SE 171 76 Stockholm, Sweden. [50]Institute of Chemistry, Universidade de Sao Paulo, Sao Paulo, Brazil. [51]The Hamilton Institute, Maynooth, National University of Ireland, Maynooth, Co. Kildare, Ireland. [52]Department of Systems Biology and Bioinformatics, University of Rostock, 18051 Rostock, Germany. [53]Faculty of Medicine, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands. [54]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Campus Limpertsberg, 162a, avenue de la Faiencerie, L-1511, Luxembourg. [55]Department of Genetics, University of Leicester, Adrian Building, University Road, Leicester, LE1 7RH, UK. [56]European Institute for Systems Biology and Medicine, HLA and Medicine, Jean Dausset Laboratory, St Louis Hospital, INSERM U940, Paris, France. [57]European Institute for Systems Biology and Medicine, Pulmonary Division, Albert Michallon University Hospital, La Tronche, France. [58]Fundamental and Applied Bioenergetics, INSERM U1055, Joseph Fourier University, Grenoble, France. [59]Centre for Systems Biomedicine, Jiao-Tong University, Shanghai, China. [60]European Institute for Systems Biology and Medicine, Claude Bernard University, Lyon, France. [61]Functional Genomics and Systems Biology for Health, CNRS Institute of Biological Sciences, Villejuif, France.

Published: 6 July 2011

References

1. **Council conclusions "Innovative approaches for chronic diseases in public health and healthcare systems".** Council of the European Union 3053rd Employment, Social Policy Health and Consumer Affairs Council Meeting, Brussels, 7 December 2010. [http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/lsa/118254.pdf]
2. **2008-2013 Action plan for the global strategy for the prevention and control of non communicable diseases. Prevent and control cardiovascular diseases, cancers, chronic respiratory diseases, diabetes** [http://www.who.int/nmh/Actionplan-PC-NCD-2008.pdf]
3. de Rijk MC, Tzourio C, Breteler MM, Dartigues JF, Amaducci L, Lopez-Pousa S, Manubens-Bertran JM, Alpérovitch A, Rocca WA: **Prevalence of parkinsonism and Parkinson's disease in Europe: the EUROPARKINSON Collaborative Study. European Community Concerted Action on the Epidemiology of Parkinson's disease.** *J Neurol Neurosurg Psychiatry* 1997, **62**:10-15.
4. Beaglehole R, Horton R: **Chronic diseases: global action must match global evidence.** *Lancet* 2010, **376**:1619-1621.
5. Alwan A, Maclean DR, Riley LM, d'Espaignet ET, Mathers CD, Stevens GA, Bettcher D: **Monitoring and surveillance of chronic non-communicable diseases: progress and capacity in high-burden countries.** *Lancet* 2010, **376**:1861-1868.
6. Narayan KM, Ali MK, Koplan JP: **Global noncommunicable diseases--where worlds meet.** *N Engl J Med* 2010, **363**:1196-1198.
7. **World Health Statistics 2010 report** [http://www.who.int/whr/en/index.html]
8. **Essential Medicines. WHO Model List (revised March 2008)** [http://www.who.int/medicines/publications/essentialmedicines/en/]
9. Cruz AA, Bousquet PJ: **The unbearable cost of severe asthma in underprivileged populations.** *Allergy* 2009, **64**:319-321.
10. Mayosi BM, Flisher AJ, Lalloo UG, Sitas F, Tollman SM, Bradshaw D: **The burden of non-communicable diseases in South Africa.** *Lancet* 2009, **374**:934-947.
11. Busse R, Blümel M, Scheller-Kreinsen D, Zentner A: *Tackling Chronic Disease in Europe. Strategies, Interventions and Challenges.* Berlin: WHO; 2010.
12. Barabasi AL, Gulbahce N, Loscalzo J: **Network medicine: a network-based approach to human disease.** *Nat Rev Genet* 2011, **12**:56-68.
13. Christensen K, Doblhammer G, Rau R, Vaupel JW: **Ageing populations: the challenges ahead.** *Lancet* 2009, **374**:1196-1208.
14. van Weel C, Schellevis FG: **Comorbidity and guidelines: conflicting interests.** *Lancet* 2006, **367**:550-551.
15. Valderas JM, Starfield B, Sibbald B, Salisbury C, Roland M: **Defining comorbidity: implications for understanding health and health services.** *Ann Fam Med* 2009, **7**:357-363.
16. Vogeli C, Shields AE, Lee TA, Gibson TB, Marder WD, Weiss KB, Blumenthal D: **Multiple chronic conditions: prevalence, health consequences, and implications for quality, care management, and costs.** *J Gen Intern Med* 2007, **22 Suppl 3**:391-395.
17. Spinetti G, Kraenkel N, Emanueli C, Madeddu P: **Diabetes and vessel wall remodelling: from mechanistic insights to regenerative therapies.** *Cardiovasc Res* 2008, **78**:265-273.
18. Haahtela T: **Allergy is rare where butterflies flourish in a biodiverse environment.** *Allergy* 2009, **64**:1799-1803.
19. Jackson FL: **Ethnogenetic layering (EL): an alternative to the traditional race model in human variation and health disparity studies.** *Ann Hum Biol* 2008, **35**:121-144.
20. Simeoni U, Barker DJ: **Offspring of diabetic pregnancy: long-term outcomes.** *Semin Fetal Neonatal Med* 2009, **14**:119-124.
21. Barker DJ: **Coronary heart disease: a disorder of growth.** *Horm Res* 2003, **59 Suppl 1**:35-41.
22. Bousquet J, Jacot W, Yssel H, Vignola AM, Humbert M: **Epigenetic inheritance of fetal genes in allergic asthma.** *Allergy* 2004, **59**:138-147.
23. Svanes C, Sunyer J, Plana E, Dharmage S, Heinrich J, Jarvis D, de Marco R, Norbäck D, Raherison C, Villani S, Wjst M, Svanes K, Antó JM: **Early life origins of chronic obstructive pulmonary disease.** *Thorax* 2010, **65**:14-20.
24. Thornburg KL, Shannon J, Thuillier P, Turker MS: ***In utero* life and epigenetic predisposition for disease.** *Adv Genet* 2010, **71**:57-78.
25. Rook GA: **The hygiene hypothesis and the increasing prevalence of chronic inflammatory disorders.** *Trans R Soc Trop Med Hyg* 2007, **101**:1072-1074.
26. Gluckman PD, Hanson MA, Mitchell MD: **Developmental origins of health and disease: reducing the burden of chronic disease in the next generation.** *Genome Med* 2010, **2**:14.
27. Frison EA, Smith IF, Johns T, Cherfas J, Eyzaguirre PB: **Agricultural biodiversity, nutrition, and health: making a difference to hunger and nutrition in the developing world.** *Food Nutr Bull* 2006, **27**:167-179.
28. Knip M, Virtanen SM, Seppä K, Ilonen J, Savilahti E, Vaarala O, Reunanen A, Teramo K, Hämäläinen AM, Paronen J, Dosch HM, Hakulinen T, Akerblom HK; Finnish TRIGR Study Group: **Dietary intervention in infancy and later signs of beta-cell autoimmunity.** *N Engl J Med* 2010, **363**:1900-1908.
29. Lock K, Smith RD, Dangour AD, Keogh-Brown M, Pigatto G, Hawkes C, Fisberg RM, Chalabi Z: **Health, agricultural, and economic effects of adoption of healthy diet recommendations.** *Lancet* 2010, **376**:1699-1709.
30. Marteau TM, French DP, Griffin SJ, Prevost AT, Sutton S, Watkinson C, Attwood S, Hollands GJ: **Effects of communicating DNA-based disease risk estimates on risk-reducing behaviours.** *Cochrane Database Syst Rev* 2010:CD007275.
31. Wipfli H, Samet JM: **Global economic and health benefits of tobacco control: part 2.** *Clin Pharmacol Ther* 2009, **86**:272-280.
32. Torres-Duque C, Maldonado D, Perez-Padilla R, Ezzati M, Viegi G: **Biomass fuels and respiratory diseases: a review of the evidence.** *Proc Am Thorac Soc* 2008, **5**:577-590.
33. Khoury MJ, Gwinn M, Ioannidis JP: **The emergence of translational epidemiology: from scientific discovery to population health impact.** *Am J Epidemiol* 2010, **172**:517-524.
34. Marmot M: **Achieving health equity: from root causes to fair outcomes.** *Lancet* 2007, **370**:1153-1163.
35. Kivimäki M, Shipley MJ, Ferrie JE, Singh-Manoux A, Batty GD, Chandola T, Marmot MG, Smith GD: **Best-practice interventions to reduce socioeconomic inequalities of coronary heart disease mortality in UK: a prospective occupational cohort study.** *Lancet* 2008, **372**:1648-1654.
36. **The World Health Report 2008 – primary health care (now more than ever)** [http://www.who.int/whr/2008/en/index.html]
37. Starfield B, Lemke KW, Bernhardt T, Foldes SS, Forrest CB, Weiner JP: **Comorbidity: implications for the importance of primary care in 'case' management.** *Ann Fam Med* 2003, **1**:8-14.
38. Campbell SM, McDonald R, Lester H: **The experience of pay for performance in English family practice: a qualitative study.** *Ann Fam Med* 2008, **6**:228-234.
39. Stange KC: **A science of connectedness.** *Ann Fam Med* 2009, **7**:387-395.
40. Carrier E, Gourevitch MN, Shah NR: **Medical homes: challenges in translating theory into practice.** *Med Care* 2009, **47**:714-722.
41. Butte AJ: **Medicine: the ultimate model organism.** *Science* 2008, **320**:325-327.
42. Vestbo J, Rennard S: **Chronic obstructive pulmonary disease biomarker(s) for disease activity needed--urgently.** *Am J Respir Crit Care Med* 2010, **182**:863-864.
43. National Heart, Lung and Blood Institute: *Expert Panel Report 3: Guidelines for the Diagnosis and Management of Asthma. National Asthma Education and Prevention Program.* Washington DC: US Department of Health and Human Services; 2007.
44. Vijan S: **Type 2 diabetes.** *Ann Intern Med* 2010, **152**:ITC31-15.
45. Bousquet J, Mantzouranis E, Cruz AA, Aït-Khaled N, Baena-Cagnani CE, Bleecker ER, Brightling CE, Burney P, Bush A, Busse WW, Casale TB, Chan-Yeung M, Chen R, Chowdhury B, Chung KF, Dahl R, Drazen JM, Fabbri LM, Holgate ST, Kauffmann F, Haahtela T, Khaltaev N, Kiley JP, Masjedi MR,

Mohammad Y, O'Byrne P, Partridge MR, Rabe KF, Togias A, van Weel C, *et al.*: **Uniform definition of asthma severity, control, and exacerbations: document presented for the World Health Organization Consultation on Severe Asthma.** *J Allergy Clin Immunol* 2010, **126**:926-938.
46. Pare G, Moqadem K, Pineau G, St-Hilaire C: **Clinical effects of home telemonitoring in the context of diabetes, asthma, heart failure and hypertension: a systematic review.** *J Med Internet Res* 2010, **12**:e21.
47. Légaré F, Ratté S, Stacey D, Kryworuchko J, Gravel K, Graham ID, Turcotte S: **Interventions for improving the adoption of shared decision making by healthcare professionals.** *Cochrane Database Syst Rev* 2010:CD006732.
48. Collins RE, Wright AJ, Marteau TM: **Impact of communicating personalized genetic risk information on perceived control over the risk: a systematic review.** *Genet Med* 2011, **13**:273-277.
49. Reeves S, Zwarenstein M, Goldman J, Barr H, Freeth D, Koppel I, Hammick M: **The effectiveness of interprofessional education: key findings from a new systematic review.** *J Interprof Care* 2010, **24**:230-241.
50. **Ottawa Charter for Health Promotion First International Conference on Health Promotion Ottawa, 21 November 1986.** WHO/HPR/HEP/95.1, 1986. [http://www.who.int/healthpromotion/conferences/previous/ottawa/en/]
51. Hood L, Heath JR, Phelps ME, Lin B: **Systems biology and new technologies enable predictive and preventative medicine.** *Science* 2004, **306**:640-643.
52. Haahtela T, Tuomisto LE, Pietinalho A, Klaukka T, Erhola M, Kaila M, Nieminen MM, Kontula E, Laitinen LA: **A 10 year asthma programme in Finland: major change for the better.** *Thorax* 2006, **61**:663-670.
53. Chan BC, Perkins D, Wan Q, Zwar N, Daniel C, Crookes P, Harris MF: **Finding common ground? Evaluating an intervention to improve teamwork among primary health-care professionals.** *Int J Qual Health Care* 2010, **22**:519-524.
54. Auffray C, Chen Z, Hood L: **Systems medicine: the future of medical genomics and healthcare.** *Genome Med* 2009, **1**:2.
55. Price N, Edelman L, Lee I, Yoo H, Hwang D, Carlson G, *et al.*: *Genomic and Personalized Medicine: from Principles to Practice.* Edited by Ginsburg G, Willard H. New York: Elsevier; 2008
56. Auffray C, Balling R, Benson M, Bertero M, Byrne H, Cascante M, Colding-Jörgensen M, De Pauw E, Fabbri LM, Foulkes T, Goryanin I, Harrison D, Henney A, Hoeveler A, Iris F, Kyriakopoulou C, Klingmüller U, Kolch W, Lahesmaa R, Lemberger T, Lévi F, Lichtenberg H, Lotteau V, Mayer B, Mialhe A, Mulligan B, Rozman D, Siest G, Swinton J, Ueffing M, *et al.*: **From Systems Biology to Systems Medicine, European Commission, DG Research, Directorate of Health. Brussels 14-15 June 2010.** Workshop report; 2010. [ftp://ftp.cordis.europa.eu/pub/fp7/health/docs/final-report-systems-medicine-workshop_en.pdf]
57. Hood L, Friend S: **Predictive, personalized, preventative, participatory cancer medicine.** *Nat Rev Clin Oncol* 2011, in press.
58. Auffray C, Charron D, Hood L: **Predictive, preventive, personalized and participatory medicine: back to the future.** *Genome Med* 2010, **2**:57.
59. Burke W, Burton H, Hall AE, Karmali M, Khoury MJ, Knoppers B, Meslin EM, Stanley F, Wright CF, Zimmern RL: **Extending the reach of public health genomics: what should be the agenda for public health in an era of genome-based and "personalized" medicine?** *Genet Med* 2010, **12**:785-791.
60. Manolio TA, Bailey-Wilson JE, Collins FS: **Genes, environment and the value of prospective cohort studies.** *Nat Rev Genet* 2006, **7**:812-820.
61. Scheffer M, Bascompte J, Brock WA, Brovkin V, Carpenter SR, Dakos V, Held H, van Nes EH, Rietkerk M, Sugihara G: **Early-warning signals for critical transitions.** *Nature* 2009, **461**:53-59.
62. Hwang D, Lee IY, Yoo H, Gehlenborg N, Cho JH, Petritis B, Baxter D, Pitstick R, Young R, Spicer D, Price ND, Hohmann JG, Dearmond SJ, Carlson GA, Hood LE: **A systems approach to prion disease.** *Mol Syst Biol* 2009, **5**:252.
63. Muskulus M, Slats AM, Sterk PJ, Verduyn-Lunel S: **Fluctuations and determinism of respiratory impedance in asthma and chronic obstructive pulmonary disease.** *J Appl Physiol* 2010, **109**:1582-1591.
64. Frey U, Brodbeck T, Majumdar A, Taylor DR, Town GI, Silverman M, Suki B: **Risk of severe asthma episodes predicted from fluctuation analysis of airway function.** *Nature* 2005, **438**:667-670.
65. Bousquet J, Anto J, Auffray C, Akdis M, Cambon-Thomsen A, Keil T, Haahtela T, Lambrecht BN, Postma DS, Sunyer J, Valenta R, Akdis CA, Annesi-Maesano I, Arno A, Bachert C, Ballester F, Basagana X, Baumgartner U, Bindslev-Jensen C, Brunekreef B, Carlsen KH, Chatzi L, Crameri R, Eveno E, Forastiere F, Garcia-Aymerich J, Guerra S, Hammad H, Heinrich J, Hirsch D, *et al.*: **MeDALL (Mechanisms of the Development of ALLergy): an integrated approach from phenotypes to systems medicine.** *Allergy* 2011, **66**:596-604.
66. Agustí A, Sobradillo P, Celli B: **Addressing the complexity of chronic obstructive pulmonary disease: from phenotypes and biomarkers to scale-free networks, systems biology, and p4 medicine.** *Am J Respir Crit Care Med* 2011, **183**:1129-1137.
67. Kuhn KA, Knoll A, Mewes HW, Schwaiger M, Bode A, Broy M, Daniel H, Feussner H, Gradinger R, Hauner H, Höfler H, Holzmann B, Horsch A, Kemper A, Krcmar H, Kochs EF, Lange R, Leidl R, Mansmann U, Mayr EW, Meitinger T, Molls M, Navab N, Nüsslin F, Peschel C, Reiser M, Ring J, Rummeny EJ, Schlichter J, Schmid R, Wichmann HE, Ziegler S: **Informatics and medicine--from molecules to populations.** *Methods Inf Med* 2008, **47**:283-295.
68. Ullman-Cullere MH, Mathew JP: **Emerging landscape of genomics in the electronic health record for personalized medicine.** *Hum Mutat* 2011, **32**:512-516.
69. Maojo V, de la Calle G, Martin-Sanchez F, Diaz C, Sanz F: **INFOBIOMED: European Network of Excellence on Biomedical Informatics to support individualised healthcare.** *AMIA Annu Symp Proc* 2005:1041.
70. Maojo V, Martin-Sanchez F: **Bioinformatics: towards new directions for public health.** *Methods Inf Med* 2004, **43**:208-214.
71. **BBMRI during the transition phase** [http://www.bbmri.eu/]
72. **BioSHaRE** [http://www.p3g.org/bioshare/]
73. Dudley JT, Schadt E, Sirota M, Butte AJ, Ashley E: **Drug discovery in a multidimensional world: systems, patterns, and networks.** *J Cardiovasc Transl Res* 2010, **3**:438-447.
74. Sarkar IN: **Biomedical informatics and translational medicine.** *J Transl Med* 2010, **8**:22.
75. Broadhurst D, Kell D: **Statistical strategies for avoiding false discoveries in metabolomics and related experiments.** *Metabolomics* 2006, **2**:171-196.
76. Auffray C, Adcock IM, Chung KF, Djukanovic R, Pison C, Sterk PJ: **An integrative systems biology approach to understanding pulmonary diseases.** *Chest* 2010, **137**:1410-1416.
77. Oresic M, Simell S, Sysi-Aho M, Näntö-Salonen K, Seppänen-Laakso T, Parikka V, Katajamaa M, Hekkala A, Mattila I, Keskinen P, Yetukuri L, Reinikainen A, Lähde J, Suortti T, Hakalax J, Simell T, Hyöty H, Veijola R, Ilonen J, Lahesmaa R, Knip M, Simell O: **Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes.** *J Exp Med* 2008, **205**:2975-2984.
78. Bougneres P, Valleron AJ: **Causes of early-onset type 1 diabetes: toward data-driven environmental approaches.** *J Exp Med* 2008, **205**:2953-2957.
79. Szalma S, Koka V, Khasanova T, Perakslis ED: **Effective knowledge management in translational medicine.** *J Transl Med* 2010, **8**:68.
80. Gröne O, Garcia-Barbero M: **Integrated care. A position paper of the WHO European office for integrated health care services.** *Int J Integr Care* 2001, **1**:e21.
81. **UK Medical Research Council strategy "Research Changing Lives"** [http://www.mrc.ac.uk/About/Strategy/StrategicPlan2009-2014/index.htm]
82. Tapp H, Dulin M: **The science of primary health-care improvement: potential and use of community-based participatory research by practice-based research networks for translation of research into practice.** *Exp Biol Med (Maywood)* 2010, **235**:290-299.
83. Gofin J, Foz G: **Training and application of community-oriented primary care (COPC) through family medicine in Catalonia, Spain.** *Fam Med* 2008, **40**:196-202.
84. Katon WJ, Lin EH, Von Korff M, Ciechanowski P, Ludman EJ, Young B, Peterson D, Rutter CM, McGregor M, McCulloch D: **Collaborative care for patients with depression and chronic illnesses.** *N Engl J Med* 2010, **363**:2611-2620.
85. Ninot G, Moullec G, Desplan J, Prefaut C, Varray A: **Daily functioning of dyspnea, self-esteem and physical self in patients with moderate COPD before, during and after a first inpatient rehabilitation program.** *Disabil Rehabil* 2007, **29**:1671-1678.
86. O'Connor AM, Bennett CL, Stacey D, Barry M, Col NF, Eden KB, Entwistle VA, Fiset V, Holmes-Rovner M, Khangura S, Llewellyn-Thomas H, Rovner D: **Decision aids for people facing health treatment or screening decisions.** *Cochrane Database Syst Rev* 2009:CD001431.
87. Sidall C, Kjaeserud G, Dziworski W, Przywara B, Xavier A: *Healthy Ageing: Keystone for a Sustainable Europe. EU Health Policy in the Context of Demographic Change: Discussion Paper of the Services of DG SANCO, DG ECFIN and DG EMPL.* Edited by HaCPD-G. European Commission; 2007.
88. Koh HK, Oppenheimer SC, Massin-Short SB, Emmons KM, Geller AC, Viswanath K: **Translating research evidence into practice to reduce health disparities: a social determinants approach.** *Am J Public Health* 2010, **100 Suppl 1**:S72-80.

89. Marmot M, Friel S, Bell R, Houweling TA, Taylor S: **Closing the gap in a generation: health equity through action on the social determinants of health.** *Lancet* 2008, **372**:1661-1669.
90. Kenny NP, Sherwin SB, Baylis FE: **Re-visioning public health ethics: a relational perspective.** *Can J Public Health* 2010, **101**:9-11.
91. Bousquet J, Schünemann HJ, Zuberbier T, Bachert C, Baena-Cagnani CE, Bousquet PJ, Brozek J, Canonica GW, Casale TB, Demoly P, Gerth van Wijk R, Ohta K, Bateman ED, Calderon M, Cruz AA, Dolen WK, Haughney J, Lockey RF, Lötvall J, O'Byrne P, Spranger O, Togias A, Bonini S, Boulet LP, Camargos P, Carlsen KH, Chavannes NH, Delgado L, Durham SR, Fokkens WJ, *et al.*: **Development and implementation of guidelines in allergic rhinitis – an ARIA-GA2LEN paper.** *Allergy* 2010, **65**:1212-1221.
92. Boyd CM, Darer J, Boult C, Fried LP, Boult L, Wu AW: **Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance.** *JAMA* 2005, **294**:716-724.
93. Ait-Khaled N, Enarson DA, Bissell K, Billo NE: **Access to inhaled corticosteroids is key to improving quality of care for asthma in developing countries.** *Allergy* 2007, **62**:230-236.
94. Beran D, McCabe A, Yudkin JS: **Access to medicines versus access to treatment: the case of type 1 diabetes.** *Bull World Health Organ* 2008, **86**:648-649.
95. Zhu C: **Science-based health care.** *Science* 2010, **327**:1429.
96. **Innovative Medicines Initiative** [http://www.imi.europa.eu]
97. **The NIH Common Fund** [http://nihroadmap.nih.gov/]
98. Auffray C: **Sharing knowledge: a new frontier for public-private partnerships in medicine.** *Genome Med* 2009, **1**:29.
99. Sundewall J, Swanson RC, Betigeri A, Sanders D, Collins TE, Shakarishvili G, Brugha R: **Health-systems strengthening: current and future activities.** *Lancet* 2011, **377**:1222-1223.
100. Swanson RC, Bongiovanni A, Bradley E, Murugan V, Sundewall J, Betigeri A, Nyonator F, Cattaneo A, Harless B, Ostrovsky A, Labonté R: **Toward a consensus on guiding principles for health systems strengthening.** *PLoS Med* 2010, **7**:e1000385.
101. Bousquet J, Khaltaev N: *Global Surveillance, Prevention and Control of Chronic Respiratory Diseases. A comprehensive approach.* Geneva: Global Alliance against Chronic Respiratory Diseases, World Health Organization; 2007.
102. Alleyne G, Stuckler D, Alwan A: **The hope and the promise of the UN Resolution on non-communicable diseases.** *Global Health* 2010, **6**:15.
103. **Healthy aging. Improving and extending quality of life among older Americans. Center for Disease Control and Prevention** [http://www.cdc.gov/chronicdisease/resources/publications/aag/aging.htm]
104. Gordon RS Jr: **An operational classification of disease prevention.** *Public Health Rep* 1983, **98**:107-109.

doi:10.1186/gm259
Cite this article as: Bousquet J, *et al.*: Systems medicine and integrated care to combat chronic noncommunicable diseases. *Genome Medicine* 2011, **3**:43.

Genomics 99 (2012) 1–9

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

A prioritization analysis of disease association by data-mining of functional annotation of human genes

Takayuki Taniya [a,b], Susumu Tanaka [c], Yumi Yamaguchi-Kabata [b,1], Hideki Hanaoka [d], Chisato Yamasaki [a,b], Harutoshi Maekawa [a,e], Roberto A. Barrero [f], Boris Lenhard [g], Milton W. Datta [h], Mary Shimoyama [i], Roger Bumgarner [j], Ranajit Chakraborty [k,l], Ian Hopkinson [m], Libin Jia [n], Winston Hide [o], Charles Auffray [p], Shinsei Minoshima [q], Tadashi Imanishi [b,*], Takashi Gojobori [b,r,**]

[a] Japan Biological Information Research Center, Japan Biological Informatics Consortium, AIST Bio-IT Research Building 7F, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan
[b] Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan
[c] Tokyo Metropolitan Institute of Medical Science, 2-1-6 Kamikitazawa, Setagaya-ku, Tokyo 156-8506, Japan
[d] University of Tokyo, Laboratory of Plant Biotechnology, Biotechnology Research Center, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan
[e] C's Lab Co., Ltd., Yamate Bldg. 2F, 7-5-1 Hamamatsu-cho, Minato-ku, Tokyo 101-0041, Japan
[f] Centre for Comparative Genomics, Murdoch University, South Street, WA 6150, Australia
[g] Center for Genomics and Bioinformatics, Karolinska Institute, Berzelius väg 35, SE-171 77 Stockholm, Sweden
[h] Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, United Hospital, 333 N. Smith Ave, Minneapolis, MN, USA
[i] Rat Genome Database and Human and Molecular Genetics Center, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, USA
[j] Center for Expression Arrays, Department of Microbiology, University of Washington, 262B Rosen, Seattle, WA 98195, USA
[k] Center for Computational Genomics, Institute of Investigative Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA
[l] Department of Forensic and Investigative Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA
[m] Cardiovascular Genetics, British Heart Foundation Laboratories, Royal Free and University College Medical School, Rayne Building, University Street, London WC1E 6JJ, UK
[n] National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA
[o] South African National Bioinformatics Institute, University Western Cape, Bellville 7535, South Africa
[p] Array s/IMAGE, Genexpress, Functional Genomics and Systems Biology for Health, LGN, UMR 7091-CNRS and Pierre and Marie Curie University, Paris VI, Villejuif, France
[q] Department of Medical Photobiology, Medical Photonics Research Center, Hamamatsu University School of Medicine, 1-20-1 Handayama, Higashi-ku, Hamamatsu, Shizuoka 431-3192, Japan
[r] National Institute of Genetics, Center of Information Biology and DNA Data Bank of Japan, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

ARTICLE INFO

Article history:
Received 1 March 2011
Accepted 6 October 2011
Available online 14 October 2011

Keywords:
Disease
Rheumatoid arthritis
Prostate cancer
Data-mining
Gene function
Discriminant analysis

ABSTRACT

Complex diseases result from contributions of multiple genes that act in concert through pathways. Here we present a method to prioritize novel candidates of disease-susceptibility genes depending on the biological similarities to the known disease-related genes. The extent of disease-susceptibility of a gene is prioritized by analyzing seven features of human genes captured in H-InvDB. Taking rheumatoid arthritis (RA) and prostate cancer (PC) as two examples, we evaluated the efficiency of our method. Highly scored genes obtained included *TNFSF12* and *OSM* as candidate disease genes for RA and PC, respectively. Subsequent characterization of these genes based upon an extensive literature survey reinforced the validity of these highly scored genes as possible disease-susceptibility genes. Our approach, Prioritization ANalysis of Disease Association (PANDA), is an efficient and cost-effective method to narrow down a large set of genes into smaller subsets that are most likely to be involved in the disease pathogenesis.

1. Introduction

Finding disease-causing genes is one of the major issues in the human genome studies. The use of linkage analysis and positional cloning techniques has led to the identification of genes involved in numerous Mendelian genetic disorders [1–4]. However, finding causative genes for complex disorders is more challenging. Multiple genes contribute to the pathogenesis of common disorders and each gene may contribute to a different degree. Therefore, association studies to detect genes contributing to complex diseases are designed by recruiting large numbers of subjects and genotyping many polymorphic markers. The possible susceptible loci identified by association studies usually contain dozens of candidate disease genes. Further studies identifying causative variants and characterizing biological processes of disease-onset mechanism require additional substantial resources, and can be impractical when examining dozens of genes. Even if genome-wide

* Corresponding author.
** Correspondence to: T. Gojobori, Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan.
E-mail addresses: t.imanishi@aist.go.jp (T. Imanishi), tgojobor@genes.nig.ac.jp (T. Gojobori).
[1] Present address: Center for Genomic Medicine, RIKEN, 1-7-22 Suehirocho, Tsurumi, Yokohama 230-0045, Japan.

0888-7543/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2011.10.002

association studies (GWAS) are conducted with an enough number of polymorphic markers, a fraction of markers showing statistically significant associations may be false positive. In finding genes involved in the disease pathogenesis, it is not easy to distinguish biological relevance from the false positives in studies with a large number of markers. Thus, a method to prioritize candidate disease genes based on a priori biological knowledge is useful to narrow down candidate genes to identify causative genes of common diseases.

Cataloguing human genes [5–7] and annotations enabled us to analyze biological information of human genes in various ways. One important application has been to prioritize possible candidate genes of disease implication by evaluating similarity or interactions between genes that have not only sequence homology but also share functional information including biological pathways. Recent studies showed that related phenotypes share common genetic basis and susceptibility genes [8–10] because the proteins involved in the pathogenesis are likely to interact together [11,12] in a few biological pathways. Although there are previous studies on prioritizing disease-related genes by use of biological information [13–20], using various kinds of biological information of human genes would be useful. Therefore, we developed a method to prioritize candidate genes for common diseases by utilizing biological information of human genes. Our method depends on two assumptions: (1) genes related to a particular disease often have common inherent structural and functional properties, and (2) known causative genes of a particular disease and novel candidate genes for the disease may share specific biological pathways or subcellular locations of gene products. The analysis starts with collection and curation of known related genes for the target disease. We analyze enrichment of biological functions in known related genes for the target disease, by using biological terms of functional domains from InterPro (www.ebi.ac.uk/interpro/), Enzyme Commission (EC) numbers (www.chem.qmul.ac.uk/iubmb/enzyme/), biological pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (www.genome.ad.jp/kegg/) [21], and Gene Ontology (GO) (www.geneontology.org/) [22]. Then we examine all the other genes one by one whether a gene is biologically closer to the known related genes for the target disease than to the other genes by a subsequent discriminant analysis. Finally, we obtain a prioritized list of candidate genes for the disease.

We applied our approach to rheumatoid arthritis (RA) and prostate cancer (PC) in order to evaluate its efficiency. RA is the most common disabling autoimmune disease, affecting approximately 0.3–1% of the population [23]. While the etiology and pathogenesis of RA are not completely understood, a previous report has identified candidate genes and loci that may be related to RA [24]. Another common disease, PC, is the most commonly diagnosed male malignancy, and is the second leading cause of male cancer mortality in the world [25]. PC has long been known to cluster in families [26], and both segregation and linkage analyses have identified specific prostate cancer susceptibility loci and candidate genes [27,28]. Thus, both RA and PC represent complex genetic diseases for which the identification of causative disease genes would have a dramatic impact on public health.

This method, which we implemented by a system, Prioritization ANalysis for Disease Association (PANDA), provides prioritized candidates of disease-related genes that are useful for further studies to identify of new causative genes. The resulting output of our approach is a prioritized list of candidate genes for the diseases that represent potential disease biomarkers and perhaps even potential targets for therapies.

2. Results

2.1. Prioritization analysis of candidate genes for RA and PC

Known disease-related genes for RA and PC were retrieved from OMIM database (www.ncbi.nlm.nih.gov/Omim) [29] and subsequently checked in Entrez Gene (www.ncbi.nlm.nih.gov/gene) by using MeSH terms (www.nlm.nih.gov/mesh/) for the two diseases (see Methods and Fig. 1). After manual curation, we selected 139 genes as known RA-related genes and 296 genes as known PC-related genes (Supplemental Tables 1–2). These genes were used as 'training sets' to analyze all other genes in the H1-REFSEQ DB (an in-house database of all human genes containing 14,959 human genes, see Methods) for likeliness of disease susceptibility. Biological information of all human genes was retrieved from H-InvDB, and seven biological features (paralogy, InterPro, EC number, biological pathways from KEGG, and three categories of Gene Ontology, see Methods) of human genes were used to examine biological similarity of a gene to a group of known related genes for the target disease (group 1) or another group of all the other genes (group 2). Then we examined whether a gene tested is closer to the group 1 than to the group 2 by comparing the Mahalanobis distance between a gene and the group 1 (MD1) with another distance between the gene and the group 2 (MD2). We calculated the ratio of MD2 to MD1 (PANDA score) for each gene tested. To narrow down more promising candidate genes for the target diseases, we used a higher threshold; the average values of the PANDA score for the known related genes (21.2 for RA and 14.1 for PC). As a result, 526 genes were detected as candidate genes for RA and 609 genes for PC (Table 1).

2.2. Candidate genes on genomic regions of interest

Although putative susceptible loci for RA and PC have been identified by linkage analysis, association study, comparative genomic hybridization and chromosomal transfer, the actual genes involved in the disease have not been identified yet for many of these genomic regions. Therefore, a localization of a candidate gene on one of these genomic regions of interest (GROI) provides additional support for its potential role in the disease. To select candidate disease genes in GROIs, we searched the GROIs in OMIM for RA and PC, and found nine GROIs for RA and 18 GROIs for PC (listed in Supplemental Tables 3–4). Then we selected the candidate genes that were localized within these GROIs, and obtained 56 candidate genes for RA and 63 for PC.

2.3. Highly scored genes for rheumatoid arthritis

To further narrow down the most plausible candidate genes for RA, we ranked the 56 candidates on the GROIs according to the discriminant analysis with the Mahalanobis distances (Supplemental Table 3). We inspected the highly scored genes, surveyed literature and found that these genes included biologically reasonable candidate genes. For example, tumor necrosis factor ligand superfamily member 12 (*TNFSF12*) showed the third highest score in 17p13 (Table 2A) which is one of the GROIs with RA [24]. The *TNFSF12* gene shares three GO annotations (tumor necrosis factor receptor, immune response and membrane) with other known RA-related genes, including *TNF* (tumor necrosis factor), *HLA-DRB1* (major histocompatibility complex, class II, DR beta 1) and toll-like receptor 2 (*TLR2*) (Table 2B). The *TNFSF12* gene is expressed in macrophages, which infiltrates the synovial membrane to form the inflammatory pannus that is characteristic of RA invading joint cartilage and destroying the underlying bone. TNFSF12 induces activation of matrix metallopeptidase 9 through nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (NFKB1) pathway [30] that controls the incidence of collagen-based arthritis in mice through modulation of inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase beta (IKBKB) [31,32]. In addition, a molecule in the NFKB1 pathway, encoded by nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor-like 1(NFKBIL1), was reported as a candidate RA-susceptibility gene, from an evidence of a di-allelic polymorphism in the promoter region of the gene [33]. Biological annotation of TNFSF12 shows functional similarities with some of the known RA-related genes. Thus, our result suggests that *TNFSF12* may be involved in the pathogenesis of RA.

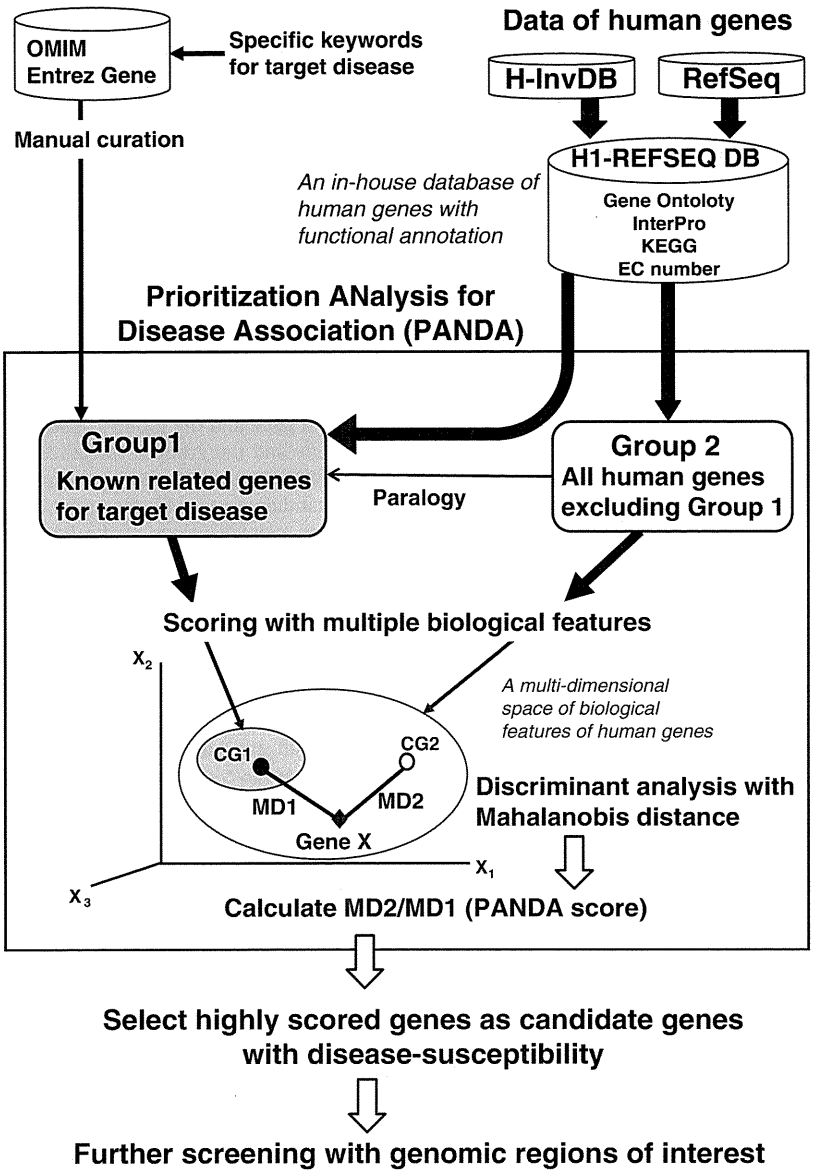

Fig. 1. Analysis pipeline with the Prioritization ANalysis of Disease Association (PANDA) system. Known related genes for the target disease (rheumatoid arthritis and prostate cancer in this study) were retrieved from OMIM and Entrez Gene, manually annotated and used as training set (group 1). We examined the proportion of biological terms in the training set and that in all human genes (H1-REFSEQ DB, see Methods). Based on different proportions of biological functions between the training set and all human genes, scores for multiple parameters were given to the all human genes. Candidate disease-related genes were prioritized by the discriminant analysis with multiple parameters. We calculated the Mahalanobis distances (MD1) between the position of a gene tested (open diamond) and the center of gravity of the group 1 (CG1, open circle) with multiple parameters. Similarly the Mahalanobis distances (MD2) between each gene and the center of gravity of the group 2 (CG2, filled circle) were calculated. Then we calculated the ratio of MD2 to MD1 (PANDA score) for each gene. Using the average value of the PANDA score for the known disease genes as a threshold, we considered any gene to be a candidate gene of disease-susceptibility if the ratio was greater than the threshold.

Table 1
Summary of prioritization analysis of disease-susceptibility genes for RA and PC.

Disease relation	Predicted similarity to known disease genes	RA	PC
Known	–	139	296
Unknown	Yes[a]	526	609
	No	14,294	14,054

Candidate genes with disease-susceptibility for rheumatoid arthritis (RA) and prostate cancer (PC) were prioritized by analyzing biological information of human genes.
We analyzed total 14959 human genes in H1-REFSEQ DB (an in-house database of human genes) including known related genes for the target diseases as the training sets. Each gene was tested one by one whether the gene is more similar to the known disease genes than the others by data-mining and discriminant analysis (see Methods). The numbers of genes for each category are shown.
[a] Genes whose PANDA scores (MD2/MD1) were higher than the threshold (the median in the training set).

2.4. Highly scored genes for prostate cancer

Similarly, we ranked the 63 candidate genes for PC (Supplemental Table 4) that were located in the 18 GROIs. One GROI for PC, 22q12, has been reported to be associated with PC by a combined genome-wide linkage scan [34]. The highly scored genes in 22q12 included Oncostatin M (*OSM*) with the highest PANDA score (30.4, Table 3A). OSM, an interleukin 6 (IL6)-type cytokine, induces activation of the androgen receptor (AR) in the absence of androgen [35]. *OSM* shares four GO annotations (cytokine activity, immune response, regulation of cell growth, extracellular) with known PC-related genes, including *IL6*, *IL8*, *IGFBP3* and *INS* (Table 3B). Interestingly, AR and IL6 have been implicated as relevant molecules for PC, and were present in the known PC-related genes as the training set. Thus, OSM is a candidate

Table 2A
Highly scored genes within 17p13 in relation to RA.

Gene symbol	Gene name	Transcript	PANDA score (MD2/MD1)	No. of GO features[a]	No. of other features[b]
CHRNE	Cholinergic receptor, nicotinic, epsilon polypeptide	NM_000080	85.3	3	0
CHRNB1	Cholinergic receptor, nicotinic, beta polypeptide 1 (muscle)	BC011371	85.3	3	0
TNFSF12	Tumor necrosis factor (ligand) superfamily, member 12	NM_172089	62.1	3	0
NLGN2	Neuroligin 2	AB037787	16.2	2	2
MYH10	Myosin, heavy polypeptide 10, non-muscle	AK026977	14.7	1	2
AURKB	Aurora kinase B	BC000442	13.6	2	3
ALOX12P2	Arachidonate 12-lipoxygenase pseudogene 2	AL832768	11.2	0	2
ALOX15	Arachidonate 15-lipoxygenase	BC029032	11.2	2	3
KIF1C	Kinesin family member 1C	AB014606	10.5	2	2

Highly scored genes were ranked by the PANDA score (MD2/MD1).
[a] Number of GO features of the gene (molecular function, biological process, and cellular component).
[b] Number of non-GO features of the gene for four features (amino acid sequence similarity, InterPro, EC number, and KEGG pathways).

Table 2B
Scores for *TNFSF12* in relation to RA.

Parameter	Functional annotation	ID	Base relevance score	Proportion in gene set		Score to *TNFSF12*[a]
				Training set	All human genes	
GO molecular function	Tumor necrosis factor receptor binding	GO:0005164 (6)[b]	2.5	0.0102	0.0011	134.85
GO biological process	Immune response	GO:0006955 (7)	1.77	0.2342	0.0168	173.16
GO cellular component	Membrane	GO:0016020 (4)	2.03	0.2714	0.2704	8.14

Among the seven parameters, details for the three concepts of GO terms are shown. The score of paralogy was 0 because *TNFSF12* did not have a sequence similarity to any known RA-related gene. The scores of InterPro, EC number and KEGG pathway were also 0.
[a] Scores were calculated by Eqs. (2) and (3).
[b] The level in a nested hierarchical vocabulary is shown in parenthesis.

molecule that is involved in the disease onset of prostate cancer, possibly through its modulation of IL6 related pathways.

2.5. Sensitivity of predicting disease-related genes

In order to evaluate the sensitivity of our approach for predicting disease genes, we examined whether the method correctly detects one of the known related genes for the target disease, which was replaced in a group of genes tested. This is an application of the leave-one-out cross-validation (LOOCV) [36], and we sequentially removed one of the known related genes for the target disease from the training set and placed it in the group of genes to be tested. Then, we calculated a new MD2/MD1 score (PANDA score), and checked whether the PANDA score of the removed gene was greater than 1. As a result, 126 of 139 (90.6%) known RA-related genes had scores greater than 1 (Supplemental Table 5). In the same way, 264 of 296 (89.2%) known PC-related genes showed higher scores than 1 (Supplemental Table 6).

2.6. Contribution of the data type to prioritize candidates of disease gene

We used the seven biological features of human genes to calculate Mahalanobis distance between a gene and a set of genes. However, it is not clear whether larger numbers of parameters increase the power of prediction and what data type is more effective than the other data. To see how the number of parameters affects the prediction power, we examined the sensitivity to identify the known disease genes by the LOOCV with different numbers of parameters in different combinations. First, we examined the sensitivity to find a known disease gene with seven parameters and compared the results with only GO annotations (molecular function, biological process and cellular component). The sensitivities by LOOCV were 74.8% (104 of 139) for the known RA genes and 72.6% (215 of 296) for the known PC genes when only GO annotation was used (Supplemental Tables 5–6). These sensitivities were slightly lower than that with the seven parameters.

We also examined the sensitivities to find known disease genes for RA in all the 120 combinations of using 2–7 parameters (Supplemental Table 5). In identifying the known RA genes, the sensitivity was the highest when all the seven parameters were used. Three combinations of 5–6 parameters had high sensitivities as the seven parameters for the RA genes (Supplemental Table 5) when at least five parameters (InterPro, EC number and three GO concepts) were used.

The sensitivity of identifying the known PC-related genes was examined in all 120 combinations of the parameters in the same way (Supplemental Table 6). The use of all the seven parameters showed the second highest sensitivity for identifying the known PC-related genes. Three combinations of 5–6 parameters were a little more sensitive than the use of the seven parameters, and two combinations of 4 or 6 parameters had the same sensitivity as the seven parameters (Supplemental Table 6). These five combinations had at least four parameters including InterPro, GO molecular function and GO biological process in common. These results suggest that using several features including GO would be effective for prioritization of disease-susceptibility genes.

3. Discussion

In the past several years, several attempts have been made to develop methods for prioritizing potential disease-susceptibility candidates [13–15], and these are based on sequence similarities between genes and functional annotations from Gene Ontology. Perez-Iratxeta et al. developed a method based on the co-appearance of GO terms and MeSH disease terms [16,17]. In contrast, Freudenberg et al. used the clustering method to identify new disease genes [18], while Turner et al. presented prioritization of candidate genes using statistics (POCUS) based on the over-representation of the annotated function for the target disease [15]. Although these methods have been relatively successful, there are two problems in identifying disease-related genes: (1) these methods strongly depend on kind of functional information (i.e., information from GO), and (2) some frequently appearing biological terms in human genes can be "background noise" because

Table 3A
Highly scored genes within 22q12 in relation to PC.

Gene symbol	Gene name	Transcript	PANDA score (MD2/MD1)	No. of GO features[a]	No. of other features[b]
OSM	Oncostatin M	NM_020530	30.4	3	0
KREMEN1	Kringle containing transmembrane protein 1	NM_032045	22.4	0	1
GAS2L1	Growth arrest-specific 2 like 1	NM_006478	20.6	1	1
TIMP3	Tissue inhibitor of metalloproteinase 3	NM_000362	17.5	0	2
LIMK2	LIM domain kinase 2	NM_005569	17.3	2	2
PIK3IP1	HGFL gene	BC011049	16.6	0	1
ADRBK2	Adrenergic, beta, receptor kinase 2	NM_005160	16.4	2	2
CHEK2[c]	CHK2 checkpoint homolog (S. pombe)	NM_007194	15.5	2	2
PLA2G3	Phospholipase A2, group III	NM_015715	14.5	2	2
NF2	Neurofibromin 2	BC020257	11	1	2

The prioritized genes were ranked by the PANDA score (MD2/MD1).
[a] Number of GO features (0–3) of the gene (molecular function, biological process, and cellular component).
[b] Number of non-GO features (0–4) of the gene (amino acid sequence similarity, InterPro, EC number, and KEGG pathways).
[c] This gene was included in the training set in the subsequent analysis with new data.

Table 3B
Scores for *OSM* in relation to PC.

Parameter	Functional annotation	ID	Base relevance score	Proportion in gene set		Score to *OSM*[a]
				Training set	All human genes	
GO molecular function	Cytokine activity	GO:0005125 (5)[b]	2	0.0043	0.0018	23.12
GO biological process	Immune response	GO:0006955 (7)	2	0.0158	0.0168	107.05
	Regulation of cell growth	GO:0001558 (6)	2	0.0158	0.0020	
GO cellular component	Extracellular	GO:0005576 (3)	1.63	0.1348	0.0510	12.95

Among the seven parameters, details for the three concepts of GO terms are shown. The score of paralogy was 0 because *OSM* did not have a sequence similarity to any known PC-related gene. The scores of InterPro, EC number and KEGG pathway were also 0.
[a] Scores were calculated by Eqs. (2) and (3).
[b] The level in a nested hierarchical vocabulary is shown in parenthesis.

they may be common terms frequently found in disease-related genes. For instance, if the training sets include genes encoding proteins that belong to large superfamilies such as immunoglobulin-like domain containing proteins, too many members of the same superfamily may be prioritized as candidate genes for the disease. In order to overcome such possible biases of these methods, we developed a robust method in which we use multiple sources of gene annotation and leverage training sets consisting of known disease genes. Aerts et al. constructed Endeavour which prioritizes disease susceptibility candidate genes [19]. However, the problem of "background noise" in the functional terms was not resolved in their study. Using seven methods for computational prioritization of disease genes, Tiffin et al. (2006) generated a list of nine candidate genes for type 2 diabetes (T2D) common to six of the seven methods [20]. We have also identified three of the nine T2D candidate genes using PANDA system (data not shown), suggesting that our method can pick up reasonable candidate genes for multifactorial diseases.

Here, we have highlighted two examples (*TNFSF12* and *OSM*) that were selected with three GO annotation categories. These two genes, *TNFSF12* and *OSM*, had no positive scores for the other parameters (Tables 2B and 3B). The prioritized genes included another candidate gene for PC, kringle containing transmembrane protein 1 (*KREMEN1*) with the second highest PANDA score in 22q12 (Table 3A). *KREMEN1* has only two InterPro annotations that are shared with known PC-related genes including plasminogen activator urokinase (*PLAU*), plasminogen (*PLG*), hepatocyte growth factor (*HGF*), suppression of tumorigenicity 14 (*ST14*), and neuropilin 1 (*NRP1*) (data not shown). This suggests that a simultaneous usage of various features of human genes with GO annotation is effective for prioritization of disease-related genes.

To ensure that the prioritized genes are likely to be associated with the targeted disease, we surveyed subsequent reports in which a causative link to disease is shown. Interestingly, Godoy-Tundidor et al. reported that IL6 and OSM stimulate proliferation of prostate cancer cells, at least in part, through activation of the phosphatidylinositol 3-kinase signaling pathway [37]. Therefore, *OSM*, one of the predicted candidate genes for PC, may be a reasonable candidate gene. We checked if there is any newly reported gene with susceptibility to RA or PC by searching for the later releases of the public databases (OMIM, Entrez GENE, and PubMed). Then we prepared lists of known related genes for RA and PC again with new data. Between August 2004 and January 2008, 40 genes were newly described as RA-related genes in the OMIM database (Supplemental Table 7). When we conducted PANDA analysis by using this enhanced training set, eight genes of the 40 genes were selected in the training set via Entrez Gene. Three genes (*FLT1*, *IL1RN* and *EBI3*) of the remaining 32 genes were identified as candidate RA-susceptibility genes by our PANDA analysis (PANDA score > 21.2, Supplemental Table 7). However, the *LGALS3* gene encoding galectin-3 was not highly scored although some galectin superfamily members are involved in pathogenesis of RA. This may be because that the gene data for galectin proteins had little functional annotation. One of limitations of our approach is that genes with little annotation are less likely to be prioritized as candidate disease genes. The highly scored genes included *FLT1* (PANDA score = 115.5) whose gene product, FLT1, inhibits vascular endothelial growth factor (VEGF) response. The VEGF response may be involved in the pathogenesis of RA, because De Bandt et al. found that *VEGFA*, *FLT1*, and *KDR* are expressed in synovial cells from arthritic joints, using a transgenic mouse model of RA and antibodies to these three proteins [38]. Another gene prioritized was *IL1RN* (PANDA score = 33.8) whose gene product, IL1RN, inhibits IL1R binding by IL1-alpha and IL1-beta. Because these two cytokines (IL1-alpha and IL1-beta) are involved in both immune response and inflammatory response [39], *IL1RN* may be associated in the pathogenesis for RA. Recently, new susceptible loci for RA have been detected by GWAS [40–43], and many of them are genes involved in immune and inflammatory responses. Thus our analysis may have worked to predict biologically reasonable candidate genes.

We also searched for newly added PC-related genes in the OMIM database between August 2004 and January 2008, and found 57 additional PC-related genes that had not been reported in 2004. Then we conducted PANDA analysis again with the enhanced training set (Supplemental Table 8). Among these 57 genes, 17 genes were selected in the training set via Entrez Gene, and two genes, HNF1 homeobox B (*HNF1B*) and Eph receptor B2 (*EPHB2*), of the remaining 40 genes were identified as candidate genes by our analysis (in which PANDA score was larger than 14.1, a threshold value in this analysis). In particular, *EPHB2* showed the second highest score (PANDA score = 40.4) among the genes that are located at 1p36.1. This gene is of great interest because Huusko et al. identified mutations affecting translation in *EPHB2* in human prostate cancer cells [44]. Although reports of new susceptible loci for PC are increasing by GWAS, analysis of gene expression, and cancer genome sequencing [45–48], results of these studies include false positives. Our approach would be useful to find biologically relevant candidate genes from results by such large-scale analyses.

Although we showed that our approach can predict reasonable candidate disease genes, the approach can be extended in the following ways. First, by including text-mining techniques, we will be able to make enhanced training sets for other complex diseases through automated processes. This will allow researchers to select training sets for PANDA analysis without detailed disease-specific knowledge or manual annotation of literature. Second, an expansion of parameters (e.g. including protein–protein interaction, gene-gene interaction and gene expression) and a customized selection of parameters would increase the flexibility of the analysis and improve the sensitivity in prioritizing disease-related genes that do not have functional similarities with known disease genes or do not have detailed functional annotations. Such developments should further allow researchers to leverage their knowledge base and experience to efficiently prioritize disease-related candidate genes.

This approach, implemented in the PANDA system, is an efficient and cost-effective method to narrow down a large set of genes that are typically identified in microarray or mapping studies, into a smaller subset most likely to be associated with the disease. In addition, the PANDA system would be a useful tool to prioritize biologically relevant candidate genes with result of GWAS, large-scale gene expression data and genome sequencing to find causative variants [49,50].

4. Methods

4.1. Data of human genes

We retrieved data of human genes from the two databases, H-InvDB [6] and RefSeq [5]. H-InvDB (version 1.0, on 15th July 2002) contained 41,118 H-Inv cDNAs, and RefSeq contained a set of 37,488 human mRNA sequences that were available on September 1st, 2003. After merging the two datasets of human genes and removing the genes having no functional information, we created the H1-REFSEQ DB containing a total of 14,959 human genes. We used representative transcripts (one transcript for one gene) instead of all genes in H1-REFSEQ DB to remove the redundancy of cDNAs due to multiple forms of alternative splicing variants. Information of gene structure and functional annotation for the 14,959 human genes was retrieved from the H-InvDB [6].

4.2. Selection and curation of known related genes for the target diseases

Known related genes for the target diseases (RA and PC in this study) were searched from the OMIM database (www.ncbi.nlm.nih.gov/Omim) [29] and Entrez Gene (www.ncbi.nlm.nih.gov/gene), filtered and curated (see below), so that we can use them as the training sets of our prioritization analysis. We used MeSH terms (www.ncbi.nlm.nih.gov/ mesh) associated with the two diseases ('arthritis, rheumatoid', and 'rheumatoid, juvenile arthritis' for RA; 'prostatic neoplasms' and 'prostatic intraepithelial neoplasia' for PC) to scan all abstracts of papers cited in OMIM and Entrez Gene. We obtained 231 genes as possible RA-related genes and 728 genes for possible PC-related genes in August 2004.

Next, the two gene sets were filtered to select genes that have a series of curated references linked by PubMed IDs. We gave scores to these known disease genes based on the associated disease MeSH terms in their PubMed abstracts. The scores (1, 2, or 3) were assigned automatically according to the number of major MeSH topics that an abstract contained; no asterisk (1), asterisk for subheading (2), or asterisk for specific disease term (3). Then we manually annotated each reference whether the relationship between the gene and the target disease was mentioned by checking phenotype-specific terms. Each abstract of the article was reviewed by two or more persons to avoid possible problems by different interpretations and annotations between persons. We did not weight the articles according to approaches, sample numbers or ethnicity when we annotated the disease susceptibilities of genes. The information we collected included some genes whose role in the pathogenesis was examined in mouse models. We did not give scores to the studies without functional validation. Functionally-related genes for RA included those involved in inflammation of synovial membrane, degeneration of synovial joints or immune response. Functionally-related genes for PC included those involved in the proliferation of cells, signal transduction and transcription factors. After discarding genes that were not clearly related to the disease from a list of disease-related genes, we obtained 139 relevant genes for RA and 296 genes for PC (Supplemental Tables 1–2).

The selected known disease-related genes were used as the training set of the prioritization analysis with the relevance scores, which were automatically given and manually checked (mentioned above). Because one known disease gene may have multiple reference articles about susceptibility to the disease, we calculated w_i, the arithmetic mean of all baseline relevance scores for a particular gene i. This w_i was used as a parameter of disease susceptibility for a known related gene for the target disease.

4.3. Analysis of biological information of known related genes for target disease and all human genes

We selected genes that are relevant to the target disease from OMIM and Entrez Gene by querying disease-specific MeSH terms (downloaded on September 5, 2003). We then used these genes as a training set to analyze biological information for prioritization of candidates of disease-related genes. To utilize biological information for the analysis, annotated biological information on human genes from H-InvDB was analyzed with respect to the enrichment of specific biological terms in the known related genes for the target disease. The levels of enrichment of biological terms were converted into scores using several formulae (see below). The paralogy with the known related genes for the target disease was also examined based on similarity of amino acid sequence by using the BLAST program [51]. We gave a higher score for a gene that had sequence similarity to the known related gene for the target disease (see below). In total, we used seven kinds of biological feature of human genes (sequence similarity, InterPro annotation, EC number, three kinds of GO terms, and KEGG pathways).

4.4. Sequence similarity to known related genes for the target disease: the paralogy score

Paralogous genes that result from gene duplications have a similarity in their sequence and may share some related biological features [52]. Therefore, a paralogous gene with known disease gene may be more likely to be involved in the disease than other genes. Based on this concept, we calculated 'paralogy score' so that we give

a higher priority to a gene when the gene has a similarity to any known related gene for the target disease. For this purpose, we searched for genes that have sequence similarity to each known related gene for the target disease. We compared amino acid sequences by using BLAST [51], and 'paralogous' gene pairs were identified with significance cutoff values of *e-100*, *80*, *60* and *40*. When a gene *i* (no previous evidence of linkage to the disease) has a similarity to known disease gene *i'*, the paralogy score $S_{p,i}$ of a gene *i* (no previous evidence of linkage to the disease) has a positive value given by the equation:

$$S_{p,\,i} = -w_{i'}(1-\log_{10}[\text{e-value}])R, \tag{1}$$

where $w_{i'}$ is the baseline relevance score for the known disease gene *i'* with which the gene *i* is similar to. We set the weight *R* ($R_{100}=1$ for the gene pairs whose similarity significance was detected with *e-100*) for the cases that sequence similarity was detected by the lower significance levels (*R* values for *e-80*, *60* and *40* are given in Supplemental Table 9). When gene *i* does not have a similarity to any known related gene for the disease, this score was set to be 0.

4.5. Enrichment of functional terms: disease gene functional score

Genes associated with a disease may be prioritized through their functional similarities to known related genes for the disease or shared physiological pathway with known related genes with the disease. Here, we have made the assumption that certain groups of genes having related functions may be over or under-represented in a group of genes related to a specific disease. To test this assumption, we compared the frequencies of biological terms between the known related genes to the target disease and all the genes in H1-REFSEQ DB. The biological terms from the following resources were analyzed; InterPro identifiers (www.ebi.ac.uk/interpro/), EC numbers (www.chem.qmul.ac.uk/iubmb/enzyme/), KEGG pathways (www.genome.ad.jp/kegg/) [21], and gene ontology (GO) (www.geneontology.org/). For the known PC-related genes, "regulation of transcription, DNA-dependent" was the most frequent among the terms in "biological processes" of GO. However, this term was also frequent in all the cDNAs in H1-REFSEQ DB. From the disease-specific perspective for PC, "synaptic vesicle endocytosis" and "anti-apoptosis", were over-represented in the PC-related genes. As the prostate is an endocrine organ and secretes fluids, vesicle transport is an important component of its function. Apoptosis has also been extensively studied in cancer cells as a survival mechanism. Based on these observations we have extended the analysis to individual genes by quantifying and ranking the potential disease genes based on their levels of association to known disease-related genes through functional similarity or linkage to common physiological pathways. To carry out these comparisons we have utilized functional terms associated with genes, including InterPro identifiers, EC numbers, and KEGG pathway IDs, and expressed the proportion of each term for genes in the disease group compared with the frequencies among all genes in H1-REFSEQ DB. For a given functional term *j* in gene annotation, a score of relevance to the target disease was calculated by using the ratio of the proportion of genes with the term in known disease genes to the proportion of genes with the term in all human genes. This ratio was multiplied by the base score of disease relevance of the term (w_j) based on curation of the known disease genes (see above). Therefore, the score, S_{fj}, was calculate by the equation

$$S_{f,\,j} = w_j \times P_{d,\,j}/P_{a,\,j}, \tag{2}$$

where $P_{d,j}$ is the proportion of genes with the term in known disease genes, and $P_{a,j}$ is the proportion of genes with the term in all human genes. This score was used as an indicator of the specificity of a functional term in known disease genes for the target disease. For a gene whose disease relevance is unknown, the score was calculated by summing S_{fj} for all the functional terms with the gene.

4.6. Use of biological terms in gene ontology data: GO score

The gene ontology [22] is composed as "a nested hierarchical vocabulary" of biological terms in which 'general' parent terms have nested child terms that are more specific. We have taken advantage of the nested hierarchy present within the GO terminology to provide a scoring system for functional specificity based on the position of GO terms within the hierarchy. For example, cell (GO:0005623), is a first-level GO term under the 'cellular component' category, nucleus (GO:0005634), is a third-level term under the same category of "cellular component". Therefore, we gave a higher weight (P_{fj}) for a GO term *j* in a lower hierarchy. The score of GO term *j*, $S_{go,j}$, was calculated as a product of $P_{h,j}$ and S_{fj}:

$$S_{go,\,j} = P_{h,\,j} \times S_{f,\,j}. \tag{3}$$

The score of GO term for a gene was calculated as the sum of $S_{go,j}$ for all the GO terms for the three concepts (molecular function, biological processes, and cellular components) with the gene.

4.7. Prioritization of candidates of disease-related genes by discriminant analysis

Integration of biological information from human genes allowed us to evaluate similarities between a given gene and multiple sets of genes. Here, there are two groups of genes; the first group is known related genes to the target disease and the second group is all other genes. For a given gene tested, a distance between the gene and the known related genes for the target disease (group 1) was compared with another distance between the gene and all the other genes (group 2). To express the distance between a gene and the center of gravity of a group of genes, the Mahalanobis distance [34] was calculated. For each gene from the group 2, we examined whether the gene is closer to the group 1 than to the group 2 by using Mahalanobis distances and discriminant analysis. The Mahalanobis distance for a given gene *i* and the group 1 (MD1) was calculated as:

$$MD1 = \sqrt{(X_i-\mu_1)'Cov_1^{-1}(X_i-\mu_1)}\,, \tag{4}$$

where X_i is a vector for gene *i* consisting of scores of multiple parameters (seven parameters in this analysis), μ_1 is a vector of mean values in the group 1, and Cov_1^{-1} is an inverse of the covariance matrix of the observed values of the seven parameters for the group 1. The Mahalanobis distance between a query gene and the group 2 (MD2) was calculated in a same way. Then we judged whether a gene is closer to the group 1 (known related genes for the target disease) than the group 2 (all the other genes). Because we consider a gene showing smaller MD1 than MD2 to be a disease-susceptibility candidate, we calculated the ratio of MD2 to MD1 (PANDA score). We also calculated the PANDA score for each of known related genes to the target diseases (the training set) in the same way, and the average score in the training set was calculated. Using the average score for the training set as a threshold, we considered any gene to be a candidate gene of disease-susceptibility if the score was greater than the threshold. Our results obtained by the prioritization analysis of disease association (PANDA) are available at http://www.h-invitational.jp/panda/app.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2011.10.002.

Acknowledgments

We thank Drs. Peter Tonellato, Arek Kasprzyk, Teruyoshi Hishiki, Craig Gough, Makoto Shimada, and Naoki Nagata for their helpful discussion and comments on this study. We also thank all the members of the H-Invitational consortium and all the staff of JBIRC for construction of H-InvDB and the PANDA system. This work is financially supported by the Ministry of Economy, Trade and Industry of Japan (METI), the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT), the Japan Biological Informatics Consortium (JBiC), and National Institute of Advanced Industrial Science and Technology (AIST).

References

[1] A.L. Beaudet, 1998 ASHG presidential address. Making genomic medicine a reality, Am. J. Hum. Genet. 64 (1999) 1–13.
[2] J. Bell, The new genetics in clinical practice, BMJ 316 (1998) 618–620.
[3] J.P. Evans, C. Skrzynia, W. Burke, The complexities of predictive genetic testing, BMJ 322 (2001) 1052–1056.
[4] K. Finkler, C. Skrzynia, J.P. Evans, The new genetics and its consequences for family, kinship, medicine and medical genetics, Soc. Sci. Med. 57 (2003) 403–412.
[5] D.R. Maglott, K.S. Katz, H. Sicotte, K.D. Pruitt, NCBI's LocusLink and RefSeq, Nucleic Acids Res. 28 (2000) 126–128.
[6] T. Imanishi, T. Itoh, Y. Suzuki, C. O'Donovan, S. Fukuchi, K.O. Koyanagi, R.A. Barrero, T. Tamura, Y. Yamaguchi-Kabata, M. Tanino, K. Yura, S. Miyazaki, K. Ikeo, K. Homma, A. Kasprzyk, T. Nishikawa, M. Hirakawa, J. Thierry-Mieg, D. Thierry-Mieg, J. Ashurst, L. Jia, M. Nakao, M.A. Thomas, N. Mulder, Y. Karavidopoulou, L. Jin, S. Kim, T. Yasuda, B. Lenhard, E. Eveno, Y. Suzuki, C. Yamasaki, J. Takeda, C. Gough, P. Hilton, Y. Fujii, H. Sakai, S. Tanaka, C. Amid, M. Bellgard, F. Bonaldo Mde, H. Bono, S.K. Bromberg, A.J. Brookes, E. Bruford, P. Carninci, C. Chelala, C. Couillault, S.J. de Souza, M.A. Debily, M.D. Devignes, I. Dubchak, T. Endo, A. Estreicher, E. Eyras, K. Fukami-Kobayashi, G.R. Gopinath, E. Graudens, Y. Hahn, M. Han, Z.G. Han, K. Hanada, H. Hanaoka, E. Harada, K. Hashimoto, U. Hinz, M. Hirai, T. Hishiki, I. Hopkinson, S. Imbeaud, H. Inoko, A. Kanapin, Y. Kaneko, T. Kasukawa, J. Kelso, P. Kersey, R. Kikuno, K. Kimura, B. Korn, V. Kuryshev, I. Makalowska, T. Makino, S. Mano, R. Mariage-Samson, J. Mashima, H. Matsuda, H.W. Mewes, S. Minoshima, K. Nagai, H. Nagasaki, N. Nagata, R. Nigam, O. Ogasawara, O. Ohara, M. Ohtsubo, N. Okada, T. Okido, S. Oota, M. Ota, T. Ota, T. Otsuki, D. Piatier-Tonneau, A. Poustka, S.X. Ren, N. Saitou, K. Sakai, S. Sakamoto, R. Sakate, I. Schupp, F. Servant, S. Sherry, R. Shiba, N. Shimizu, M. Shimoyama, A.J. Simpson, B. Soares, C. Steward, M. Suwa, M. Suzuki, A. Takahashi, G. Tamiya, H. Tanaka, T. Taylor, J.D. Terwilliger, P. Unneberg, V. Veeramachaneni, S. Watanabe, L. Wilming, N. Yasuda, H.S. Yoo, M. Stodolsky, W. Makalowski, M. Go, K. Nakai, T. Takagi, M. Kanehisa, Y. Sakaki, J. Quackenbush, Y. Okazaki, Y. Hayashizaki, W. Hide, R. Chakraborty, K. Nishikawa, H. Sugawara, Y. Tateno, Z. Chen, M. Oishi, P. Tonellato, R. Apweiler, K. Okubo, L. Wagner, S. Wiemann, R.L. Strausberg, T. Isogai, C. Auffray, N. Nomura, T. Gojobori, S. Sugano, Integrative annotation of 21,037 human genes validated by full-length cDNA clones, PLoS Biol. 2 (2004) e162.
[7] C. Yamasaki, K. Murakami, J. Takeda, Y. Sato, A. Noda, R. Sakate, T. Habara, H. Nakaoka, F. Todokoro, A. Matsuya, T. Imanishi, T. Gojobori, H-InvDB in 2009: extended database and data mining resources for human genes and transcripts, Nucleic Acids Res. 38 (2010) D626–D632.
[8] A. Rzhetsky, D. Wajngurt, N. Park, T. Zheng, Probing genetic overlap among complex human phenotypes, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 11694–11699.
[9] K.I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, A.L. Barabasi, The human disease network, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 8685–8690.
[10] R. Karns, G. Zhang, N. Jeran, D. Havas-Augustin, S. Missoni, W. Niu, S.R. Indugula, G. Sun, Z. Durakovic, N.S. Narancic, P. Rudan, R. Chakraborty, R. Deka, Replication of genetic variants from genome-wide association studies with metabolic traits in an island population of the Adriatic coast of Croatia. Eur J. Hum. Genet. 19 341–346.
[11] M. Oti, M.A. Huynen, H.G. Brunner, Phenome connections, Trends Genet. 24 (2008) 103–106.
[12] M. Oti, B. Snel, M.A. Huynen, H.G. Brunner, Predicting disease genes using protein–protein interactions, J. Med. Genet. 43 (2006) 691–698.
[13] L. Franke, H. van Bakel, L. Fokkens, E.D. de Jong, M. Egmont-Petersen, C. Wijmenga, Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes, Am. J. Hum. Genet. 78 (2006) 1011–1025.
[14] E.A. Adie, R.R. Adams, K.L. Evans, D.J. Porteous, B.S. Pickard, SUSPECTS: enabling fast and effective prioritization of positional candidates, Bioinformatics 22 (2006) 773–774.
[15] F.S. Turner, D.R. Clutterbuck, C.A. Semple, POCUS: mining genomic sequence annotation to predict disease genes, Genome Biol. 4 (2003) R75.
[16] C. Perez-Iratxeta, P. Bork, M.A. Andrade, Association of genes to genetically inherited diseases using data mining, Nat. Genet. 31 (2002) 316–319.
[17] C. Perez-Iratxeta, M. Wjst, P. Bork, M.A. Andrade, G2D: a tool for mining genes associated with disease, BMC Genet. 6 (2005) 45.
[18] J. Freudenberg, P. Propping, A similarity-based method for genome-wide prediction of disease-relevant human genes, Bioinformatics (Oxford, England) 18 (Suppl. 2) (2002) S110–S115.
[19] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, Y. Moreau, Gene prioritization through genomic data fusion, Nat. Biotechnol. 24 (2006) 537–544.
[20] N. Tiffin, E. Adie, F. Turner, H.G. Brunner, M.A. van Driel, M. Oti, N. Lopez-Bigas, C. Ouzounis, C. Perez-Iratxeta, M.A. Andrade-Navarro, A. Adeyemo, M.E. Patti, C.A. Semple, W. Hide, Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes, Nucleic Acids Res. 34 (2006) 3067–3081.
[21] M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (2000) 27–30.
[22] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat. Genet. 25 (2000) 25–29.
[23] D.T. Felson, Arthritis and Allied Conditions: A Textbook of Rheumatology, Williams & Wilkins, 1996.
[24] D. Jawaheer, M.F. Seldin, C.I. Amos, W.V. Chen, R. Shigeta, C. Etzel, A. Damle, X. Xiao, D. Chen, R.F. Lum, J. Monteiro, M. Kern, L.A. Criswell, S. Albani, J.L. Nelson, D.O. Clegg, R. Pope, H.W. Schroeder Jr., S.L. Bridges Jr., D.S. Pisetsky, R. Ward, D.L. Kastner, R.L. Wilder, T. Pincus, L.F. Callahan, D. Flemming, M.H. Wener, P.K. Gregersen, Screening the genome for rheumatoid arthritis susceptibility genes: a replication study and combined analysis of 512 multicase families, Arthritis Rheum. 48 (2003) 906–916.
[25] A. Jemal, R. Siegel, E. Ward, T. Murray, J. Xu, M.J. Thun, Cancer statistics, 2007, CA Cancer J. Clin. 57 (2007) 43–66.
[26] G. Morganti, L. Gianferrari, A. Cresseri, G. Arrigoni, G. Lovati, Clinico-statistical and genetic research on neoplasms of the prostate, Acta Genet. Stat. Med. 6 (1956) 304–305.
[27] B.S. Carter, T.H. Beaty, G.D. Steinberg, B. Childs, P.C. Walsh, Mendelian inheritance of familial prostate cancer, Proc. Natl. Acad. Sci. U. S. A. 89 (1992) 3367–3371.
[28] J.R. Smith, D. Freije, J.D. Carpten, H. Gronberg, J. Xu, S.D. Isaacs, M.J. Brownstein, G.S. Bova, H. Guo, P. Bujnovszky, D.R. Nusskern, J.E. Damber, A. Bergh, M. Emanuelsson, O.P. Kallioniemi, J. Walker-Daniels, J.E. Bailey-Wilson, T.H. Beaty, D.A. Meyers, P.C. Walsh, F.S. Collins, J.M. Trent, W.B. Isaacs, Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome-wide search, Science 274 (1996) 1371–1374.
[29] A. Hamosh, A.F. Scott, J. Amberger, C. Bocchini, D. Valle, V.A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, Nucleic Acids Res. 30 (2002) 52–55.
[30] A. De Ketelaere, L. Vermeulen, J. Vialard, I. Van De Weyer, J. Van Wauwe, G. Haegeman, I. Moelans, Involvement of GSK-3beta in TWEAK-mediated NF-kappaB activation, FEBS Lett. 566 (2004) 60–64.
[31] Y. Okazaki, T. Sawada, K. Nagatani, Y. Komagata, T. Inoue, S. Muto, A. Itai, K. Yamamoto, Effect of nuclear factor-kappaB inhibition on rheumatoid fibroblast-like synoviocytes and collagen induced arthritis, J. Rheumatol. 32 (2005) 1440–1447.
[32] H. Okamoto, T. Iwamoto, S. Kotake, S. Momohara, H. Yamanaka, N. Kamatani, Inhibition of NF-kappaB signaling by fenofibrate, a peroxisome proliferator-activated receptor-alpha ligand, presents a therapeutic strategy for rheumatoid arthritis, Clin. Exp. Rheumatol. 23 (2005) 323–330.
[33] K. Okamoto, S. Makino, Y. Yoshikawa, A. Takaki, Y. Nagatsuka, M. Ota, G. Tamiya, A. Kimura, S. Bahram, H. Inoko, Identification of I kappa BL as the second major histocompatibility complex-linked susceptibility locus for rheumatoid arthritis, Am. J. Hum. Genet. 72 (2003) 303–312.
[34] J. Xu, L. Dimitrov, B.L. Chang, T.S. Adams, A.R. Turner, D.A. Meyers, R.A. Eeles, D.F. Easton, W.D. Foulkes, J. Simard, G.G. Giles, J.L. Hopper, L. Mahle, P. Moller, T. Bishop, C. Evans, S. Edwards, J. Meitz, S. Bullock, Q. Hope, C.L. Hsieh, J. Halpern, R.N. Balise, I. Oakley-Girvan, A.S. Whittemore, C.M. Ewing, M. Gielzak, S.D. Isaacs, P.C. Walsh, K.E. Wiley, W.B. Isaacs, S.N. Thibodeau, S.K. McDonnell, J.M. Cunningham, K.E. Zarfas, S. Hebbring, D.J. Schaid, D.M. Friedrichsen, K. Deutsch, S. Kolb, M. Badzioch, G.P. Jarvik, M. Janer, L. Hood, E.A. Ostrander, J.L. Stanford, E.M. Lange, J.L. Beebe-Dimmer, C.E. Mohai, K.A. Cooney, T. Ikonen, A. Baffoe-Bonnie, H. Fredriksson, M.P. Matikainen, T. Tammela, J. Bailey-Wilson, J. Schleutker, C. Maier, K. Herkommer, J.J. Hoegel, W. Vogel, T. Paiss, F. Wiklund, M. Emanuelsson, E. Stenman, B.A. Jonsson, H. Gronberg, N.J. Camp, J. Farnham, L.A. Cannon-Albright, D. Seminara, A combined genomewide linkage scan of 1,233 families for prostate cancer-susceptibility genes conducted by the international consortium for prostate cancer genetics, Am. J. Hum. Genet. 77 (2005) 219–229.
[35] Z. Culig, H. Steiner, G. Bartsch, A. Hobisch, Interleukin-6 regulation of prostate cancer cell growth, J. Cell. Biochem. 95 (2005) 497–505.
[36] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, New York, 1979.
[37] S. Godoy-Tundidor, I.T. Cavarretta, D. Fuchs, M. Fiechtl, H. Steiner, K. Friedbichler, G. Bartsch, A. Hobisch, Z. Culig, Interleukin-6 and oncostatin M stimulation of proliferation of prostate cancer 22Rv1 cells through the signaling pathways of p38 mitogen-activated protein kinase and phosphatidylinositol 3-kinase, Prostate 64 (2005) 209–216.
[38] M. De Bandt, M.H. Ben Mahdi, V. Ollivier, M. Grossin, M. Dupuis, M. Gaudry, P. Bohlen, K.E. Lipson, A. Rice, Y. Wu, M.A. Gougerot-Pocidalo, C. Pasquier, Blockade of vascular endothelial growth factor receptor I (VEGF-RI), but not VEGF-RII, suppresses joint destruction in the K/BxN model of rheumatoid arthritis, J. Immunol. 171 (2003) 4853–4859.
[39] R. Horai, A. Nakajima, K. Habiro, M. Kotani, S. Nakae, T. Matsuki, A. Nambu, S. Saijo, H. Kotaki, K. Sudo, A. Okahara, H. Tanioka, T. Ikuse, N. Ishii, P.L. Schwartzberg, R. Abe, Y. Iwakura, TNF-alpha is crucial for the development of autoimmune arthritis in IL-1 receptor antagonist-deficient mice, J. Clin. Invest. 114 (2004) 1603–1611.

[40] A. Suzuki, R. Yamada, X. Chang, S. Tokuhiro, T. Sawada, M. Suzuki, M. Nagasaki, M. Nakayama-Hamada, R. Kawaida, M. Ono, M. Ohtsuki, H. Furukawa, S. Yoshino, M. Yukioka, S. Tohma, T. Matsubara, S. Wakitani, R. Teshima, Y. Nishioka, A. Sekine, A. Iida, A. Takahashi, T. Tsunoda, Y. Nakamura, K. Yamamoto, Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis, Nat. Genet. 34 (2003) 395–402.

[41] The Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, Nature 447 (2007) 661–678.

[42] E.A. Stahl, S. Raychaudhuri, E.F. Remmers, G. Xie, S. Eyre, B.P. Thomson, Y. Li, F.A. Kurreeman, A. Zhernakova, A. Hinks, C. Guiducci, R. Chen, L. Alfredsson, C.I. Amos, K.G. Ardlie, A. Barton, J. Bowes, E. Brouwer, N.P. Burtt, J.J. Catanese, J. Coblyn, M.J. Coenen, K.H. Costenbader, L.A. Criswell, J.B. Crusius, J. Cui, P.I. de Bakker, P.L. De Jager, B. Ding, P. Emery, E. Flynn, P. Harrison, L.J. Hocking, T.W. Huizinga, D.L. Kastner, X. Ke, A.T. Lee, X. Liu, P. Martin, A.W. Morgan, L. Padyukov, M.D. Posthumus, T.R. Radstake, D.M. Reid, M. Seielstad, M.F. Seldin, N.A. Shadick, S. Steer, P.P. Tak, W. Thomson, A.H. van der Helm-van Mil, I.E. van der Horst-Bruinsma, C.E. van der Schoot, P.L. van Riel, M.E. Weinblatt, A.G. Wilson, G.J. Wolbink, B.P. Wordsworth, C. Wijmenga, E.W. Karlson, R.E. Toes, N. de Vries, A.B. Begovich, J. Worthington, K.A. Siminovitch, P.K. Gregersen, L. Klareskog, R.M. Plenge, Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci, Nat. Genet. 42 (2010) 508–514.

[43] Y. Kochi, Y. Okada, A. Suzuki, K. Ikari, C. Terao, A. Takahashi, K. Yamazaki, N. Hosono, K. Myouzen, T. Tsunoda, N. Kamatani, T. Furuichi, S. Ikegawa, K. Ohmura, T. Mimori, F. Matsuda, T. Iwamoto, S. Momohara, H. Yamanaka, R. Yamada, M. Kubo, Y. Nakamura, K. Yamamoto, A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility, Nat. Genet. 42 (2010) 515–519.

[44] P. Huusko, D. Ponciano-Jackson, M. Wolf, J.A. Kiefer, D.O. Azorsa, S. Tuzmen, D. Weaver, C. Robbins, T. Moses, M. Allinen, S. Hautaniemi, Y. Chen, A. Elkahloun, M. Basik, G.S. Bova, L. Bubendorf, A. Lugli, G. Sauter, J. Schleutker, H. Ozcelik, S. Elowe, T. Pawson, J.M. Trent, J.D. Carpten, O.P. Kallioniemi, S. Mousses, Nonsense-mediated decay microarray analysis identifies mutations of EPHB2 in human prostate cancer, Nat. Genet. 36 (2004) 979–983.

[45] M. Yeager, N. Orr, R.B. Hayes, K.B. Jacobs, P. Kraft, S. Wacholder, M.J. Minichiello, P. Fearnhead, K. Yu, N. Chatterjee, Z. Wang, R. Welch, B.J. Staats, E.E. Calle, H.S. Feigelson, M.J. Thun, C. Rodriguez, D. Albanes, J. Virtamo, S. Weinstein, F.R. Schumacher, E. Giovannucci, W.C. Willett, G. Cancel-Tassin, O. Cussenot, A. Valeri, G.L. Andriole, E.P. Gelmann, M. Tucker, D.S. Gerhard, J.F. Fraumeni Jr., R. Hoover, D.J. Hunter, S.J. Chanock, G. Thomas, Genome-wide association study of prostate cancer identifies a second risk locus at 8q24, Nat. Genet. 39 (2007) 645–649.

[46] J.S. Witte, Prostate cancer genomics: towards a new understanding, Nat. Rev. 10 (2009) 77–82.

[47] R. Takata, S. Akamatsu, M. Kubo, A. Takahashi, N. Hosono, T. Kawaguchi, T. Tsunoda, J. Inazawa, N. Kamatani, O. Ogawa, T. Fujioka, Y. Nakamura, H. Nakagawa, Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population, Nat. Genet. 42 (2010) 751–754.

[48] M.F. Berger, M.S. Lawrence, F. Demichelis, Y. Drier, K. Cibulskis, A.Y. Sivachenko, A. Sboner, R. Esgueva, D. Pflueger, C. Sougnez, R. Onofrio, S.L. Carter, K. Park, L. Habegger, L. Ambrogio, T. Fennell, M. Parkin, G. Saksena, D. Voet, A.H. Ramos, T.J. Pugh, J. Wilkinson, S. Fisher, W. Winckler, S. Mahan, K. Ardlie, J. Baldwin, J.W. Simons, N. Kitabayashi, T.Y. MacDonald, P.W. Kantoff, L. Chin, S.B. Gabriel, M.B. Gerstein, T.R. Golub, M. Meyerson, A. Tewari, E.S. Lander, G. Getz, M.A. Rubin, L.A. Garraway, The genomic complexity of primary human prostate cancer, Nature 470 (2011) 214–220.

[49] J. Chen, E.E. Bardes, B.J. Aronow, A.G. Jegga, ToppGene Suite for gene list enrichment analysis and candidate gene prioritization, Nucleic Acids Res. 37 (2009) W305–W311.

[50] A.R. Pico, I.V. Smirnov, J.S. Chang, R.F. Yeh, J.L. Wiemels, J.K. Wiencke, T. Tihan, B.R. Conklin, M. Wrensch, SNPLogic: an interactive single nucleotide polymorphism selection, annotation, and prioritization system, Nucleic Acids Res. 37 (2009) D803–D809.

[51] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[52] A. Baudot, B. Jacq, C. Brun, A scale of functional divergence for yeast duplicated genes revealed from analysis of the protein–protein interaction network, Genome Biol. 5 (2004) R76.

D38–D42 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 22 November 2011*
doi:10.1093/nar/gkr994

The DNA Data Bank of Japan launches a new resource, the DDBJ Omics Archive of functional genomics experiments

Yuichi Kodama, Jun Mashima, Eli Kaminuma, Takashi Gojobori, Osamu Ogasawara, Toshihisa Takagi, Kousaku Okubo and Yasukazu Nakamura*

Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization for Information and Systems, Yata, Mishima 411-8510, Japan

Received September 7, 2011; Revised October 17, 2011; Accepted October 18, 2011

ABSTRACT

The DNA Data Bank of Japan (DDBJ; http://www.ddbj.nig.ac.jp) maintains and provides archival, retrieval and analytical resources for biological information. The central DDBJ resource consists of public, open-access nucleotide sequence databases including raw sequence reads, assembly information and functional annotation. Database content is exchanged with EBI and NCBI within the framework of the International Nucleotide Sequence Database Collaboration (INSDC). In 2011, DDBJ launched two new resources: the 'DDBJ Omics Archive' (DOR; http://trace.ddbj.nig.ac.jp/dor) and BioProject (http://trace.ddbj.nig.ac.jp/bioproject). DOR is an archival database of functional genomics data generated by microarray and highly parallel new generation sequencers. Data are exchanged between the ArrayExpress at EBI and DOR in the common MAGE-TAB format. BioProject provides an organizational framework to access metadata about research projects and the data from the projects that are deposited into different databases. In this article, we describe major changes and improvements introduced to the DDBJ services, and the launch of two new resources: DOR and BioProject.

INTRODUCTION

Since 1987, the DNA Data Bank of Japan (DDBJ; http://www.ddbj.nig.ac.jp) has been operating public resources for biological information at the National Institute of Genetics (NIG) by providing submission, archive, search, download and analysis services. The objective of DDBJ is to support and promote the sharing and use of biological data as a public resource. The traditional DDBJ archive has collected annotated nucleotide sequences by collaborating with the EMBL-Bank at the European Bioinformatics Institute (EBI) and GenBank at the National Center for Biotechnology Information (NCBI) as partners in the International Nucleotide Sequence Database Collaboration (INSDC) (1). In the late 1990s, automated capillary platforms were introduced to many sequencing centres, which began to produce raw data at the genomic scale. Researchers soon recognized the importance of this raw data for the qualitative assessment of sequences and high-level re-analysis such as re-base-calling and re-assembly. DDBJ therefore installed the Trace Archive (http://trace.ddbj.nig.ac.jp/dta) to store the DNA sequence chromatograms (traces), base-calls and quality estimates from single-pass reads of various large-scale sequencing projects. More recently, massively parallel new generation sequencing platforms are producing biological sequencing data in unprecedented amounts. To ensure the wider scientific community can access and use these enormous amounts of data for scientific discovery and application, DDBJ established the 'Sequence Read Archive' (DRA; http://trace.ddbj.nig.ac.jp/dra) in 2008 to store and disseminate raw next-generation sequencing data (2). The traditional DDBJ and the DDBJ Trace and Sequence Read Archives share all submitted public data with the other members of INSDC.

This year, DDBJ established two new resources: the 'DDBJ Omics Archive' (DOR; http://trace.ddbj.nig.ac.jp/dor) and BioProject (http://trace.ddbj.nig.ac.jp/bioproject). The new-generation sequencing platforms produce data from gene expression, epigenomics and variation studies. In these functional genomics studies, the new-generation platforms are replacing the microarray because of their much greater resolving power and accuracy. These functional genomics data have been archived by the Gene Expression Omnibus (GEO) at

*To whom correspondence should be addressed. Tel: +81 55 981 6859; Fax: +81 55 981 6889; Email: yanakamu@genes.nig.ac.jp

NCBI (3) and by the ArrayExpress at EBI (4). DOR is a resource of high-throughput experimental data for functional genomics generated by sequencing as well as using microarray platforms. DOR accepts functional genomics experiments in the MAGE-TAB format used by ArrayExpress. All public DOR data are exchanged with ArrayExpress.

As new-generation technologies significantly increase sequencing throughput, data from a single project can be dispersed across more than one archival database. The DDBJ BioProject resource provides an overview of diverse types of projects and organizes the data from the projects deposited into the archival databases maintained by members of INSDC. The DDBJ submission portal 'D-way' was developed to be a single submission account to enable users to submit their data to all relevant DDBJ archival databases (traditional DDBJ, Omics Archive, Trace Archive, Sequence Read Archive and BioProject).

The public analysis service of new generation sequences is provided by the 'DDBJ Read Annotation Pipeline' (http://p.ddbj.nig.ac.jp) (2). This pipeline processes the uploaded raw sequence reads on cloud computers and supports data submissions to the DDBJ archival databases. DDBJ is developing into a more comprehensive and integrated public domain resource for biological research by providing archival databases and analysis services.

DDBJ ARCHIVAL DATABASES

DDBJ traditional assembled sequence archive

Here we describe the development of the DDBJ database as a member of INSD during the course of 2011, which is appended to last year's report (2). Between July 2010 and June 2011, 18 296 211 entries/13 576 228 536 bp were released from INSD as core traditional nucleotide flat files, except for whole-genome shotgun (WGS), mass sequence for genome annotation (MGA) and third party annotation (TPA) files (5). DDBJ contributed 12.7% of the entries and 10% of the base pairs added to the core nucleotide data of INSD during this period. Most of the nucleotide data were submitted by Japanese researchers; the rest came from China, Korea, Taiwan and other countries. DDBJ has also continually distributed sequence data related to patent applications transferred from the Japan Patent Office (JPO, http://www.jpo.go.jp) and the Korean Intellectual Property Office (KIPO, http://www.kipo.go.kr/en). In addition to the core nucleotide data, DDBJ has released a total of 2 228 313 WGS entries, 35 271 312 MGA entries, 660 TPA entries, 6374 TPA-WGS entries and 1272 TPA-CON entries as of 15 July 2011.

Noteworthy large-scale data released from DDBJ are listed in Table 1. The genome data of the vase tunicate (*Ciona intestinalis*) has been re-assembled and re-annotated from the first draft sequences. The first draft genome of *C. intestinalis* was published in 2002 by the international genome sequencing project (6). Main contributors to the sequencing project were from the US Department of Energy Joint Genome Institute and Kyoto University. The international collaboration for the *C. intestinalis* genome was discontinued and the genome sequences reported in INSD remained as the first version. However, the genome data were updated by the Kyoto University team independently of the sequencing collaboration (7). Since the updated version of the *C. intestinalis* genome has been used as a standard in the ascidian research community, it would be useful for researchers to share the updated genome data through INSD. Since the Kyoto University team submitted the updated genome data, DDBJ accepted and released the

Table 1. List of large-scale data released by DDBJ from July 2010 to June 2011

Type	Organism	Accession number [number of entries, number of base pairs (bp)]
GSS	Fission yeast (*S. pombe*)	FT321169-FT434719 (113 551 entries, 55 932 252)
EST	Fission yeast (*S. pombe*)	FY072959-FY174037 (101 079 entries, 51 492 242)
EST	Kale (*Brassica oleracea* var. *acephala*)	DK455950-DK575153 (119 204 entries, 93 446 677)
Metagenome	Uncultivated thermophilic archaeon (*Candidatus Caldiarchaeum subterraneum*)	Fosmid clones: AB201309, AP011633, AP011650, AP011675, AP011689, AP011708, AP011723, AP011724, AP011727, AP011745, AP011751, AP011786, AP011796, AP011826-AP011902 (90 entries, 3 092 718) Scaffold CON: BA000048 (1 entry, 1 680 938)
GSS	Uncultivated thermophilic archaeon (*Candidatus Caldiarchaeum subterraneum*)	AG993735-AG999698 (5964 entries, 3 194 765)
Genome	*Jatropha curcas*	HTG: AP011961-AP011977 (17 entries, 1 356 476) WGS: BABX01000001-BABX01150417 (150 417 entries, 285 858 490)
GSS	Mouse (*Mus musculus domesticus*)	GA000001-GA131507 (131 507 entries, 130 569 356)
GSS	Rice (*Oryza sativa* Japonica Group)	FT872362-FT932077 (59 716 entries, 26 152 462)
Full length cDNA	Domesticated barley (*Hordeum vulgare* subsp. *vulgare*)	AK353559-AK377172 (23 614 entries, 40 190 236)
GSS	African rice (*Oryza glaberrima*)	FT434720-FT872361 (437 642 entries, 206 317 130)
Genome (Chromosome 3H)	Domesticated barley (*Hordeum vulgare* subsp. *vulgare* cv. Haruna Nijo)	BACC01000001-BACC01008583 (8583 entries, 28 015 997)
Genome (Re-assemble and re-annotation)	Vase tunicate (*Ciona intestinalis*)	TPA-WGS: EAAA01000001-EAAA01006374 (6374 entries, 112 163 512) TPA-scaffold CON: HT000001-HT001272 (1272 entries, 115 226 814)
EST	Tammar wallaby (*Macropus eugenii*)	FY469875-FY736474 (265 835 entries, 212 531 387)

updated version as its first case of TPA re-assemble. TPA re-assemble was agreed among INSD partners, GenBank, EMBL-Bank and DDBJ, as a strategy to accept some important genomes that were re-assembled by researchers other than the original submitters. Moreover, DDBJ has released the following: Expressed sequence tags (EST) and genome survey sequences (GSS) of fission yeast (*Schizosaccharomyces pombe*) submitted by Osaka City University, Japan; the kale (*Brassica oleracea* var. *acephala*) EST submitted by Kirin Holdings Company, Ltd, Japan; Metagenome and GSS of an uncultivated thermophilic archaeon (*Candidatus Caldiarchaeum subterraneum*) submitted by Japan Agency for Marine-Earth Science and Technology, Japan; the biofuel crop (*Jatropha curcas*) genome submitted by Kazusa DNA Research Institute, Japan; the mouse (*Mus musculus domesticus*) GSS submitted by the RIKEN BioResource Center, Japan; GSS of Japanese rice (*Oryza sativa* Japonica Group) and African rice (*O. glaberrima*) and full length cDNAs of domesticated barley (*Hordeum vulgare* subsp. *vulgare*) submitted by National Institute of Agrobiological Sciences, Japan; WGS derived from chromosome 3H of domesticated barley (*H. vulgare* subsp. *vulgare* cv. Haruna Nijo) submitted by Okayama University, Japan; the tammar wallaby (*Macropus eugenii*) EST submitted by National Institute of Genetics, Japan.

DDBJ Sequence Read Archive

DRA accepts raw sequencing data submissions from new-generation sequencing platforms (8). New submitters should contact DRA for the creation of a D-way submission account and a secure data upload area. Sequencing data files should then be uploaded into the secure data upload area. Submitters can prepare and submit all their information, including study, experiment, sample, run and analysis metadata associated with the sequencing data by using the web-based submission tool 'MetaDefine'. In the MetaDefine intuitive interface, users can create metadata by simply entering the necessary pieces of information into the appropriate fields and can revise the content according to the feedback given by the metadata validation. Data search and download services have also been included in DRA (http://trace.ddbj.nig.ac.jp/DRASearch). Users can search all the public SRA data of DDBJ/EBI/NCBI by accession numbers, several other parameters (e.g. organism and study type) and free-text keywords. The sequencing data can be downloaded by FTP in either the SRA toolkit format or fastq format via ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra. DRA works closely with several large-sequencing projects. For example, in the Genome Science Project (http://www.genome-sci.jp), DRA serves as a data hub to support data flow from sequencing centres to participant researchers and the DDBJ Read Annotation Pipeline. DRA continues to support large-scale and collaborative projects by coordinating data flow, submission and analysis.

DDBJ Omics Archive

DDBJ has supported the Center for Information Biology Gene Expression Database (CIBEX; http://cibex.nig.ac.jp), which has collected more than 3700 hybridizations from microarray experiments since 2004 (9). DDBJ has participated in the Functional Genomics Data (FGED; http://www.mged.org) society which aims to assure that the investment in functional genomics data generates maximum public benefit by defining minimum information specifications for reporting data and standards for biological research data. CIBEX accepts microarray data compliant with the Minimum Information About a Microarray Experiment (MIAME) guideline developed by FGED, which is required to interpret and verify the results of an experiment (10). Many scientific journals require the deposition of MIAME-compliant microarray data to GEO at NCBI and ArrayExpress at EBI or CIBEX before publication of the relevant paper. In functional genomics research areas, new-generation sequencers play a key role in producing data by measuring the abundance of DNA and RNA molecules. The shift from microarrays to new generation sequencing platforms for functional genomics investigations has resulted in much greater resolving power and accuracy for such experiments. To accommodate the increasing volume of high-throughput functional genomics data from sequencing- and microarray-based technologies, DDBJ established a new resource called 'DDBJ Omics Archive' (DOR; http://trace.ddbj.nig.ac.jp/dor). DOR accepts functional genomics data (e.g. gene expression, small RNA profiling and epigenetics, etc.) in the MAGE-TAB format specified by FGED and used by ArrayExpress. DOR encourages microarray and sequencing data submissions compliant with MIAME and Minimum Information about a high-throughput SeQuencing Experiment (MINSEQE) guidelines, respectively. Submitters need to prepare the MAGE-TAB submission files by using spreadsheet applications or third-party tools until our own submission tool is released. DOR issues internationally recognized accession numbers in the E-DORD-n format to the submitted experiments. Public data are exchanged between ArrayExpress and DOR. Once new-generation sequencing data sets are deposited in DOR, the raw sequencing data are submitted to DRA with the SRA XML metadata generated from the supplied MAGE-TAB files. CIBEX will stop accepting new submissions once DOR becomes fully operational. Access to the data that have been deposited into CIBEX will be provided. The CIBEX data will be exported to DOR on the request of the data owner.

DDBJ BioProject

As new high-throughput sequencing technologies significantly increase the volume of data that can be generated, distinct types of data (e.g. genome, transcriptome and targeted locus data) from a single project can be deposited into different archival databases. The BioProject resource is a redesigned, expanded, replacement of the NCBI Genome Project resource. The redesign adds tracking of several data elements including more precise information

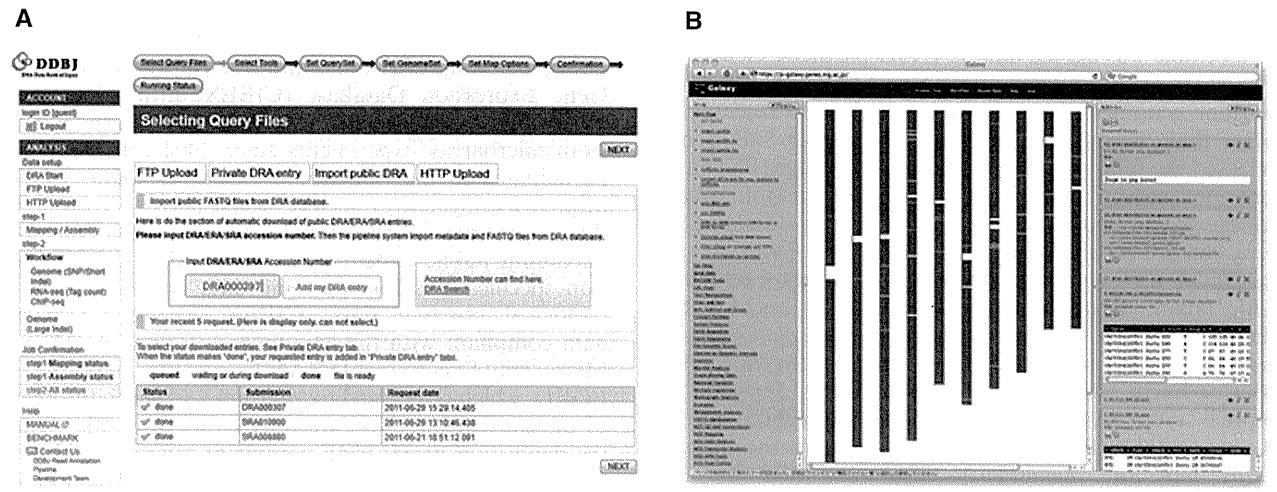

Figure 1. (**A**) DRA query import panel of DDBJ Pipeline. (**B**) Generated image for genomic locations of SNPs.

about a project's scope, material, objectives, funding source and general relevance categories. DDBJ BioProject issues accession numbers with the prefix 'PRJD' to the submitted projects and exchanges the released data with other members of INSDC. Data submitted to INSDC-associated databases cross-reference to the BioProject identifier to support navigation between the project and the project's datasets. BioProject enables users to retrieve a project's datasets across multiple archival databases. Related projects can be grouped under an umbrella project. An umbrella project may group projects that are part of a single collaborative effort but represent distinct studies that differ in methodology, sample material or resulting data type. The definition of a set of related data, a 'project', is very flexible and supports the need to define a complex project and various distinct sub-projects. Registration for a DDBJ BioProject accession is encouraged for the types of projects that result in submission of a very large volume of data, submissions from multiple members of a collaboration or submissions to multiple archival databases. A BioProject accession is required for submissions of microbial and eukaryotic genomes. At DDBJ, BioProject unites the records in the traditional DDBJ archive, Sequence Read Archive, Trace Archive and Omics Archive.

DDBJ ANALYSIS SERVICES

DDBJ Read Annotation Pipeline

The DDBJ Read Annotation Pipeline (DDBJ Pipeline, http://p.ddbj.nig.ac.jp) analyzes and annotates raw new-generation sequencing data and generates a template metadata file to support submission of analyzed data to the DDBJ archival databases (2). DDBJ Pipeline is a cloud computing-based analysis system in which users can access NIG supercomputers through a web application with a graphical user interface. DDBJ Pipeline consists of two processes: a basic process for reference sequence mapping and *de novo* assembly, and a successive analytical process for structural and functional annotations. In the basic process, popular mapping and assembly tools are available, and reference sequences can be retrieved by their INSDC accession numbers from the DDBJ database using SOAP access (11). Major enhanced functions are as follows: (i) DRA data import, (ii) New annotation workflows in the analytical process, (iii) Deletion policy of query and result data, (iv) MD5 checksum for download files and (v) Usage statistics information. DDBJ Pipeline implemented an automatic loader which loads DRA data from the public DDBJ FTP server in FASTQ file formats for query (Figure 1A). When loading is complete, the Pipeline system sends an e-mail notification to users. New workflows, genomic contig annotation, SNP detection (Figure 1B) and RNA-seq transcript annotation have been implemented in the analytical process. To save disk space for data storage, we set a policy to store query and result data up to 30 days. To verify the download process, MD5 checksums have been added to compressed download files. Usage statistics of job submissions and web access are displayed on the website.

ADDITIONAL INFORMATION

More information is available on the DDBJ website at http://www.ddbj.nig.ac.jp. News is delivered by really simple syndication (RSS), Twitter and mail magazines.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of all members of DDBJ for data collection, annotation and release and for software development. In particular, we thank Dr Hideki Nagasaki, Dr Satoshi Saruhashi, Takako Mochizuki, Naoko Sakamoto and Natsuko Sakakura for consultations on DRA and Pipeline, and Prof. Hideaki

Sugawara and Prof. Yoshio Tateno for support in the form of database maintenance and INSD collaboration.

FUNDING

Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) (management expense grant for National University Cooperation, to DDBJ); MEXT Grant-in-Aid for Scientific Research on Innovative Areas 'Genome Science' (to DDBJ Sequence Read Archive and DDBJ Pipeline, partial). Funding for open access charge: The DDBJ management expenses grant (from MEXT).

Conflict of interest statement. None declared.

REFERENCES

1. Karsch-Mizrachi,I., Nakamura,Y. and Cochrane,G. (2012) The international nucleotide sequence database collaboration. *Nucleic Acids Res.*, **40**, D33 D37.
2. Kaminuma,E., Kosuge,T., Kodama,Y., Aono,H., Mashima,J., Gojobori,T., Ogasawara,O., Okubo,K., Takagi,T. and Nakamura,Y. (2011) DDBJ progress report. *Nucleic Acids Res.*, **39**, D22 D27.
3. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets 10 years on. *Nucleic Acids Res.*, **39**, D1005 D1010.
4. Parkinson,H., Sarkans,U., Kolesnikov,N., Abeygunawardena,N., Burdett,T., Dylag,M., Emam,I., Farne,A., Hastings,E., Holloway,E. *et al.* (2011) ArrayExpress update an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **39**, D1002 D1004.
5. Cochrane,G., Bates,K., Apweiler,R., Tateno,Y., Mashima,J., Kosuge,T., Mizrachi,I.K., Schafer,S. and Fetchko,M. (2006) Evidence standards in experimental and inferential INSDC Third Party Annotation data. *OMICS*, **10**, 105 113.
6. Dehal,P., Satou,Y., Campbell,R.K., Chapman,J., Degnan,B., De Tomaso,A., Davidson,B., Di Gregorio,A., Gelpke,M., Goodstein,D.M. *et al.* (2002) The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science*, **298**, 2157 2167.
7. Satou,Y., Mineta,K., Ogasawara,M., Sasakura,Y., Shoguchi,E., Ueno,K., Yamada,L., Matsumoto,J., Wasserscheid,J., Dewar,K. *et al.* (2008) Improved genome assembly and evidence-based global gene model set for the chordate *Ciona intestinalis*: new insight into intron and operon populations. *Genome Biol.*, **9**, R152.
8. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54 D56.
9. Kodama,Y., Kaminuma,E., Saruhashi,S., Ikeo,K., Sugawara,H., Tateno,Y. and Nakamura,Y. (2010) Biological databases at DNA Data Bank of Japan in the era of next-generation sequencing technologies. *Adv. Exp. Med. Biol.*, **680**, 125 135.
10. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365 371.
11. Kwon,Y., Shigemoto,Y., Kuwana,Y. and Sugawara,H. (2009) Web API for biology with a workflow navigation system. *Nucleic Acids Res.*, **37**, W11 W16.

JVI
Journals.ASM.org

Production of Infectious Chimeric Hepatitis C Virus Genotype 2b Harboring Minimal Regions of JFH-1

Asako Murayama,[a] Takanobu Kato,[a] Daisuke Akazawa,[a] Nao Sugiyama,[a] Tomoko Date,[a] Takahiro Masaki,[a] Shingo Nakamoto,[b] Yasuhito Tanaka,[c] Masashi Mizokami,[d] Osamu Yokosuka,[b] Akio Nomoto,[e]* and Takaji Wakita[a]

Department of Virology II, National Institute of Infectious Diseases, Shinjuku-ku, Tokyo, Japan[a]; Department of Medicine and Clinical Oncology, Graduate School of Medicine, Chiba University, Chuo, Chiba, Japan[b]; Department of Virology and Liver Unit, Nagoya City University Graduate School of Medical Sciences, Kawasumi, Mizuho, Nagoya, Japan[c]; The Research Center for Hepatitis and Immunology, National Center for Global Health and Medicine, Ichikawa, Chiba, Japan[d]; and Department of Microbiology, Graduate School of Medicine, University of Tokyo, Bunkyo-ku, Tokyo, Japan[e]

To establish a cell culture system for chimeric hepatitis C virus (HCV) genotype 2b, we prepared a chimeric construct harboring the 5′ untranslated region (UTR) to the E2 region of the MA strain (genotype 2b) and the region of p7 to the 3′ UTR of the JFH-1 strain (genotype 2a). This chimeric RNA (MA/JFH-1.1) replicated and produced infectious virus in Huh7.5.1 cells. Replacement of the 5′ UTR of this chimera with that from JFH-1 (MA/JFH-1.2) enhanced virus production, but infectivity remained low. In a long-term follow-up study, we identified a cell culture-adaptive mutation in the core region (R167G) and found that it enhanced virus assembly. We previously reported that the NS3 helicase (N3H) and the region of NS5B to 3′ X (N5BX) of JFH-1 enabled replication of the J6CF strain (genotype 2a), which could not replicate in cells. To reduce JFH-1 content in MA/JFH-1.2, we produced a chimeric viral genome for MA harboring the N3H and N5BX regions of JFH-1, combined with a JFH-1 5′ UTR replacement and the R167G mutation (MA/N3H+N5BX-JFH1/R167G). This chimeric RNA replicated efficiently, but virus production was low. After the introduction of four additional cell culture-adaptive mutations, MA/N3H+N5BX-JFH1/5am produced infectious virus efficiently. Using this chimeric virus harboring minimal regions of JFH-1, we analyzed interferon sensitivity and found that this chimeric virus was more sensitive to interferon than JFH-1 and another chimeric virus containing more regions from JFH-1 (MA/JFH-1.2/R167G). In conclusion, we established an HCV genotype 2b cell culture system using a chimeric genome harboring minimal regions of JFH-1. This cell culture system may be useful for characterizing genotype 2b viruses and developing antiviral strategies.

Hepatitis C virus (HCV) is a major cause of chronic liver disease (5, 13), but the lack of a robust cell culture system to produce virus particles has hampered the progress of HCV research (2). Although the development of a subgenomic replicon system has enabled research into HCV RNA replication (15), infectious virus particle production has not been possible. Recently, an HCV cell culture system was developed using a genotype 2a strain, JFH-1, cloned from a fulminant hepatitis patient (14, 29, 32), thereby allowing investigation of the entire life cycle of this virus. However, several groups of investigators have reported genotype- and/or strain-dependent effects of some antiviral reagents (6, 17) and neutralizing antibodies (7, 25). Therefore, efficient virus production systems using various genotypes and strains are indispensable for HCV research and the development of antiviral strategies.

The JFH-1 strain is the first HCV strain that can efficiently produce HCV particles in HuH-7 cells (29). Other strains can replicate and produce infectious virus by HCV RNA transfection, but the efficiency is far lower than that of JFH-1 (24, 31). In the case of replication-incompetent strains, chimeric virus containing the JFH-1 nonstructural protein coding region is useful for analyses of viral characteristics (6, 9, 14, 23, 30, 31).

In this study, we developed a genotype 2b chimeric infectious virus production system using the MA strain (accession number AB030907) (19) harboring minimal regions of JFH-1 and cell culture-adaptive mutations that enhance infectious virus production.

MATERIALS AND METHODS

Cell culture. Huh7.5.1 cells (a kind gift from Francis V. Chisari) (32) and Huh7-25 cells (1) were cultured at 37°C in Dulbecco's modified Eagle's medium containing 10% fetal bovine serum under 5% CO_2 conditions. For follow-up study, RNA-transfected cells were passaged every 2 to 5 days depending on cell status.

Full-length genomic HCV constructs. Plasmids used in the analysis of genomic RNA replication were constructed based on pJFH1 (29) and pMA (19). For convenience, an EcoRI recognition site was introduced upstream of the T7 promoter region of pMA by PCR, and an XbaI recognition site was introduced at the end of the 3′ untranslated region (UTR). To construct MA/JFH-1, the EcoRI-BsaBI (nucleotides [nt] 1 to 2570; 5′ UTR to E2) fragment of pMA was substituted into pJFH1 (Fig. 1A). Replacement of the 5′ UTR was performed by exchanging the EcoRI-AgeI (nt 1 to 159) fragment. A point mutation in the core region (R167G) was introduced into MA chimeric constructs by PCR using the following primers: sense, 5′-TTA TGC AAC GGG GAA TTT ACC CGG TTG CTC T-3′; antisense, 5′-GGT AAA TTC CCC GTT GCA TAA TTT ATC CCG TC-3′. G167R substitution in the JFH-1 construct was performed by PCR using the following primers: sense, 5′-ATT ATG CAA CAA GGA ACC TAC CCG GTT TCC C-3′; antisense, 5′-GGT AGG TTC CTT GTT GCA TAA TTA ACC CCG TC-3′. Point mutations (L814S, R1012G, T1106A, and V1951A) were introduced into MA chimeric constructs by PCR using the following primers: L814S, 5′-GCT TAC GCC TCG GAC GCC GCT GAA CAA GGG G-3′ (sense) and 5′-AGC GGC GTC CGA GGC GTA AGC CTG CTG CGG C-3′ (antisense); R1012G, 5′-GAG GCT AGG TGG

Received 13 June 2011 Accepted 23 November 2011
Published ahead of print 7 December 2011
Address correspondence to Takaji Wakita, wakita@nih.go.jp.
* Present address: Institute of Microbial Chemistry, Shinagawa-ku, Tokyo, Japan.

doi:10.1128/JVI.05386-11

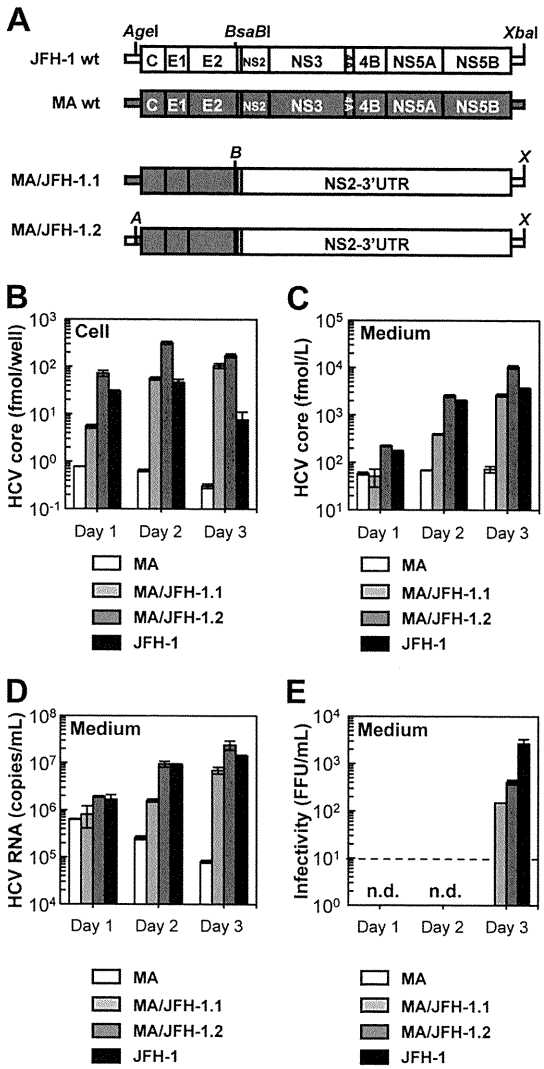

FIG 1 Replication and virus production by MA/JFH-1 chimeras in Huh7.5.1 cells. (A) Schematic structures of JFH-1, MA, and two MA/JFH-1 chimeras (MA/JFH-1.1 and MA/JFH-1.2). The junction of JFH-1 and MA in the 5′ UTR is an AgeI site, and the junction of MA and JFH-1 in the NS2 region is a BsaBI site. A, AgeI; B, BsaBI; X, XbaI. (B to E) Chimeric HCV RNA replication in Huh7.5.1 cells. HCV core protein level in cells (B) and culture medium (C) and HCV RNA levels in medium (D) and infectivity of culture medium (E) from HCV RNA-transfected Huh7.5.1 cells are shown. Ten micrograms of HCV RNA was transfected into Huh7.5.1 cells, and cells and culture medium were harvested on days 1, 2, and 3. n.d., not determined. Assays were performed three times independently, and data are presented as means ± standard deviation. Dashed line indicates detection limit. wt, wild type.

GGA AGT TCT GCT CGG CCC T-3′ (sense) and 5′-AGA ACT TCC CCA CCT AGC CTC GCG GAA ACC G-3′ (antisense); T1106A, 5′-CAG ATG TAC GCC AGC GCA GAG GGG GAC CTC-3′ (sense) and 5′-CTG CGC TGG CGT ACA TCT GGG TGA CTG GTC-3′ (antisense); and V1951A, 5′-GTG ACG CAG GCG TTA AGC TCA CTC ACA ATT ACC-3′ (sense) and 5′-TGA GCT TAA CGC CTG CGT CAC GCG CAG CGA G-3′ (antisense). To construct the MA chimeric virus harboring minimal regions of JFH-1 (MA/N3H+N5BX-JFH1), ClaI (nt 3930), EcoT22I (nt 5294), and BsrGI (nt 7782) recognition sites were introduced into pMA by site-directed mutagenesis. The 5′ UTR (EcoRI-AgeI), the region of the NS3 helicase (N3H; ClaI-EcoT22I), and the region of NS5B to 3′ X (N5BX; BsrGI-XbaI) were then replaced with the corresponding regions from JFH-1.

RNA synthesis, transfection, and determination of infectivity. RNA synthesis and transfection were performed as described previously (12, 22). Determination of infectivity was also performed as described previously, with infectivity expressed as the number of focus-forming units per milliliter (FFU/ml) (12, 22). When necessary, culture medium was concentrated 20-fold in Amicon Ultra-15 spin columns (100-kDa molecular-weight-cutoff; Millipore, Bedford, MA) in order to determine infectivity.

Quantification of HCV core protein and HCV RNA. In order to estimate the concentration of HCV core protein in culture medium, we performed a chemiluminescence enzyme immunoassay (Lumipulse II HCV core assay; Fujirebio, Tokyo, Japan) in accordance with the manufacturer's instructions. HCV RNA from harvested cells or culture medium was isolated using an RNeasy Mini RNA kit (Qiagen, Tokyo, Japan) or QiaAmp Viral RNA Minikit (Qiagen), respectively. Copy number of HCV RNA was determined by real-time quantitative reverse transcription-PCR (qRT-PCR), as described previously (28).

HCV sequencing. Total RNA in culture supernatant was extracted with Isogen-LS (Nippon Gene Co., Ltd., Tokyo, Japan). cDNA was synthesized using Superscript III Reverse Transcriptase (Invitrogen, Carlsbad, CA). cDNA was subsequently amplified with LA *Taq* DNA polymerase (TaKaRa, Shiga, Japan). Four separate PCR primer sets were used to amplify the fragments of nt 130 to 2909, 2558 to 5142, 4784 to 7279, and 7081 to 9634 covering the entire open reading frame and part of the 5′ UTR and 3′ UTR of the MA strain. Sequences of amplified fragments were determined directly.

Immunostaining. Infected cells were cultured on Multitest Slides (MP Biomedicals, Aurora, OH) and were fixed in acetone-methanol (1:1, vol/vol) for 15 min at −20°C. After a blocking step, infected cells were visualized with anti-core protein antibody (clone 2H9) (29) and Alexa Fluor 488 goat anti-mouse IgG (Invitrogen), and nuclei were visualized with 4′,6′-diamidino-2-phenylindole (DAPI).

Assessment of interferon sensitivity. Two micrograms of *in vitro* transcribed RNA was transfected into 3×10^6 Huh7.5.1 cells. Four hours after transfection, cells were placed in fresh medium or medium containing 0.1, 1, 10, 100, and 1,000 IU/ml of interferon α-2b (Intron A; Schering-Plough Corporation, Osaka, Japan). Culture medium was then harvested on day 3, and HCV core levels in the cells and in the medium were measured.

Statistical analysis. Significant differences were evaluated by Student's *t* test. A *P* value of <0.05 was considered significant.

RESULTS

Transient replication and production of 2b/2a chimeric virus. We first tested whether the MA strain (genotype 2b) (19) was able to replicate and produce infectious virus in cultured cells. When the *in vitro* transcribed RNA of MA was transfected into Huh7.5.1 cells, a highly HCV-permissive cell line, replication and virus production were not observed (Fig. 1A to C). We then tested whether 2b/2a chimeric RNA harboring the structural region (5′ UTR to E2) of the MA strain and the nonstructural region (p7 to 3′ UTR) of JFH-1 (Fig. 1A, MA/JFH-1.1) was able to replicate in the cells. After MA/JFH-1.1 RNA transfection, time-dependent accumulation of core protein in the cells (Fig. 1B) and culture medium (Fig. 1C) was observed, indicating that MA/JFH-1.1 RNA was able to replicate in the cells autonomously. HCV RNA levels in the medium were determined by qRT-PCR, and time-dependent increases in HCV RNA level were also observed (Fig. 1D). Infectious virus production was observed on day 3, but infectivity was 17.6-fold lower than that of JFH-1 (Fig. 1E).

In order to improve the level of infectious virus production, we tested another chimeric construct, MA/JFH-1.2, which contained an additional MA-to-JFH-1 replacement of the 5′ UTR (Fig. 1A),

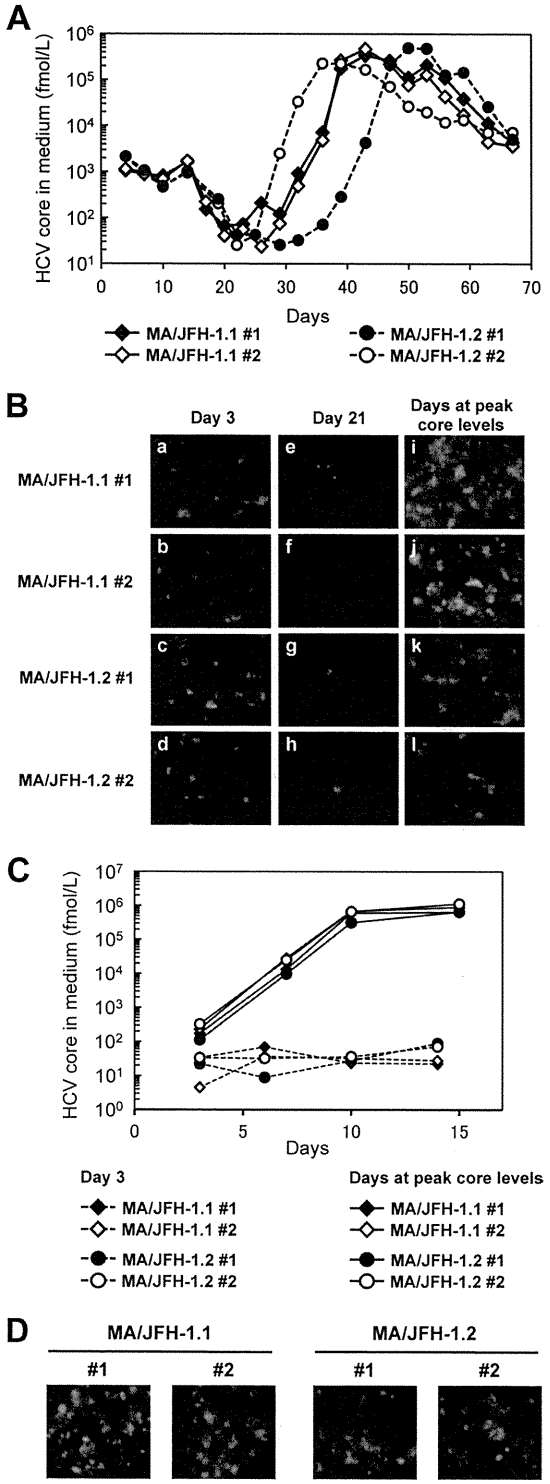

FIG 2 Long-term culture of MA/JFH-1.1 and MA/JFH-1.2 RNA-transfected cells. Ten micrograms of HCV RNA was transfected into Huh7.5.1 cells, and cells were passaged every 2 to 5 days, depending on cell status. Culture medium was collected after every passage, and HCV core protein levels were measured. Transfection was performed twice for each chimeric RNA (1 and 2 for each construct). (A) HCV core protein levels in culture medium from MA/JFH-1.1 and MA/JFH-1.2 RNA-transfected cells. (B) Immunostained cells at 3 days after transfection (a to d), at 21 days after transfection (e to h), and at the time of peak core levels (days 42 to 49). Infected cells were visualized with anti-core protein antibody (green), and nuclei were visualized with DAPI (blue). (C) Infection of naïve cells by culture medium at an MOI of 0.001. (D) Immunostained cells at 15 days after infection with medium at peak core protein levels (Fig. 2A) at an MOI of 0.001. Infected cells were visualized with anti-core antibody (green), and nuclei were visualized with DAPI (blue).

TABLE 1 HCV core protein levels and infectivity in culture medium immediately after RNA transfection (day 3) and after long-term culture (days 35 to 49)

Sample period and virus	Sample no.	Day no.[a]	HCV core (fmol/liter)	Infectivity (FFU/ml)
After transfection				
MA/JFH-1.1	1	3	1.06×10^3	5.00×10^1
	2	3	1.14×10^3	5.70×10^1
MA/JFH-1.2	1	3	2.14×10^3	7.30×10^1
	2	3	2.15×10^3	9.30×10^1
After long-term culture				
MA/JFH-1.1	1	42	3.38×10^5	1.62×10^5
	2	42	4.70×10^5	3.23×10^5
MA/JFH-1.2	1	35	2.27×10^5	1.61×10^5
	2	49	4.93×10^5	3.27×10^5

[a] For the long-term culture, the days are those of peak core protein levels.

as a 5′ UTR replacement from J6CF (genotype 2a) to JFH-1 enhanced virus production of chimeric J6CF virus harboring the region of NS2 to 3′ X of JFH-1 (J6/JFH-1) (A. Murayama et al., unpublished data). The core protein accumulation levels with MA/JFH-1.2 RNA-transfected cells were higher than those with MA/JFH-1.1 ($P < 0.05$) (Fig. 1B). Similarly, core protein and HCV RNA levels in the medium of MA/JFH-1.2 RNA-transfected cells were higher than those of MA/JFH-1.1 ($P < 0.05$) (Fig. 1C and D). Infectivity on day 3 was also higher than with MA/JFH-1.1 ($P < 0.05$) (Fig. 1E), indicating that the 5′ UTR of JFH-1 enhanced virus production. However, infectivity of medium from MA/JFH-1.2 RNA-transfected cells on day 3 remained 6.4-fold lower than that of JFH-1 although HCV RNA levels in the medium were similar to those of JFH-1 (Fig. 1D and E).

These results indicate that 2b/2a chimeric RNA is able to replicate autonomously in Huh7.5.1 cells and produce infectious virus although infectivity remains lower than that of JFH-1.

Assembly-enhancing mutation in core region introduced during long-term culture. Because MA/JFH-1.1 and MA/JFH-1.2 replicated efficiently but produced small amounts of infectious virus, we performed long-term culture of these RNA-transfected cells in order to examine whether these chimeric RNAs would continue replicating and producing infectious virus over the long term. We prepared two RNA-transfected cell lines for each construct (MA/JFH-1.1 and MA/JFH-1.2) as both of these replicated and produced infectious virus at different levels.

Immediately after transfection, core protein levels and infectivity in culture medium were low (1.06×10^3 to 2.15×10^3 fmol/liter and 5.00×10^1 to 9.30×10^1 FFU/ml, respectively) (Fig. 2A and Table 1) although a considerable number of core protein-positive cells were observed by immunostaining (Fig. 2B, frames a to d). Subsequently, core protein levels in the culture medium decreased gradually (Fig. 2A), and core protein-positive cells were rare (Fig. 2B, frames e to h). However, at 30 to 40 days