D. & Moore, J. Origin of HIV type 1 in colonial French Equatorial Africa? *AIDS Res. Hum. Retrov.* **16,** 5–8 (2000).
24. Taubenberger, J. K. *et al.* Characterization of the 1918 influenza virus polymerase genes. *Nature* **437,** 889–893 (2005).
25. Tumpey, T. M. *et al.* Characterization of the reconstructed 1918 Spanish Influenza pandemic virus. *Science* **310,** 77–80 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Wertheim and M. Sanderson for computational assistance, and L. Jewel for providing the Canadian control specimen. The NIH/NIAID and the David and Lucile Packard Foundation funded the research.

Author Contributions M.W., D.E.T., S.M.W. and M.T.P.G. designed the study. M.G., T.H., K.K. and M.T.P.G. performed digestion and extraction, PCR, quantitative PCR, cloning and sequencing experiments. M.T.P.G., M.G. and M.B. optimized DNA/RNA isolation methods and designed PCR assays. D.E.T., J.-J.M., E.V.M., J.-M.M.K. and R.M.K. organized and provided samples. M.W. analysed the data, performed the phylogenetic analyses, and wrote the paper. S.M.W. contributed to the analyses and writing. All authors discussed the results and commented on the manuscript.

Author Information The sequences reported in this study have been deposited in GenBank under accession numbers EU580739–EU580854 and EU589211–EU589236. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.W. (worobey@email.arizona.edu).

METHODS

Archival samples. Each individual block carried an original paper identification number permanently embedded in the paraffin. Laboratory books listed the corresponding identification numbers sequentially, and included the patient's age, sex, department of hospitalization, tissue type and date of sampling. Block identification number, sampling date and tissue type were transcribed onto an Excel spreadsheet, and the blocks were indexed, transferred into plastic boxes and photographed.

The results of the quantitative RT–PCR assay indicated that the integrity of the RNA preserved in these 27 samples ranged from moderate to undetectable, a range typical of Bouin's-fixed specimens[3]. The human RNA found in the 1960 lymph node biopsy sample that was found to be HIV-1-RNA-positive was of relatively good quality. The C_t values (quantitative PCR data available from the authors on request) were as low or lower (better) than more recent (1980–1990) paraffin-embedded tissues that have yielded short HIV-1 RNA amplicons[3].

Three formalin-fixed paraffin-embedded necropsy specimens were obtained: a Canadian patient who died in 1997 (CAN97); a Congolese woman who died in Belgium in 1981 and who was retrospectively identified as an AIDS patient (BE81); and a Congolese man who died in Belgium in 1985 (BE85). The latter two cases were presumably infected in Zaire (now the Democratic Republic of the Congo). The phylogenetic reconstruction shows that their viral sequences are most closely related to modern sequences from the Democratic Republic of the Congo, whereas the Canadian specimen yielded a subtype B sequence, as predicted (Figs 1 and 2 and Supplementary Figs 1 and 2).

RNA isolation and reverse transcription. Between 5 and 10 microtome sections, 5–10 μm in thickness, or an approximately equivalent amount of tissue shaved from each block with a disposable scalpel blade, were used for each digestion and extraction, as described[3]. Rigorous attention was given to preventing cross-contamination between samples by cleaning the outer surface of each block with a bleach solution, using fresh microtome/scalpel blades for each sectioning of each block, discarding the first few (exposed-surface) sections, and by performing the work in a room physically isolated from any human or HIV-1 PCR-product DNA. A 48-h digestion period (24 h at 65 °C, 24 h at 75 °C) was used. Post extraction nucleic acids were eluted into 100 μl elution buffer AE and stored frozen at −80 °C until required for analyses.

Reverse transcription was performed simultaneously for the *gag*, *pol* and human *B2M* RNA fragments; *env* fragments 3–10; and *env* fragments 11–14 (Supplementary Table 1). This was performed with SuperScript III used according to the manufacturer's instructions. The protocol was as described[3] except that alternating 50 °C and 55 °C incubation periods of 30 min were used for a total of 6 h.

Amplification, cloning and DNA sequencing. The cDNA was PCR amplified in 25-μl reactions, using 0.1 μl Platinum Taq HiFi enzyme (Invitrogen), 250 μM DNTP mix, 2 mM $MgSO_4$, 1× PCR buffer, 0.4 μM per primer, and 2 μl cDNA for the *gag* and *pol* reactions or 1 μl for the *env* ones, with annealing temperatures of 60 °C (*gag*, 60 cycles) or 55 °C (*pol*, 50 cycles; *env*, 55 cycles). Full details are available from the authors on request.

After amplification, the PCR-product DNA was visualized by agarose gel electrophoresis and then purified using Zymoclean DNA Clean and Concentrator-25 spin tubes (Zymo Research Corporation). PCR-product DNA was inserted into vector pCR 2.1-TOPO using the TOPO TA Cloning Kit (Invitrogen). The University of Arizona Genomic Analysis and Technology Core Facility resolved the DNA sequence of the vector inserts on an Applied Biosystems 3730xl DNA analyser using ABI Big Dye 3.1 chemistry (Applied Biosystems). Nearly identical protocols were followed for the independent replication of the DRC60 results at Northwestern University.

Alignments. We downloaded the 2006 full-length HIV-1 sequence alignment from the Los Alamos National Laboratories HIV sequence database[26]. We retained only non-recombinant HIV-1 group M A–K subtype sequences (excluding G) and removed sequences suspected a priori of unusual evolutionary dynamics (such as those associated with the intravenous drug user epidemic in Eastern Europe and those with *nef* deletions, both of which exhibit abnormally slow evolutionary rates). We also reduced the size of the subtype B and C clades, which are heavily over-sampled relative to the others, by keeping only the first 5 sequences from any year/country pair and then randomly removing sequences until the sample size was similar to that of the other subtypes. This procedure left a total of 156 sequences. We then manually aligned the consensus sequence from the 12 regions amplified from DRC60, plus the 4 regions available for ZR59, to the full-length sequences. These short regions (Fig. 1a and Supplementary Table 1) were then concatenated into an alignment 994 nucleotides in length. The four *env* fragments from DRC60 that overlapped with available data from ZR59 were concatenated into an alignment 163 nucleotides in length. Matching alignments with DRC60 and ZR59 removed were also constructed. All the alignments are available from the authors on request.

MrBayes analyses. We used a general time-reversible nucleotide substitution model with gamma-distributed rate heterogeneity among sites and performed four independent runs of 20 million steps, sampling every 2,000 steps. Examination of the MCMC samples with Tracer v1.4 (ref. 9) indicated convergence and adequate mixing of the Markov chain with estimated sample sizes in the thousands. We discarded the first 2 million steps from each run as burn-in, and combined the resulting MCMC samples for subsequent estimation of posteriors. The 50% majority rule consensus tree (Supplementary Fig. 1) is shown rooted on the branch identified by the rooted-tree method in BEAST v1.4.7 (ref. 9), described below; however, the group M rooting was not relevant to any dating analysis. We also estimated phylogenies using the same data set under neighbour-joining and maximum likelihood methods and the same substitution model. The DRC60 sequences fell in the same topological position as with the BMCMC methods, with short root-to-tip genetic distances, consistent with the MrBayes results (Supplementary Fig. 1). All data and trees available from the authors on request.

We used the posterior tree sample to test the hypothesis that the terminal nodes of the DRC60 sequences were closer to the inferred A/A1 ancestral node by calculating the proportion of sampled trees where the A/A1 node-to-tip distances were smaller for these sequences than for the three modern sequences from the DRC in the same clade (Fig. 1b).

To predict the date of sampling on the basis of the phylogenetic properties of the DRC60 sequences, we also plotted the branch lengths (A/A1 node to tips) against the time of sampling for all A/A1 sequences excluding DRC60 and calculated the best fit for the linear regression of genetic divergence against the year of sampling of the viruses[27]. We calculated the mean and 95% HPD of the predicted sampling date of each DRC60 consensus sequence on the basis of its node-to-tip distance and the inferred regression line calculated for each of 100 trees sampled by MrBayes.

Bayesian MCMC inference of phylogeny using BEAST v1.4.7. We used the Bayesian methods described previously[9,10], which allow for the co-estimation of phylogeny and divergence times under a 'relaxed' molecular clock model, as implemented in BEAST v1.4.7 (ref. 9). All analyses were performed under an uncorrelated lognormal relaxed molecular clock model, using a general time-reversible nucleotide substitution model with heterogeneity among sites modelled with a gamma distribution. We investigated each demographic model (constant population, exponential growth, expansion growth, logistic growth) as well as a Bayesian skyline plot coalescent tree prior[13], a general, non-parametric prior that enforces no particular demographic history. We used a piecewise linear skyline model with 10 groups. We then compared the marginal likelihoods for each model using Bayes factors estimated in Tracer v1.4 as described[12,15]. Bayes factors represent the ratio of the marginal likelihoods of the models being compared. A large ratio can indicate that one model is a significantly better fit to the data than another. We assessed the strength of the evidence that the best-fit model was superior to the others as described[15].

For each analysis, two independent runs of 50 million steps were performed. Examination of the MCMC samples with Tracer v1.4 indicated convergence and adequate mixing of the Markov chains, with estimated sample sizes in the hundreds or thousands. After inspection with Tracer, we discarded an appropriate number of steps from each run as burn-in, and combined the resulting MCMC tree samples for subsequent estimation of posteriors. We summarized the MCMC samples using the maximum clade credibility topology found with TreeAnnotator v1.4.7 (ref. 9), with branch length depicted in years (median of those branches that were present in at least 50% of the sampled trees; Fig. 2). The Bayesian skyline plot was reconstructed using the posterior tree sample and Tracer v1.4.

26. Leitner, T. *et al. HIV Sequence Compendium* ⟨http://www.hiv.lanl.gov⟩ (Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, 2005).
27. Drummond, A., Pybus, O. G. & Rambaut, A. Inference of viral evolutionary rates from molecular sequences. *Adv. Parasitol.* **54**, 331–358 (2003).

LETTERS

Individual differences in non-verbal number acuity correlate with maths achievement

Justin Halberda[1], Michèle M. M. Mazzocco[1,2] & Lisa Feigenson[1]

Human mathematical competence emerges from two representational systems. Competence in some domains of mathematics, such as calculus, relies on symbolic representations that are unique to humans who have undergone explicit teaching[1,2]. More basic numerical intuitions are supported by an evolutionarily ancient approximate number system that is shared by adults[3–6], infants[7] and non-human animals[8–13]—these groups can all represent the approximate number of items in visual or auditory arrays without verbally counting, and use this capacity to guide everyday behaviour such as foraging. Despite the widespread nature of the approximate number system both across species and across development, it is not known whether some individuals have a more precise non-verbal 'number sense' than others. Furthermore, the extent to which this system interfaces with the formal, symbolic maths abilities that humans acquire by explicit instruction remains unknown. Here we show that there are large individual differences in the non-verbal approximation abilities of 14-year-old children, and that these individual differences in the present correlate with children's past scores on standardized maths achievement tests, extending all the way back to kindergarten. Moreover, this correlation remains significant when controlling for individual differences in other cognitive and performance factors. Our results show that individual differences in achievement in school mathematics are related to individual differences in the acuity of an evolutionarily ancient, unlearned approximate number sense. Further research will determine whether early differences in number sense acuity affect later maths learning, whether maths education enhances number sense acuity, and the extent to which tertiary factors can affect both.

Behavioural, neuropsychological and brain imaging techniques show that a signature of the approximate number system (ANS) is its imprecision[2–13]. Unlike exact verbal counting, the ANS produces numerical representations that grow increasingly imprecise as a linear function of the target array, with larger quantities represented less precisely than smaller quantities. This imprecision is expressed as a Weber fraction that indexes the amount of error in the underlying mental representation of any numerosity[3–5]. On average, the Weber fraction of adults is approximately 0.11, yielding successful non-verbal discrimination of arrays differing by as little as a 9:10 ratio[5,14]. Here we address whether there are significant individual differences in ANS acuity, and also whether these differences correlate with individual differences in symbolic maths achievement.

We examined 64 14-yr-old children with normal development whose performance in a variety of mathematical and more general cognitive tasks had been measured longitudinally, starting in kindergarten[15]. We tested for correlations between the current ANS acuity of the subjects and their past achievement in symbolic maths, while controlling for a wide range of other variables. Each subject's ANS acuity was assessed by psychophysical modelling of performance on a simple more/less judgement task similar to those used previously with infants and non-human animals. On each trial, subjects saw spatially intermixed blue and yellow dots presented on a computer screen too rapidly (200 ms) to serially count (Fig. 1a)[16]. Subjects indicated which colour was more numerous by key press and verbal response. The ratio between the two sets varied randomly among 1:2, 3:4, 5:6 and 7:8, with between 5 and 16 dots in each set. The colour of the more numerous set varied randomly, and half of the trials were area-controlled to ensure that responses were on the basis of the

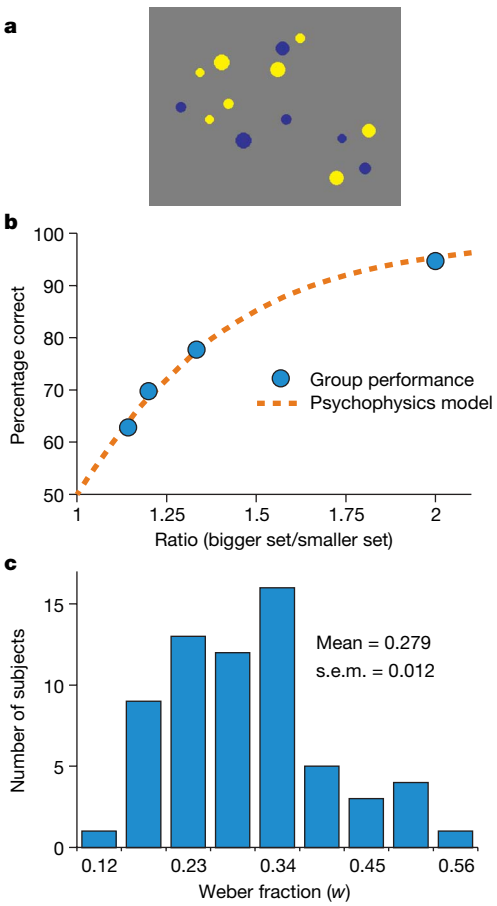

Figure 1 | Method and group performance. **a**, A representation of the trial from the numerical discrimination task. **b**, Group performance and modelled best-fit for all trials in the numerical discrimination task. **c**, Histogram of w, the acuity of the ANS, for the sample ($n = 64$), as determined by the psychophysical model for each subject.

[1]Johns Hopkins University, Ames Hall, 3400 North Charles Street, Baltimore, Maryland 21218, USA. [2]Kennedy Krieger Institute, 3825 Greenspring Avenue, Painter Building, Top Floor, Baltimore, Maryland 21211, USA.

number of dots and not on the total dot area (see Supplementary Information). Subjects participated in two sessions of 10 practice trials and 40 test trials each, totalling 80 test trials (approximately 10 min of testing per subject).

Collapsing across subjects, numerical discrimination improved as the ratio between the presented numerosities increased, in accord with Weber's law and with previous investigations of the ANS[3–9] (Fig. 1b). This gradual improvement in performance as a function of ratio was modelled using classical psychophysical tools to determine the group Weber fraction (see Methods and Supplementary Information). This returned a value of 0.265 for the group Weber fraction (w) with an R^2 value of 0.995, suggesting that there is very high agreement between this psychophysical model of the ANS and the behavioural data (Fig. 1b). Next, we used this same method to fit each individual subject's data and thereby determine each subject's Weber fraction. This showed surprisingly large variation in the ANS acuity (w), ranging from 0.119 to 0.567 (Fig. 1c). The Weber fractions of subjects can also be translated into more intuitive whole numbers that show the ratio that would result in 75% correct performance. Using this translation, some subjects could discriminate numerical ratios as fine as 9:10 ($w = 0.11$) whereas others had difficulty with ratios finer than 2:3 ($w = 0.5$; mean subject $w \approx 4{:}5$).

A question to address is whether these individual differences in ANS acuity (w) predict individual differences in symbolic maths achievement. Each of our subjects was tested annually from kindergarten to sixth grade (ages 5–11) on a battery of standardized and investigator-designed measures. This longitudinal assessment of mathematical, verbal and other cognitive abilities provides a unique opportunity to detect any enduring correlations between ANS acuity and symbolic maths ability while controlling for other factors. Each year (ages 5–11), symbolic maths ability was assessed using the 'test of early mathematical ability, second edition' (TEMA-2)[17] and/or the 'Woodcock–Johnson revised calculation subtest' (WJ-Rcalc)[18], yielding an age-referenced standardized score for each subject. We found that the ANS acuity (w) of subjects correlated with symbolic maths performance in every year tested (from kindergarten to sixth grade) for both of the standardized maths tests, as summarized in Table 1. ANS acuity in ninth grade retrospectively predicted the symbolic maths performance of individual students from as early as kindergarten, a 9-yr time span. The linear correlations of ANS acuity (w) with symbolic maths achievement (TEMA-2 and WJ-Rcalc) for the third grade are shown in Fig. 2a, b.

A further question to address was whether the correlation between ANS acuity and symbolic maths achievement was due to individual differences in more general cognitive or performance factors. In the third grade (when subjects were approximately aged 8) we administered several non-numerical standardized tests including measures of rapid lexical access for colour names (rapid automatic naming, RAN-colour)[19] and full-scale IQ (Wechsler abbreviated scale of intelligence, WASI-full)[20]. The RAN-colour is an appropriate control for our task because it measures the reaction time to identify the colours of 50 stimuli quickly; rapid colour naming is precisely the behaviour required by our ANS acuity assessment. The WASI-full IQ test acts as a control for general intelligence. WASI-full and RAN-colour did not correlate with one another in our sample ($P = 0.699$), making them largely orthogonal for purposes of linear regressions with ANS acuity. To examine the relationship of ANS acuity and symbolic maths achievement while controlling for other variables, two separate linear regressions were performed with ANS acuity (w) as the dependent variable and performance on either the TEMA-2 or the WJ-Rcalc test, and WASI-full and RAN-colour as independent variables. These showed that ANS acuity (w) correlated with symbolic maths achievement in the third grade even with rapid lexical access and general intelligence controlled for (Table 2).

To assess the strength of the correlation between ANS acuity (w) and symbolic maths achievement further, we performed extra linear regressions between w (measured at age 14) and an even broader range of standardized test scores obtained when subjects were in the third grade. These 16 measures controlled for the widest possible range of behavioural, cognitive and intelligence factors in our sample including many factors promoted as predictors of mathematical ability (for example, visual–spatial reasoning, working memory)[21–25]. ANS acuity (w) significantly correlated with symbolic maths achievement (measured in the third grade) for both TEMA-2 and WJ-Rcalc performance, with all 16 measures controlled for ($r_p^2 = 0.167$ and 0.200, respectively, where p represents partial correlation). In contrast, no other measure correlated with ANS acuity when symbolic maths performance and other variables were controlled for (Table 3). This means that success on tests of symbolic mathematics throughout the school years

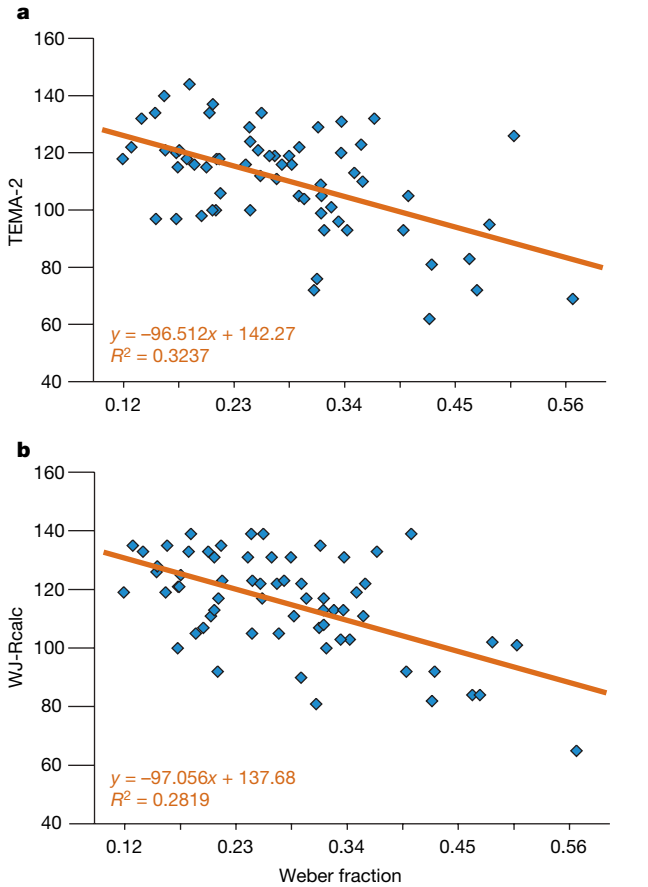

Figure 2 | Regressions. **a**, **b**, Linear regression of the standard score for each subject on the TEMA-2 test (**a**) or on the WJ-Rcalc test (**b**) of symbolic maths achievement and the acuity of the ANS (w). For TEMA-2 and WJ-Rcalc, higher numbers indicate better performance, whereas for the Weber fraction, lower numbers indicate better performance.

Table 1 | Correlation of ANS acuity (w) with symbolic maths achievement

Grade	TEMA-2 R^2	t d.f. = 62	P	WJ-Rcalc R^2	t d.f. = 62	P
Kindergarten	0.137	3.134	0.003	0.127	2.959	0.004
First	0.140	3.171	0.002	0.326	5.480	8×10^{-7}
Second	0.238	4.399	4×10^{-5}	–	–	–
Third	0.324	5.448	9×10^{-7}	0.282	4.933	6×10^{-6}
Fourth	–	–	–	0.248	4.518	3×10^{-5}
Fifth	–	–	–	0.117	2.866	0.006
Sixth	–	–	–	0.251	4.564	2×10^{-5}

ANS acuity (w) measured in ninth grade retroactively correlated with symbolic maths achievement. R^2 values represent the proportion of the variance in symbolic maths achievement that is explained by ANS acuity. R^2 values >0.25 are considered large in behavioural science and are generally viewed as having large practical significance. t values represent the distance, measured in units of standard error, between the obtained correlation and the null hypothesis of no correlation. P values represent the probability of obtaining the observed correlation in a sample of data by random chance when there is truly no relation in the population.