Variants in the SP110 gene are associated with genetic susceptibility to tuberculosis in West Africa

Kerrie Tosh*, Sarah J. Campbell*, Katherine Fielding†, Jackson Sillah‡, Boubacar Bah§, Per Gustafson¶, Kebba Manneh‖, Ida Lisse¶, Giorgio Sirugo‡, Steve Bennett†**, Peter Aaby¶, Keith P. W. J. McAdam†, Oumou Bah-Sow§, Christian Lienhardt‡††, Igor Kramnik‡‡, and Adrian V. S. Hill*§§

*Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, United Kingdom; †London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom; ‡Medical Research Council Laboratories, Fajara, The Gambia; §Programme National de Lutte Anti-Tuberculeuse, BP 634, Conakry, Republic of Guinea; ¶Projecto de Saude Bandim, Danish Epidemiology Service Centre, Statens Serum Institut, Bissau, Guinea; ‖National Tuberculosis/Leprosy Control Programme, Banjul, The Gambia; ††Institut de Recherche pour le Développement, Route des Peres Maristes, Hann, BP1386, Dakar, Senegal; and ‡‡Department of Immunology and Infectious Diseases, Harvard School of Public Health, 667 Huntington Avenue, Boston, MA 02115

Communicated by Barry R. Bloom, Harvard School of Public Health, Boston, MA, April 28, 2006 (received for review November 17, 2005)

The *sst1* locus has been identified in a mouse model to control resistance and susceptibility of *Mycobacterium tuberculosis* infection. Subsequent studies have now identified *Ipr1* (intracellular pathogen resistance 1) to be the gene responsible. *Ipr1* is encoded within the *sst1* locus and is expressed in the tuberculosis lung lesions and macrophages of *sst1*-resistant, but not *sst1*-susceptible mice. We have therefore examined the closest human homologue of *Ipr1*, SP110, for its ability to control susceptibility to *M. tuberculosis* infection in humans. In a study of families from The Gambia we have identified three polymorphisms that are associated with disease. On examination of additional families from Guinea-Bissau and the Republic of Guinea, two of these associations were independently replicated. These variants are in strong linkage disequilibrium with each other and lie within a 31-kb block of low haplotypic diversity, suggesting that a polymorphism within this region has a role in genetic susceptibility to tuberculosis in humans.

association study | murine genetics | macrophage

Approximately one-third of the world's population is thought to be infected with *Mycobacterium tuberculosis*, resulting in ≈1.7 million deaths in 2001 (1). The majority of individuals infected with *M. tuberculosis* remain asymptomatic and noninfectious; however, 10% will go on to develop active disease. The factors determining an individual's risk of infection and development of active disease are multifactorial and involve host–pathogen interactions and environmental components. Several lines of evidence also indicate that host genetics has a strong role, such as monozygotic twins having a higher concordance rate for tuberculosis than dizygotic twins and clear racial differences in the risk of developing disease (2).

There are many approaches that could be taken to find genes that are involved in genetic susceptibility, including candidate gene and linkage studies, both of which have showed some success. However, it was the discovery that a single locus could mediate the susceptibility of mice to *Mycobacterium bovis* bacillus Calmette–Guérin infection that has perhaps had the biggest impact on mycobacterial genetics to date (3). The gene at this locus, identified as *Nramp1*, does not only control susceptibility to bacillus Calmette–Guérin but it is also able to mediate susceptibility to *Salmonella typhimurium*, *Leishmania donovani*, and other mycobacterial species such as *Mycobacterium lepraemurium* and *Mycobacterium intracellulare* in mice. Although, some associations between the human *NRAMP1* gene and *M. tuberculosis* infection have been detected (4) there is no evidence that this locus controls tuberculosis infection in mice. More recently, however, a study of inbred mice identified the *sst1* (susceptibility to tuberculosis 1) locus, which controls progression of tuberculosis in a lung-specific manner after infection with *M. tuberculosis* and virulent *M. bovis* strains (5). Both *sst1* and *Nramp1* are located within close proximity to each other on chromosome 1 of the mouse genome; however, the phenotypic effects of *sst1* have been found to be distinct from that of *Nramp1*.

Further fine-mapping and expression studies of the *sst1* locus have now been carried out to identify the gene responsible for controlling *M. tuberculosis* infection in mice (6). The gene identified, *Ipr1* (intracellular pathogen resistance 1), has been found to be strongly expressed in tuberculosis lung lesions and macrophages of *sst1*-resistant, but not *sst1*-susceptible mice. In addition, the *in vitro* expression of the *Ipr1* transgene in macrophages was able to reproduce the major effects seen at the *sst1* locus, such as the ability to control *M. tuberculosis* growth and switch the infected cells from necrotic to apoptotic cells.

As *Ipr1* plays a major role in the outcome of tuberculosis infection in the mouse model, we examined polymorphisms in *SP110*, the nearest homologous gene in humans, for their ability to control *M. tuberculosis* infection by using family data from three West African populations.

Results and Discussion

Using the mouse model Kramnik *et al.* (5) have found that the *sst1* locus confers susceptibility to tuberculosis. A subsequent study found that *Ipr1* was the gene responsible for this phenotype (6); therefore, we examined the closest human homologue of *Ipr1*, SP110, for its ability to control susceptibility to *M. tuberculosis* infection in humans.

A total of 27 SNPs in the SP110 gene were examined in 219 families from The Gambia (Table 1). Of these, 6 were not polymorphic and rs3948463 had a minor allele frequency of <1% so it was not included in the analysis. The remaining 20 polymorphisms were analyzed by using transmission disequilibrium testing (TDT) (Table 2), and 3 were found to be associated (rs2114592, $P = 0.02$; sp110int10, $P = 0.02$; and rs3948464, $P = 0.01$). For each of the polymorphisms it was the most common allele that was found to be transmitted more times than expected to the affected offspring. To confirm these associations in an independent set of samples, rs2114592, sp110int10, and rs3948464 were examined in an additional 99 families from the Republic of Guinea and 102 families from Guinea-Bissau (Table 3). The C allele of rs2114592, associated in The Gambia, was also associated with disease susceptibility in the Republic of Guinea and Guinea-Bissau, and when all three populations were analyzed together this was also significant ($P = 0.000005$). The C allele of r3948464 was associated with susceptibility in The Gambia. This same allele was found to be associated with disease susceptibility in the Republic of Guinea; however, the result was

Conflict of interest statement: No conflicts declared.

Abbreviations: TDT, transmission disequilibrium testing; LD, linkage disequilibrium.

**Deceased March 27, 2003.

§§To whom correspondence should be addressed. E-mail: adrian.hill@imm.ox.ac.uk.

www.pnas.org/cgi/doi/10.1073/pnas.0603340103

Table 1. The location of each SNP within the *SP110* gene (GenBank accession no. NT_005403.13)

SNP	NT_005403.13	Location in gene	Amino acid change	Base change	Comment
rs1346311	81241720	Intron 1		C/T	
rs1966555	81241267	Intron 1		A/G	
rs1427294	81237781	Intron 3		T/C	
rs3177554	81237058	Exon 4	Trp122Arg	C/T	
rs9061	81236443	Exon 5	Gly207Lys	A/G	
rs1047254	81236427	Exon 5	Glu212Gly	G/A	Not polymorphic
rs3769839	81235958	Intron 5		T/C	
rs3820974	81235622	Intron 5		T/G	
rs3769838	81235524	Exon 6	Met249Val	A/G	Not polymorphic
rs2114592	81234539	Intron 6		T/C	
rs1129425	81234008	Exon 7	Glu267Gly	A/G	Not polymorphic
rs12467	81233979	Exon 7	Gly277Arg	A/G	Not polymorphic
rs1365776	81232042	Exon 8	Gly299Arg	G/A	
rs2303540	81227053	Intron 8		G/C	
rs1469345	81220525	Intron 10		A/G	
rs930031	81218766	Intron 10		C/T	
sp110int10	81210204	Intron 10		A/G	Identified through sequencing
rs3948464	81210048	Exon 11	Leu425Ser	C/T	
rs2114591	81209902	Intron 11		T/C	
rs958978	81209319	Intron 11		A/G	
rs1427292	81204721	Intron 12		A/G	
rs957683	81202709	Intron 12		T/C	
rs959506	81202153	Intron 12		T/G	Not polymorphic
rs1804027	81201609	Exon 14	Met523Thr	C/T	
rs1004869	81198604	Intron 15		A/C	
rs2278198	81196688	Intron 16		A/G	Not polymorphic
rs3948463	81196193	Exon 17	Met579Ile	A/G	Frequency < 1%

Where the variant is nonsynonymous the amino acid changes are given.

not statistically significant in Guinea-Bissau, although the trend was in the same direction. The combined analysis of all three populations found rs3948464 to be significantly associated with disease ($P = 0.0002$). In the analysis of sp110int10 we saw that the A allele was associated with disease susceptibility in The Gambia, but this finding was not replicated in the other popu-

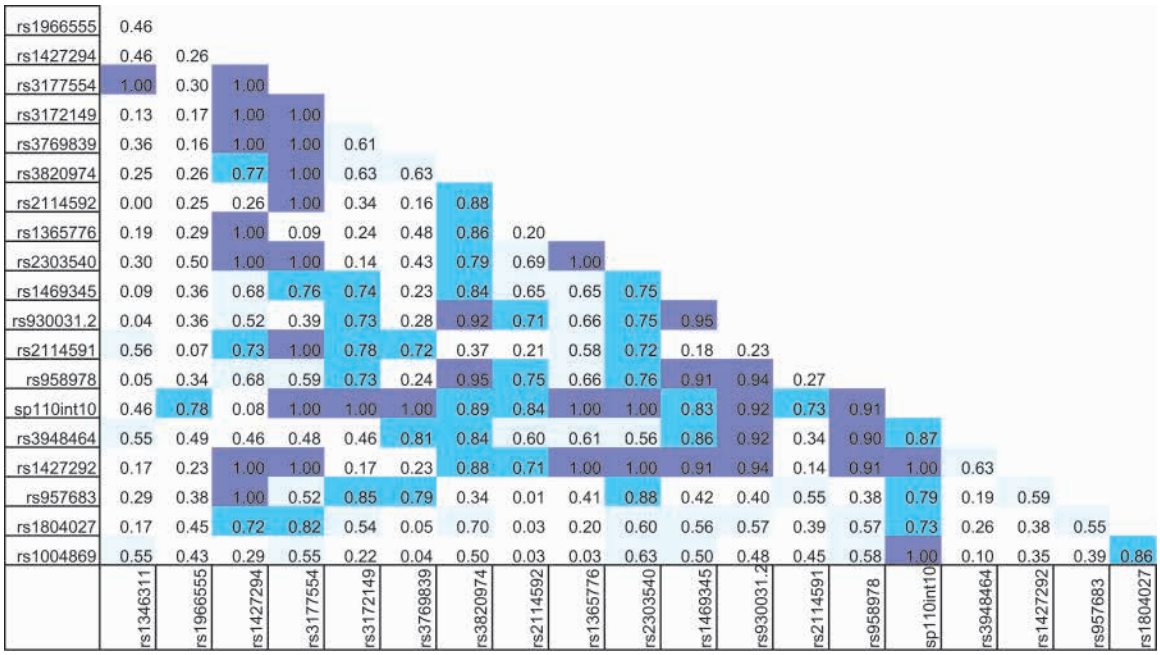

Fig. 1. Pairwise LD between SP110 markers as measured by the D′ statistic. D′ values between 0.5 and 0.7 are shaded light blue, values between 0.7 and 0.9 are medium blue, and values >0.9 are dark blue.

Table 2. TDT results of the polymorphic *SP110* polymorphisms in Gambian families

SNP	Allele	Allele frequency, %	Observed	Experiment	χ^2	*P*
rs1346311 (*N* = 207)	C	78.9	324	326.36	0.2	
	T	21.1	90	87.64		
rs1966555 (*N* = 162)	A	64.0	203	209.31	1.41	
	G	36.0	121	114.69		
rs1427294 (*N* = 172)	T	98.4	339	338.93	0	
	C	1.6	5	5.07		
rs3177554 (*N* = 195)	C	98.6	383	384.63	1.26	
	T	1.4	7	5.37		
rs9061 (*N* = 179)	G	95.5	341	340.35	0.06	
	A	4.5	17	17.65		
rs3769839 (*N* = 190)	T	80.0	305	304.15	0.03	
	C	20.0	75	75.85		
rs3820974 (*N* = 192)	G	51.0	201	196.4	0.58	
	T	49.0	183	187.6		
rs2114592 (*N* = 139)	C	86.4	251	243.80	5.49	0.02
	T	13.6	27	34.20		
rs1365776 (*N* = 189)	A	94.6	356	358.73	1.05	
	G	5.4	22	19.27		
rs2303540 (*N* = 205)	C	94.3	389	387.73	0.2	
	G	5.7	21	22.27		
rs1469345 (*N* = 195)	A	70.8	284	275.67	2.45	
	G	29.2	106	114.33		
rs930031 (*N* = 200)	C	72.5	295	290.11	0.84	
	T	27.4	105	109.89		
sp110int10 (*N* = 211)	A	96.9	414	408.73	5.44	0.02
	G	3.1	8	13.27		
rs3948464 (*N* = 199)	C	84.4	347	335.35	6.26	0.01
	T	15.6	51	62.65		
rs2114591 (*N* = 205)	C	57.8	240	253.62	0.49	
	T	42.2	170	174.38		
rs958978 (*N* = 204)	A	73.0	306	297.28	2.73	
	G	27.0	102	110.72		
rs1427292 (*N* = 200)	G	90.1	357	359.73	0.55	
	A	9.9	43	40.27		
rs957683 (*N* = 176)	C	68.1	240	241.08	0.04	
	T	31.9	112	110.92		
rs1804027 (*N* = 180)	T	80.4	293	289.87	0.45	
	C	19.6	67	70.13		
rs1004869 (*N* = 149)	C	90.0	271	267.91	0.87	
	A	10.0	27	30.09		

N = number of families in the analysis. Only $P < 0.05$ are shown.

lations, although a trend could be seen in the Republic of Guinea families. The minor allele frequency of sp110int10 ranges from 0.8% in the Republic of Guinea to 4% in Guinea-Bissau, indicating that this variant may be under different selection pressures or occur on a different genetic background in these populations.

It is not known whether any of the variants examined here are functional. SNP rs3948464 occurs in exon 11 of the gene and is a nonsynonymous change (leucine to serine). However, the alteration does not occur in the SP100 like SAND (Sp100, AIRE-1, NucP41/75, and DEAF-1/suppressin), plant homeobox, or bromodomains, and it is not conserved between the SP140 and SP110 sequences (7). Therefore, it is difficult to determine what the functional relevance of the variant might be. Many of the other variants, such as the associated SNPs sp110int10 and rs2114592, occur within intronic regions, and as SP110 isoforms are known to exist it is possible that they could have a role in alternative splicing. It is also possible that none of the associated variants are actually involved in controlling susceptibility directly, and it is another variant in the region, in linkage disequilibrium (LD) with the associated markers, which is the functional polymorphism.

To examine LD across the SP110 gene two approaches have been taken. The first was to calculate pairwise LD statistics for each of the markers (Fig. 1), and the second was to construct a map of haplotype diversity (Fig. 2). We used only the information from The Gambia as all of the polymorphisms were examined in this collection. Both types of analysis showed very similar results. All three associated SNPs in The Gambia are in strong LD with each other and lie within a 31-kb block of low haplotype diversity, suggesting that a polymorphism within this region has a role in genetic susceptibility to tuberculosis.

TDT analysis of haplotypes containing rs2114592, sp110int10, and rs3948464 was also carried out. It was found that the sp110int10 polymorphism made no contribution to the haplotypes' ability to control susceptibility or protection (data not shown). Using only rs2114592 and rs3948464 the common C/C haplotype was found to confer susceptibility in all three populations when analyzed separately and together (Table 4; all populations combined $P = 0.000005$). Although the two variants are in LD ($D' = 0.6$) they are not predictive of each other, again suggesting that the variants themselves or one occurring on the same haplotype as rs2114592 and rs3948464 are the functional variants responsible for regulating tuberculosis susceptibility in humans.

Table 3. TDT of rs2114592, sp100int10, and rs3948464 in families from the Republic of Guinea and Guinea-Bissau

Haplotype	Population	Allele	Allele frequency, %	Observed	Experiment	χ^2	*P*
rs2114592	Guinea-Bissau	C	82	152	143	9.36	0.002
	(*N* = 87)	T	18	22	31		
	Republic of Guinea	C	81	124	116.19	6.91	0.009
	(*N* = 72)	T	19	20	27.8		
	All	C	84	527	503.32	20.77	5.16E-06
	(*N* = 298)	T	16	69	92.68		
sp110int10	Guinea-Bissau	A	96	166	166.83	0.25	
	(*N* = 87)	G	4	8	7.16		
	Republic of Guinea	A	99.2	150	148.8	2.98	
	(*N* = 75)	G	0.8	0	1.19		
	All	A	97	730	724.27	3.91	0.048
	(*N* = 373)	G	3	16	21.73		
rs3948464	Guinea-Bissau	C	84	144	140.02	1.88	
	(*N* = 83)	T	16	22	25.98		
	Republic of Guinea	C	86	150	143.15	5.89	0.015
	(*N* = 83)	T	14	16	22.85		
	All	C	85	641	618.41	13.41	0.0002
	(*N* = 365)	T	15	89	111.59		

The combined analysis of families from all three West African countries is also shown (All). *N* = number of families in the analysis. Only $P < 0.05$ are shown.

SP110 is a component of the multiprotein nuclear body complex and is expressed at high levels in human peripheral blood leukocytes and the spleen and at lower levels in many other tissues such as the lung (7). The function of SP110 is not fully understood but it is thought to have a role in the differentiation of myeloid cells and function as a transcriptional coactivator (7). More recent studies have identified an isoform of SP110, SP110b, to be a transcriptional coactivator that binds to the hepatitis C core protein and negatively regulates retinoic acid receptor α-mediated transcription (8). A variant of the SP110 gene has also been found to be associated with the course of hepatitis C infection in a Japanese population (IMS-JST013416) (9). This SNP corresponds to rs1804027 in this study, with which we did not find an association. However, the two results may be a reflection of the different ethnicities and haplotype structures being examined, and it is tempting to speculate that the there is a single variant within the SP110 gene that could be acting to alter susceptibility to tuberculosis, hepatitis C, and perhaps other infectious agents.

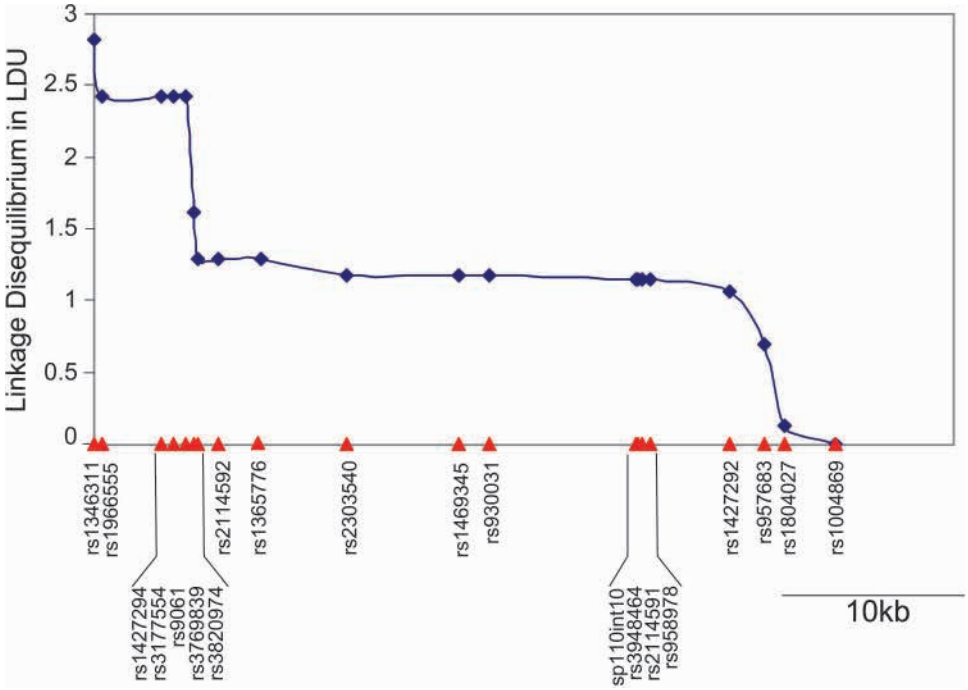

Fig. 2. LDMAP analysis of the *SP110* gene. LD maps are scaled in LD units (LDUs) against the physical map of the markers. Plateaus are a reflection of low haplotype diversity.

Table 4. TDT results of rs2114592 and rs3948464 haplotypes

Population	rs2114952/rs3948464 haplotype	Observed	Experiment	χ^2	*P*
Gambia	C/C	350.57	338.81	6.02	0.014
(*N* = 208)	T/T	39.77	49.06	5.68	0.017
Guinea-Bissau	C/C	153.74	144.09	8.04	0.005
(*N* = 91)	T/T	20.16	23.27	1.29	
Republic of Guinea	C/C	139.45	129.87	8.31	0.004
(*N* = 83)	T/T	11.34	17.67	6.49	0.01
All	C/C	643.04	612.18	20.98	4.65E-06
(*N* = 382)	T/T	71.19	90.31	12.57	0.0004

The combined analysis of The Gambia, the Republic of Guinea, and Guinea-Bissau is shown (All). *N* = number of families in the analysis. Only $P < 0.05$ are shown.

Nuclear body proteins, including SP110, are induced by IFN, suggesting they have a role in the IFN response mechanism. Mendelian susceptibility to atypical mycobacterial disease is an extremely rare group of conditions, and the genes identified to date all have involved the IFN-γ pathway (10). Studies in general populations have also identified IFN-γ polymorphisms as having a role in complex predisposition to tuberculosis (11, 12), indicating the molecule's importance in immunity to mycobacteria. In addition, Kramnik *et al.* (6) have demonstrated that *Ipr1* regulates the balance between necrosis and apoptosis of the *M. tuberculosis*-infected cells *in vitro*, and they postulate that factors such as IFN, which are produced during infection, may lead to the *Ipr1* switch in the mechanism of cell death.

Two major loci controlling mycobacterial infection in mice have been identified and subsequently found to have a role in human tuberculosis infection: NRAMP1 (4) and now SP110. With some putative tuberculosis susceptibility loci there has been difficulty in replicating associations possibly because of problems with stratification in the original case-control study. However, the family-based design used here should be able to avoid such problems of stratification. The identification of a further susceptibility locus for tuberculosis based on mouse genetics highlights the utility of using a variety of approaches for identifying genes involved in susceptibility to complex diseases.

Materials and Methods

Patient Samples. DNA samples from newly detected smear-positive pulmonary tuberculosis cases and their family members were collected from three West African countries as described (13, 14). In total, 420 index cases had more than one family member available for genotyping: 219 from The Gambia, 99 from the Republic of Guinea, and 102 from Guinea-Bissau.

Genotyping. Polymorphisms in the SP110 gene were identified from the National Center for Biotechnology Information dbSNP database (www.ncbi.nlm.nih.gov/SNP), with the exception of sp110int10, which was identified through sequencing (Table 1). SNPs were genotyped by using the Sequenom (San Diego) MassARRAY system, and the primer extension products were analyzed by using MALDI-TOF mass spectrometry (15, 16). Details of the primers used are shown in Table 5, which is published as supporting information on the PNAS web site.

Analysis. TDT was carried out on the data by using the program TRANSMIT, which is able to use information from siblings to infer missing parental data (17). Haplotypes were constructed by using GENEHUNTER (18, 19). Only the parental haplotypes were used when calculating the LD statistics with the program HAPLOXT (20). In addition, diploid data from the parents was used to define regions of low haplotypic diversity with the program LDMAP (21).

We thank the subjects and families for their participation. This work was funded by the European Commission under Contract IC18CT980375, the Wellcome Trust, the National Institutes of Health (to I.K.), and Medical Research Council support to the Tropical Epidemiology Unit at the London School of Hygiene and Tropical Medicine (to S.B. and K.F.). A.V.S.H. is a Wellcome Trust Principal Research Fellow.

1. W.H.O. (2002) *The World Health Report 2002* (W.H.O., Geneva).
2. Bellamy, R. (2003) *Genes Immun.* **4,** 4–11.
3. Vidal, S. M., Malo, D., Vogan, K., Skamene, E. & Gros, P. (1993) *Cell* **73,** 469–485.
4. Bellamy, R., Ruwende, C., Corrah, T., McAdam, K. P., Whittle, H. C. & Hill, A. V. (1998) *N. Engl. J. Med.* **338,** 640–644.
5. Kramnik, I., Dietrich, W. F., Demant, P. & Bloom, B. R. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 8560–8565.
6. Pan, H., Yan, B.-S., Rojas, M., Shebzukhov, Y. V., Zhou, H., Kobzik, L., Higgins, D. E., Daly, M. J., Bloom, B. R. & Kramnik, I. (2005) *Nature* **434,** 767–772.
7. Bloch, D. B., Nakajima, A., Gulick, T., Chiche, J. D., Orth, D., de La Monte, S. M. & Bloch, K. D. (2000) *Mol. Cell Biol.* **20,** 6138–6146.
8. Watashi, K., Hijikata, M., Tagawa, A., Doi, T., Marusawa, H. & Shimotohno, K. (2003) *Mol. Cell. Biol.* **23,** 7498–7509.
9. Saito, T., Ji, G., Shinzawa, H., Okumoto, K., Hattori, E., Adachi, T., Takeda, T., Sugahara, K., Ito, J. i., Watanabe, H., *et al.* (2004) *Biochem. Biophys. Res. Commun.* **317,** 335–341.
10. Casanova, J. L. & Abel, L. (2002) *Annu. Rev. Immunol.* **20,** 581–620.
11. Lopez-Maderuelo, D., Arnalich, F., Serantes, R., Gonzalez, A., Codoceo, R., Madero, R., Vazquez, J. J. & Montiel, C. (2003) *Am. J. Respir. Crit. Care Med.* **167,** 970–975.
12. Lio, D., Marino, V., Serauto, A., Gioia, V., Scola, L., Crivello, A., Forte, G. I., Colonna-Romano, G., Candore, G. & Caruso, C. (2002) *Eur. J. Immunogenet.* **29,** 371–374.
13. Lienhardt, C., Bennett, S., Del Prete, G., Bah-Sow, O., Newport, M., Gustafson, P., Manneh, K., Gomes, V., Hill, A. & McAdam, K. (2002) *Am. J. Epidemiol.* **155,** 1066–1073.
14. Bennett, S., Lienhardt, C., Bah-Sow, O., Gustafson, P., Manneh, K., Del Prete, G., Gomes, V., Newport, M., McAdam, K. & Hill, A. (2002) *Am. J. Epidemiol.* **155,** 1074–1079.
15. Jurinke, C., van den Boom, D., Cantor, C. R. & Koster, H. (2002) *Adv. Biochem. Eng. Biotechnol.* **77,** 57–74.
16. Jurinke, C., van den Boom, D., Cantor, C. R. & Koster, H. (2002) *Methods Mol. Biol.* **187,** 179–192.
17. Clayton, D. (1999) *Am. J. Hum. Genet.* **65,** 1170–1177.
18. Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. & Lander, E. S. (1996) *Am. J. Hum. Genet.* **58,** 1347–1363.
19. Kruglyak, L. & Lander, E. S. (1998) *J. Comput. Biol.* **5,** 1–7.
20. Abecasis, G. R. & Cookson, W. O. (2000) *Bioinformatics* **16,** 182–183.
21. Maniatis, N., Collins, A., Xu, C. F., McCarthy, L. C., Hewett, D. R., Tapper, W., Ennis, S., Ke, X. & Morton, N. E. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 2228–2233.