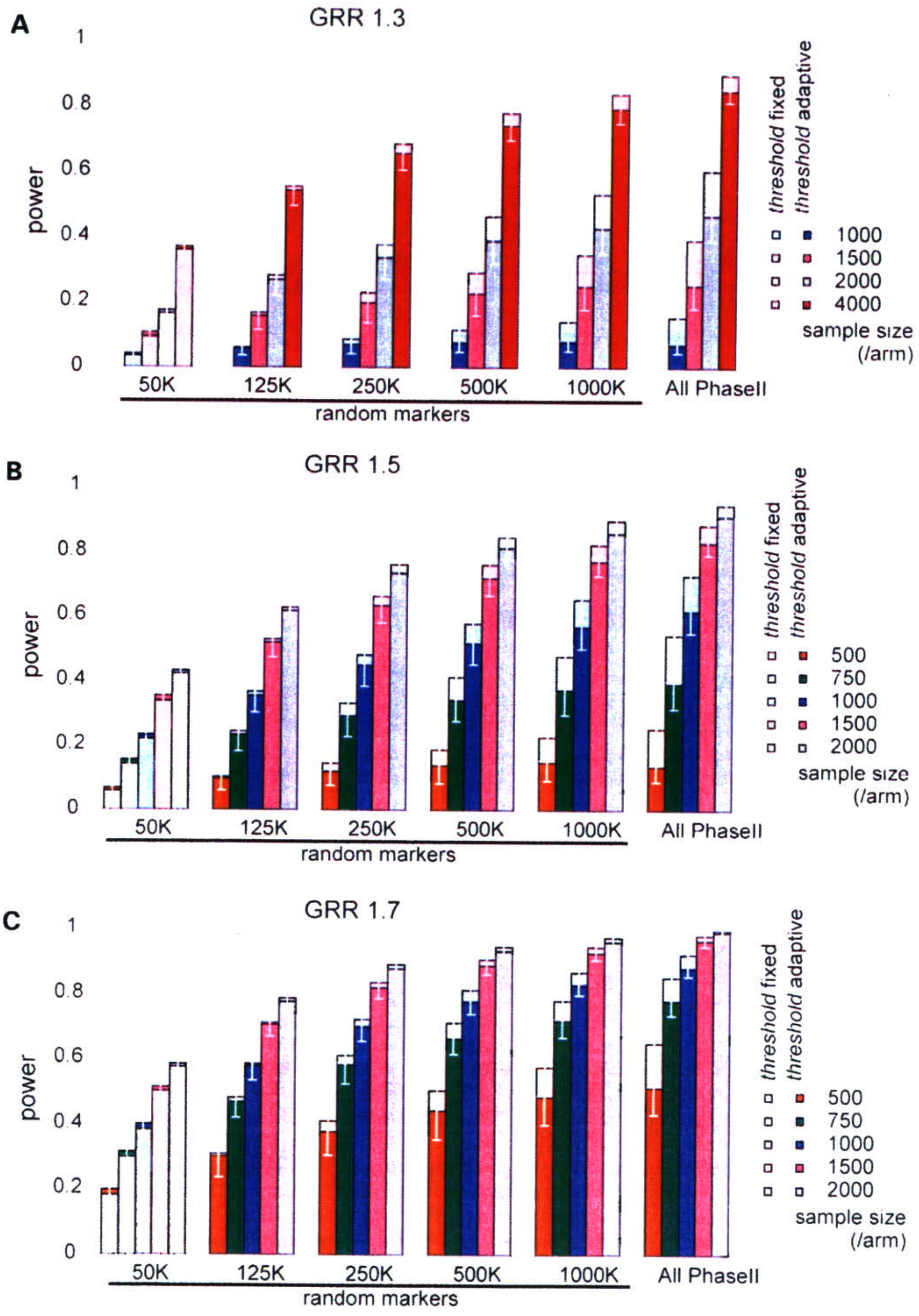

Figure 3. Genome-wide power of association studies for common causal alleles with weak to moderate genetic effects. Genome-wide power was calculated in CEU by averaging single point power for each putative causal allele over all common (MAF $\geq$ 0.05) SNPs in the Ref$^{\text{Phase II 5Kb}}$ reference set, with increasing marker and sample sizes for small to moderate GRRs (1.3–1.7) in multiplicative disease models. Power was computed using adaptive thresholds for max(χ^2) that provides a genome-wide *P*-value of 0.05 (dark columns) or using a fixed threshold ($P = 1 \times 10^{-8}$; light columns) for each marker set. The power with an adaptive threshold for a genome-wide *P*-value of 0.01 was also indicated by a lower bar within each column.

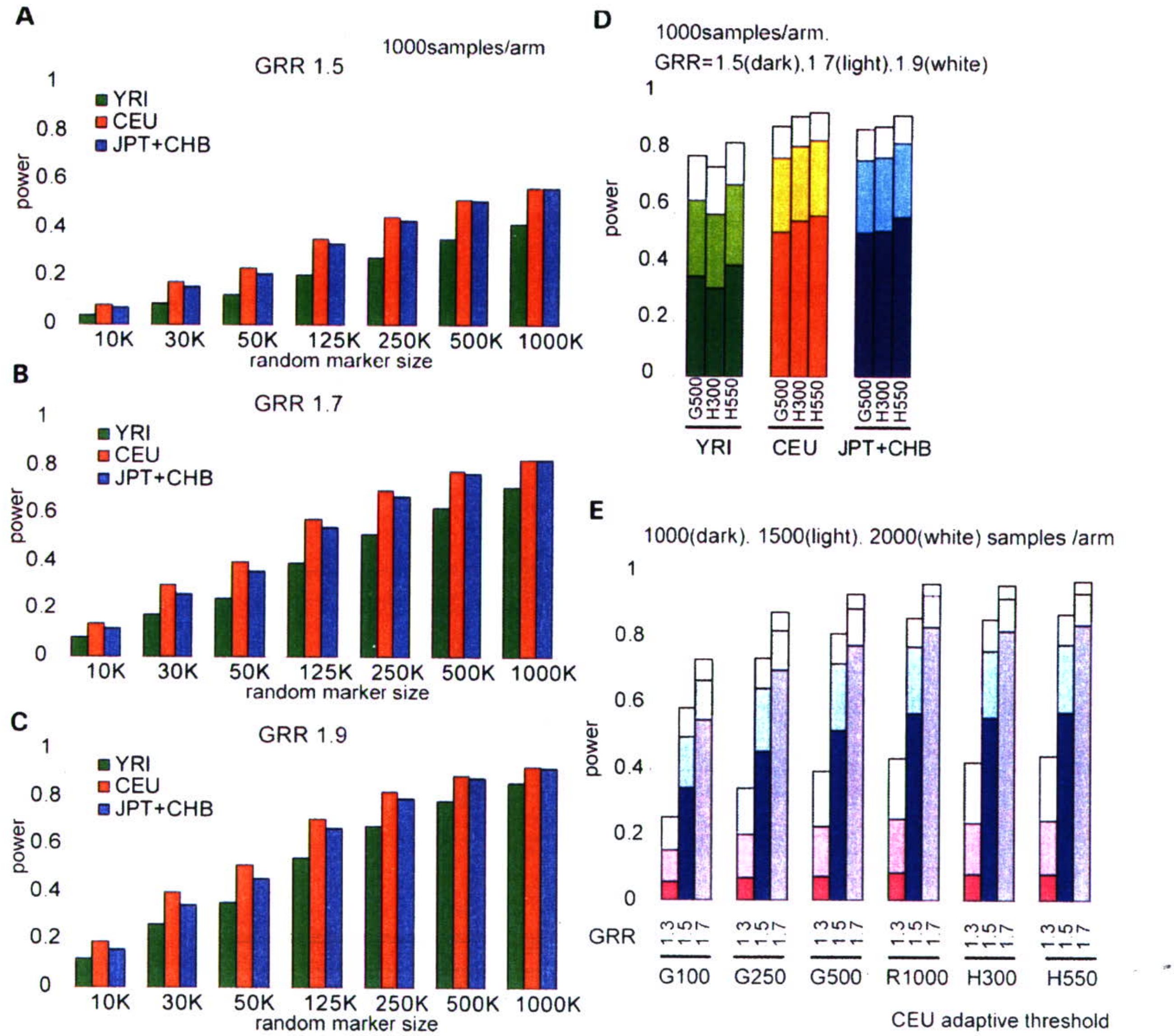

Figure 4. Comparison of power in different HapMap panels and in commercially available genotyping platforms. Genome-wide power was calculated for different HapMap panels in a variety of marker sets, including indicated numbers of randomly selected SNP markers for GRR = 1.5 (**A**), GRR = 1.7 (**B**), and GRR = 1.9 (**C**). Statistical thresholds were adjusted to provide genome-wide P-values of 0.05. Genome-wide power was also calculated for commercially available genotyping platforms in different HapMap panels (**D**) and varying sample numbers and effect sizes for CEU (**E**). The examined platforms are GeneChip® 100K (G100), GeneChip® Nsp250K (G250), GeneChip® 500K (G500), HumanHap300® (H300) and HumanHap550® (H550). Power in a random 1000K set (R1000) is shown for comparison in E.

low MAF values by any means, but also to increase the sample size (Fig. 6B and C).

Discussion

Through the current analysis, we empirically determined the size of test statistics for causal as well as null markers under varying degrees of genome coverage and realistic study parameters, and thereby demonstrated how genome-wide power is affected by the interplay between genome-coverage and other determinants. Here it is appropriate to compare the performance [power $(1 - \beta)$ or sensitivity] of the different SNP sets with their specificity (or $1 - \alpha$) being constant by applying adaptive thresholds, where α denotes genome-wide type I error probability. In addition, the power calculated in this way is directly related to false positive report probability (FPRP), which is simply expressed as $1/[1+(1-\beta)/\alpha]$, which is approximately extended to $1/[1+m(1-\beta)/\alpha]$ assuming a total of m independent causative loci having the same effect size. Note that α is a constant for all SNP sets,

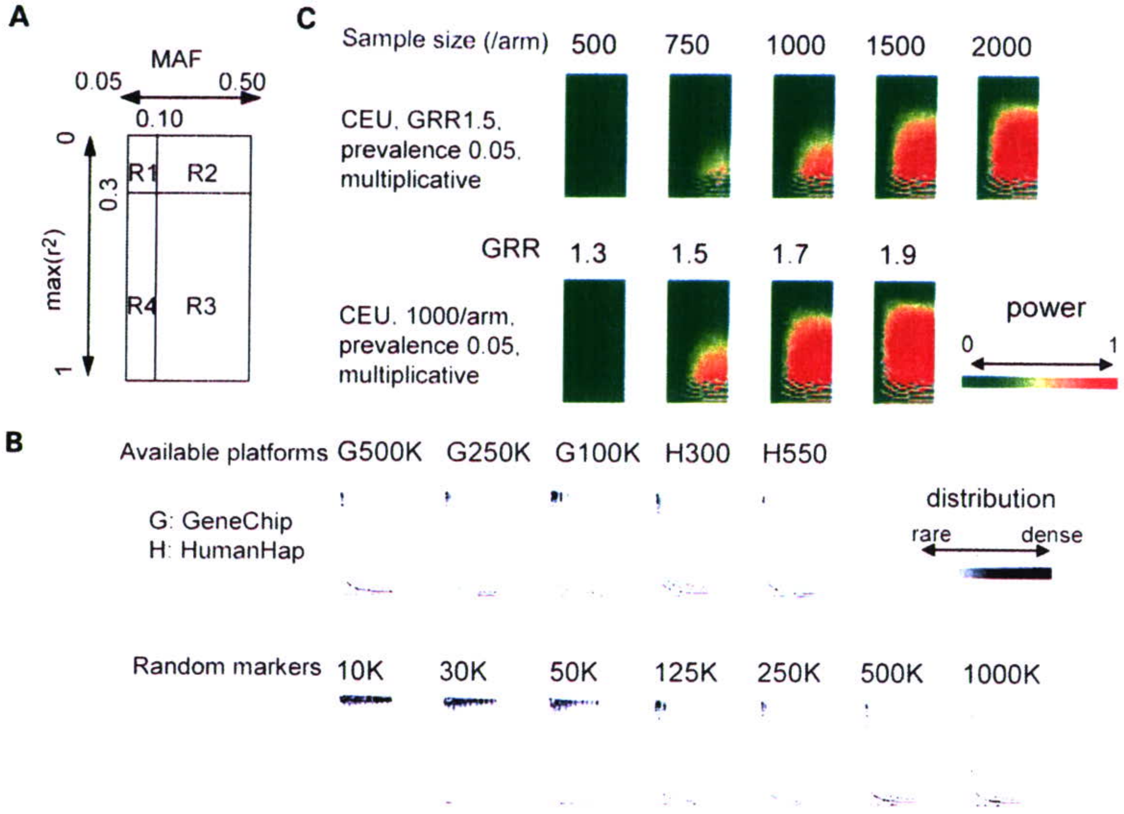

Figure 5. Impact of allele frequencies and genome coverage on genome-wide power. Reference SNPs randomly selected from the Phase II CEU set (Ref$^{\mathrm{Phase\ II\ 5K}}$) are plotted onto a panel according to their MAF and the max(r^2) within the indicated marker set, and assigned into four categories: sub-common and weakly proxied SNPs [MAF < 0.10 and max(r^2) < 0.3] SNPs (R1), common and weakly proxied SNPs (MAF ≥ 0.10 and max(r^2) < 0.3) SNPs (R2), common and strongly proxied SNPs [MAF ≥ 0.10 and max(r^2) ≥ 0.3] (R3), or sub-common and strongly proxied SNPs [MAF < 0.10 and max(r^2) ≥ 0.3] (R4). (A). Distributions of these SNPs are shown by gray-scaled density for different marker set, where the SNP distribution shifts downward as the genome coverage improves (**B**). GeneChip® 500K, 250K (NspI), 100K, HumanHap300®, and HumanHap550® are designated as G500K, G250K, G100K, H300K, and H550K, respectively. On the other hand, neglecting the collaborative capture effect, the power for SNPs with a given MAF and max(r^2) value is largely determined by GRR and sample size. Distributions of the power are color-coded for different parameter sets as indicated (**C**). Genome-wide power is roughly estimated by taking the product sum of corresponding cells in both panels.

i.e. 0.05 or 0.01. So from our simulations, readers will easily evaluate the power and FPRP expected form given SNP set, sample size and predicted effect size. As long as practical power (for example, $1 - \beta > \alpha$) is obtained, FPRP is expected to less than 0.5, which will be satisfactory for initial discovery studies.

We estimated genome-wide thresholds based on the simulations using small numbers of HapMap chromosomes. In real studies, the threshold should be determined using their own applicable data sets, where diploid, rather than phased, chromosomes could be used when enough samples are analyzed. A larger number of chromosomes should contain more numbers of rare segregating SNPs, but these rare SNPs would not increase χ^2 thresholds substantially (22).

In terms of the effective number of independent SNPs (Ne) in various marker sets, the diversity of the human genome is likely to be on the order of 1000K in CEU and the corresponding nominal P-value giving a genome-wide α error of 0.05 is 5×10^{-8}. For moderate GRRs (≦ 1.5), this threshold could be overcome with ≦ 1500 samples per arm for very common SNPs (MAF > 0.20), but for less common SNPs or those with a small genetic effect (GRR=1.1–1.2), extremely large numbers of samples will be required (Supplementary Material, Figure S8), which urges moves toward sharing typing data across multiple groups as exemplified in recent reports that identified predisposing factors with very modest genetic effects for type 2 diabetes (35–37). The diversity of our genome may not allow for detecting very rare causative alleles (<0.01) with even smaller genetic effects (i.e. GRR < 1.1) using this approach (Fig. 6D).

Under these limitations, several issues should be considered to efficiently exploit study resources and to increase the chance of finding a true association. First, for the increased genome coverage to be effectively translated into power, it needs to be accompanied by a corresponding increase in sample size. When sample numbers are small relative to the effect size, the cost of multiple testing largely offsets the expected increase in the test statistics for causal alleles with

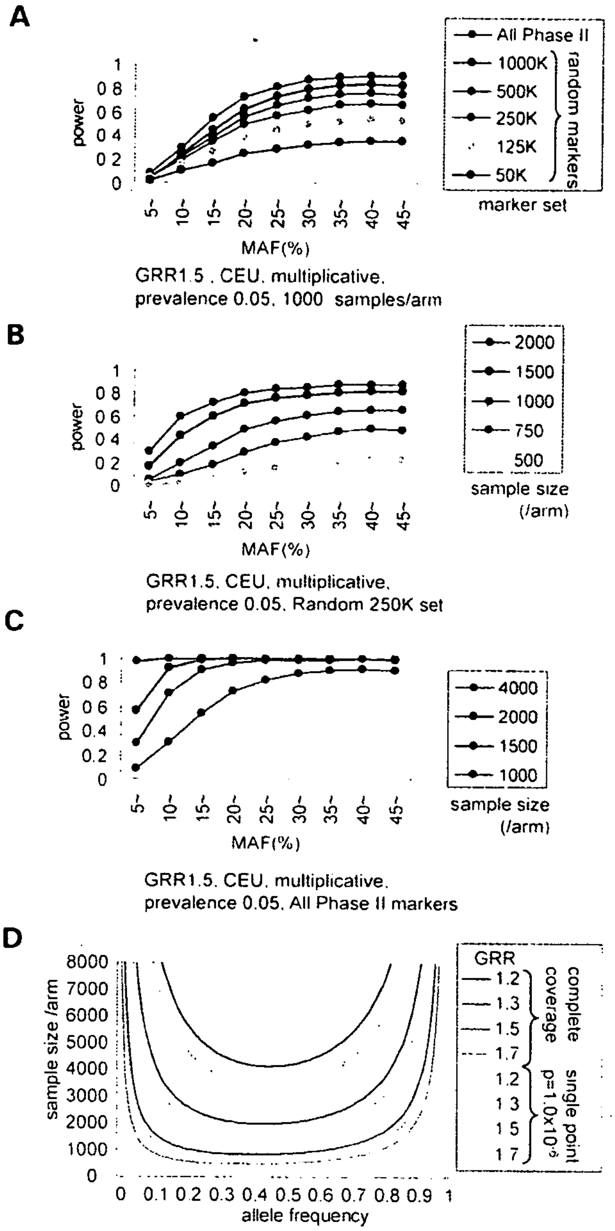

Figure 6. Effects of allele frequency on simulated power. Distribution of power on MAF in association studies are shown for varying marker sets under a constant sample size (1000 /arm) (**A**). and for varying sample sizes under a fixed marker set; GeneChip® 250K (**B**) or a hypothetical complete marker set (**C**). CEU was used for simulations with fixed GRR (1.5) and disease prevalence (0.05). The sample size that is required for detecting a causative allele with 80% power was calculated for GRRs of 1.2. 1.3. 1.5 and 1.7. assuming complete genome coverage in a multiplicative model (**D**). The significance threshold for genome-wide *P*-values of 0.05 is set assuming complete genome coverage (Ne^G=1023K. solid lines) or independent 50K markers (single point *P*-value =1×10^{-6}. Ne=50K. broken lines).

no measurable gain in power, and can even exceed the gain in causal distributions (Fig. 4). Increasing genome coverage with insufficient sample sizes would only consume resources with no substantial benefit in power. In addition, power tends to saturate in higher genome coverage and the effect of increasing the number of marker SNPs is less prominent compared to that of increasing sample sizes. In most simulated situations, more power is expected by doubling the sample size than by doubling the number of maker SNPs. For example, our simulations predict that doubling the sample size using GeneChip® Nsp 250K is almost certainly more efficient than analyzing half of the samples with both Nsp 250K+Sty 250K (Supplementary Material, Figure S9).

The tagging strategy or statistical imputation is effective for increasing genome coverage with limited numbers of marker SNPs (21,38,39), although it does not save the cost of multiple-hypothesis testing. The efficiency of generating a tag SNP set with higher genome coverage, however, is increasingly compromised. The additional gain in power becomes smaller with increasing genome coverage, while more and more effort will be required to find additional independent tag SNPs, because many SNPs are already captured by existing tag SNPs. In addition, we simulated power using 'All Phase II' set. In the sense that all references are captured through direct association, this marker set provides the ultimate coverage of the genome. Considering that modest increase of power using 'All Phase II' set compared with random 1000K set (Fig. 3), multimarker tagging presumably may not push up the power profoundly. Transferability of a tag SNP set from one population to another is also a problem. Tag SNPs for CEU are transferable to a certain degree to JPT+CEU, but they are less effective for YRI.

In any simulated scenarios, detecting SNPs with lower MAF values (0.05–0.10) is very difficult using whole genome approaches, which is especially true for SNPs with less than 0.05 MAF values. In this situation, genome coverage to capture these rare SNPs becomes definitely important, but the required increase in the sample size is greater for rare SNPs than for common ones. Effort to devising SNP sets for these rare alleles, or exhaustive multimarker tests (21,38), is not likely to be rewarding unless their genetic effects are substantially large.

MATERIALS AND METHODS

HapMap data sets

The phased genotyping data of the HapMap Phase II (release 21) were obtained from the International HapMap Project web site (http://www.hapmap.org/downloads/phasing/2006-07_Phase II/) (10). It includes the data from 60 CEU parents (120 chromosomes), 60 YRI parents (120 chromosomes) and the combined set of 45 JPT and 45 CHB unrelated individuals (180 chromosomes), and is provided in three discrete sets ('all', 'consensus', and 'phased'), of which we used the former two sets for analysis. The 'all' set contains the comprehensive data of all SNPs genotyped in each population including non-segregating sites, and the 'consensus' set consists of the intersection of 'all' sets from the three population panels. The 'all' sets contain 3755 469, 368 5205 and 3776 850 SNPs for CEU, YRI and JPT+CHB, respectively, and the 'consensus' set includes 3535 396 SNPs.

Marker sets and the references for power calculation

We generated a series of marker sets consisting of 10K, 30K, 50K, 125K, 250K, 500K and 1000K SNPs, by randomly selecting SNPs from the Phase II 'all' sets for each HapMap panel. The number of segregating SNPs in each set is denoted as Ns and shown in Table 1for CEU panel. Because the Phase II 'all' set contains most of the SNPs on commercially available platforms, including Affymetrix® GeneChip® 500K (Nsp+Sty), 250K (Nsp), 100K (Hind+Xba), Illumina® HumanHap300®, and HumanHap550® (Supplementary Material, Table S1), the intersectional SNPs of these platforms with the Phase II 'all' set were incorporated into the analysis as representative SNPs of each commercial set. Annotation files for SNPs on GeneChip® series are available from the Affymetrix® web site (http://www.affymetrix.com/products/application/whole_genome.affx). The SNP information of HumanHap® series was kindly provided by Illumina® Inc. A subset of the Phase II SNPs, referred to as 'Ref$^{\text{Phase II 5Kb}}$', was constructed and used as a reference in the calculation of genome-wide powers by randomly selecting SNPs from the 'consensus' set so that each SNP is, on average, 5 Kb apart from the adjacent SNPs. Combined SNPs from the 10 ENCODE regions, denoted as Ref$^{\text{ENCODE}}$, were used as an alternative reference set. Only common SNPs (MAF $\geq$ 0.05) were included in the power calculations as putative causal alleles.

Simulation of case-control panels under the null hypothesis and fitting simulated distributions

Null distributions in genetic association studies are considered for only vaguely defined ensembles having limited population sizes, e.g. all adult Japanese eligible for a study. To obtain asymptotic distributions, we generated 10 000 null case-control panels by randomly resampling phased autosomal chromosomes from the 'all' set of CEU, YRI and JPT+CHB. Simulations were performed with different sample numbers, i.e. 500, 750, 1000, 1500, 2000 and 4000 per single arm. For each case-control panel, the maximum χ^2 value (max(χ^2); d.f.=1) in the standard allele test was calculated for different marker sets to obtain empirical null distributions of max(χ^2).

The simulated distributions, $\Phi(\chi^2)$, were fitted to the null distribution for hypothetical Nc independent SNPs, $\varphi_{\mathrm{Nc}}(\chi^2)$, by the least squares method as follows:

$$\mathrm{Nc} = \arg\min_{\mathrm{N}} \int (\varphi_{\mathrm{N}}(\chi^2) - \Phi(\chi^2))^2 \mathrm{d}\chi^2$$

The Gnu Scientific Library was used to handle these functions.

Simulation of case-control studies and calculation of power

We consider multiplicative disease models showing a prevalence e, and assume a single causative allele whose MAF and GRR are P ($\geq$0.05) and γ, respectively. Given the penetrance for AA, Aa and aa genotypes as f_{AA}, f_{Aa}, and f_{aa}, respectively, expected genotype frequencies in the case and control panels are given as,

$$P(AA|\text{case}) = \frac{p^2 f_{AA}}{e}$$

$$P(Aa|\text{case}) = \frac{2p(1-p)f_{Aa}}{e}$$

$$P(aa|\text{case}) = \frac{(1-p)^2 f_{aa}}{e}$$

$$P(AA|\text{control}) = \frac{p^2(1-f_{AA})}{1-e}$$

$$P(Aa|\text{control}) = \frac{2p(1-p)(1-f_{Aa})}{1-e}$$

$$P(aa|\text{control}) = \frac{(1-p)^2(1-f_{aa})}{1-e}$$

where

$$e = p^2 f_{AA} + 2p(1-p)f_{Aa} + (1-p)^2 f_{aa}$$

$$f_{AA} = \gamma^2 f_{aa}, \quad f_{Aa} = \gamma f_{aa}$$

According to these allele frequencies, we generated 2000 case-control panels under the alternative hypothesis by resampling a predetermined number of phased chromosomes, and calculated max(χ^2) of the marker SNPs for each panel, where the calculations were performed only for those marker SNPs that are within 500 Kb from the putative causal SNP. The proportion of simulated case-control panels whose max(χ^2) exceeded the upper 95 or 99% point of the corresponding null distribution for that marker set was defined as the power. The genome-wide power was computed by averaging each power for all SNPs within the reference set. As the number of marker SNPs increases, up to as high as 1000K, there is a considerable chance of detecting direct associations, i.e. the causative SNP is included in the marker set. Assuming 7500K common SNPs within the human genome (17), the Phase II data set includes one-fourth (2167K common SNPs in CEU) of all the common SNPs. Based on this estimation, we excluded three-fourths of the direct associations from the calculation of genome-wide power to avoid overestimating its chance. The adjustment of direct association, however, has little influence on the results. This correction was not applied to the power calculation on the Ref$^{\text{ENCODE}}$ set, because it represents the nearly complete data set for those regions.

Computational resources

All simulations were run on the GXP clustering computer system in the Department of Information and Communication Engineering, Graduate School of Information Science, University of Tokyo.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

This work is totally indebted to the achievement of the International HapMap Consortium and we thank all the people who participated in the project. We also thank Jun Ohashi for helpful discussions.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by Research on Measures for Intractable Diseases, Health and Labor, Sciences Research Grants, Ministry of Health, Labor and Welfare, Research on Health Sciences focusing on Drug Innovation, the Japan Health Sciences Foundation, and Core Research for Evolutional Science and Technology (CREST), Japan Science and Technology Agency.

REFERENCES

1. Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516 - 1517.
2. Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.*, **22**, 139 - 144.
3. Risch, N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847 - 856.
4. Syvanen, A.C. (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat. Rev. Genet.*, **2**, 930 - 942.
5. Kennedy, G.C., Matsuzaki, H., Dong, S., Liu, W.M., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J. *et al.* (2003) Large-scale genotyping of complex DNA. *Nat. Biotechnol.*, **21**, 1233 - 1237.
6. Fan, J.B., Chee, M.S. and Gunderson, K.L. (2006) Highly parallel genomic assays. *Nat. Rev. Genet.*, **7**, 632 -644.
7. Hirschhorn, J.N. and Daly, M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, **6**, 95 - 108.
8. Wang, W.Y., Barratt, B.J., Clayton, D.G. and Todd, J.A. (2005) Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.*, **6**, 109 - 118.
9. The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789 - 796.
10. The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299 - 1320.
11. Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F. *et al.* (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.*, **29**, 233 - 237.
12. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225 - 2229.
13. Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L. and Nickerson, D.A. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, **74**, 106 - 120.
14. Halldorsson, B.V., Istrail, S. and De La Vega, F.M. (2004) Optimal selection of SNP markers for disease association studies. *Hum. Hered.*, **58**, 190 - 202.
15. Zhang, K., Qin, Z., Chen, T., Liu, J.S., Waterman, M.S. and Sun, F. (2005) HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics*, **21**, 131 - 134.
16. Ao, S.I., Yip, K., Ng, M., Cheung, D., Fong, P.Y., Melhado, I. and Sham, P.C. (2005) CLUSTAG: hierarchical clustering and graph methods for selecting tag SNPs. *Bioinformatics*, **21**, 1735 - 1736.
17. Barrett, J.C. and Cardon, L.R. (2006) Evaluating coverage of genome-wide association studies. *Nat. Genet.*, **38**, 659 - 662.
18. Pe'er, I., de Bakker, P.I., Maller, J., Yelensky, R., Altshuler, D. and Daly, M.J. (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat. Genet.*, **38**, 663 - 667.
19. Ohashi, J. and Tokunaga, K. (2001) The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers. *J. Hum. Genet.*, **46**, 478 - 482.
20. Zondervan, K.T. and Cardon, L.R. (2004) The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.*, **5**, 89 - 100.
21. de Bakker, P.I., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J. and Altshuler, D. (2005) Efficiency and power in genetic association studies. *Nat. Genet.*, **37**, 1217 - 1223.
22. Neale, B.M. and Sham, P.C. (2004) The future of association studies: gene-based analysis and replication. *Am. J. Hum. Genet.*, **75**, 353 - 362.
23. Dudbridge, F. and Koeleman, B.P. (2003) Rank truncated product of *P*-values, with application to genomewide association scans. *Genet. Epidemiol.*, **25**, 360 - 366.
24. Hoh, J. and Ott, J. (2003) Mathematical multi-locus approaches to localizing complex human trait genes. *Nat. Rev. Genet.*, **4**, 701 - 709.
25. Hoh, J., Wille, A. and Ott, J. (2001) Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res.*, **11**, 2115 - 2119.
26. Zaykin, D.V., Zhivotovsky, L.A., Westfall, P.H. and Weir, B.S. (2002) Truncated product method for combining *P*-values. *Genet. Epidemiol.*, **22**, 170 - 185.
27. De La Vega, F.M., Isaac, H., Collins, A., Scafe, C.R., Halldorsson, B.V., Su, X., Lippert, R.A., Wang, Y., Laig-Webster, M., Koehler, R.T. *et al.* (2005) The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. *Genome Res.*, **15**, 454 - 462.
28. Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G. and Chee, M.S. (2005) A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.*, **37**, 549 - 554.
29. Matsuzaki, H., Dong, S., Loi, H., Di, X., Liu, G., Hubbell, E., Law, J., Berntsen, T., Chadha, M., Hui, H. *et al.* (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods*, **1**, 109 - 111.
30. Steemers, F.J., Chang, W., Lee, G., Barker, D.L., Shen, R. and Gunderson, K.L. (2006) Whole-genome genotyping with the single-base extension assay. *Nat. Methods*, **3**, 31 - 33.
31. Tenesa, A. and Dunlop, M.G. (2006) Validity of tagging SNPs across populations for association studies. *Eur. J. Hum. Genet.*, **14**, 357 - 363.
32. de Bakker, P.I., Burtt, N.P., Graham, R.R., Guiducci, C., Yelensky, R., Drake, J.A., Bersaglieri, T., Penney, K.L., Butler, J., Young, S. *et al.* (2006) Transferability of tag SNPs in genetic association studies in multiple populations. *Nat. Genet.*, **38**, 1298 - 1303.
33. Pritchard, J.K. (2001) Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.*, **69**, 124 - 137.
34. Slager, S.L., Huang, J. and Vieland, V.J. (2000) Effect of allelic heterogeneity on the power of the transmission disequilibrium test. *Genet. Epidemiol.*, **18**, 143 - 156.
35. Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U. *et al.* (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, **316**, 1341 - 1345.
36. Saxena, R., Voight, B.F., Lyssenko, V., Burtt, N.P., de Bakker, P.I., Chen, H., Roix, J.J., Kathiresan, S., Hirschhorn, J.N., Daly, M.J. *et al.* (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, **316**, 1331 - 1336.
37. Zeggini, E., Weedon, M.N., Lindgren, C.M., Frayling, T.M., Elliott, K.S., Lango, H., Timpson, N.J., Perry, J.R., Rayner, N.W., Freathy, R.M. *et al.* (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, **316**, 1336 - 1341.
38. Lin, S., Chakravarti, A. and Cutler, D.J. (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat. Genet.*, **36**, 1181 - 1188.
39. Weale, M.E., Depondt, C., Macdonald, S.J., Smith, A., Lai, P.S., Shorvon, S.D., Wood, N.W. and Goldstein, D.B. (2003) Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *Am. J. Hum. Genet.*, **73**, 551 -565.

Expression Profiling of Immature Thymocytes Revealed a Novel Homeobox Gene That Regulates Double-Negative Thymocyte Development[1]

Masahito Kawazu,[*†] Go Yamamoto,[*] Mayumi Yoshimi,[*] Kazuki Yamamoto,[*] Takashi Asai,[*] Motoshi Ichikawa,[*] Sachiko Seo,[*] Masahiro Nakagawa,[*] Shigeru Chiba,[†] Mineo Kurokawa,[*] and Seishi Ogawa[2*‡§¶]

Intrathymic development of CD4/CD8 double-negative (DN) thymocytes can be tracked by well-defined chronological subsets of thymocytes, and is an ideal target of gene expression profiling analysis to clarify the genetic basis of mature T cell production, by which differentiation of immature thymocytes is investigated in terms of gene expression profiles. In this study, we show that development of murine DN thymocytes is predominantly regulated by largely repressive rather than inductive activities of transcriptions, where lineage-promiscuous gene expression in immature thymocytes is down-regulated during their differentiation. Functional mapping of genes showing common temporal expression profiles implicates previously uncharacterized gene regulations that may be relevant to early thymocytes development. A small minority of genes is transiently expressed in the CD44^{low}CD25$^+$ subset of DN thymocytes, from which we identified a novel homeobox gene, *Duxl*, whose expression is up-regulated by Runx1. *Duxl* promotes the transition from CD44^{high}CD25$^+$ to CD44^{low}CD25$^+$ in DN thymocytes, while constitutive expression of *Duxl* inhibits expression of TCR β-chains and leads to impaired β selection and greatly reduced production of CD4/CD8 double-positive thymocytes, indicating its critical roles in DN thymocyte development. ***The Journal of Immunology*, 2007, 179: 5335–5345.**

Intrathymic development of thymocytes from their bone marrow progenitors is a critical process for the generation of mature T cells, through which the immature thymocyte progenitors, as identified by the absence of mature T cell markers (CD4/CD8 double-negative (DN)[3] T cells), differentiate into the CD4/CD8 double-positive (DP) cells, and finally produce CD4 or CD8 single-positive T cells. Although accounting for <5% of total thymocytes in mice, the DN thymocytes undergo a dynamic developmental process that is essential for the subsequent expansion into the DP population (1). Among these DN thymocytes, the earliest chronological subset (DN1) is recognized as a CD44^{high}C-Kit$^+$CD25$^-$ population (2, 3). In the first wave of cytokine-dependent pre-T cell expansion, the DN1 cells begin to proliferate with concomitant up-regulation of CD25, giving rise to the DN2 population showing the CD44^{high}C-Kit$^+$CD25$^+$ phenotype (4). The DN2 thymocytes then start to rearrange their *TCR* genes and down-regulate the CD44 expression to generate the CD44^{low}CD25$^+$ DN3 subset (5). In the DN3 stage, thymocytes are subjected to a process called β selection and only those DN3 cells that have productively rearranged their *TCRβ* gene can survive and transit into CD44^{low}CD25$^-$ DN4 thymocytes, followed by rapid expansion into the DP population (6–8).

During these early developmental processes, immature thymocyte progenitors lose their multilineage plasticity and exclusively commit to T cell differentiation. Under appropriate conditions, the DN1 population can generate multilineage hemopoietic components in vitro (9, 10), although there still remains some controversy with regard to their potential to B lineage differentiation (11). The potential of the multilineage commitment is more restricted in the DN2 stage, where cells can still give rise to NK cells and thymic dendritic cells, but no more B cells (10, 12–14), and after the DN2/DN3 transition and the succeeding β selection, thymocytes mostly lose their potential to multilineage plasticity and totally commit to T cell lineage (10).

Because these processes are thought to take place under tightly controlled gene expression, it is of particular importance to clarify the nature of this gene regulation and the key regulators involved in that regulation. To date, a number of molecules have been identified that regulate these developmental processes (15). *Notch1*-deficient thymocytes, for example, are not able to produce T cells but differentiate into B cells (16), while *pTα, TCRβ, Lck, SLP76,* and *Lat* are shown to be indispensable for β selection and their knockout mice show a severe maturational block at the DN3/DN4 transition (17). Similarly, the DN2/DN3 transition is completely blocked in *Runx1*-deficinet mice (18, 19) and also in double knockout mice of *pTα* and *common cytokine receptor γ-chain* genes (20). In contrast, it is well anticipated that these developmental processes in DN thymocytes should involve regulation of much

*Department of Hematology and Oncology, †Cell Therapy and Transplantation Medicine, and ‡Regeneration Medicine for Hematopoiesis; §21st Century Core of Excellence (COE) program, Graduate School of Medicine, University of Tokyo, University of Tokyo Hospital, Tokyo, Japan; and ¶Core Research for Evolutional Science and Technology, Japan Science and Technology Agency

Received for publication June 5, 2007. Accepted for publication July 23, 2007.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] This work was supported by Research on Measures for Intractable Diseases, Health and Labor, Sciences Research Grants, Ministry of Health, Labor and Welfare, Research on Health Sciences focusing on Drug Innovation, the Japan Health Sciences Foundation, and Core Research for Evolutional Science and Technology, Japan Science and Technology Agency.

[2] Address correspondence and reprint requests to Dr. Seishi Ogawa, Department of Regeneration Medicine for Hematopoiesis, Graduate School of Medicine, University of Tokyo, Bunkyo-ku, Tokyo, Japan. E-mail address: sogawa-tky@umin.ac.jp

[3] Abbreviations used in this paper: DN, double negative; DP, double positive; FL, fetal liver; rh, recombinant human; shRNA, small hairpin RNA; iTCRβ, intracellular TCRβ; qPCR, quantitative PCR; NGFRt, nerve growth factor receptor; OP9-DL1, OP9-delta-like-1.

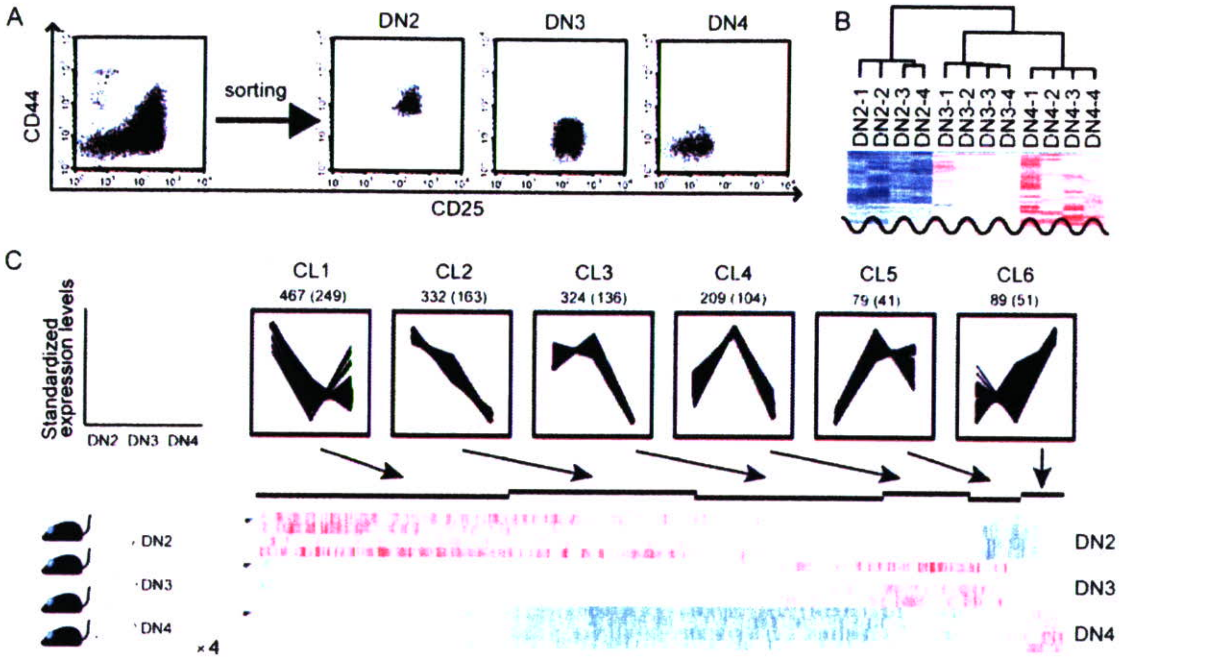

FIGURE 1. Microarray analysis and clustering of genes differentially expressed during early thymocyte development. *A*, DN2, DN3, and DN4 thymocytes from four C57BL/6 mice were FACS sorted and subjected to microarray analysis of gene expression profiles. *B*, Hierarchical clustering of 12 microarray data obtained from four independent experiments. Only branch portion is presented. *C*, One thousand five hundred differentially expressed genes were grouped into six clusters showing discrete temporal expression profiles by K-means clustering methods (*upper panels*), which are also presented in heat map (*lower panel*). The expression values for a gene across all samples were standardized to have a mean 0 and SD 1. Standardized expression levels of genes are indicated in graphs (*upper panels*) and in heat map (*lower panel*). Vertical scale of the graphs is from −1.5 to +1.5. Each column of heat map represents a gene and each row represents a sample. Red indicates high expression and blue low expression. The numbers of genes within clusters are indicated along with those for which human counterparts were identified (in the parentheses).

larger numbers of genes than those related to these known molecules. To understand the molecular mechanisms of DN thymocyte development, it may be also of use to clarify how these developmental processes are regulated in terms of their entire gene expression, to which cell differentiation is ultimately ascribed.

In the current study, we approached this issue by investigating gene expression profiles in discrete subsets of DN thymocytes under development in which DN2, DN3, and DN4 thymocytes were sorted and subjected to expression profiling analysis with high-density oligonucleotide microarrays. Clustering of differentially expressed genes among these DN thymocyte subsets demonstrated that during DN development, regulation of gene expression is predominantly repressive rather than inductive, in which multiple lineage-affiliated genes expressed in immature thymocytes are down-regulated during the course of DN thymocyte development. Functional mapping of clustered genes also revealed a possible involvement of previously uncharacterized functional gene regulations in thymocyte development. Finally, we identified a novel homeobox gene (*Duxl*), transiently expressed in DN3 thymocytes. We showed that *Duxl* is induced by Runx1 and regulates DN thymocyte development by promoting the DN2/DN3 transition, while deregulated expression of *Duxl* resulted in impaired β selection and severely compromised production of DP thymocytes, indicating its critical roles in DN thymocyte production.

Materials and Methods

Cell sorting and RNA extraction

All Abs used for cell sorting were purchased from BD Pharmingen. Thymocytes were harvested from 5- to 6-wk-old female C57BL/6 mice. Four independent cell sortings were performed and four mice were sacrificed for each experiment. Before cell sorting, CD4$^+$ cells and CD8$^+$ cells were depleted using the MACS LD system (Miltenyi Biotec). The remaining fraction was stained with anti-CD44 and anti-CD25 Abs conjugated to FITC or PerCP-Cy5.5, respectively, and also with PE-conjugated Abs to CD4, CD8, CD3, NK1.1, and TCRγδ and sorted using a FACSAria cell sorter (BD Biosciences). DN2, DN3, and DN4 subsets were identified as FITC$^+$PE$^-$PerCP-Cy5.5$^+$, FITC$^-$PE$^-$PerCP-Cy5.5$^+$, and FITC$^-$PE$^-$PerCP-Cy5.5$^-$ populations, respectively. For expression analysis of various hemopoietic lineages, mononuclear cells were separated from a single-cell suspension of bone marrow of 5- to 6-wk-old female C57BL/6 mice by centrifugation on a Histopaque-1083 (Sigma-Aldrich). c-kit$^+$ cells were obtained by positive selection for the c-kit Ag with MACS magnetic beads. The remaining fraction was stained with FITC-conjugated Ab to Ter119, PerCP-conjugated Ab to B220, and PE-conjugated Abs to Mac1, and Ter119$^+$ fraction, B220$^+$ fraction, and Mac1$^+$ fraction were sorted. B220+ splenocytes and CD3+ splenocytes were collected by positive selection of splenocytes for the B220 Ag or CD3 Ag with MACS magnetic beads. RNA was extracted from sorted cells using an RNeasy Mini kit (Qiagen) according to the manufacturer's instruction.

Microarray experiments

Biotin-labeled cRNA probes were prepared using a Two-Cycle cDNA Synthesis Kit (Affymetrix). Following fragmentation, biotin-labeled cRNA was hybridized to the Mouse Genome 430 2.0 Array (Affymetrix) for 16 h at 45°C as recommended by the manufacturer. Washing was performed using an automated fluidics workstation, and the array was immediately scanned with GeneChip Scanner 3000 7G. Expression data were extracted from image files produced on Affymetrix GeneChip Operating software 1.0 (GCOS). The absolute detection call (present, absent, or marginal) for each probe set was determined on GCOS. Normalization and expression value calculation were performed using a DNA-Chip Analyzer (www.dchip.org) (21). The invariant set normalization method (22) was used to normalize arrays at probe cell level to make them comparable and the model-based method (22) was used for computing expression values. These expression levels were attached with SEs as measurement accuracy, which were subsequently used to compute 90% confidence intervals of fold changes in two group comparisons (22). The lower confidence bounds of fold changes

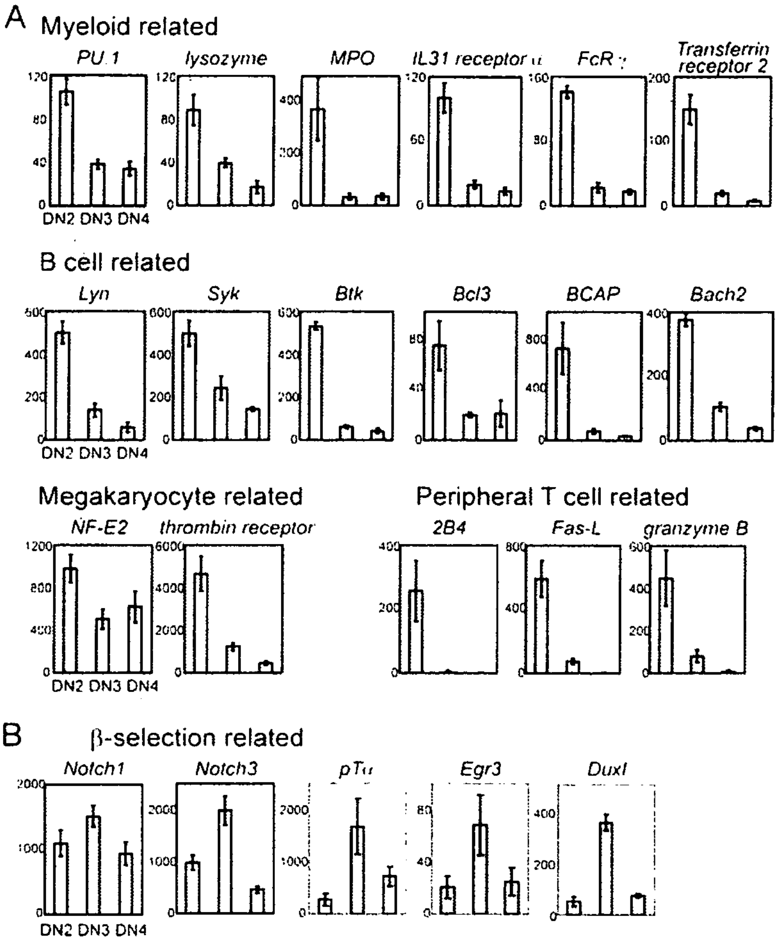

FIGURE 2. Expression patterns of lineage-affiliated genes and genes involved in β selection. Signal intensities on microarray for representative genes affiliated with various hemopoietic lineages are shown in *A* and for those relevant to β selection in *B* by their mean values (±SD), where the outlier data were excluded from the calculation. Expression data for *Duxl* are also presented in *B*.

were conservative estimate of the real fold changes. Differentially expressed probe sets were identified as those sets whose mean signals showed >1.5-fold difference between DN2 and DN3 and DN3 and DN4. The probability of false discoveries with this threshold was calculated by random permutations as described previously (23). Raw microarray data can be found at http://www.ncbi.nlm.nih.gov/geo/. GEO accession no. GSE7784.

Clustering and pathway analysis

One thousand five hundred differentially expressed genes that were identified as described above were clustered using a K-means (24) server on Gene Expression Pattern Analysis Suite (http://gepas.bioinfo.cnio.es/) (25, 26). Molecular network of each cluster was analyzed by KeyMolnet software, which was developed by the Institute of Medicinal Molecular Design. Inc. (IMMD) (27). Known molecular data were curated by IMMD and the obtained gene list of each cluster was combined with this software and shown as molecular networks. Significance of a determined pathway to obtained networks was determined. To ascertain whether any molecular pathways determined by IMMD annotate a relation between molecules in the networks at a frequency greater than that would be expected by chance, this software calculates a *p* value using the hypergeometric distribution as described previously (28).

Cloning of Duxl cDNA and construction of retrovirus plasmid

cDNAs of Duxl and FLAG-tagged Duxl having *Not*I and *Xho*I sites on their 5′ and 3′ terminus, respectively, were PCR amplified from template cDNA prepared from total thymic RNA using TaKaRa LA *Taq* (Takara Bio) with the following sets of primers: 5′-AAAAGCGGCCGCATCGA TACCATGGAGCTGAGCTGCAGTACT-3′ (for Duxl sense), 5′-AAAA GCGGCCGCATCGATACCATGGACTACAAGGACGACGATGACAA GATGGAGCTGAGCTGCAGTACT-3′ (for FLAG-tagged Duxl sense), and 5′-AAAACTCGAGCTACGGAGTTTGGTGTGCTT-3′ (for the common antisense for both). Each PCR product was digested with *Not*I and *Xho*I and cloned into the *Not*I-*Xho*I site located at the 5′ upstream of IRES-GFP or IRES-NGFRt of the pGCDNsam (a gift from Dr. H. Nakauchi, University of Tokyo, Tokyo, Japan) retrovirus vector. Nucleotide sequences of these plasmids were confirmed by resequencing. Retrovirus was prepared by transfecting the PlatE (a gift from Dr. T. Kitamura, University of Tokyo) cell line with each construct.

Quantitative PCR analysis

Total cellular RNA was converted into cDNAs by reverse transcriptase (Superscript III: Invitrogen Life Technologies) with random primers. cDNAs were amplified in triplicate for 40 cycles at 95°C for 15 s and 60°C for 60 s using an Applied Biosystems PRISM 7000 Sequence Detection System according to the manufacturer's instructions. Predesigned TaqMan primer and probe sets for 1110051B16Rik (Mm00841823_m1: Applied Biosystems) (see Fig. 4), for PU.1 (Mm00488140_m1), and for 18S rRNA (no. 4308329: Applied Biosystems) were used for the assay. PCR amplification of GAPDH, 1110051B16Rik (see Fig. 5*B*) and Rag1 was performed using Platinum SYBR Green qPCR SuperMix-UDG with ROX (Invitrogen Life Technologies) and the following primer sequences: GAPDH (forward, 5′-GAATCTACTGGAGTCTTCACC-3′; reverse, 5′-GTCATGAGCCCTTCCACGATGC-3′), 1110051B16Rik (forward, 5′-GGGAAAACTGGCTCAACAA-3′; reverse, 5′-GTGTTCTGTCCTGG GTCTGG-3′), Rag1 (forward, 5′-CTGAAGCTCAGGGTAGACGG-3′; reverse, 5′-CAACCAAGCTGCAGACATTC-3′). Significant PCR fluorescent signals were normalized for each sample to a PCR fluorescent signal obtained using GAPDH or 18S rRNA as control.

Coculture of fetal liver (FL) cells with OP9-delta-like-1 (OP9-DL1) stromal cells

OP9-DL1 is a bone marrow stromal cell line that expresses a Notch ligand, Delta-like 1, and supports development of DP thymocytes from FL-derived hemopoietic progenitors (29) and was provided by Dr. J. C. Zúñiga-Pflücker (University of Toronto, Toronto, Canada). In our OP9-DL1 assay, FL cells were harvested from E14.5 embryos and cultured on OP9-DL1 cells in combination with retroviral gene transfer as previously described (19). Briefly, mononuclear cells were separated from FL cells of C57BL/6 mice. In brief, 5 × 10^4 mononuclear cells were cultured on confluent OP9-DL1 cells in flat-bottom 24-well culture plates with 500 ml of MEM (Invitrogen Life Technologies) supplemented with 20% FCS, penicillin/streptomycin, and 5 ng/ml recombinant human (rh) IL-7 (Techne Laboratories). After 5 or 6 days of culture, 5 × 10^4 cells were passed onto newly prepared OP9-DL1 cells in the presence of 5 ng/ml rhIL-7, and retrovirus infection was performed using polybrene (final concentration, 8 mg/ml), followed by another 5 or 6 days of culture. In brief, 5 × 10^4 cells were again passed onto newly prepared OP9-DL1 cells and cultured for another 5 or 6 days, but in rhIL-7-free culture medium.

PCR for TCRβ rearrangement

PCR for TCRβ rearrangements was performed as described elsewhere (30) on DNA isolated from FL cells cultured on OP9-DL1 using the following primers: Dβ2, GTAGGCACCTGTGGGGAAGAAACT; Jβ2, TGAGAGCTGTCTCCTACTATCGATT; and Vβ5.1, GTCCAACAGTTTGATGACTATCAC. After 40 cycles of amplification (10 s at 98°C, 2 min at 68°C), PCR products were separated on a 4% agarose gel.

RNA interference

The vector backbone was RNAi-Ready pSIREN-RetroQ-ZsGreen (BD Clontech). The RNA interference target sequences were GGAGCAGGATAAACCTAGA (sequence 1), GACTGATATTCTAATTGAA (sequence 2), and GTTCCAGACTGATATTCTA (sequence 3). The small hairpin RNA (shRNA) were designed by a shRNA design algorithm, which was developed by Dr. M. Miyagishi (31). Oligonucleotides used for construction were GATCCGGGGTAGGATAAACTTAGAACGTGTGCTGTCCGTTCTAGGTTTATCCTGCTCCTTTTTACGCGTG (oligonucleotide 1, sense), AATTCACGCGTAAAAAGGAGCAGGATAAACCTAGAACGGACAGCACACGTTCTAAGTTTATCCTACCCCG (oligonucleotide 1, antisense), GATCCGATTGATGTTCTAGTTGAAACGTGTGCTGTCCGTTTCAATTAGAATATCAGTCTTTTTACGCGTG (oligonucleotide 2, sense), AATTCACGCGTAAAAAGACTGATATTCTAATTGAAACGGACAGCACACGTTTCAACTAGAACATCAATCG (oligonucleotide 2, antisense), GATCCGTTTCAGATTGATGTTCTAACGTGTGCTGTCCGTTAGAATATCAGTCTGGAACTTTTTACGCGTG (oligonucleotide 3, sense), and AATTCACGCGTAAAAAGTTCCAGACTGATATTCTAACGGACAGCACACGTTAGAACATCAATCTGAAACG (oligonucleotide 3, antisense). Retrovirus was prepared by transfecting the PlatE cell line with the knockdown vectors.

Results

Clustering of differentially expressed genes during DN thymocyte development

The DN2, DN3, and DN4 populations were FACS sorted from DN thymocytes harvested from four C57BL/6 mice and analyzed by an Affymetrix Mouse Genome 430 2.0 Array for gene expression (Fig. 1*A*). Four independent experiments were performed using 16 mice. After normalizing the array signals using an invariant gene set (22), we extracted differentially expressed probe sets whose signals showed ≥1.5-fold difference between DN2 and DN3 or between DN3 and DN4. With this threshold, 1,901 probe sets (or 1,500 nonredundant genes) were extracted as "differentially expressed" from a total of 27,330 probes (16,131 genes) expressed in either of the three DN subsets, with a false discovery rate of 0.007 as determined by random permutation tests (23). In hierarchical clustering, the four independent array data sets for each subset were correctly clustered into the same clusters, validating the reproducibility across the experiments (Fig. 1*B*). In contrast, the K-means clustering of the 1,500 differentially expressed genes identified six clusters, CL1–CL6, showing discrete temporal expression profiles: genes expressed higher in DN2 and down-regulated in DN3 (CL1), those expressed higher in DN2 and gradually down-regulated in DN3 and DN4 (CL2), those expressed in DN2 and DN3 but down-regulated in DN4 (CL3), those only transiently expressed in DN3 (CL4), those expressed both in DN3 and DN4 (CL5), and those showing low expression in DN2 and DN3 and up-regulated in DN4 (CL6) (Fig. 1*C*). The lists of genes in these clusters were presented in supplementary Table S1, a–f.[4]

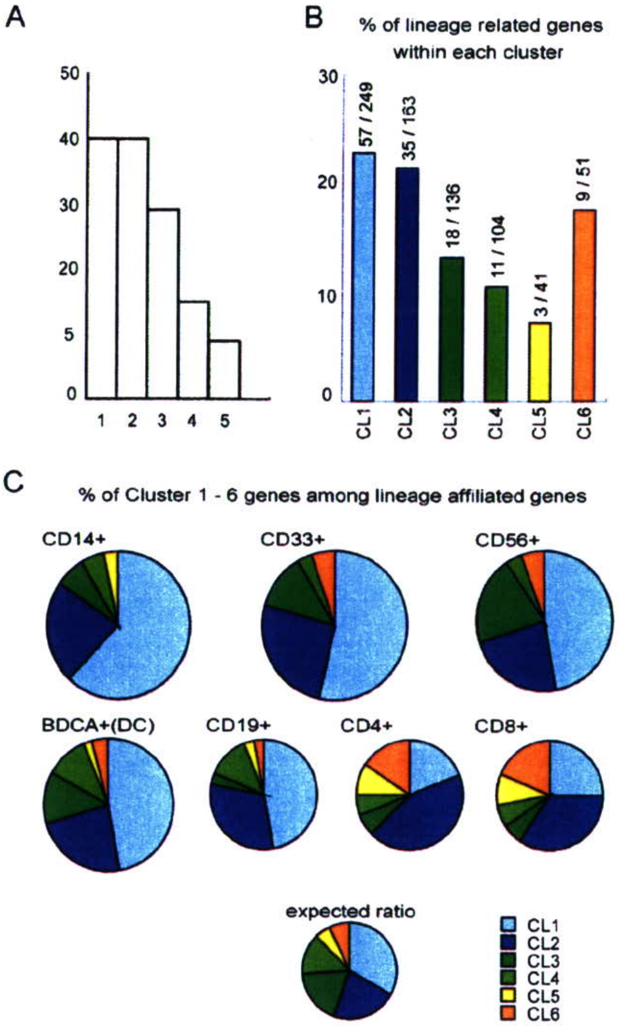

FIGURE 3. Temporal expression profiles of hemopoietic lineage-affiliated genes. *A*, Histogram of the number of lineage-affiliated genes with regard to the number of lineages to which each gene is assigned. *B*, Percentages of lineage-affiliated genes within each cluster showing discrete temporal expression profiles. Absolute numbers of lineage-affiliated genes and all genes in each cluster are shown above the bar. *C*, Number and composition of different lineage-affiliated genes with regard to clusters. Area of each circular graph represents the number of genes affiliated with each lineage.

Predominantly repressive gene regulation in DN thymocytes and their lineage-promiscuous gene expression

With regard to the temporal profiles of gene expression in DN thymocytes, our first note is that 1123 (74.9%) of the 1500 differentially expressed genes are initially expressed in DN2 but eventually down-regulated during the course of DN thymocyte development (CL1–3), while only 79 (5.27%) and 89 (5.93%) genes are up-regulated in DN3 and DN4 (CL5 and CL6), respectively (Fig. 1*C*). The other set of genes (CL4) are transiently expressed in DN3 but have low expression in DN2 and DN4. Thus, gene regulation during DN thymocyte development is largely repressive rather than inductive. Our next note was that these down-regulated genes

[4] The online version of this article contains supplemental material.

Table I. *Significantly relevant pathways in each cluster*

Rank	Name	p
CL1		
1	Ets (Spi subfamily) regulation	3.84×10^{-9}
2	IL-4 signaling	1.35×10^{-8}
3	IgG signaling	1.39×10^{-4}
4	Integrin signaling	6.94×10^{-4}
5	c-kit signaling	1.47×10^{-3}
6	G protein (G12/13) signaling	1.55×10^{-3}
7	Ets (TCF subfamily) regulation	3.11×10^{-3}
8	Rho family signaling	8.30×10^{-3}
9	CD44 signaling	8.60×10^{-3}
CL2		
1	RUNX regulation	2.19×10^{-7}
2	Endoplasmic reticulum regulation	9.72×10^{-5}
3	Platelet-derived growth factor signaling	4.94×10^{-4}
4	Prolactin signaling	7.09×10^{-4}
5	PI3K signaling	1.23×10^{-3}
6	γ-Aminobutyric acid signaling	1.23×10^{-3}
7	IL-2 signaling	1.78×10^{-3}
8	CCR10 signaling	4.66×10^{-3}
9	G protein (Gq/11) signaling	5.87×10^{-3}
10	Human growth factor signaling	9.55×10^{-3}
CL3		
1	NFAT regulation	9.67×10^{-8}
2	AP-1 regulation	3.12×10^{-5}
3	Caspase signaling	3.59×10^{-3}
4	PPARα regulation	5.24×10^{-3}
CL4		
1	TCRαβ signaling	2.35×10^{-6}
2	Ets (Ets family) regulation	1.40×10^{-5}
3	Integrin signaling	1.87×10^{-3}
4	NF-κB regulation	7.58×10^{-3}
5	Notch signaling	8.23×10^{-3}
CL5 and 6		
1	Integrin signaling	1.67×10^{-6}
2	TCRαβ signaling	7.23×10^{-6}
3	Ets (Ets family) regulation	3.17×10^{-5}
4	RB/E2F regulation	2.64×10^{-4}
5	TCF regulation	4.53×10^{-4}

contain a variety of genes whose expression is relatively characteristic to one or more hemopoietic lineages (lineage affiliated), including *NF-E2* and *thrombin receptor* (megakaryocytes), *PU.1*, *lysozyme*, and *myeloperoxidase* (myeloid lineages), *Lyn*, *Syk*, *Btk*, *Bcl3*, *BCAP*, and *Bach2* (B cells), and *2B4*, *Fas ligand*, and *granzyme B* (mature T cells) (Fig. 2*A*). This suggested that "promiscuous" expression of multiple lineage-affiliated genes in early thymocyte progenitors and their repression during the course of their commitment to mature T cells could be one of the characteristics of immature thymocytes. To confirm this in more detail, we identified those genes whose expression was thought to characterize a variety of mature hemopoietic lineages and tracked their expression levels during the course of DN thymocyte development. Because no comprehensive gene expression database was available for mice, we did this using a human tissue expression database at the Genomics Institute of the Novartis Research Foundation web site (http://symatlas.gnf.org/SymAtlas/) (32) by translating mouse genes into their human counterparts. A gene is considered to be affiliated with a lineage if its human counterpart shows >10 times higher expression in that lineage than their median expression among 79 different tissues. Among the 1500 differentially expressed genes, the human counterparts were uniquely identified for 744 (49.6%) genes (Fig. 1*C*), of which 133 (17.9%) satisfied the above criteria for being affiliated with one or more hemopoietic cell lineages (Fig. 3*A*), and lists of genes affiliated with respective lineages are presented in supplementary Table S2, a–g.[4] Among the 133 lineage-affiliated genes, 110 (83%) are expressed in DN2 and down-regulated during the course of the DN thymocyte development (CL1–CL3), accounting for 20% of 548 down-regulated genes assigned to CL1–3 (Fig. 3*B*). More genes are rapidly down-regulated during the DN2/DN3 transition (CL1), whereas as indicated from the relatively high CL3 component, down-regulation of genes affiliated with NK cells seems to occur more slowly (Fig. 3*C*). Although most of these down-regulated genes are affiliated with lineages other than T cells, many T cell-affiliated genes are also prematurely expressed in immature thymocytes and undergo down-regulation before they are definitely expressed later in thymocyte maturation (Figs. 2*A* and 3*C*). Only 12 (9%) lineage-affiliated genes were newly up-regulated until the DN4 stages and mostly related to T cells (Fig. 3, *B* and *C*). Note that our criteria for lineage relatedness may be too conservative, since many of genes presumed to be specific to mature T cells, such as *2B4*, *Fas ligand*, and *granzyme B*, were not extracted as T cell-related genes.

Pathway analysis of gene clusters during DN thymocytes development

The functional features of genes showing discrete expressional time courses are of interest to understand the regulation of DN development. To address this, we statistically explored possible functional links among genes within each cluster by evaluating a probability that a given set of genes on the known functional pathway was grouped together by chance within that cluster, showing a similar temporal expression profile using KeyMolnet software

Table II. *List of genes in CL4 with presumed DNA-binding capacity*

Entrez Gene Identification	Gene Title	GO Molecular Function Description
26972	Sporulation protein, meiosis-specific, SPO11 homolog (*Saccharomyces cerevisiae*)	DNA binding, DNA topoisomerase (ATP-hydrolyzing) activity, ATP binding, hydrolase activity
18131	Notch gene homolog 3 (*Drosophila*)	DNA binding, transcription factor activity, receptor activity, calcium ion binding, protein binding
72693	Zinc finger, CCHC domain containing 12	Nucleic acid binding, zinc ion binding, metal ion binding
278672	RIKEN cDNA 1110051B16 gene	DNA binding, transcription factor activity, sequence-specific DNA binding
13655	Early growth response 3	Nucleic acid binding, DNA binding, zinc ion binding, metal ion binding
210104	cDNA sequence BC043301	Zinc ion binding, metal ion binding, nucleic acid binding
320790	Chromodomain helicase DNA binding protein 7	—
67344	Tctex1 domain containing 1	—
15463	HIV-1 Rev-binding protein	DNA binding, zinc ion binding, metal ion binding
272382	SpiB transcription factor (Spi1/PU.1 related)	DNA binding, transcription factor activity, sequence-specific DNA binding, transcription factor activity

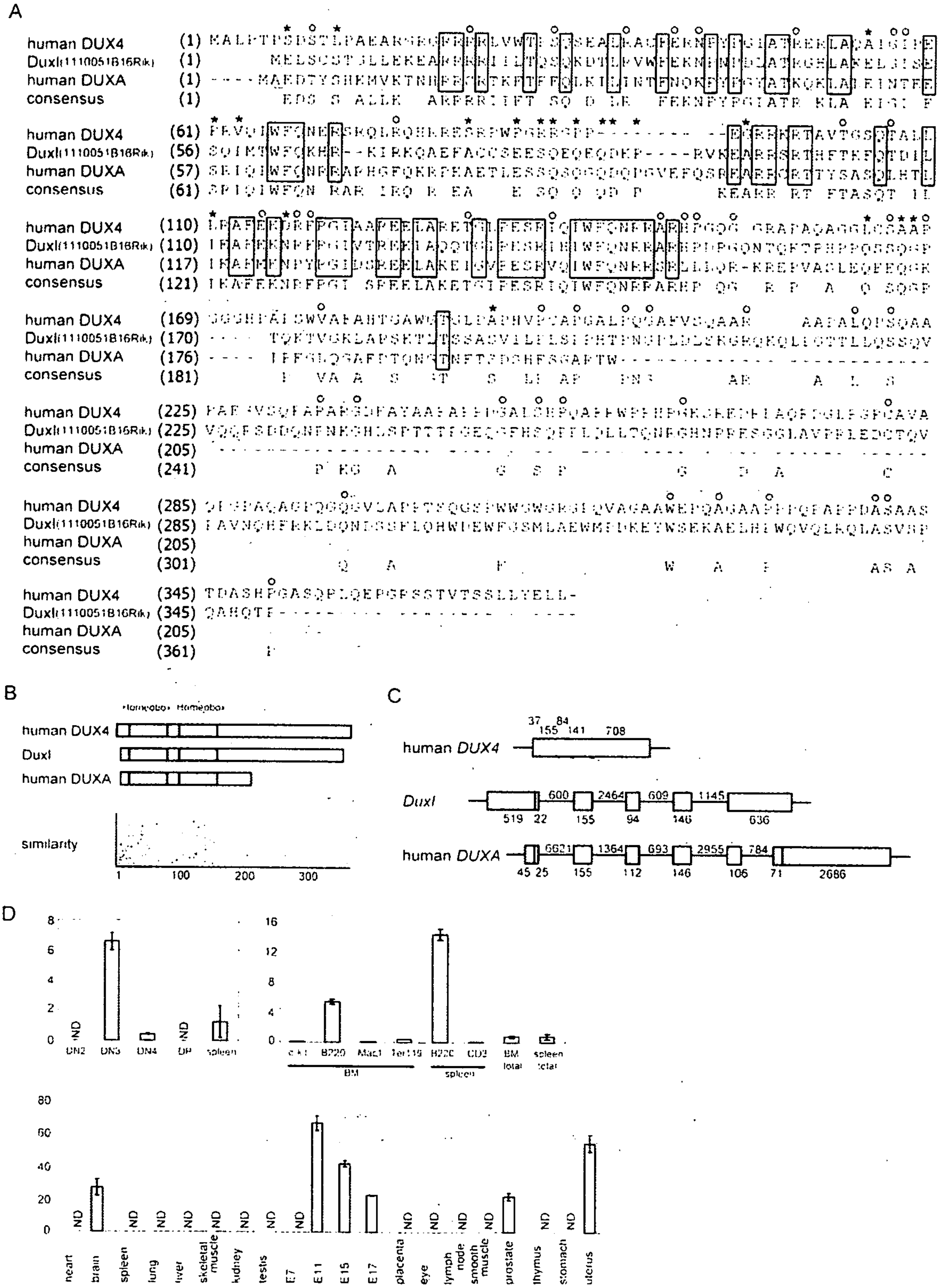

FIGURE 4. Structure of Duxl and similarity to its candidate orthologs. *A*, Homology of Duxl to its putative human orthologs (DUXA and DUX4) in amino acid sequences in which boxes indicate the identical amino acids among the three genes, asterisks indicate the common amino acids between Duxl and DUXA, and open circles indicate the common amino acids between Duxl and DUX4. *B*, Structure of predicted Duxl protein with its similarity to DUXA and DUX4. *C*, Gene structures of *Duxl* in mouse and *DUXA* and *DUX4* in humans. *D*, Distribution of *Duxl* expression in various tissues and cell lineages as determined by qPCR. The amount of transcript of *Duxl* was normalized to the amount of 18S rRNA in each tissue/population and is shown relative to levels in the total splenocytes in the *upper right panel*. Data shown are the average ± SD from triplicate samples.

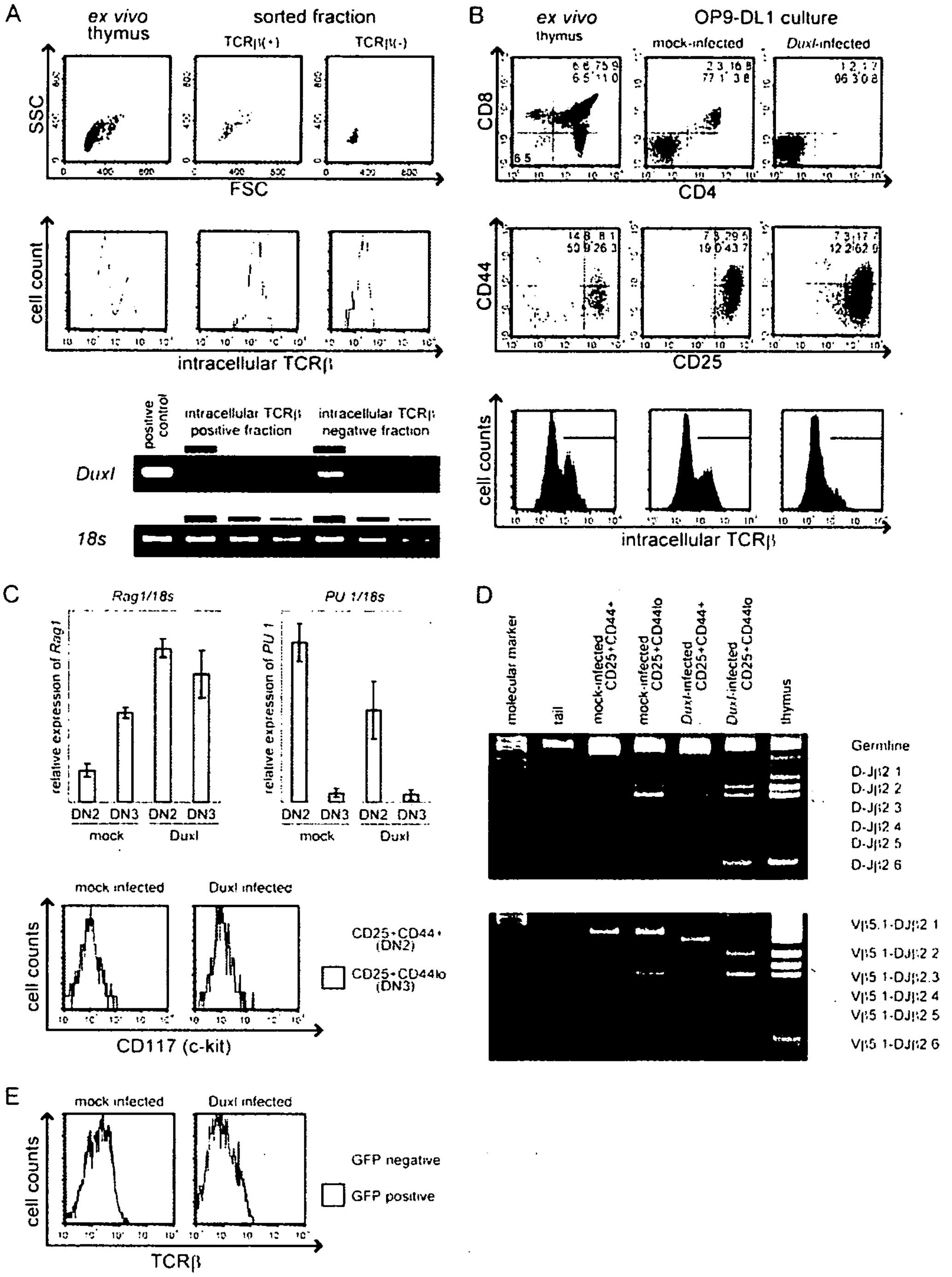

FIGURE 5. Effect of *Duxl* transduction on thymocyte development in OP9-DL culture. *A*, DN3 cells from normal mice were FACS sorted according to the expression of iTCRβ chain (*upper two panels*) and expression of *Duxl* in iTCRβ^+ and iTCRβ^- cells was examined by qPCR analysis. Resultant PCR products were electrophoresed on 4% agarose gel (*lower two panels*). *B*, *Duxl* was transduced into FL cells and the development into thymocytes was examined by FACS analysis for their expression of CD4/CD8 (*upper panels*), CD25/CD44 (*middle panels*), and iTCRβ chains (*bottom panels*) in ex vivo thymus (*left*), mock-infected cultured FL cells (*middle*), and *Duxl*-transduced cultured FL cells (*right*). *C*, Expression levels of *Rag1* and *PU.1* were examined by qPCR using RNA isolated from the sorted GFP$^+$CD25$^+$CD44$^+$ fraction and GFP$^+$CD25$^+$CD44^{low} fraction of mock- or *Duxl*-transduced FL cells cultured on OP9-DL1 (*upper panels*). The amount of transcript of *Rag1* and *PU.1* was normalized to the amount of 18S rRNA in each population and is shown as relative values. Data shown are the average ± SD from triplicate samples. Expression of CD117 was analyzed with FACS (*lower panels*). *D*, TCRβ rearrangement status was analyzed by PCR using DNA isolated from the sorted GFP$^+$CD25$^+$CD44$^+$ fraction and GFP$^+$CD25$^+$CD44^{low} fraction of mock- or *Duxl*-transduced FL cells cultured on OP9-DL1. *E*, *Duxl* was overexpressed in FL cells or AKR1 cells using retrovirus vector, and the expression levels of TCRβ in the GFP$^+$ fraction were examined by FACS.

(IMMD) (27). Because of the small numbers of genes within the clusters CL5 and CL6, both clusters were analyzed after being combined together. In this pathway analysis, several sets of functionally related genes or pathways were extracted from each cluster. Table I shows a list of pathways extracted in high significance values ($p < 0.01$). For examples, genes involved in c-kit signaling

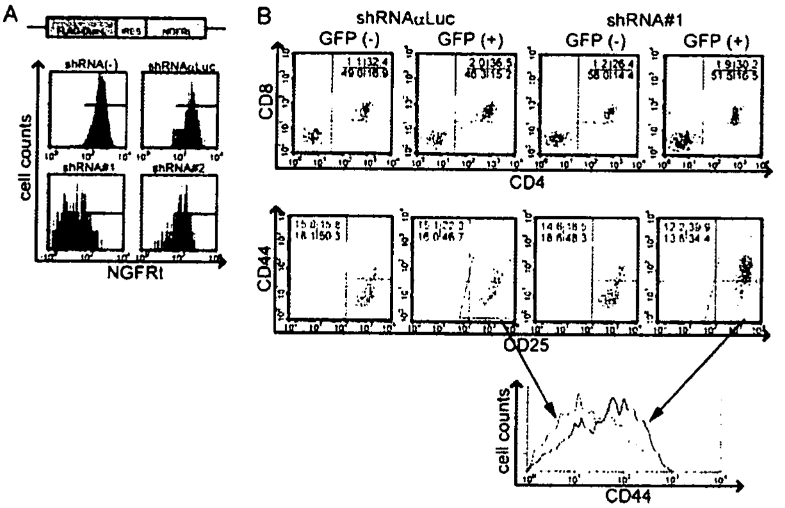

FIGURE 6. Effect of *Duxl* knockdown on expression levels of CD44. *A*, NIH3T3 cells stably expressing FLAG-Duxl-IRES-NGFRt (*upper left*) were infected with retroviruses encoding shRNA directed against a luciferase sequence (αLuc; *upper right*) or shRNA against a DuxL sequence #1 (*lower left*) and sequence #2 (*lower right*). Transduction of shRNA was tracked by GFP. NGFRt expression on GFP^{+} cells was analyzed by flow cytometry. *B*, FL cells infected with Luc-shRNA retroviruses (αLuc) or Duxl-shRNA retroviruses (sequence #1) were cultured on OP9-DL1 cells and expression of CD4/CD8 (*upper panels*) and CD25/CD44 (*lower panels*) was analyzed after 17 days. Dot plots are gated on GFP^{-}CD4^{-}CD8^{-} cells or GFP^{+}CD4^{-}CD8^{-} cells.

(p = 0.00147) and CD44 signaling (p = 0.0086) were extracted from the CL1 cluster, reflecting the fact that *c-kit* and *CD44* are down-regulated in DN3 and also indicating that the pathways extracted from CL1 are expected to be inactivated in DN3. Other pathways that are known to undergo dynamic regulations during DN thymocyte development were also extracted from different clusters, including the Runx1 pathway from CL2 (gradually down-regulated), NFAT and AP-1 pathways from CL3 (down-regulated in DN4), TCRαβ, NF-κB, and Notch pathways from CL4 (transiently activated in DN3), and Ets and TCRαβ pathways from CL5/CL6 (up-regulated in DN3 or DN4). The other pathways that were extracted with high significance values but have not been previously implicated in T cell development include those related to platelet-derived growth factor, γ-aminobutyric acid, and peroxisome proliferator-activated receptor α.

Expression and structure of 110051B16Rik *or* Duxl *gene*

In view of clarifying gene regulations that operate during thymocyte development, of particular interest are genes included in CL4, because they show a unique temporal expression profile of transient induction or up-regulation in DN3 and contain key genes for the thymocyte development or β selection, including *pre-TCRα*, *SpiB*, *Egr3*, and *Notch3* (33–37) (Fig. 2*B*). Especially, we were interested in genes that were thought to be involved in transcriptional regulations (Table II). Among these genes, we focused on a gene, *1100051B16Rik*, which encodes a putative transcription factor with unknown functions (Table II and Fig. 4*A*). It has an open reading frame of 1050 nt and the predicted protein shares structural features and sequence similarities with the families of double homeobox proteins that are thought to have rapidly diverged during evolution (Fig. 4*B*) (38). Although currently no definite human ortholog is uniquely identified due to incomplete sequence homology to human sequences, it shows the highest similarity with human DUXA or DUX4 in their amino acids sequence and gene structure with DUXA (Fig. 4, *B* and *C*), and thus, was named as *Duxl* (*Dux* in lymphoid lineage). In hemopoietic compartments, expression of *Duxl* is largely restricted to DN3 thymocytes and B220-positive B cells in bone marrow and spleen, implicating their functional roles in both subsets of lymphocytes, although it is also expressed in embryos from mid to late gestation, as well as in other non-hemopoietic adult organs, including brain, prostate, and uterus (Fig. 4*D*).

Function of Duxl *in thymocyte development*

To get an insight into the role of *Duxl* in thymocyte development, we first examined its expression among the subpopulations of DN3 thymocytes by FACS sorting DN3 cells into two populations according to their expression of intracellular TCRβ (iTCRβ) chains (Fig. 5*A*, *upper panel*) (8, 39). In quantitative PCR (qPCR) analysis, *Duxl* transcripts were detected in the iTCRβ^{-} fraction (DN3a) but not in the iTCRβ^{+} fraction (DN3b), indicating that the *Duxl* expression is largely restricted to the DN3a fraction (Fig. 5*A*, *lower panel*).

To explore the effect of constitutively expressed *Duxl* on DN thymocyte differentiation, we transduced the *Duxl* cDNA into mouse FL cells using a bicistronic retrovirus vector with a marker GFP cDNA in the second cistron, and the *Duxl*-transduced FL cells were then assayed on the OP9-Delta-1 (OP9-DL1) culture system that can mimic intrathymic development of DP thymocytes from the FL hemopoietic progenitors (29). As previously described, the mock-infected FL cells generated substantial numbers of GFP^{+} DP thymocytes after 15 days of culture on OP9-DL1 (Fig. 5*B*, *upper central panel*) (19). In contrast, the number of GFP^{+} DP cells was dramatically reduced in the culture from the *Duxl*-transduced FL cells (Fig. 5*B*, *upper right panel*). Although the total cell numbers were not significantly different between both cultures, the GFP^{+} cells from the *Duxl*-transduced culture generated an increased proportion of the CD25^{+}CD44^{low} cells, compared with mock-infected cells (Fig. 5*B*, *middle panels*) that produced a similar proportion of CD25^{+}CD44^{low} cells as GFP^{-} cells (data not shown). The increased proportion of CD25^{+}CD44^{low} in *Duxl*-transduced culture was accompanied with a reduced DN2 population. The GFP^{+}CD25^{+}CD44^{low} cells in the *Duxl*-transduced FL cells had a Thy1^{+} CD11c^{-} Mac1^{-}NK1.1^{-}B220^{-} phenotype. We extracted total RNAs from mock-infected DN2 and DN3 and Duxl-introduced DN2 and DN3 cells cultured on OP9-DL1 and analyzed their expression of *Rag1*, *PU.1*, and c-kit using real-time PCR and flow cytometry, respectively. In Duxl-introduced DN2 and DN3 cells, expression of *Rag1* was enhanced and expression of *PU.1* was reduced, whereas no difference was observed in c-kit expression between mock-infected and Duxl-introduced cells (Fig. 5*C*). Therefore, Duxl is supposed to promote the DN2/DN3 transition in some respects, while not in others.

Mock-infected and *Duxl*-transduced CD25^{+}CD44^{low} cells were sorted and their genomic DNA was analyzed by PCR for TCRβ

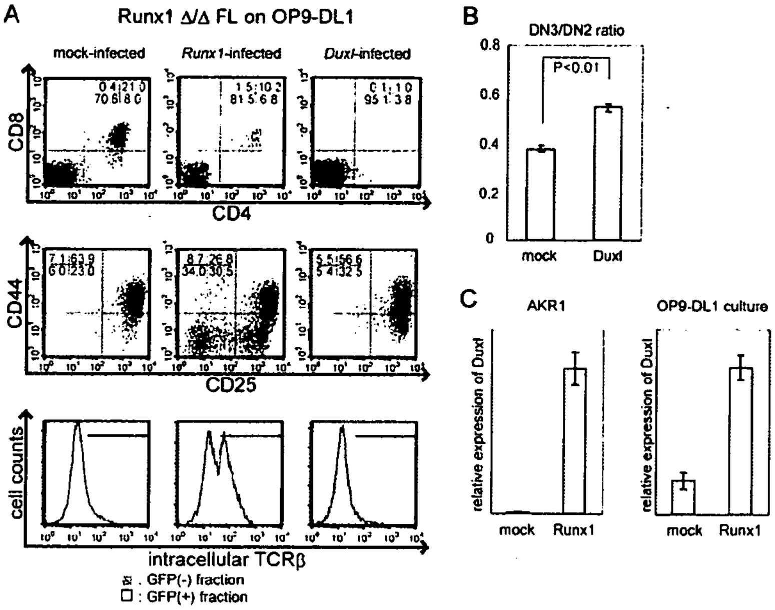

FIGURE 7. Effect of *Duxl* transduction on *Runx1*-deficient FL cells cultured on OP9-DL1. *A*, Expression of CD4/CD8 (*upper panels*), CD25/CD44 (*middle panels*), and iTCRβ (*lower panels*, solid line: GFP^{+}, gray shade: GFP^{-}) was examined in OP9-DL1 culture of FL cells harvested from the conditional knockout mice for the *Runx1* gene, in which two floxed alleles of *Runx1* are deleted by Cre recombinase induced from the *Lck* promoter. FL cells were transfected with mock virus (*left*) or with retrovirus expressing *Runx1* (*middle*) or *Duxl* (*right*). Representative FACS analyses are shown. *B*, Average DN3:DN2 ratio of mock-infected and *Duxl*-introduced cells in three independent cultures are shown with ±SD. The two groups were compared using Student's *t* test. *C*, *Runx1* was overexpressed in FL cells or AKR1 cells using retrovirus vector, and the expression levels of *Duxl* in the GFP^{+} fraction were examined by qPCR. The amount of transcript of *Duxl* was normalized to the amount of GAPDH RNA in each population and is shown as relative values. Data shown are the average ± SD from triplicate samples.

rearrangement. Interestingly, although the fraction of iTCRβ-positive cells in *Duxl*-transduced DN cells was substantially reduced compared with that in control DN cells (Fig. 5*B*, *lower panels*), the extent of DJβ and V-DJβ rearrangement in *Duxl*-transduced CD25^{+}CD44$^{\text{low}}$ cells was comparable to those of control cells or rather slightly accelerated compared with control cells (Fig. 5*D*), indicating that Duxl actively represses expression of rearranged *TCRβ* genes. Indeed, expression of surface TCRβ in the mouse thymoma cell line AKR1, which exhibits the DP phenotype (CD4^{+}CD8^{+}TCR$^{\text{int}}$), was reduced after *Duxl* was introduced using retroviral vector (Fig. 5*E*). Taken together, these data suggest that the increased CD25^{+}CD44$^{\text{low}}$ population in the *Duxl*-transduced cells show a DN3a-like phenotype and that constitutive expression of Duxl promotes the DN2/DN3 transition but compromises the process of β selection.

To further investigate the role of Duxl in DN2/DN3 transition, we knocked down the expression of *Duxl* in FL cells using RNA interference, in which FL cells were transfected with a retrovirus that produces shRNA directed against *Duxl* and examined for their development in the OP9-DL1 culture. Among three different shRNAs (sequences 1–3), only sequence 1 shRNA showed significant RNA interference as confirmed by the reduced cell surface expression of a truncated form of nerve growth factor receptor (NGFRt) in NIH3T3, where the NGFRt was translated from a bicistronic message of the *Duxl-IRES-NGFRt* (Fig. 6*A*). When transduced with sequence 1 shRNA, FL cells showed a reduced DN2/DN3 transition in OP9-DL1 culture as determined by the higher levels of CD44 expression in GFP^{+}CD25^{+} cells compared with GFP^{-}CD25^{+} cells within the same culture or with GFP^{+}CD25^{+} cells in the *αLuc* (mock)-infected culture (Fig. 6*B*).

It was of particular interest to explore a possible functional link between Duxl and Runx1, because we showed that Runx1 is essential for the DN2/DN3 transition (18). Thus, we tested whether Duxl can rescue the phenotype of Runx1 deficiency in developing thymocytes by transducing *Duxl* into FL cells from *Runx1*-deficient mice in OP9-DL1 culture. Although mock-transduced *Runx1*-deficinet FL cells showed the defective DN2/DN3 transition with impaired down-regulation of CD44 (19), *Duxl* transduction clearly increased the CD44$^{\text{low}}$CD25^{+} population (Fig. 7, *A* and *B*), although the latter population still did not express intracellular TCRβ chains (Fig. 7*A*, *lower panels*). Thus, *Duxl* can induce down-regulation of CD44 in a *Runx1*-independent manner. We further examined whether *Duxl* is induced by Runx1 or not, where *Runx1* was transduced into normal FL cells in OP9-DL1 culture or an AKR1 cell line, and expression of *Duxl* was measured by quantitative PCR 72 h after the transduction. In both experiments, *Runx1*-transduced GFP^{+} cells showed a marked increase in *Duxl* expression (Fig. 7*C*). These results indicated that Runx1 promotes the DN2/DN3 transition by, at least in part, regulating expression of *Duxl*, although it is not clear whether this regulation is direct or indirect.

Discussion

Several groups have investigated the gene expression profiles of developing thymocytes (40–43). Despite the differences in analytical methods, panels of genes to be examined, microarray used, thymocyte subpopulation analyzed, or species, the reported expression pattern of some representative genes such as *pTα*, *Nothc3*, *Egr3*, or *SpiB* were reproduced in our study (15, 40, 42). Furthermore, our study revealed a new aspect that was not described so far and adds to knowledge of the development of thymocytes. Intrathymic development of thymocytes is a dynamic process, during which immature thymocytes progressively commit to mature T cell differentiation accompanied by explosive expansions of their population. As such, we initially expected that this process be driven by induction of a large number of regulatory genes. Unexpectedly, however, our expression profiling clearly showed that gene regulation during early thymocyte development is characterized by mostly repressive activities of transcription, where 90% of the differentially expressed genes were eventually down-regulated. This means that as thymocytes differentiate from their progenitors, they become to use more limited sets of genes at least before differentiating into DP cells.

Among these down-regulated genes, a notable subset is a group of genes that show lineage-promiscuous expression. According to our somewhat arbitrary criteria, these genes account for a substantial proportion (20%) of down-regulated genes in DN2. In other words, immature thymocytes in the DN2 subset simultaneously express genes from different lineage-affiliated programs. Such lineage-promiscuous states were previously implicated for multipotent hemopoietic progenitors based on the observation for limited numbers of lineage-affiliated genes (44–46). We confirmed this for DN thymocytes by analyzing expression of a large number of genes and also evaluated their temporal changes during the course of their development. Of note is that mature T cell-related genes expressed in DN2 are also transiently repressed during DN2 to DN4 progression. Because our criteria of >10 times higher expression than the median could be too strict, more DN2 genes undergoing down-regulation thereafter could be explained under this framework.

The interpretation of this lineage-promiscuous gene expression in very immature thymocytes is elusive, but it may be speculated that these immature cells are conditioned or primed before their differentiation into particular lineages and that these "primed" states are represented by expression of multiple lineage-affiliated genes. In this scenario, the lineage-promiscuous expression of DN thymocytes could be related to plasticity of these cells. DN2 thymocytes are committed to T lineage to a certain extent but still can give rise to other lineages such as NK cells (10, 12–14), which might reflect the observation that NK cell-affiliated genes were more slowly down-regulated during DN thymocyte development (Fig. 3*C*). Alternatively, lineage-promiscuous gene expression could be explained by the heterogeneous lineage potential at the level of cellular complexity. Because CD44^{+}LCD25^{low} DN1 cells are composed of multiple subsets with distinct differentiation capacity that could be identified with additional c-kit and CD24 staining (11), DN1 cells express a large number of lineage-promiscuous genes down-regulated at the DN1/DN2 transition (40). This may be also the case with DN2 cells we have sorted without c-kit and CD24 staining in this study. Although, the observation that differentiation capacity of DN1 cells with lysozyme expression was similar to that of DN1 cells without lysozyme expression (47) strongly supports the former interpretation, it is necessary to examine the gene expression profiles of each single DN cell, as well as more precise sorting with c-kit and CD24 staining, to know which interpretation is more accurate.

Computational mapping of differentially expressed genes showing similar temporal profiles (CL1–CL6) to known functional pathways seems to be effective to extract the relevant pathways in DN thymocyte development, because it successfully extracted well-characterized pathways in developing thymocytes, such as those related to *Runx1*, *TCRβ*, *NF-κB*, *NFAT*, and *Notch* genes. In addition, it also implicated the presence of several previously unknown pathways that might be regulated during thymocyte development. Although further evaluations are required, finding of these potentially relevant pathways will provide valuable clues to the exploring molecular mechanisms of regulation of DN thymocyte development.

During DN thymocyte development, only a minority of genes was newly induced, among which those included in the CL4 cluster were of our particular interest, because they were transiently expressed within a window in DN3 and contained several well-known pathways and genes that are critical for thymocyte development. From a few candidate genes within this cluster that are potentially involved in transcriptional regulations, we identified a novel double homeobox gene, named *Duxl*, which is transiently expressed in DN3a and could have critical roles in regulation of DN thymocyte development. When constitutively expressed in FL cells in OP9-DL1 culture, Duxl enhances the proportion of CD44^{low}CD25^{+} thymocytes that mimic the phenotype of DN3a cells with severely reduced DP cells, while the knockdown of Duxl in FL cells impaired the proper DN2/DN3 transition. Thus, Duxl is thought to play an important role in promoting the DN2/DN3 transition in the development of DN thymocytes.

In early thymocyte development, Duxl has several functional similarities to SpiB, in that it is highly expressed in DN3a as well as B cells, and that normal thymocyte development is impaired when constitutively expressed or knocked down in early thymocytes (33). These observations may implicate a possible functional link between SpiB and Duxl, which should be addressed in further analysis. For example, SpiB also regulates B cell differentiation, implicating that Duxl is also involved in development of B cells.

Of particular interest is the finding that *Duxl* is induced by Runx1 expression and can partially rescue the *Runx1*-deficient phenotype with regard to the down-regulation of cell surface CD44, indicating that Duxl is one of the effecter molecules involved in the DN2/DN3 transition that is regulated by Runx1. The detailed molecular mechanism of the DN2/DN3 transition is largely unknown because only limited numbers of mice strains, namely, *Runx1*-deficient mice (18) and *pTα/common cytokine receptor γ-chain* double knock-out mice (20), exhibit the maturational block between the DN2 and the DN3 stage. However, considering the fact that during the transition from the DN2 to the DN3 stage, the *TCRβ* gene is rearranged and αβT cell, γδT cell, and NK cell lineages begin to diverge (10, 48, 49), the DN2/DN3 transition, on which our findings could shed light, is supposed to be a crucial developmental step.

On the other hand, the interpretation of the severely reduced production of DP cells associated with constitutive expression of *Duxl* in FL cells is complicated in a context of its physiological roles. Since the *Duxl*-transduced FL cells seem to show a maturational block at the transition between DN3a and DN3b, during which Duxl undergoes down-regulation in vivo, it may be postulated that *Duxl* being normally down-regulated in this step is important for the β selection and subsequent DP thymocyte production. We may safely conclude that the precise regulation of Duxl expression is essential for DP thymocytes to be normally generated in OP9 culture, but its physiological functions in β selection and DP thymocyte generation are still elusive. To address this, more sophisticated experimental approaches with precisely targeted expression of *Duxl* in vivo should be required.

In conclusion, through the gene expression profiling of chronologically discrete subsets of DN thymocytes, we demonstrated the predominantly repressive gene regulation during DN thymocyte development along with its implication in lineage-promiscuous gene expression in immature thymocytes. Among the gene cluster showing transient expression in DN3, we identified *Duxl*, a novel double homeobox gene, that is induced by Runx1 and involved in regulation of DN thymocyte development.

Acknowledgments

We thank J. C. Zúñiga-Pflücker for OP9-DL1 stromal cells, H. Nakauchi for pGCDNsam retroviral vector, M. Miyagishi for designing shRNA, and T. Kitamura for PlatE packaging cells.

Disclosures

The authors have no financial conflict of interest.

References

1. Ceredig, R., and T. Rolink. 2002. A positive look at double-negative thymocytes. *Nat. Rev. Immunol.* 2: 888–897.
2. Pearse, M., L. Wu, M. Egerton, A. Wilson, K. Shortman, and R. Scollay. 1989. A murine early thymocyte developmental sequence is marked by transient expression of the interleukin 2 receptor. *Proc. Natl. Acad. Sci. USA* 86: 1614–1618.
3. Godfrey, D. I., J. Kennedy, T. Suda, and A. Zlotnik. 1993. A developmental pathway involving four phenotypically and functionally distinct subsets of CD3^{-}CD4^{-}CD8^{-} triple-negative adult mouse thymocytes defined by CD44 and CD25 expression. *J. Immunol.* 150: 4244–4252.

4. Penit, C., B. Lucas, and F. Vasseur. 1995. Cell expansion and growth arrest phases during the transition from precursor (CD4$^-$8$^-$) to immature (CD4$^+$8$^+$) thymocytes in normal and genetically modified mice. *J. Immunol.* 154: 5103–5113.
5. Godfrey, D. I., J. Kennedy, P. Mombaerts, S. Tonegawa, and A. Zlotnik. 1994. Onset of TCR-β gene rearrangement and role of TCR-β expression during CD3$^-$CD4$^-$CD8$^-$ thymocyte differentiation. *J. Immunol.* 152: 4783–4792.
6. Mombaerts, P., A. R. Clarke, M. A. Rudnicki, J. Iacomini, S. Itohara, J. J. Lafaille, L. Wang, Y. Ichikawa, R. Jaenisch, M. L. Hooper, et al. 1992. Mutations in T-cell antigen receptor genes α and β block thymocyte development at different stages. *Nature* 360: 225–231.
7. Shinkai, Y., S. Koyasu, K. Nakayama, K. M. Murphy, D. Y. Loh, E. L. Reinherz, and F. W. Alt. 1993. Restoration of T cell development in RAG-2-deficient mice by functional TCR transgenes. *Science* 259: 822–825.
8. Hoffman, E. S., L. Passoni, T. Crompton, T. M. Leu, D. G. Schatz, A. Koff, M. J. Owen, and A. C. Hayday. 1996. Productive T-cell receptor β-chain gene rearrangement: coincident regulation of cell cycle and clonality during development in vivo. *Genes Dev.* 10: 948–962.
9. Allman, D., A. Sambandam, S. Kim, J. P. Miller, A. Pagan, D. Well, A. Meraz, and A. Bhandoola. 2003. Thymopoiesis independent of common lymphoid progenitors. *Nat. Immunol.* 4: 168–174.
10. Schmitt, T. M., M. Ciofani, H. T. Petrie, and J. C. Zúñiga-Pflücker. 2004. Maintenance of T cell specification and differentiation requires recurrent notch receptor-ligand interactions. *J. Exp. Med.* 200: 469–479.
11. Porritt, H. E., L. L. Rumfelt, S. Tabrizifard, T. M. Schmitt, J. C. Zúñiga-Pflücker, and H. T. Petrie. 2004. Heterogeneity among DN1 prothymocytes reveals multiple progenitors with different capacities to generate T cell and non-T cell lineages. *Immunity* 20: 735–745.
12. Ikawa, T., H. Kawamoto, S. Fujimoto, and Y. Katsura. 1999. Commitment of common T/natural killer (NK) progenitors to unipotent T and NK progenitors in the murine fetal thymus revealed by a single progenitor assay. *J. Exp. Med.* 190: 1617–1626.
13. Wu, L., C. L. Li, and K. Shortman. 1996. Thymic dendritic cell precursors: relationship to the T lymphocyte lineage and phenotype of the dendritic cell progeny. *J. Exp. Med.* 184: 903–911.
14. Manz, M. G., D. Traver, T. Miyamoto, I. L. Weissman, and K. Akashi. 2001. Dendritic cell potentials of early lymphoid and myeloid progenitors. *Blood* 97: 3333–3341.
15. David-Fung, E. S., M. A. Yui, M. Morales, H. Wang, T. Taghon, R. A. Diamond, and E. V. Rothenberg. 2006. Progression of regulatory gene expression states in fetal and adult pro-T-cell development. *Immunol. Rev.* 209: 212–236.
16. Radtke, F., A. Wilson, G. Stark, M. Bauer, J. van Meerwijk, H. R. MacDonald, and M. Aguet. 1999. Deficient T cell fate specification in mice with an induced inactivation of Notch1. *Immunity* 10: 547–558.
17. Michie, A. M., and J. C. Zúñiga-Pflücker. 2002. Regulation of thymocyte differentiation: pre-TCR signals and β-selection. *Semin. Immunol.* 14: 311–323.
18. Ichikawa, M., T. Asai, T. Saito, G. Yamamoto, S. Seo, I. Yamazaki, T. Yamagata, K. Mitani, S. Chiba, H. Hirai, et al. 2004. AML-1 is required for megakaryocytic maturation and lymphocytic differentiation, but not for maintenance of hematopoietic stem cells in adult hematopoiesis. *Nat. Med.* 10: 299–304.
19. Kawazu, M., T. Asai, M. Ichikawa, G. Yamamoto, T. Saito, S. Goyama, K. Mitani, K. Miyazono, S. Chiba, S. Ogawa, et al. 2005. Functional domains of Runx1 are differentially required for CD4 repression, TCRβ expression, and CD4/8 double-negative to CD4/8 double-positive transition in thymocyte development. *J. Immunol.* 174: 3526–3533.
20. Di Santo, J. P., I. Aifantis, E. Rosmaraki, C. Garcia, J. Feinberg, H. J. Fehling, A. Fischer, H. von Boehmer, and B. Rocha. 1999. The common cytokine receptor γ chain and the pre-T cell receptor provide independent but critically overlapping signals in early αβ T cell development. *J. Exp. Med.* 189: 563–574.
21. Li, C., and W. H. Wong. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* 98: 31–36.
22. Li, C., and W. H. Wong. 2001. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* 2: 0032.0031–0032.0011.
23. Tusher, V. G., R. Tibshirani, and G. Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* 98: 5116–5121.
24. de Hoon, M. J., S. Imoto, J. Nolan, and S. Miyano. 2004. Open source clustering software. *Bioinformatics* 20: 1453–1454.
25. Herrero, J., F. Al-Shahrour, R. Diaz-Uriarte, A. Mateos, J. M. Vaquerizas, J. Santoyo, and J. Dopazo. 2003. GEPAS: A web-based resource for microarray gene expression data analysis. *Nucleic Acids Res.* 31: 3461–3467.
26. Herrero, J., J. M. Vaquerizas, F. Al-Shahrour, L. Conde, A. Mateos, J. S. Diaz-Uriarte, and J. Dopazo. 2004. New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res.* 32: W485–W491.
27. Sato, H., S. Ishida, K. Toda, R. Matsuda, Y. Hayashi, M. Shigetaka, M. Fukuda, Y. Wakamatsu, and A. Itai. 2005. New approaches to mechanism analysis for drug discovery using DNA microarray data combined with KeyMolnet. *Curr. Drug Discov. Technol.* 2: 89–98.
28. Boyle, E. I., S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock. 2004. GO: Termfinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics* 20: 3710–3715.
29. Schmitt, T. M., and J. C. Zúñiga-Pflücker. 2002. Induction of T cell development from hematopoietic progenitor cells by delta-like-1 in vitro. *Immunity* 17: 749–756.
30. Anderson, S. J., K. M. Abraham, T. Nakayama, A. Singer, and R. M. Perlmutter. 1992. Inhibition of T-cell receptor β-chain gene rearrangement by overexpression of the non-receptor protein tyrosine kinase p56^{lck}. *EMBO J.* 11: 4877–4886.
31. Miyagishi, M., and K. Taira. 2003. Strategies for generation of an siRNA expression library directed against the human genome. *Oligonucleotides* 13: 325–333.
32. Su, A. I., T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* 101: 6062–6067.
33. Lefebvre, J. M., M. C. Haks, M. O. Carleton, M. Rhodes, G. Sinnathamby, M. C. Simon, L. C. Eisenlohr, L. A. Garrett-Sinha, and D. L. Wiest. 2005. Enforced expression of Spi-B reverses T lineage commitment and blocks β-selection. *J. Immunol.* 174: 6184–6194.
34. Carleton, M., M. C. Haks, S. A. Smeele, A. Jones, S. M. Belkowski, M. A. Berger, P. Linsley, A. M. Kruisbeek, and D. L. Wiest. 2002. Early growth response transcription factors are required for development of CD4$^-$CD8$^-$ thymocytes to the CD4$^+$CD8$^+$ stage. *J. Immunol.* 168: 1649–1658.
35. Miyazaki, T. 1997. Two distinct steps during thymocyte maturation from CD4$^-$CD8$^-$ to CD4$^+$CD8$^+$ distinguished in the early growth response (Egr)-1 transgenic mice with a recombinase-activating gene-deficient background. *J. Exp. Med.* 186: 877–885.
36. Xi, H., R. Schwartz, I. Engel, C. Murre, and G. J. Kersh. 2006. Interplay between RORγt, Egr3, and E proteins controls proliferation in response to pre-TCR signals. *Immunity* 24: 813–826.
37. Bellavia, D., A. F. Campese, A. Vacca, A. Gulino, and I. Screpanti. 2003. Notch3, another Notch in T cell development. *Semin. Immunol.* 15: 107–112.
38. Booth, H. A., and P. W. Holland. 2007. Annotation, nomenclature, and evolution of four novel homeobox genes expressed in the human germ line. *Gene* 387: 7–14.
39. Taghon, T., M. A. Yui, R. Pant, R. A. Diamond, and E. V. Rothenberg. 2006. Developmental and molecular characterization of emerging β- and γδ-selected pre-T cells in the adult mouse thymus. *Immunity* 24: 53–64.
40. Tabrizifard, S., A. Olaru, J. Plotkin, M. Fallahi-Sichani, F. Livak, and H. T. Petrie. 2004. Analysis of transcription factor expression during discrete stages of postnatal thymocyte differentiation. *J. Immunol.* 173: 1094–1102.
41. Huang, Y. H., D. Li, A. Winoto, and E. A. Robey. 2004. Distinct transcriptional programs in thymocytes responding to T cell receptor, Notch, and positive selection signals. *Proc. Natl. Acad. Sci. USA* 101: 4936–4941.
42. Dik, W. A., K. Pike-Overzet, F. Weerkamp, D. de Ridder, E. F. de Haas, M. R. Baert, P. van der Spek, E. E. Koster, M. J. Reinders, J. J. van Dongen, et al. 2005. New insights on human T cell development by quantitative T cell receptor gene rearrangement studies and gene expression profiling. *J. Exp. Med.* 201: 1715–1723.
43. Hoffmann, R., L. Bruno, T. Seidl, A. Rolink, and F. Melchers. 2003. Rules for gene usage inferred from a comparison of large-scale gene expression profiles of T and B lymphocyte development. *J. Immunol.* 170: 1339–1353.
44. Akashi, K., D. Traver, T. Miyamoto, and I. L. Weissman. 2000. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* 404: 193–197.
45. Hu, M., D. Krause, M. Greaves, S. Sharkis, M. Dexter, C. Heyworth, and T. Enver. 1997. Multilineage gene expression precedes commitment in the hemopoietic system. *Genes Dev.* 11: 774–785.
46. Nutt, S. L., B. Heavey, A. G. Rolink, and M. Busslinger. 1999. Commitment to the B-lymphoid lineage depends on the transcription factor Pax5. *Nature* 401: 556–562.
47. Laiosa, C. V., M. Stadtfeld, H. Xie, L. de Andres-Aguayo, and T. Graf. 2006. Reprogramming of committed T cell progenitors to macrophages and dendritic cells by C/EBPα and PU.1 transcription factors. *Immunity* 25: 731–744.
48. Ciofani, M., G. C. Knowles, D. L. Wiest, H. von Boehmer, and J. C. Zúñiga-Pflücker. 2006. Stage specific and differential notch dependency at the αβ and γδ T lineage bifurcation. *Immunity* 25: 105–116.
49. Prinz, I., A. Sansoni, A. Kissenpfennig, L. Ardouin, M. Malissen, and B. Malissen. 2006. Visualization of the earliest steps of γδ T cell development in the adult thymus. *Nat. Immunol.* 7: 995–1003.

Genome-Wide, High-Resolution Detection of Copy Number, Loss of Heterozygosity, and Genotypes from Formalin-Fixed, Paraffin-Embedded Tumor Tissue Using Microarrays

Sharoni Jacobs,[1] Ella R. Thompson,[2,3] Yasuhito Nannya,[4] Go Yamamoto,[4] Raji Pillai,[1] Seishi Ogawa,[5] Dione K. Bailey,[1] and Ian G. Campbell[2,3]

[1]Affymetrix, Inc., Santa Clara, California; [2]Victorian Breast Cancer Research Consortium Cancer Genetics Laboratory, Peter MacCallum Cancer Centre, East Melbourne, Victoria, Australia; [3]Department of Pathology, University of Melbourne, Parkville, Victoria, Australia; and Departments of [4]Hematology/Oncology and [5]Regeneration Medicine for Hematopoiesis, University of Tokyo, Tokyo, Japan

Abstract

Formalin-fixed, paraffin-embedded (FFPE) material tends to yield degraded DNA and is thus suboptimal for use in many downstream applications. We describe an integrated analysis of genotype, loss of heterozygosity (LOH), and copy number for DNA derived from FFPE tissues using oligonucleotide microarrays containing over 500K single nucleotide polymorphisms. A prequalifying PCR test predicted the performance of FFPE DNA on the microarrays better than age of FFPE sample. Although genotyping efficiency and reliability were reduced for FFPE DNA when compared with fresh samples, closer examination revealed methods to improve performance at the expense of variable reduction in resolution. Important steps were also identified that enable equivalent copy number and LOH profiles from paired FFPE and fresh frozen tumor samples. In conclusion, we have shown that the Mapping 500K arrays can be used with FFPE-derived samples to produce genotype, copy number, and LOH predictions, and we provide guidelines and suggestions for application of these samples to this integrated system. [Cancer Res 2007;67(6):2544–51]

Introduction

The challenges associated with DNA derived from formalin-fixed, paraffin-embedded (FFPE) samples have prevented widespread application of FFPE DNA to many of the technologies available for high-quality DNA, although some options with lower genomic coverage are available (1–3). In this study, we show the feasibility and limitations of a genome-wide assessment of genotype, loss of heterozygosity (LOH), and copy number using FFPE DNA on the Affymetrix Mapping 500K array set, which includes the Mapping 250K Nsp Array and the Mapping 250K Sty Array (Santa Clara, CA). These arrays use a process termed whole-genome sampling analysis (WGSA; ref. 4), in which genomic DNA is digested and ligated to adaptors. A subset of digested fragments are then PCR amplified in a complexity reduction step before hybridization to the arrays. PCR proved to be the critical step when processing FFPE samples.

We compared several extraction methods to determine which protocol provides FFPE DNA most suitable for array analysis and found that a PCR-based assessment of DNA quality predicted the downstream performance of FFPE DNA samples better than age of FFPE sample. We identified a necessity for (*a*) *in silico* compensation against fragment size bias and (*b*) a fragment size filter during analysis of FFPE samples. We tested our new guidelines for FFPE DNA qualification and analysis on archival samples of various tissue types, storage times, and location sources. Quality of FFPE DNA varied but the methods outlined by this study enabled prediction of performance. These results show that FFPE DNA can be suitable for a combined study of genotype, LOH, and copy number on a whole-genome scale.

Materials and Methods

Sample selection and DNA extraction. Five primary endometrioid ovarian cancers were selected without screening for the initial portion of this study. For each sample set, normal lymphocytic DNA, fresh tumor tissue, and FFPE tissue were analyzed. Samples were collected between 1993 and 1999 as part of a larger study of ovarian cancer in women living in and around Southampton, United Kingdom (5). At the time of collection, DNA was extracted from blood samples and fresh tumor biopsies were snap frozen in liquid nitrogen. A portion of each frozen tumor biopsy was sectioned to assess the proportion of tumor. For samples 526T and 594T, microdissection was done (6) to obtain DNA with a >80% tumor component. DNA was extracted from the fresh frozen tissue using a salt chloroform method (7).

In 2002, a portion of each frozen tumor biopsy was formalin fixed and paraffin embedded as described previously (8), with all tumors fixed in 10% neutral buffered formalin for <24 h at room temperature. At the time of DNA extraction, the FFPE tumors had been embedded in paraffin blocks for 3 years. Five sections (10 μm) were deparaffinized twice in xylene (5 min) and rehydrated in 100%, 90%, and 70% ethanol (1 min each). The sections were stained with hematoxylin (4 min) and washed with water (1 min), acid alcohol (10 s), water (1 min), Scott's tap water (1 min), and water (1 min). The sections were then stained with eosin (3 min), rinsed with water (10 s), and dehydrated in 70%, 90%, and 100% ethanol (30 s each). Tumor cells were manually microdissected under a dissecting microscope as described previously (6) to obtain high-purity (>80%) tumor DNA. The tumor component for sample 594 was high enough that it was not stained or microdissected. DNA was extracted from the five endometrioid FFPE tissues using a modified Qiagen protocol (Valencia, CA; described below). Following DNA extraction from FFPE tissue, a salt precipitation DNA cleanup was done as described in the Affymetrix GeneChip Mapping Assay Manuals.

For the study of independent sample sets, DNA was extracted from FFPE tissue from 17 breast tumors and 8 colorectal tumors. FFPE blocks were collected from 11 pathology laboratories and ranged in age from 1 to 17 years. The formalin fixation and paraffin embedding protocols used for these tissues are not known but are likely to be quite varied. For breast

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

S. Jacobs and E.R. Thompson contributed equally to this work.

Conflict of Interest Statement: S. Jacobs, R. Pillai, and D.K. Bailey are employees of Affymetrix, Inc.

Requests for reprints: Sharoni Jacobs, Affymetrix, Inc., 3420 Central Expressway, Santa Clara, CA 95051. Phone: 408-731-5880; Fax: 408-481-0435; E-mail: sharoni_jacobs@affymetrix.com.

doi:10.1158/0008-5472.CAN-06-3597

tumors, 10 μm sections were deparaffinized, stained with H&E, and manually microdissected (described above). The colorectal tumors were not stained or microdissected due to their high tumor component. DNA was extracted from breast and colorectal tissues (described below), and as before, a salt precipitation DNA cleanup was done.

The collection and use of tissues for this study were approved by the appropriate institutional ethics committees.

Trial of DNA extraction methods for FFPE tissue. Five DNA extraction methods were trialed using whole 20 μm sections from three FFPE blocks. The methods that were compared were the MagneSil Genomic Fixed Tissue System (Promega,[6] Madison, WI), ChargeSwitch gDNA Micro Tissue kit (Invitrogen,[7] Carlsbad, CA), PureGene (Gentra Systems,[8] Minneapolis, MN), DNeasy Tissue kit (Qiagen[9]), and a phenol/chloroform extraction. With the exception of the DNeasy Tissue kit and phenol/chloroform, the extractions were done according to the manufacturer's instructions. The extractions done with the DNeasy Tissue kit and with phenol/chloroform both were modified to include an initial incubation at 95°C for 15 min followed by 5 min at room temperature as described previously (9), before being digested with proteinase K for 3 days at 56°C in a rotating oven with periodic mixing and fresh enzyme added each 24 h. A salt precipitation was done on DNA from all five extraction methods.

DNA quality assessment and preparation. The extracted DNA was quantified using UV spectroscopy at 260 nm. Random amplified polymorphic DNA-PCR (RAPD-PCR; ref. 10) was done to assess the quality of DNA and maximum fragment lengths as described previously using 50, 5, or 0.5 ng DNA (11). Qiagen HotStarTaq was used, with 0.4 units per reactions (Qiagen[9]). Products were visualized with ethidium bromide on a 3% gel.

Preparation and application of DNA to the mapping arrays. Matched fresh and FFPE samples were analyzed on the Affymetrix GeneChip Human Mapping 10K v2 Xba Array and 50K Xba Array and prepared using the Mapping 10K v2 Assay kit and the Mapping 100K Assay kit (Affymetrix)[10] The only exception to the manufacturer's protocol was that 10 cycles were added to the PCR cycling conditions for each FFPE sample.

Matched fresh tumor, FFPE tumor, and normal samples were assayed using the Mapping 250K Nsp Assay kit and the Mapping 250K Sty Assay kit[10] and hybridized to the 250K arrays. The 500K assay was done according to the manufacturer's protocol, beginning with 250 ng DNA. Ninety micrograms of PCR product were fragmented and labeled, using additional PCRs when necessary for FFPE breast and colorectal samples.

Data analysis. Genotype calls were produced using the dynamic model algorithm (12) by the Affymetrix GeneChip Genotyping Analysis Software version 4.0. A stringent P value cutoff threshold of 0.26 was used. Concordance was determined by calculating the number of single nucleotide polymorphisms (SNP) that gave the same call in both fresh frozen and FFPE DNA from the same tumor and dividing this number by the total number of SNPs that were called in both samples.

LOH predictions were produced using dChipSNP software (dChip2005_f4 version[11]; ref. 13). LOH values were inferred using the Hidden Markov Model and restricting to SNPs on fragment sizes ≤700 bp.

Copy number estimates for ovarian tumor samples using 500K data were determined by pairing tumor and matching normal samples in CNAG_v2.0.[12] Nonpaired, nonmatching references were used for copy number prediction of 10K and 50K data. Log 2 ratios were imported into Spotfire DecisionSite (Spotfire,[13] Somerville, MA) and the Affymetrix Integrated Genome Browser for visualization and comparison. Copy number estimates for breast and colon FFPE tumors were done using data from 48 HapMap samples (available online[10]) as a reference.

Estimated inter-SNP mean and median distances after exclusion of fragment sizes >700 bp were determined by first calculating the distance between all SNPs on each chromosome. Distances were then sorted per chromosome in descending order and the largest distances (representing centromeres) were removed for each chromosome, except for the acrocentric chromosomes 13 to 15 and 21 to 22.

Pearson (linear) correlations were calculated in Partek Genomics Suite (Partek,[14] St. Louis, MO).

Microsatellite analysis. Nine microsatellite markers were used to assess LOH at three loci: chromosome 1q (D1S2816, D1S413, and D1S1726), chromosome 7p (D7S691, D7S670, and D7S2506), and chromosome 14q (D14S1011, D14S258, and D14S1002). Regions were selected where array-based LOH analysis showed discordant LOH results for fresh and FFPE-derived DNA. The forward primer was labeled with a 5′-fluorescent dye (FAM or HEX). The samples were analyzed using a 3130 Genetic Analyzer (Applied Biosystems,[15] Foster City, CA) with POP7 polymer. An assessment of LOH was done using GeneMapper version 3.7. LOH was scored by calculation of the ratio of tumor DNA peaks (T1/T2) compared with that in the normal DNA to give a relative ratio (T1/T2)/(N1/N2). A ratio of 0 indicates complete allele loss and a ratio of 1 indicates no LOH. A ratio of <0.5 was scored as indicative of LOH.

Results

DNA extraction from FFPE tissue. Five DNA extraction methods (phenol/chloroform, Qiagen DNeasy Tissue kit, Invitrogen ChargeSwitch, Promega MagneSil, and Gentra PureGene) were tested on consecutive sections from different FFPE ovarian tumor biopsies. Phenol/chloroform and modified Qiagen protocols (see Materials and Methods) provided the highest DNA yield as determined by UV spectroscopy; these yields were 2.2 times more than the average yield from any of the other three extraction protocols (Fig. 1*A*). RAPD-PCR, which uses primers of 10 bps to produce a ladder of amplicons, was also done to assess both amplification efficiency and maximum product size for each extraction protocol (11). Compared with DNA extracted from fresh lymphocytes, the FFPE-derived DNA from all extraction methods yielded consistently smaller PCR fragments, with a maximum reliable size of ~800 bp (Fig. 1*A*). Phenol/chloroform and modified Qiagen extractions produced more intense and consistent PCR fragments across dilutions, suggesting that products were relatively free of contaminant inhibitors (Fig. 1*A*). DNA extracted with these two methods was processed through the PCR step of the Mapping 50K Xba Assay to further assess amplification efficiency. In this test, the modified Qiagen extraction provided a slightly higher PCR yield on average than the phenol/chloroform method (21.4 μg compared with 19.2 μg) and was therefore chosen for DNA extraction from FFPE tissues in this study.

Mapping 500K array performance. Five matched sets; each containing (*a*) nontumor, non-FFPE lymphocytic DNA, (*b*) fresh frozen ovarian tumor DNA, and (*c*) FFPE ovarian tumor DNA; were assessed for performance on the Mapping 500K arrays. All five FFPE samples had been stored for 3 years and provided average RAPD-PCR maximum amplicon sizes from 526 to 800 bp. During the PCR step of the Mapping assay, amplification products from all five FFPE tumors were concentrated <700 bp, a fragment size range that was reduced compared with non-FFPE samples (Fig. 1*B*). Decreased yield from the Mapping PCRs (Table 1) accompanied the decrease in amplicon size distributions. FFPE samples produced

[6] http://www.promega.com
[7] http://www.invitrogen.com
[8] http://www.gentra.com
[9] http://www.qiagen.com
[10] http://www.affymetrix.com
[11] http://www.dchip.org
[12] http://www.genome.umin.jp/
[13] http://www.spotfire.com

[14] http://www.partek.com
[15] http://www.appliedbiosystems.com

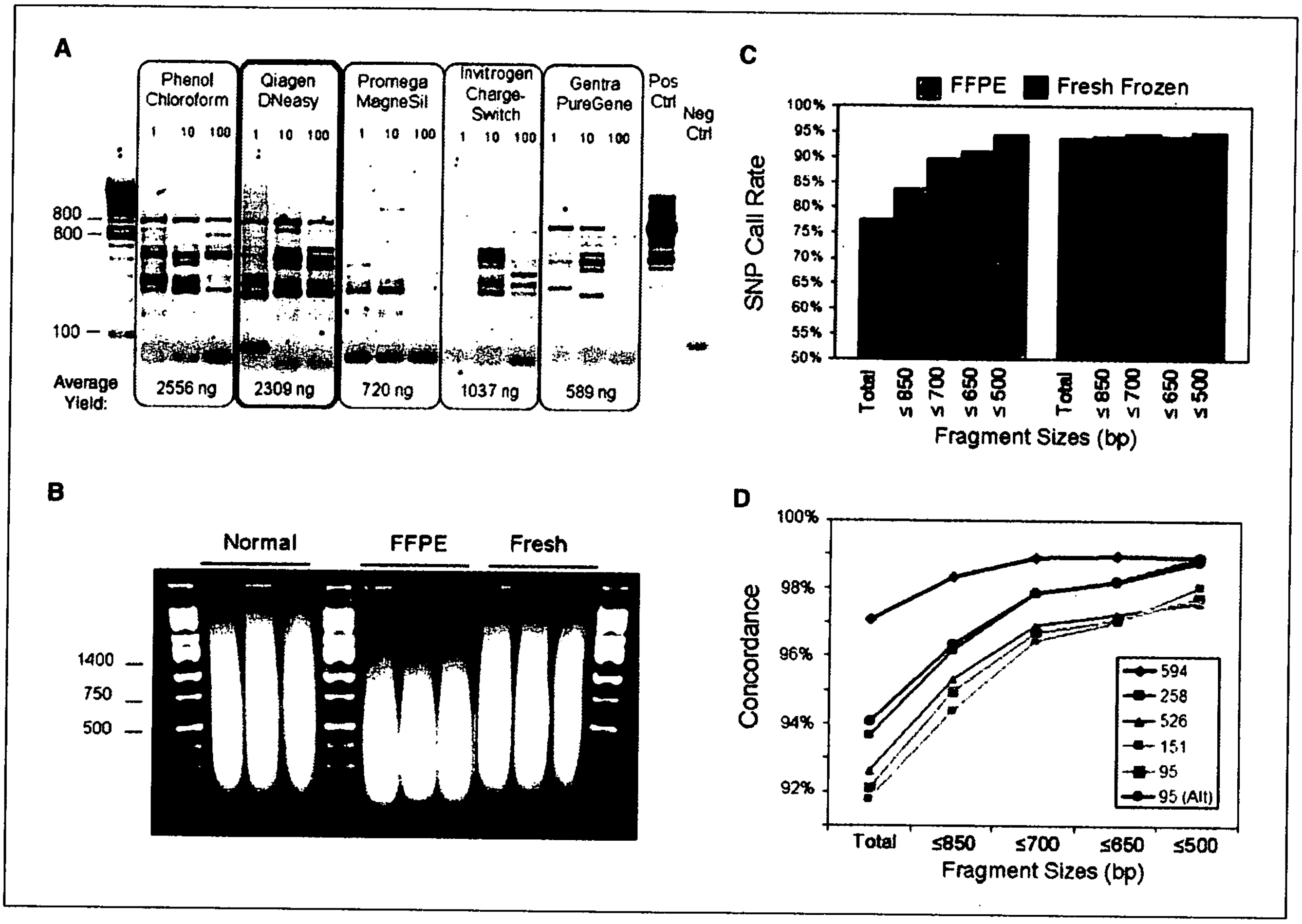

Figure 1. Performance of different FFPE DNA extraction methods and the Affymetrix GeneChip Mapping 500K assay. *A*, visualization of RAPD-PCR products on a 3% agarose gel comparing the undiluted DNA extraction (1), a 1:10 dilution of input DNA (10), and a 1:100 dilution of DNA (100) from one FFPE tissue (047) using five different extraction methods. The maximum fragment size in the extracted FFPE DNA samples reached 1,100 bp although only with sample dilution. The maximum reproducible fragment was 800 bp. DNA yield per extraction method is listed below. *B*, visualization of the PCR products during the Mapping 500K assay reveals a downshift in the distribution of fragment size, which is specific to the FFPE samples. *C*, SNP call rates are reduced in FFPE samples, but SNPs on smaller fragments are genotyped with equal efficiency from fresh and paraffin samples. The size dependence for higher call rates is specific to the FFPE samples. *D*, concordance between fresh frozen and matching FFPE samples is incrementally increased with fragment size selectivity, with larger dips in accuracy for sizes >700 bp. Exclusion of some regions (chromosomes 1q, 7p, 15, and 16q) shown to be genetically different between 95 fresh and FFPE samples causes an upshift in concordance for this sample (95 Alt versus 95).

63 to 83 μg PCR products for the Mapping 250K Nsp Array, whereas all non-FFPE samples produced >90 μg.

The assay was continued using 90 μg PCR product as the manual instructs or the total PCR yield when this was less than assay requirements. Importantly, the protocol was otherwise never modified. Normal and fresh tumor samples gave typical SNP call rates, with an average of 94.5% and 93.5%, respectively. These call rates are lowered due to application of a strict confidence score threshold ($P \leq 0.26$; the default threshold is $P \leq 0.33$). In contrast, FFPE samples achieved an overall average call rate of 79.84% and 75.17% for Nsp and Sty, respectively (Table 1). These decreased call rates are consistent with the poor amplification of larger fragments during PCR. Exclusion of SNPs on larger fragments significantly increased the call rates, such that incrementally more stringent fragment size restrictions incrementally increased call rates (Fig. 1*C*). In fact, stringent fragment size restrictions produced similar call rates between fresh frozen and FFPE samples, indicating that the Mapping 500K is well suited for FFPE DNA and identifying the limiting factor as the size of amplicons produced from the degraded DNA.

Concordance of genotype calls between paired FFPE and fresh frozen ovarian tumor DNA samples was examined to determine the reliability of genotypes from FFPE DNA. It is important to note that tumor heterogeneity lead to confirmed genuine differences in genomic content between matched FFPE and fresh frozen DNA, which would lower these concordance rates. Average overall concordance between FFPE and fresh frozen samples from the same tumor was 93.6%. Exclusion of the larger fragments increased concordance such that all SNPs located on fragment sizes ≤700 bp displayed an average of 97.4% concordance (Fig. 1*D*). Exclusion of several regions (chromosomes 1q, 7p, 15, and 16q) displaying heterogeneity between fresh frozen and paraffin sample 95 increased the concordance by >2% (Fig. 1*D*). These high rates of concordance, despite shown genetic differences between paired samples, underscore the reliability and reproducibility of genotype calls produced using FFPE-derived DNA samples with this