facilitate the identification of allelic imbalances such as copy-neutral LOH in the absence of a paired normal DNA reference.

The aberrations in chromosomes 1q, 2, 8, and 20 have been noted as the most commonly occurring aberrations in all previous reports,[21,22] as well as in the present study. In the present study, the most frequently detected aberrations were gains in chromosomes 1q and 2 (or 2q), observed in approximately 50% of the cases.

Trisomy in chromosome 1q is a well-known alteration in HBL.[34] Similar 1q imbalances have also been described in other pediatric neoplastic disorders such as lymphoma,[35] Wilms' tumor,[36] and sarcoma,[37] indicating that these aberrations are related to tumor progression. The candidate genes in 1q included the *NTRK1*, *ABL2*, *CD34*, *DAP3* (death receptor protein-3), and caspase-3 genes.[38] The anomalies in chromosome 2, which almost always result in gains in 2q, are also common in HBL. These imbalances are also commonly found in embryonal rhabdomyosarcoma and other pediatric tumors related to BWS. Translocation involving the *PAX3* gene located in 2q35 has been suggested to play a crucial role in the pathogenesis of alveolar rhabdomyosarcoma.[39] Based on this, a genetic link has been suggested between HBL and alveolar rhabdomyosarcoma. The role of the *PAX3* gene in the pathogenesis of HBL is yet to be determined. Additionally, the 2q24–32 region contains several genes that may also have an oncogenic potential. These include a serine/threonine kinase receptor, *ITRAF*, *FRZB*, a secreted antagonist of WNT signaling, and BRCA1-associated RING domain 1 (*BARD1*) genes. However, no specific gene has been identified in the previous,[21,22] and present studies.

The losses in chromosomes 4q and 11q were comprehensively observed. In hepatocellular carcinoma (HCC) cells, Wong *et al.* demonstrated a growth advantage following the loss in the 4q arm.[40] In HCC, 4q21–q22 and 4q35 have been identified as commonly deleted regions, and allelic losses in 4q35 have been associated with a larger tumor size and an aggressive histological tumor type.[41] Previous studies have not reported a significant correlation between HBL with loss in the distal 4q arm and prognosis, but the underlying oncogenic event might be due to the loss of a gene on the distal 4q arm.

Many minimal regions of amplification and deletion were detected using high-density SNP arrays, although homozygous deletion was not identified in any sample. The SNP loci located in 7q34 and 14q11.2 were found to be highly amplified in sporadic HBL samples. The candidate genes at these loci are *EphB6*, *DAD1*, and BCL-like 2 (*BCL2L2*) genes that encode the proteins associated with the execution of cell apoptosis. Gains as well as high amplifications in this region have not been reported previously; however, such an observation will be of particular interest for the discovery of oncogenes involved in the pathogenesis of HBL.

The UPD regions were identified in five of the 17 samples. This is chiefly important because UPD is being particularly considered as a possible mechanism of tumor initiation. During tumorigenesis, UPD is believed to arise due to a mitotic recombination caused by a rare crossover event during mitotic cell division. The products of mitotic recombination are the regions of the genome exhibiting UPD, and both the genomic regions originate from the same parent. We could identify a common UPD on chromosome 11p that is reminiscent of BWS with paternal UPD; in this case, the loss of function of the 11p15 maternal alleles through various mechanisms may be the critical event associated with tumorigenesis and BWS.[42] BWS is a neonatal overgrowth syndrome that predisposes an individual to cancer,[6] and the importance of the maternally active locus in chromosome 11p15 in tumorigenesis is supported by the finding that the loss of imprinted allele and paternal duplication leads to tissue overgrowth and subsequent tumor development. Methylation analysis was performed for the four HBL samples having UPD within 11p15, and hypermethylation of *H19* DMR was detected in all four HBL samples. Because *H19* DMR was hypermethylated on the paternal allele and hypomethylated on the maternally expressed allele in humans, we consider that the UPD within 11p15 was of paternal origin.

Two candidate genes, namely, *IGF2* and *H19*, are located within the telomeric region of chromosome 11p15.5 and have opposite imprinting patterns.[43] In the majority of human tissues, *IGF2* is expressed only from the paternal allele, whereas *H19* is transcribed only from the maternal allele. *H19* is an untranslated gene but has been suggested to function as a tumor suppressor.[44] In fetal and adult organs, the transcriptionally silent *H19* allele was extensively hypermethylated throughout the entire gene and its promoter. On the maternally expressed *H19* allele, *H19* DMR is unmethylated and can bind to the CTCF protein. On the paternal *H19* allele, *H19* DMR is highly methylated. This not only prevents the expression of the imprinted paternal *H19* alleles but also blocks the binding of the CTCF protein.[43] In general, the outcome of UPD with losses of the 11p15 maternal alleles in HBL is the same as that of the loss of imprinting on the inactivated, imprinted, and maternally expressed genes in BWS. Weksberg *et al.* proposed a dual pathway model for tumor development in BWS, wherein methylation defects at *H19* and/or *IGF2* in 11p15 were found to play a role in Wilms' and HBL tumorigenesis.[45] The combined loss of expressions in various 11p15-imprinted genes may contribute to tumorigenesis.

In the present study, we identified that the expression patterns of *IGF2* and *H19* were opposite between genes with and without the UPD in 11p15. This difference in the expression patterns might influence the clinical features of HBL. Further prospective studies are required to reveal any potential correlations between specific LOH and clinical outcomes.

In summary, the analysis of LOH and CN alterations using the SNP microarray in HBL samples revealed significant areas of allelic imbalance. We hypothesize that UPD, in addition to allelic imbalance, constitutes a novel genetic mechanism involved in tumorigenesis. Therefore, detailed characterizations such as functional studies should be conducted to elucidate the significance of the regions detected in this study, many of which may contain the candidate tumor suppressor genes and oncogenes involved in the pathogenesis of HBL.

Acknowledgments

This work was supported in part by a Grant-in-Aid for Cancer Research from the Ministry of Health, Labour and Welfare of Japan, a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and a grant from the 21st Century COE Program from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1 Roebuck DJ, Perilongo G. Hepatoblastoma: an oncological review. *Pediatr Radiol* 2006; **36**: 183–6.

2 Tiao GM, Bobey N, Allen S *et al.* The current management of hepatoblastoma: a combination of chemotherapy, conventional resection, and liver transplantation. *J Pediatr* 2005; **146**: 204–11.

3 Schnater JM, Kohler SE, Lamers WH, von Schweinitz D, Aronson DC. Where do we stand with hepatoblastoma? A review. *Cancer* 2003; **98**: 668–78.

4 Ikeda H, Matsuyama S, Tanimura M. Association between hepatoblastoma and very low birth weight: a trend or a chance? *J Pediatr* 1997; **130**: 557–60.

5 Hughes LJ, Michels VV. Risk of hepatoblastoma in familial adenomatous polyposis. *Am J Med Genet* 1992; **43**: 1023–5.

6 DeBaun MR, Tucker MA. Risk of cancer during the first four years of life in children from The Beckwith–Wiedemann Syndrome Registry. *J Pediatr* 1998; **132**: 398–400.

7 Fukuzawa R, Hata J, Hayashi Y, Ikeda H, Reeve AE. Beckwith–Wiedemann syndrome-associated hepatoblastoma: wnt signal activation occurs later in

tumorigenesis in patients with 11p15.5 uniparental disomy. *Pediatr Dev Pathol* 2003; **6**: 299–306.

8 Little MH, Thomson DB, Hayward NK, Smith PJ. Loss of alleles on the short arm of chromosome 11 in a hepatoblastoma from a child with Beckwith–Wiedemann syndrome. *Hum Genet* 1988; **79**: 186–9.

9 Albrecht S, von Schweinitz D, Waha A, Kraus JA, von Deimling A, Pietsch T. Loss of maternal alleles on chromosome arm 11p in hepatoblastoma. *Cancer Res* 1994; **54**: 5041–4.

10 Oda H, Imai Y, Nakatsuru Y, Hata J, Ishikawa T. Somatic mutations of the APC gene in sporadic hepatoblastomas. *Cancer Res* 1996; **56**: 3320–3.

11 Lengauer C, Kinzler KW, Vogelstein B. Genetic instabilities in human cancers. *Nature* 1998; **396**: 643–9.

12 Yeh YA, Rao PH, Cigna CT, Middlesworth W, Lefkowitch JH, Murty VV. Trisomy 1q, 2, and 20 in a case of hepatoblastoma: possible significance of 2q35-q37 and 1q12-q21 rearrangements. *Cancer Genet Cytogenet* 2000; **123**: 140–3.

13 Nagata T, Mugishima H, Shichino H *et al.* Karyotypic analyses of hepatoblastoma. Report of two cases and review of the literature suggesting chromosomal loci responsible for the pathogenesis of this disease. *Cancer Genet Cytogenet* 1999; **114**: 42–50.

14 Sainati L, Leszl A, Stella M *et al.* Cytogenetic analysis of hepatoblastoma: hypothesis of cytogenetic evolution in such tumors and results of a multicentric study. *Cancer Genet Cytogenet* 1998; **104**: 39–44.

15 Tonk VS, Wilson KS, Timmons CF, Schneider NR. Trisomy 2, trisomy 20, and del (17p) as sole chromosomal abnormalities in three cases of hepatoblastoma. *Genes Chromosomes Cancer* 1994; **11**: 199–202.

16 Park JP, Ornvold KT, Brown AM, Mohandas TK. Trisomy 2 and 19, and tetrasomy 1q and 14 in hepatoblastoma. *Cancer Genet Cytogenet* 1999; **115**: 86–7.

17 Balogh E, Swanton S, Kiss C, Jakab ZS, Secker-Walker LM, Olah E. Fluorescence in situ hybridization reveals trisomy 2q by insertion into 9p in hepatoblastoma. *Cancer Genet Cytogenet* 1998; **102**: 148–50.

18 Sainati L, Leszl A, Surace C, Perilongo G, Rocchi M, Basso G. Fluorescence in situ hybridization improves cytogenetic results in the analysis of hepatoblastoma. *Cancer Genet Cytogenet* 2002; **134**: 18–20.

19 Surace C, Leszl A, Perilongo G, Rocchi M, Basso G, Sainati L. Fluorescent in situ hybridization (FISH) reveals frequent and recurrent numerical and structural abnormalities in hepatoblastoma with no informative karyotype. *Med Pediatr Oncol* 2002; **39**: 536–9.

20 Parada LA, Limon J, Iliszko M *et al.* Cytogenetics of hepatoblastoma: further characterization of 1q rearrangements by fluorescence in situ hybridization: an international collaborative study. *Med Pediatr Oncol* 2000; **34**: 165–70.

21 Hu J, Wills M, Baker BA, Perlman EJ. Comparative genomic hybridization analysis of hepatoblastomas. *Genes Chromosomes Cancer* 2000; **27**: 196–201.

22 Weber RG, Pietsch T, von Schweinitz D, Lichter P. Characterization of genomic alterations in hepatoblastomas. A role for gains on chromosomes 8q and 20 as predictors of poor outcome. *Am J Pathol* 2000; **157**: 571–8.

23 Janne PA, Li C, Zhao X *et al.* High-resolution single-nucleotide polymorphism array and clustering analysis of loss of heterozygosity in human lung cancer cell lines. *Oncogene* 2004; **23**: 2716–26.

24 Huang J, Wei W, Zhang J *et al.* Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum Genomics* 2004; **1**: 287–99.

25 Peiffer DA, Le JM, Steemers FJ *et al.* High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* 2006; **16**: 1136–48.

26 Zhao X, Li C, Paez JG *et al.* An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res* 2004; **64**: 3060–71.

27 Nannya Y, Sanada M, Nakazaki K *et al.* A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res* 2005; **65**: 6071–9.

28 Yamamoto G, Nannya Y, Kato M *et al.* Highly sensitive method for genomewide detection of allelic composition in nonpaired, primary tumor specimens by use of affymetrix single-nucleotide-polymorphism genotyping microarrays. *Am J Hum Genet* 2007; **81**: 114–26.

29 Wong KK, Tsang YT, Shen J *et al.* Allelic imbalance analysis by high-density single-nucleotide polymorphic allele (SNP) array with whole genome amplified DNA. *Nucleic Acids Res* 2004; **32**: e69.

30 Trask BJ. Fluorescence in situ hybridization: applications in cytogenetics and gene mapping. *Trends Genet* 1991; **7**: 149–54.

31 Herman JG, Graff JR, Myohanen S, Nelkin BD, Baylin SB. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc Natl Acad Sci USA* 1996; **93**: 9821–6.

32 Li LC, Dahiya R. MethPrimer: designing primers for methylation PCRs. *Bioinformatics* 2002; **18**: 1427–31.

33 Koch A, Waha A, Hartmann W *et al.* Elevated expression of Wnt antagonists is a common event in hepatoblastomas. *Clin Cancer Res* 2005; **11**: 4295–304.

34 Douglass EC, Green AA, Hayes FA, Etcubanas E, Horowitz M, Wilimas JA. Chromosome 1 abnormalities: a common feature of pediatric solid tumors. *J Natl Cancer Inst* 1985; **75**: 51–4.

35 Kaneko Y, Variakojis D, Kluskens L, Rowley JD. Lymphoblastic lymphoma: cytogenetic, pathologic, and immunologic studies. *Int J Cancer* 1982; **30**: 273–9.

36 Kaneko Y, Kondo K, Rowley JD, Moohr JW, Maurer HS. Further chromosome studies on Wilms' tumor cells of patients without aniridia. *Cancer Genet Cytogenet* 1983; **10**: 191–7.

37 Nilsson M, Meza-Zepeda LA, Mertens F, Forus A, Myklebost O, Mandahl N. Amplification of chromosome 1 sequences in lipomatous tumors and other sarcomas. *Int J Cancer* 2004; **109**: 363–9.

38 Kissil JL, Kimchi A. Assignment of death associated protein 3 (DAP3) to human chromosome 1q21 by in situ hybridization. *Cytogenet Cell Genet* 1997; **77**: 252.

39 Turc-Carel C, Lizard-Nacol S, Justrabo E, Favrot M, Philip T, Tabone E. Consistent chromosomal translocation in alveolar rhabdomyosarcoma. *Cancer Genet Cytogenet* 1986; **19**: 361–2.

40 Wong N, Lai P, Lee SW *et al.* Assessment of genetic changes in hepatocellular carcinoma by comparative genomic hybridization analysis: relationship to disease stage, tumor size, and cirrhosis. *Am J Pathol* 1999; **154**: 37–43.

41 Bando K, Nagai H, Matsumoto S *et al.* Identification of a 1-cM region of common deletion on 4q35 associated with progression of hepatocellular carcinoma. *Genes Chromosomes Cancer* 1999; **25**: 284–9.

42 Koufos A, Hansen MF, Copeland NG, Jenkins NA, Lampkin BC, Cavenee WK. Loss of heterozygosity in three embryonal tumours suggests a common pathogenetic mechanism. *Nature* 1985; **316**: 330–4.

43 Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* 2000; **405**: 486–9.

44 Zhang Y, Shields T, Crenshaw T, Hao Y, Moulton T, Tycko B. Imprinting of human H19: allele-specific CpG methylation, loss of the active allele in Wilms tumor, and potential for somatic allele switching. *Am J Hum Genet* 1993; **53**: 113–24.

45 Weksberg R, Nishikawa J, Caluseriu O *et al.* Tumor development in the Beckwith–Wiedemann syndrome is associated with a variety of constitutional molecular 11p15 alterations including imprinting defects of KCNQ1OT1. *Hum Mol Genet* 2001; **10**: 2989–3000.

doi: 10.1111/j.1349-7006.2007.00710.x

Human Molecular Genetics, 2007, Vol. 16, No. 20 **2494–2505**
doi:10.1093/hmg/ddm205
Advance Access published on July 31, 2007

Evaluation of genome-wide power of genetic association studies based on empirical data from the HapMap project

Yasuhito Nannya[1,2,4], **Kenjiro Taura**[3], **Mineo Kurokawa**[1], **Shigeru Chiba**[2] **and Seishi Ogawa**[2,4,*]

[1]Department of Hematology/Oncology, [2]Department of Cell Therapy and Transplantation Medicine, Graduate School of Medicine and [3]Department of Information and Communication Engineering, Graduate School of Information Science, University of Tokyo, Tokyo 113-8655, Japan and [4]Core Research for Evolutional Science and Technology, Japan Science and Technology Agency, Saitama 332-0012, Japan

Received April 7, 2007; Revised and Accepted July 22, 2007

With recent advances in high-throughput single nucleotide polymorphism (SNP) typing technologies, genome-wide association studies have become a realistic approach to identify the causative genes that are responsible for common diseases of complex genetic traits. In this strategy, a trade-off between the increased genome coverage and a chance of finding SNPs incidentally showing a large statistics becomes serious due to extreme multiple-hypothesis testing. We investigated the extent to which this trade-off limits the genome-wide power with this approach by simulating a large number of case-control panels based on the empirical data from the HapMap Project. In our simulations, statistical costs of multiple hypothesis testing were evaluated by empirically calculating distributions of the maximum value of the χ^2 statistics for a series of marker sets having increasing numbers of SNPs, which were used to determine a genome-wide threshold in the following power simulations. With a practical study size, the cost of multiple testing largely offsets the potential benefits from increased genome coverage given modest genetic effects and/or low frequencies of causal alleles. In most realistic scenarios, increasing genome coverage becomes less influential on the power, while sample size is the predominant determinant of the feasibility of genome-wide association tests. Increasing genome coverage without corresponding increase in sample size will only consume resources without little gain in power. For common causal alleles with relatively large effect sizes [genotype relative risk ≥1.7], we can expect satisfactory power with currently available large-scale genotyping platforms using realistic sample size (∼1000 per arm).

INTRODUCTION

Genome-wide association studies have been proposed as a strategy to identify genetic factors with small to moderate genetic effects in the development of human diseases, as typically assumed for a common disease common variant (CDCV) model (1). In this strategy, a disease-associated locus is identified through single nucleotide polymorphisms (SNPs) that show 'significantly' different allele frequencies between affected (cases) and unaffected (controls) individuals, and a large number of SNPs are tested for association in an attempt to realistically identify such SNPs (2,3). Although only a theoretical perspective a decade ago (1), with the unprecedented advance in large-scale genotyping technologies (4–6), it has now become a realistic approach to exploring the genetic basis of human disease (7,8). In addition, recent efforts in the International HapMap Project to understand the genetic diversity among human populations (9,10) have greatly contributed to clarifying the extent to which the number of marker SNPs could be reduced to achieve given genome coverage, or how much genome coverage can be obtained with a given marker SNP set by optimally 'tagging' untyped SNPs based on the linkage disequilibrium (LD) observed in the human genome (11–16).

*To whom correspondence shóuld be addressed to: Department of Cell Therapy and Transplantation Medicine, The 21st Century COE Program, Graduate School of Medicine, University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8655, Japan. Tel: +81 358008741; Fax: +81 358046261; Email: sogawa-tky@umin.ac.jp

Meanwhile, the major interest of the most researchers, who plan genetic association studies, would be the practical success rates in such attempts and their efficient study designs, rather than mere genome coverage (17,18), because increase in genome coverage might not be linearly translated into gain in power (19,20). In addition, the more SNPs are genotyped to achieve better genome coverage, the higher hurdle is imposed for a target allele to be detected.

This dilemma, known as the trade-off between increased genome coverage and the consequent inflation of null statistics due to extreme multiple testing, is a unique feature of genetic association studies, and is best described by considering the distributions of test statistics for markers truly associated with a causative allele ('causal distribution') and for all other markers ('null distribution') (21). Regardless of the properties of the causative SNP and whether one or more tagging strategies are used, the null distribution for a given marker set depends on its genome coverage in the study population. In particular, the null distribution with complete genome coverage is related to the overall diversity of the human genome and should substantially shift to the right (7,8,22). On the other hand, for a given disease model, the size of the test statistic expected for the causative SNPs is limited by the number of samples to be analyzed, once they are directly captured by one or more marker SNPs. After all, the feasibility of genome-wide association studies, or the required sample size to obtain realistic power, is determined by the overall diversity of the human genome, or given restricted study resources, the diversity of the human genome determines the property of disease-associated SNPs that can be detected with this approach.

Our questions are, therefore, how diverse is the human genome in view of conducting genome-wide association studies, how much power could be obtained to identify causative SNPs given that diversity and how the typical study parameters affects the power in that situation? To answer these questions, we need to evaluate both null and causal distributions in a quantitative manner. Because both distributions intrinsically depend on the LD structure within N (typically $>\sim 10^{5-6}$) interrelated marker SNPs and the particular location of causative SNPs within the genome, they cannot be calculated in an algebraic manner, but need to be estimated based on the observed data of human genome variations (10,21). So we approach these issues by extensively simulating a large number of case-control panels under both null and alternative scenarios based on the data from the International HapMap Consortiums (9,10), and assess the feasibility and efficient designs of whole genome association studies by estimating the genome-wide power that would be obtained using this genetic approach under varying study conditions.

RESULTS

Estimation of null distributions of the maximum χ^2 statistics

In considering the issue of multiple testing in genetic association studies, it is convenient to evaluate the maximum value of the χ^2 statistic [$\max(\chi^2)$] in all the marker SNPs that are truly unrelated to the causative SNP (21). Different statistics can be used (23–26), but the power calculated for this statistic, i.e. the probability of $\max(\chi^2)$ indicating a true association, will provide a reasonable bottom line to discuss the feasibility of typical genetic association studies (21). When all N marker SNPs are independent, the null distribution for $\max(\chi^2)$ is given as

$$\varphi_N(\chi^2) = \frac{d}{d\chi^2}\{\phi(\chi^2)^N\},$$

where $\phi(\chi^2)$ is the cumulative density function of the χ^2 distribution (d.f. = 1). However, since SNPs in real marker sets are variably degenerated due to the presence of LD between adjacent SNPs, we empirically estimated the distribution of $\max(\chi^2)$ for a series of marker sets by simulating 10 000 null case-control panels, where each panel was generated by randomly resampling phased chromosomes from the HapMap data sets, and $\max(\chi^2)$ was calculated for each case-control panel. Although the number of resampled chromosomes for each case-control panel (i.e. the sample size) does not significantly affect the distributions (data not shown), there arises some concern about the possibility of underestimating the null distributions due to resampling from very limited numbers of chromosomes, because the latter procedure could restrict the freedom of allelic segregation within the same chromosome. To address this issue, we progressively divided the whole genome into larger numbers of sub-blocks consisting of 10 000 to 10 SNPs in the HapMap Phase II set, and resampled these sub-blocks to simulate distributions of $\max(\chi^2)$. Reducing the mean block size down to 7.1 kb, these divisions allow for greater freedom of allelic segregation, but does not significantly affect the $\max(\chi^2)$ distributions until the resampled block size becomes smaller than the mean LD length (27), indicating that our simulations are not likely to substantially underestimate the null distributions (Supplementary Material, Figure S1).

Figure 1 A shows the simulated null distributions in the CEU panel for varying numbers of randomly selected SNPs ('correlated' SNP sets). The number of segregating or polymorphic markers contained in each random set is designated as Ns. The theoretical distribution for the same numbers (Ns) of 'independent' SNPs, $\varphi_{Ns}(\chi^2)$, is also provided (Fig. 1B). The null distribution increases as the number of randomly selected SNP markers increases, and in a random 1000K set containing 681K segregating SNPs, the threshold χ^2 value that provides a genome-wide *P*-value of 0.05 or 0.01 becomes as large as 27.6 or 30.5, respectively. On the other hand, reflecting the growing inter-marker LD intensity, the empirical distributions gradually deviate from the theoretical ones, $\varphi_{Ns}(\chi^2)$'s, for increasing Ns within the corresponding marker sets, underscoring the importance of considering inter-marker LD to avoid overestimation of the statistical threshold for multiple testing, especially for higher marker density.

Evaluation of the inter-marker LD

The intensity of the inter-marker LD in a given marker set is more simply evaluated by fitting the simulated distribution to a theoretical one for independent Nc makers, $\varphi_{Nc}(\chi^2)$ (see Methods). Irrespective of marker sets, fitting is finely

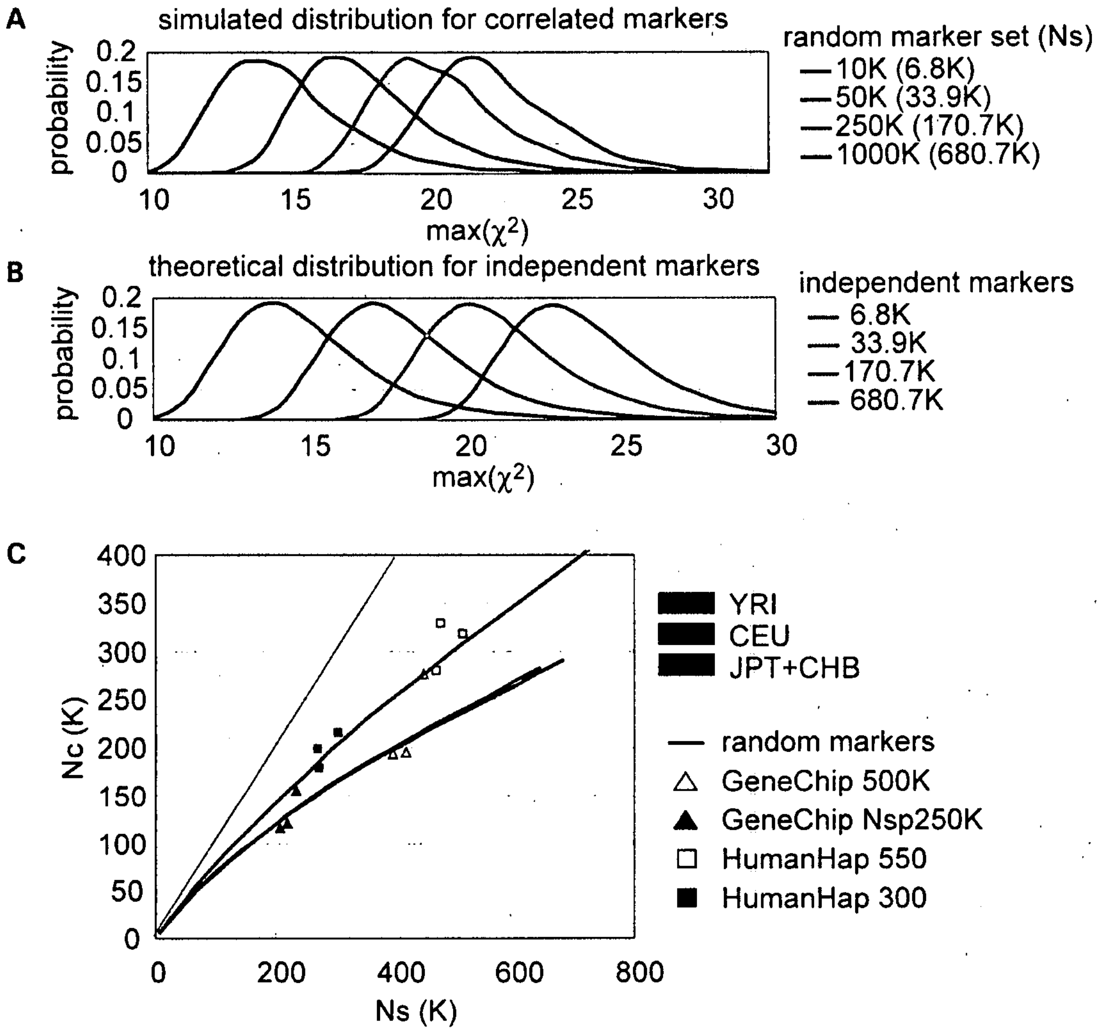

Figure 1. Null distributions of max(χ^2) and the effective number of independent SNPs (Nc) for various marker sets. Distributions of max(χ^2) for all null SNPs (null distributions) were simulated for increasing numbers of randomly selected SNP markers in the CEU panel. Ten thousand null panels, each consisting of 1000 cases and 1000 controls, were generated for the indicated marker sets by randomly resampling phased autosomal chromosomes from the HapMap Phase II data in CEU (**A**). Theoretical null distributions corresponding to each SNP set, $\varphi_{Ns}(\chi^2)$, were calculated assuming all Ns segregating SNPs therein are independent (**B**). The effective numbers of hypothetical independent SNPs (Nc) were estimated by fitting simulated null distributions to theoretical ones for Nc independent SNPs, $\varphi_{Nc}(\chi^2)$, for the indicated SNP sets, and are plotted against the number of segregating SNPs of the corresponding marker set (Ns) for different HapMap panels (**C**).

performed except in the vicinity of the maximal points (Supplementary Material, Figure S2). In particular, the distribution in extreme χ^2 values is satisfactorily approximated to provide a rough estimate of the nominal P-value for given genome-wide thresholds as confirmed by the concordance of the upper p point in the simulated distribution with the upper p/Nc point in the χ^2 distribution (d.f. = 1) (Bonferroni) (Table 1). In this formulation, it is reasonable to regard Nc as the number of hypothetical independent SNPs equivalent to the corresponding marker set, where the null distribution for a large number of mutually degenerated SNPs is described by an integer and the mean intensity of the inter-marker LD is measured through the Nc/Ns ratio.

Nc values were calculated for a variety of randomly selected SNP marker sets and plotted against the number of segregating SNP markers therein (Fig. 1C). As the Phase II data contain most of the SNPs in commercially available platforms, including Affymetrix® GeneChip® and Illumina® HumanHap® arrays (28–30), Nc values were also evaluated for these platforms (Supplemental Material, Table S1). Note that the numbers of segregating SNP markers varies among different HapMap panels, even though the same numbers of SNPs are randomly selected for each panel (Supplementary Material, Figure S3). Figure 1C illustrates how the degree of degeneration within marker SNPs increases in different HapMap panels as more marker SNPs are selected.

Table 1. Size of null distributions of max(χ^2) in various marker sets in the CEU panel

Platform	Ns	Nc	Fold degeneration	$P = 0.05$			$P = 0.01$		
				Nominal P^a	Actual^b	Bonferroni^c	Nominal P^a	Actual^b	Bonferroni^c
Random 10K	6.8K	6K	1.1	7.99×10^{-6}	19.94	19.86	1.57×10^{-6}	23.06	22.95
Random 30K	20.6K	17K	1.2	2.86×10^{-6}	21.91	21.85	5.73×10^{-7}	25.00	24.95
Random 50K	33.9K	27K	1.3	1.76×10^{-6}	22.84	22.74	4.01×10^{-7}	25.69	25.84
Random 125K	85.1K	60K	1.4	7.39×10^{-6}	24.51	24.28	1.56×10^{-7}	27.51	27.39
Random 250K	170.7K	105K	1.6	4.52×10^{-7}	25.46	25.36	9.04×10^{-8}	28.57	28.47
Random 500K	340.4K	179K	1.9	2.45×10^{-7}	26.64	26.39	5.39×10^{-8}	29.57	29.50
Random 1000K	680.7K	290K	2.3	1.48×10^{-7}	27.62	27.32	3.41×10^{-8}	30.46	30.44
GeneChip 500K	417.8K	196K	2.1	2.05×10^{-7}	26.99	26.56	4.94×10^{-8}	29.74	29.68
GeneChip Nsp250K	219.4K	120K	1.8	3.69×10^{-7}	25.85	25.62	7.94×10^{-8}	28.82	28.73
GeneChip 100K	101.3K	62K	1.6	7.75×10^{-7}	24.42	24.34	1.38×10^{-7}	27.75	27.45
HumanHap 300	305.1K	215K	1.4	2.18×10^{-7}	26.87	26.74	4.06×10^{-8}	30.12	29.86
HumanHap 550	513.8K	318K	1.6	1.41×10^{-7}	27.71	27.50	2.90×10^{-8}	30.77	30.62
HapMap Phase II	2557.4K	603K	4.2	7.09×10^{-8}	29.04	28.74	1.48×10^{-8}	32.08	31.86
ENCODE 7 regions	7.7K	1.3K	5.8						

[a] Nominal P-value to reach given experiment-wide significance obtained from actual distribution.
[b] The upper $1-P$ point of the actual null distribution.
[c] The argument of χ^2 distribution (d.f.=1) for cumulative density $1 - P/\text{Nc}$.

For example, 681K segregating SNPs within a random 1000K set in the CEU panel are equivalent to independent 290K SNPs, indicating that in this panel, these SNPs are degenerated 2.3-fold. On the other hand, the degeneration in 1000K random markers is reduced to 1.8-fold for the YRI panel, as expected from the lower inter-marker LD for this panel compared to that of CEU.

The SNPs on the Affymetrix® GeneChip® mapping array sets are degenerated to the same degree as random SNP sets, reflecting the fact that the SNPs on GeneChip® platforms are virtually randomly selected. In contrast, the SNPs on the Illumina® HumanHap300 are selected by efficiently tagging the HapMap Phase I SNPs in CEU, in which redundant SNPs are effectively eliminated (28). As a result, degeneration in the HumanHap300 is substantially reduced compared to the corresponding random marker sets. In CEU, Nc for this 305.1K segregating SNP set (215K Nc) exceeds that for 417.8K segregating SNPs on GeneChip® 500K set (196K), as predicted by the higher genome coverage of the former set (see Table 1 and Supplementary Material, Figure S4). The tagging for CEU also increases the Nc in JPT+CHB, suggesting that tagging in one panel is also effective to a certain degree for another (31,32). The tagging seems to be less efficient in YRI, because the Nc value of HumanHap300® in YRI is less deviated from that of the random marker set with a corresponding Ns. In HumanHap550®, more tag SNPs are selected from YRI, which contributes to the relative increase in Nc for this marker set compared to that for the corresponding random marker SNP set.

Estimation of Nc for common SNPs in complete genome coverage

It is particularly interesting to calculate the Nc values for the ENCODE regions, in which human variations have been most densely explored. Currently 10 regions have been extensively genotyped in the ENCODE Project (http://www.hapmap.org/downloads/encode1.html.en), of which we used 7 regions that had been randomly chosen from the genome. A total of 7741, 9832 and 7396 SNPs are segregated in these seven ENCODE regions, and they are equivalent to 1340 (5.8-fold), 2580 (3.8-fold), and 1460 (5.1-fold) hypothetical independent SNPs, in the CEU, YRI, and JPT+CHB panels, respectively. Assuming the entire genome shows the similar LD intensity to that in the seven ENCODE regions on average, the Nc values for common SNPs in complete genome coverage (Nc^G) are roughly estimated to be 1971K (YRI), 1023K (CEU), and 1115K (JPT+CHB) (Table 2), although the values would be much more inflated if rare polymorphisms [minor allele frequency (MAF) <0.01], many of which could not be found in the HapMap panels, are taken into consideration. Nc/Nc^G could also be used as another indicator of genome coverage of a given marker set.

Causal distribution of max(χ^2)

In view of power estimation, our next interest was the expected size of causal distributions relative to that of the inflated null distributions under varying disease/study parameters that affect the former distributions. To illustrate this, we simulated causal distributions of max(χ^2) for representative CEU alleles assumed to be causative (Fig. 2). Two thousand case-control panels were generated for each simulation, in which phased HapMap SNPs within 500 Kb around the causative locus were randomly resampled assuming a multiplicative model with varying genotype relative risks (GRRs) and the max(χ^2) was calculated for the resampled marker SNPs on GeneChip® 500K. Prevalence of the trait was set to 0.05. While the χ^2 threshold for genome-wide p of 0.05 could inflate from 19.9 for the random 10K set (6K Nc; semi-solid line) to as high as 29.8 for complete genome coverage (1023K Nc^G; dotted lines), these costs of multiple testing are acceptable when LD capture of the causative SNP by one or more markers with high correlation coefficient (r^2) can create large causal distributions with practical sample sizes (Fig. 2D–F), i.e. when the causal allele is common

Table 2. The number of corresponding independent markers

	ENCODE[a]	Whole genome[b]	All Phase II[c]
YRI	2580	1971K	1049K
CEU	1340	1023K	603K
JPT + CHB	1460	1115K	632K

[a] Nc values calculated for combined SNPs from seven regions.
[b] Nc of ENCODE regions are extrapolated to the entire genome.
[c] Nc of all SNPs in the HapMap Phase II.

(MAF > 0.2) and has a large GRR (>1.7) (Fig. 2A, D and G). In contrast, in the case where the causal allele with smaller MAF value (<0.2) or with a modest to weak GRR (<1.5) is to be detected, the trade-off between increased chance to capture the allele with higher r^2 using more markers and the accompanying cost of multiple testing can offset the power to varying degrees (Fig. 2A–C, G–I). The effect of 'collaborative' capture, i.e. the probability of detecting an association by one of the multiple surrounding marker SNPs other than the SNPs showing max(r^2), creates measurable gain in causal distributions and overall power, but does not essentially influence the above observations (Supplementary Material, Figure S5).

Estimation of genome-wide power

Based on the above consideration, we estimated the genome-wide power in genetic association studies for common (MAF ≧ 0.05) causal alleles with weak to moderate genetic effects. To do this, after assuming all the common SNPs in the human genome being equally causative, we used two sets of SNPs, the Ref$^{\text{ENCODE}}$ and the Ref$^{\text{Phase II 5Kb}}$ sets (see Methods), as references that are considered as random sampling from the entire SNPs. For each putative causative SNP, we simulated case-control panels as described in the previous section, and calculated the single point power as the proportion of simulated panels whose max(χ^2) exceeded a predetermined χ^2 threshold corresponding to a genome-wide $P = 0.01$ or 0.05 for each marker set. For genome-wide power, each single point power was averaged for all common SNPs within the reference set. For the Ref$^{\text{Phase II 5Kb}}$ set, over-representation of the direct association was adjusted based on the estimated genome coverage of the Phase II data set (see Methods). Figure 3 shows the genome-wide power in the CEU panel that was calculated for the Ref$^{\text{Phase II 5Kb}}$ for moderate to small effect sizes (i.e. GRR ≦ 1.7) assuming various parameter values. The calculation on the Ref$^{\text{ENCODE}}$ set provides a largely equivalent estimation of the power (Supplementary Material, Figure S6), although the power is expected to be less reliable for smaller marker sets, reflecting their poor representation of the genome.

Under strong genetic effects (GRR ≧ 2.0) and large sample sizes (≧ 1500/arm), the power tends to saturate as the number of randomly selected SNPs increases (≧ 250K), because most of the common SNPs would be already captured by one or more marker SNPs with enough r^2 (Supplementary Material, Figure S4), and the capture causes large shifts of causal distributions to the extent that the cost of multiple testing is trivial (Fig. 2). On the other hand, when causative SNPs with weak to moderate genetic effects are detected with insufficient sample numbers, causal distributions cannot exceed large thresholds resulting from extreme multiple testing, even though more and more SNPs are captured by strong LD. With increasing effect size and sample number, the genome coverage is less influential except for smaller numbers of marker SNPs (<250K). The power gain obtained with increased genome-coverage tends to be offset by the increased cost of multiple testing. After all, in most scenarios, genome coverage is less influential on power when ≧ 250K random markers or equivalent tag SNPs are used. In contrast, the effect of sample numbers is predominant. To detect weak genetic effects (GRR ≦ 1.3), the number of samples becomes critical. More than 4000 samples per arm will be required, but the requirement of genome coverage is not substantially increased when more than 250K randomly selected SNPs or their equivalents are used (Fig. 3A). Given a higher genetic effect, this dependence on sample size is dramatically ameliorated, but the genome coverage remains less influential.

Power in different HapMap panels and in commercially available platforms

Power is significantly reduced in YRI compared to CEU and JPT+CHB for any marker set (Fig. 4A–C). The lower power in YRI is mainly due to the lower 'relative' genome coverage of the marker set (Nc/Nc$^{\text{G}}$), rather than the higher cost of type I errors in this population.

The Illumina® HumanHap® series are commercially available platforms that incorporate the tagging theory, in which marker SNPs were selected to efficiently tag the CEU SNPs in the Phase I data set. Tagging seems to be effective, since HumanHap300® in the Ref$^{\text{Phase II 5Kb}}$ set shows slightly higher power than the GeneChip® 500K in CEU, although the power is slightly biased by the higher representation of the Phase I SNPs in the Ref$^{\text{Phase II 5Kb}}$ set (Fig. 4D). HumanHap300® shows comparable power to that of GeneChip® 500K, but the power of HumanHap300® is significantly reduced in YRI. In HumanHap550®, more tag SNPs from YRI and JPT+CHB were added to HumanHap300®, the power is more improved in YRI and in JPT+CHB, but the power is also increased to a lesser degree in CEU reflecting a transferability of tag SNPs between CEU and JPT+CHB. The power of various commercially available platforms with various sample sizes are shown in Figure 4E (adaptive threshold) and in Supplementary Material, Figure S7 (fixed threshold). Genome coverage and power of HumanHap550® in the CEU are comparable to those of the random 1000K set (Supplementary Material, Figure S4), an equivalent to Human SNP Array 6.0® that is planned by Affymetrix® (Fig. 4E). Nevertheless, and in spite of the significant difference in cost, the gain of power in HumanHap550® is not so prominent. Also note that the power calculation for HumanHap550® could be slightly biased by using the subset of the Phase II SNPs as a reference.

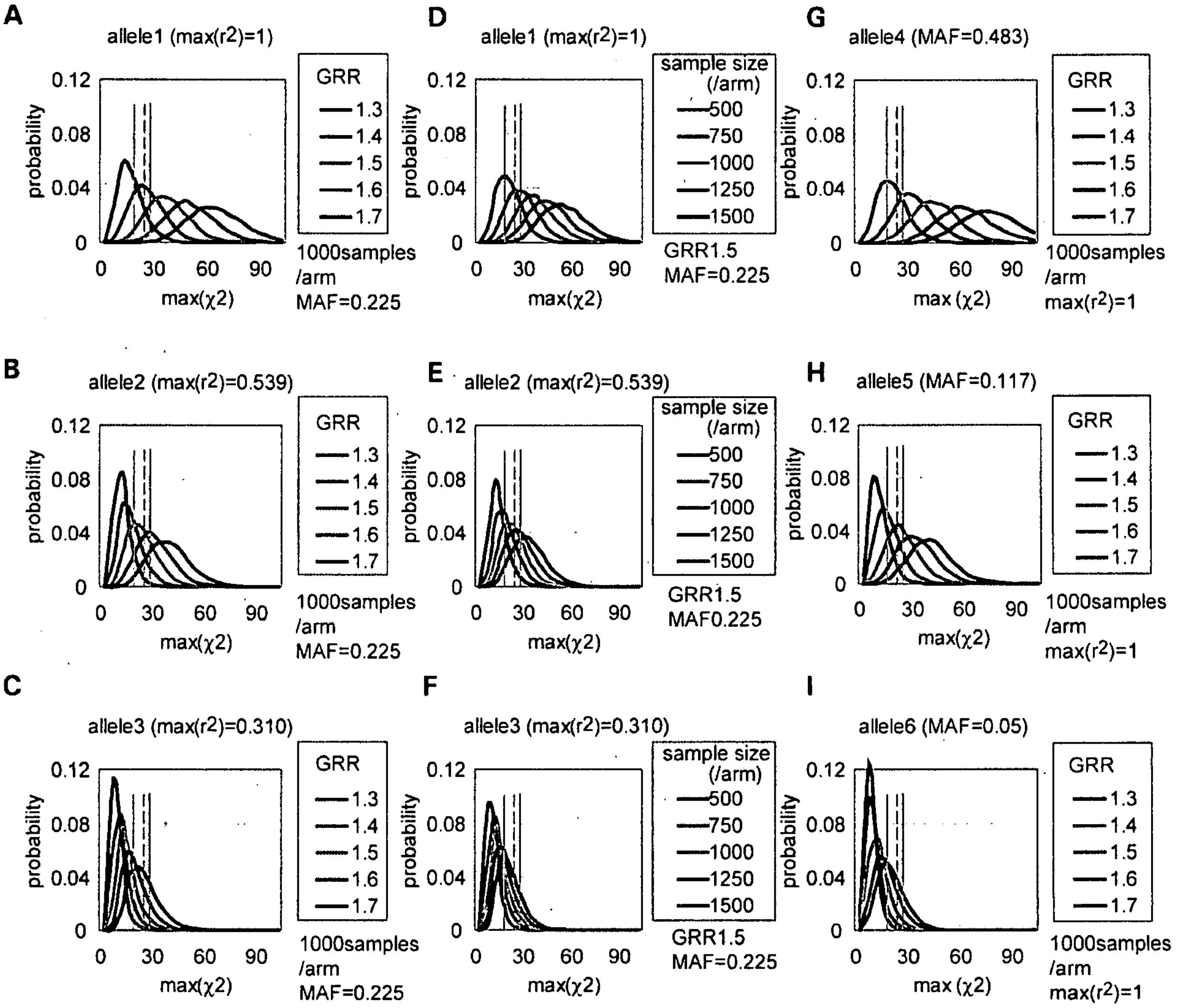

Figure 2. Enhancement of causal distributions by various parameters. Combined effects of LD [in max(r^2)] and effect size (in GRR) on causal distributions under constant sample size (1000/arm) and MAF value (0.225) (**A–C**), LD and sample size under constant effect size (GRR = 1.5) and MAF value (0.225) (**D–F**), and MAF and effect size under constant sample size (1000/arm) and LD [max(r^2) = 1.0] (**G–I**), are illustrated based on the simulations for six representative CEU alleles analyzed on GeneChip® 500K [rs9782915 in (A and D); rs7543006 in (B and E); rs731030 in (C and F); rs6603803 in (G); rs3052 in (H); rs1307490 in (I)]. Thresholds for genome-wide *P*-value of 0.05 are indicated for random 10K (solid lines), GeneChip 500K (dashed lines), and complete genome coverage (dotted lines), corresponding to Nc values of 6K, 196K, and 1023K (Nc^G), respectively. Effects of collaborative capture by nearby markers are incorporated, but they are generally small (Supplementary Material, Figure S5).

Power depends on allele frequencies of causative alleles

Power strongly depends on MAF of causative alleles, and detecting rare causative alleles is very difficult (Fig. 2) (8,20) for two reasons. First, rare variants are difficult to capture in high r^2 values. With currently available platforms (GeneChip® 500K or HumanHap550®), most SNPs with more than 0.10 MAF values are captured in high r^2, which could be effectively detected in high power given moderate GRRs (≧ 1.5) and sample size (≧ 1000/arm) (Fig. 5). In contrast, capturing rare causal SNPs (MAF < 0.10) requires many more marker SNPs or their combinations than capturing common SNPs at the more cost of multiple hypothesis testing. Second, even when captured in high r^2 with one or more marker SNPs, associations with these rare SNPs are more difficult to detect than those with common SNPs (Fig.5). In common diseases, the existence of multiple phenocopy variants would further compromise detection (multiple rare variants) (33,34). Thus, regardless of genome coverage, power is consistently lower for less common SNPs (Fig. 6A and C). To detect rare causative SNPs, we need not only to invest in genotyping large numbers of marker SNPs with

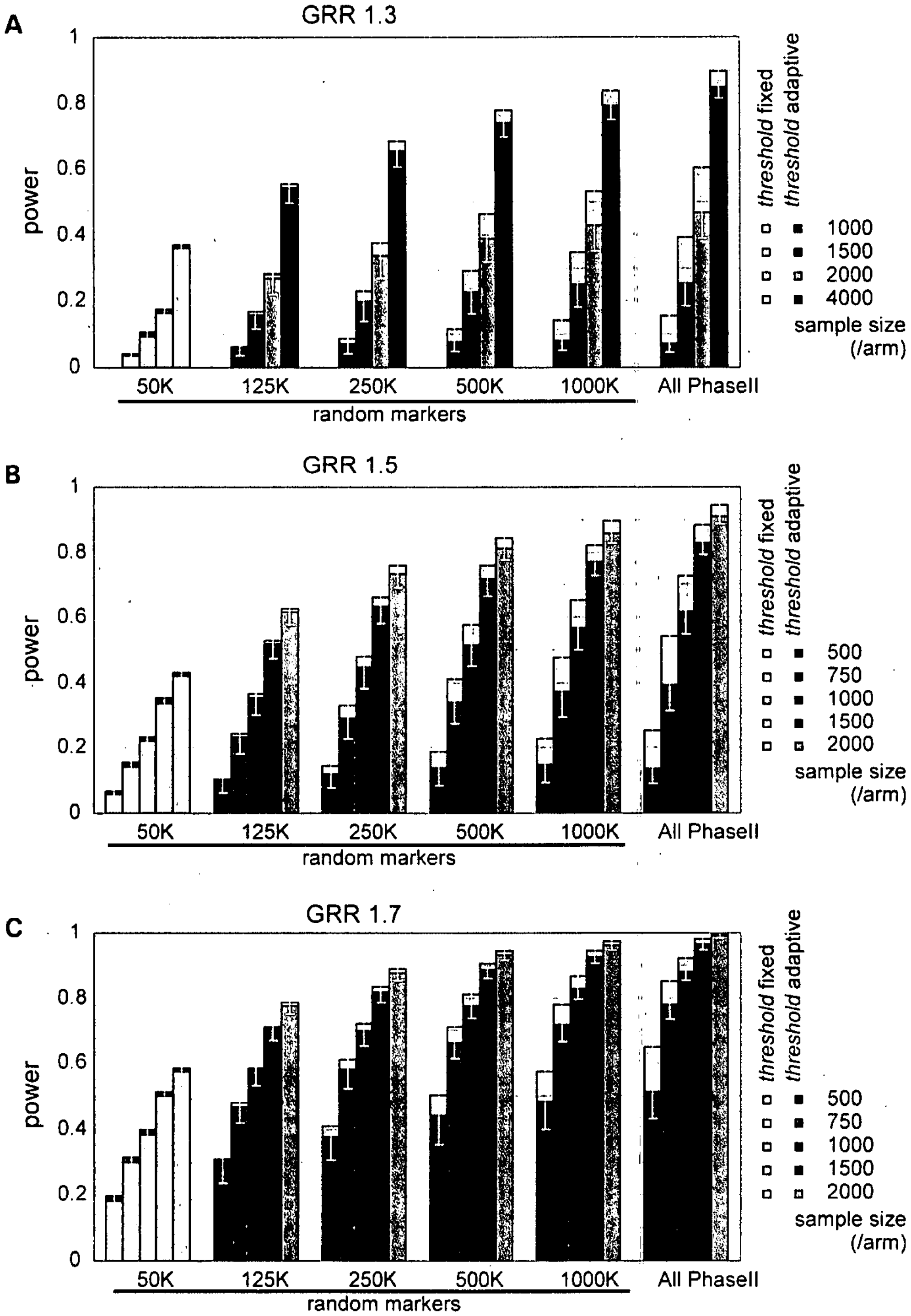

Figure 3. Genome-wide power of association studies for common causal alleles with weak to moderate genetic effects. Genome-wide power was calculated in CEU by averaging single point power for each putative causal allele over all common (MAF ≧ 0.05) SNPs in the Ref $^{\text{Phase II 5Kb}}$ reference set, 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 ($P = 1 \times 10^{-6}$; 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.

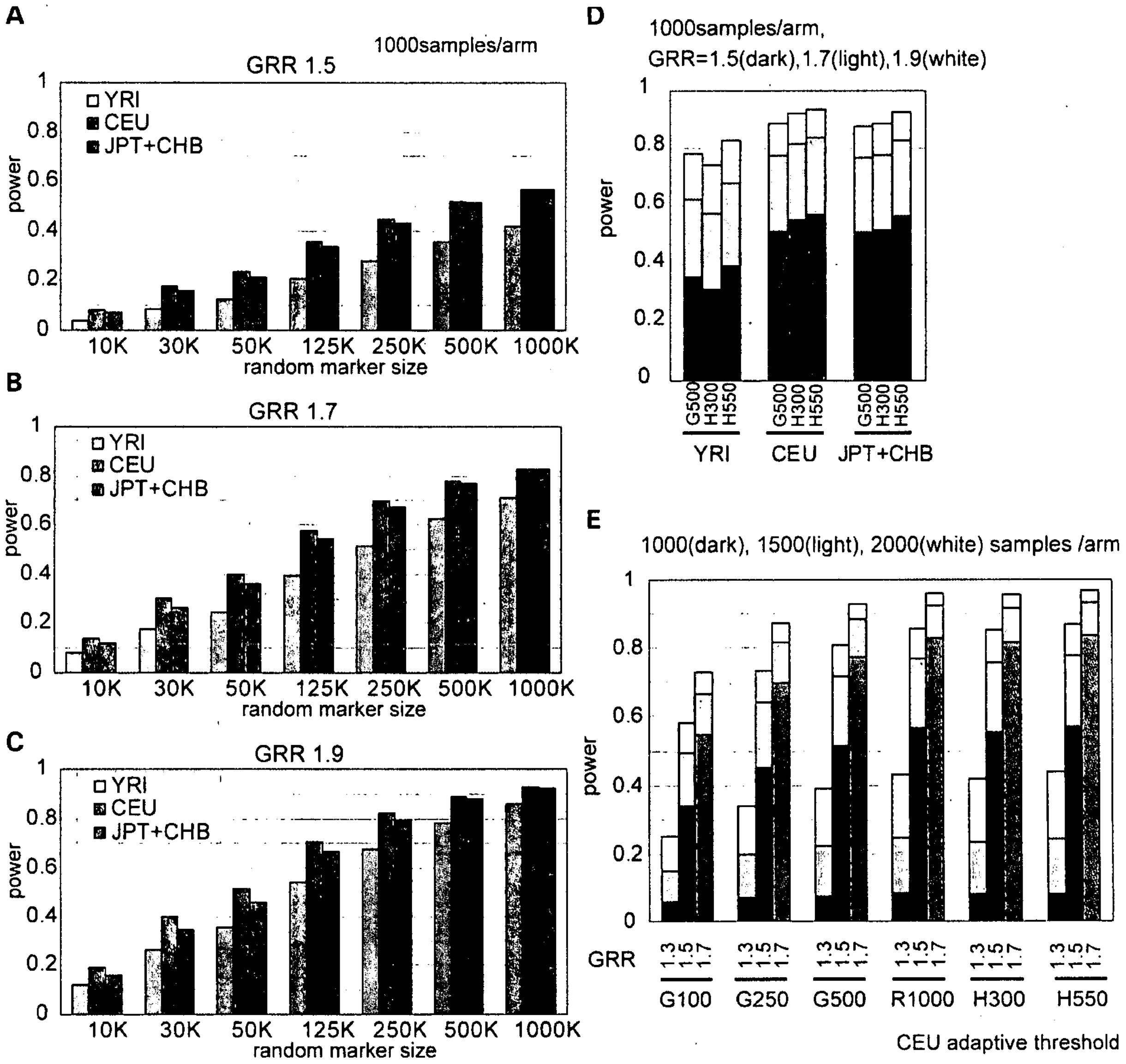

Figure 4. Comparison of power in different HapMap panels and in commercially available genotyping platforms. Genome-wide power was calculated for different HapMap panels in a variety of marker sets, including indicated numbers of randomly selected SNP markers for GRR=1.5 (**A**), GRR=1.7 (**B**), and 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,

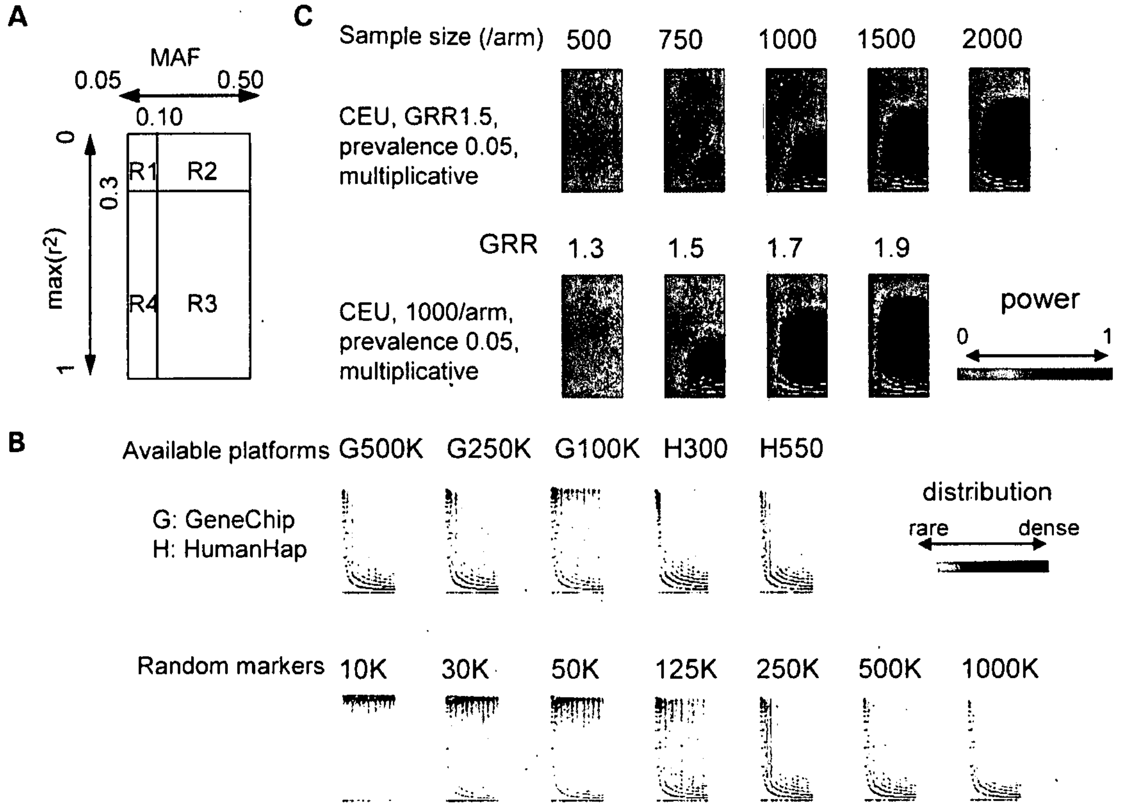

Figure 5. Impact of allele frequencies and genome coverage on genome-wide power. Reference SNPs randomly selected from the Phase II CEU set ($\mathrm{Ref}^{\mathrm{Phase\ II\ 5Kb}}$) are plotted onto a panel according to their MAF and the max(r^2) within the indicated marker set, and assigned into four categories; sub-common and weakly proxied SNPs [MAF < 0.10 and max(r^2) < 0.3] SNPs (R1), common and weakly proxied SNPs (MAF $\geqq 0.10$ and max(r^2) < 0.3) SNPs (R2), common and strongly proxied SNPs [MAF $\geqq 0.10$ and max(r^2) $\geqq 0.3$] (R3), or sub-common and strongly proxied SNPs [MAF < 0.10 and max(r^2) $\geqq 0.3$] (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 (Nc) 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 ($\leqq 1.5$), this threshold could be overcome with $\leqq 1500$ 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

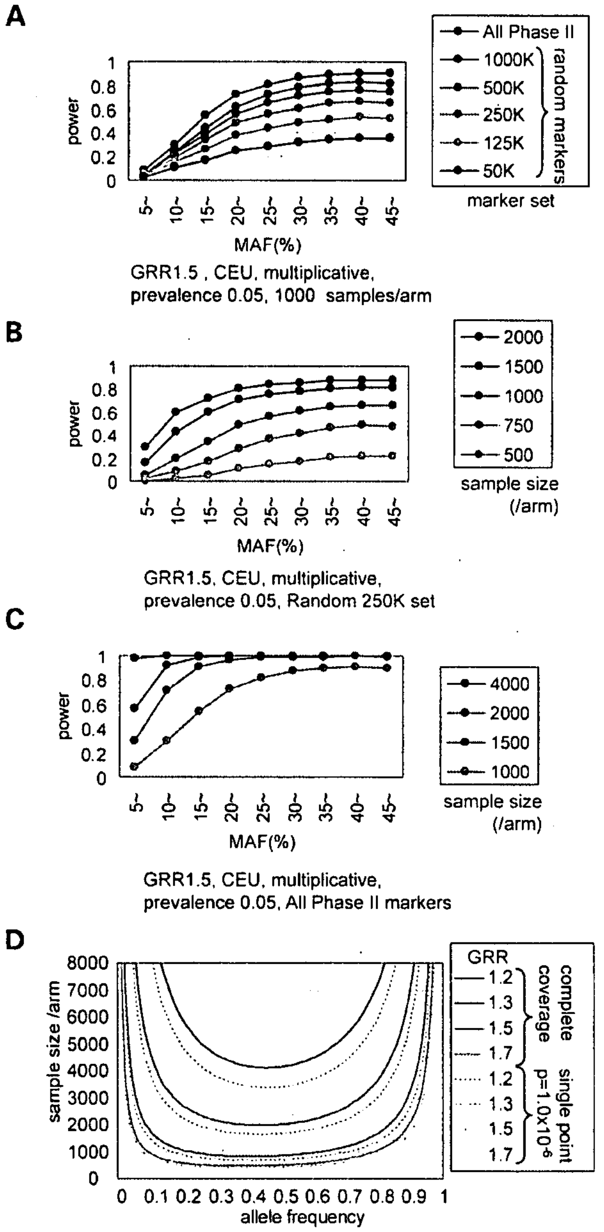

Figure 6. Effects of allele frequency on simulated power. Distribution of power on MAF in association studies are shown for varying marker sets under a constant sample size (1000 /arm) (**A**), and for varying sample sizes under a fixed marker set; GeneChip® 250K (**B**) or a hypothetical complete marker 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 (Nc^{G}=1023K, solid lines) or independent 50K markers (single point P-value =1 × 10^{-6}, Nc=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 ≥ 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 ($\max(\chi^2)$; d.f.=1) in the standard allele test was calculated for different marker sets to obtain empirical null distributions of $\max(\chi^2)$.

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:

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

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 (≥ 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 $\max(\chi^2)$ 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 $\max(\chi^2)$ 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.

Journal of Pathology
J Pathol 2007; **212**: 269–277
Published online 30 April 2007 in Wiley InterScience
(www.interscience.wiley.com) **DOI:** 10.1002/path.2180

Original Paper

Single nucleotide polymorphism array analysis of chromosomal instability patterns discriminates rectal adenomas from carcinomas

EH Lips,[1] EJ de Graaf,[5] RAEM Tollenaar,[2] R van Eijk,[1] J Oosting,[1] K Szuhai,[3] T Karsten,[6] Y Nanya,[7] S Ogawa,[7] CJ van de Velde,[2] PHC Eilers,[4] Tom van Wezel[1] and H Morreau[1]*

[1] *Department of Pathology, Leiden University Medical Centre, The Netherlands*
[2] *Department of Surgery, Leiden University Medical Centre, The Netherlands*
[3] *Department of Molecular Cell Biology, Leiden University Medical Centre, The Netherlands*
[4] *Department of Medical Statistics, Leiden University Medical Centre, The Netherlands*
[5] *IJsseland Hospital, The Netherlands*
[6] *Reinier de Graaf Gasthuis, The Netherlands*
[7] *Department of Regeneration Medicine for Haematopoiesis, Graduate School of Medicine, University of Tokyo, Japan*

**Correspondence to:*
Dr H Morreau, Department of Pathology L1-Q, Leiden University Medical Centre, PO Box 9600, 2300 RC Leiden, The Netherlands.
E-mail: J.Morreau@lumc.nl

No conflicts of interest were declared.

Received: 22 November 2006
Revised: 13 February 2007
Accepted: 22 March 2007

Abstract

Total mesorectal excision (TME) is the standard treatment for rectal cancer, while transanal endoscopic microsurgery (TEM) is a recently introduced surgical approach for the treatment of rectal adenomas. Incorrect preoperative staging before TEM is a problem. To identify genetic changes that might correlate with tumour stage and could lead to optimized treatment selection we performed a genome-wide chromosomal instability search in a homogeneous, clinical cohort of rectal tumours. 78 rectal tumours during different clinical stages were analysed with 10K single nucleotide polymorphism (SNP) arrays. Logistic regression was performed to build a quantitative model of specific chromosomal aberrations. Overall, most cases (95%) had one or more chromosomal aberrations. We observed a clear correlation between the total number of aberrations and the different tumour stages. Specifically, the chromosomal events: gain of 8q22–24, 13q and 20q, and loss of 17p and 18q12–22, were far more abundant in carcinoma than in adenoma. In adenoma fractions from cases with a carcinoma (infiltrating at least in the submucosa), twice the amount of such 'malignant aberrations' was observed, compared to pure adenomas. Furthermore, combined aberrations such as gain of 13q and loss of 18q were only found in adenomatous fractions of carcinomas and not in benign lesions. Based on these five genomic events associated with carcinoma, a clear distinction between adenoma and carcinoma tissue could be made. These data should be validated further in order that they may be used in preoperative staging of rectal tumours.

Keywords: **rectal cancer; rectal adenoma; SNP array; chromosomal instability; chromosomal aberrations; loss of heterozygosity; genomic profiling; transanal endoscopic microsurgery; total mesorectal excision**

Introduction

Total mesorectal excision (TME) and preoperative radiation is the standard treatment for rectal cancer in most North and West European countries [1]. However, this treatment results in significant functional morbidity [2]. Recently, transanal endoscopic microsurgery (TEM) has been developed. This sparing technique for the local resection of rectal adenomas results in minimal mortality and morbidity [3–5]. In spite of adequate preoperative staging, carcinomas (invasive in at least the submucosa) can be found after TEM treatment, with an increased risk of local recurrence and lymph node metastases. Reported risks of local recurrence vary considerably: between 10 and 67% (reviewed in [4,6–10]), with more recent studies reporting lower risks of between 4.3% and 11% [11–14]. Recent experimental evidence states that local treatment of T1 carcinomas by TEM may be possible [10,14]. As T1 rectal tumours have a chance of 12% of lymph node metastasis [15], preoperative knowledge on the aggressive behaviour of a tumour is important.

One of the characteristics of human cancer is chromosomal instability [16]. For rectal cancer, this is the predominant characteristic [17]. The common patterns of chromosomal instability in colorectal cancer include gains on chromosomes 13q and 20q, loss of 17p and 18q and loss of heterozygosity (LOH) of chromosomes 5q, 8p, 17p, and 18q [18–24]. The patterns of chromosomal losses and gains, and copy number neutral LOH can accurately and simultaneously be studied from the same sample with single nucleotide polymorphism (SNP) arrays [25–27]. This combined analysis offers the advantage of detecting extra genomic events, which would have been disregarded with other techniques [28,29]. Combined with standard clinicopathological variables, these chromosomal aberrations could serve as accurate biomarkers [30,31] and improve preoperative staging of rectal tumours. For colorectal cancer, associations were found between prognosis and changes of chromosomes 8 and 18q [30]; between high-grade dysplasia and gain of chromosome 7 and 20 and loss on 17p and 18q [31] and gain of chromosome 8q23 with lymph node metastases [32]. However, owing to methodological differences and heterogeneity in study populations, there is no conclusive evidence on the prognostic significance of commonly implicated regions [33]. Furthermore, left- and right-sided colon cancer clearly differ in their aetiology, clinical behaviour, pathological features and genetic abnormalities [34–38].

To mark the progression from adenoma to carcinoma in a homogeneous clinical cohort of rectal tumours, we studied genome-wide copy number changes and LOH using SNP arrays. After extensive validation, the identified patterns of aberrations might complement the current criteria for treatment selection.

Material and methods

Samples

Seventy eight snap frozen rectal adenomas and carcinomas were collected: 43 samples from patients from the IJsselland Hospital or Reinier de Graaf Hospital in The Netherlands who had undergone TEM treatment, and 35 samples from a Dutch multicentre TME trial [1]. None of these patients received radiotherapy or other adjuvant therapy. Leucocyte DNA was available for 11 cases. All samples were reviewed by a pathologist (HM), dysplasia was scored, and the tumour percentage was assessed (50–80%). Intramucosal carcinomas were considered as adenomas with high-grade dysplasia, instead of invasive carcinomas [39,40]. The local medical ethics committee approved the study.

Copy number and LOH analysis

Tumours were microdissected in a cryostat through removal of surrounding healthy tissue. Twenty 30 μm sections were cut from each tumour. To guide microdissection, a 4 μm section was cut and haematoxylin and eosin stained, before the first section, and after the tenth and twentieth section, and assessed for the presence of adenoma or carcinoma tissue, or a mixture of both. DNA was isolated with the Genomic Wizard kit, according to the manufacturer (Promega, Madison, WI, USA). Leucocyte DNA was obtained as described [41]. DNA quality was checked on a 1% agarose gel.

DNA from leucocytes and fresh frozen tumours was hybridized to GeneChip Mapping 10K 2.0 arrays (Affymetrix, Inc, Santa Clara, CA, USA) at the Leiden Genome Technology Centre (http://www.lgtc.nl, accessed 2 April 2007), as previously described [42]. Genotypes were scored with the GDAS software (Affymetrix).

For normalization we used in-house normal reference samples consisting of 11 leucocyte DNAs from TEM samples and 11 unrelated controls. Copy number and LOH data were calculated as follows:

(a) Cell files of the tumour samples were normalized with the 22 reference samples and analysed in dCHIP SNP [43]. Raw copy number data were exported and smoothed with an R-script computing a median curve and lower and upper quartile curves that indicate the spread of the data [44]. When this 50% interval fell entirely under the $n = 2$ diploid baseline, it was called a loss and when it fell above this line it was called a gain. LOH calls were obtained with the Hidden Markov Model, which was implemented in dCHIP SNP.
(b) Additionally, non-allele-specific analysis was performed in the Copy Number Analyser for Affymetrix GeneChip Mapping arrays (CNAG) [45]. The best combination of normal references was computed by CNAG.
(c) Allele-specific analysis with CNAG was performed for the 11 TEM cases with corresponding leucocyte DNA as described [45].

All data were imported in an Access database (Microsoft, Redmond, WA, USA) and the average value per chromosome sub-band was calculated. Average size of a chromosome sub-band was 2.96 Mb (range 0.002–15.02 Mb). Only aberrant sub-band regions identified by all analysis methods were considered, and three different events were discerned: loss, gain, and copy number neutral LOH.

Array comparative genomic hybridization (array-CGH)

Array-CGH was performed for four control samples, including three from a previous study [46], as previously described [47].

Illumina SNP arrays

We used Illumina SNP arrays, which are suitable for LOH detection and copy number analysis of formalin-fixed, paraffin-embedded (FFPE) tissue [46,48]. Briefly, DNA extracted from FFPE tissue samples was prepared according to the Goldengate assay in combination with linkage mapping panel version 4 and hybridized to Illumina BeadArrays. Copy number aberrations and LOH were determined using the in-house generated "beadarraySNP" R-package.

Statistics

One way ANOVA and Bonferroni *post hoc* tests were performed on the square root of the number of aberrations in order to stabilize the variance. χ^2 Tests were performed to test significance between groups for specific loss and gain events. Physical loss and copy number neutral LOH were considered as identical events in this analysis. For all analyses, p-values <0.05 were considered significant. All analyses were performed in SPSS 12.

For logistic regression analysis, a model was built in Matlab, in which each chromosome and each group is characterized by specific odds of occurrence of events. The model is $\gamma_{ct} = \alpha_c\ \beta_t$, where γ_{ct} indicates the odds for the combination of chromosome c and tumour group t, α_c the odds for chromosome c, and β_t the odds for tumour group t. This model supposes that there is no interaction between chromosomes and groups. It also assumes that, if the odds for one chromosome (over the groups) are known, the odds for other chromosomes, and similarly groups (over the chromosomes), are known as well. The use of logarithms ($a_c = log\alpha_c$ and $b_t = log\beta_t$) leads to an additive model for the log-odds, which can be fitted through logistic regression.

Table 1. Summary of clinical and pathological data of 78 tumour samples

	A/A	A/C	AC/C	C/C	C/C (N+)
Treatment					
TEM	21	8	7	5	2
TME	2	3	2	17	11
Sex (M/F)	12/11	5/6	5/4	10/12	10/3
Age (years), mean	69	70	66	65	62
Dysplasia (adenoma)					
Low	12	7			
High	11	4			
Stage (carcinoma)					
T1		8	4	10	
T2		4	4	12	12
T3			1		1
Size (cm), mean	6.2	4.8	4.2	3.1	5.3

A/A = adenomas; A/C = tumours consisting of adenoma and carcinoma tissue, from which we analysed the adenoma fraction; AC/C = tumours consisting of adenoma and carcinoma tissue, from which we analysed a mixture of adenoma and carcinoma tissue; C/C = carcinomas without lymph node metastasis; C/C (N+) = primary tumours of cases with lymph node metastasis.

Results

Sample description

Genomic DNA from 78 snap frozen rectal tumour samples from 77 different patients was hybridized to 10K SNP arrays in order to determine copy number changes and allelic imbalances. The samples analysed consisted of adenomas and carcinomas. Subsequently, five subcategories were defined on the basis of tissue fraction analysed and tumour stage (Table 1 and Supplementary Table 1, available online at http://www.interscience.wiley.com/jpages/0022-3417/suppmat/path.2180.html). The adenomas were subdivided into cases consisting of only adenoma tissue in the resection (A/A) and adenoma fractions of cases with a carcinoma focus infiltrating at least in the submucosa (A/C). The carcinomas were subdivided into three groups: tumour fractions consisting of a mixture of adenoma and carcinoma tissue (AC/C), carcinomas without (C/C), and carcinomas with lymph node metastasis (C/C (N+)). In a single case, A140, both the adenoma and carcinoma fractions were analysed separately.

SNP array data analysis and validation by array-CGH

Copy number and LOH profiles from the SNP array data were generated with two algorithms, dCHIP SNP and CNAG, and compared with array-CGH in four control cases. Figure 1 shows an example. Physical loss of chromosome 4q is detected by quantile smoothing of dCHIP data, CNAG raw and smoothed data, and array-CGH (Figures 1A–C). Allele-specific analysis shows two different LOH mechanisms: one allele of chromosome 5q is lost in case A608, while the other is retained, indicating LOH due to physical loss. In contrast, one allele was deleted from sample A609, while the other allele was amplified, resulting in copy number neutral LOH (Figure 1D). In the four control cases, all gains and physical losses identified with the SNP array analysis were concordant with array-CGH data. Additionally, the SNP arrays identified regions of copy number neutral LOH on chromosome 5q, 9, 12, and 17p.

Number of chromosomal aberrations in different groups

An overview of the extent of chromosomal instability in all different tumour samples was generated. Therefore, the number of aberrant chromosome subbands per sample was counted. All but four of 78 cases showed one or more chromosomal aberrations (95%) (Supplementary Table 1, available online at http://www.interscience.wiley.com/jpages/0022-3417/suppmat/path.2180.html, and Figure 2). Overall, an average for each sample of 65 sub-bands showing physical loss, 81 sub-bands showing gain, and 29 sub-bands showing copy number neutral LOH was detected. A significant increase (ANOVA, $p < 0.001$) in aberrations was found for the more aggressive tumours. Significant differences were found when adenoma tissue (A/A and A/C) was compared with carcinoma tissue (C/C and C/C (N+)) (68 and 100, respectively, versus 281 and 290 mean number of aberrations, $p < 0.001$).

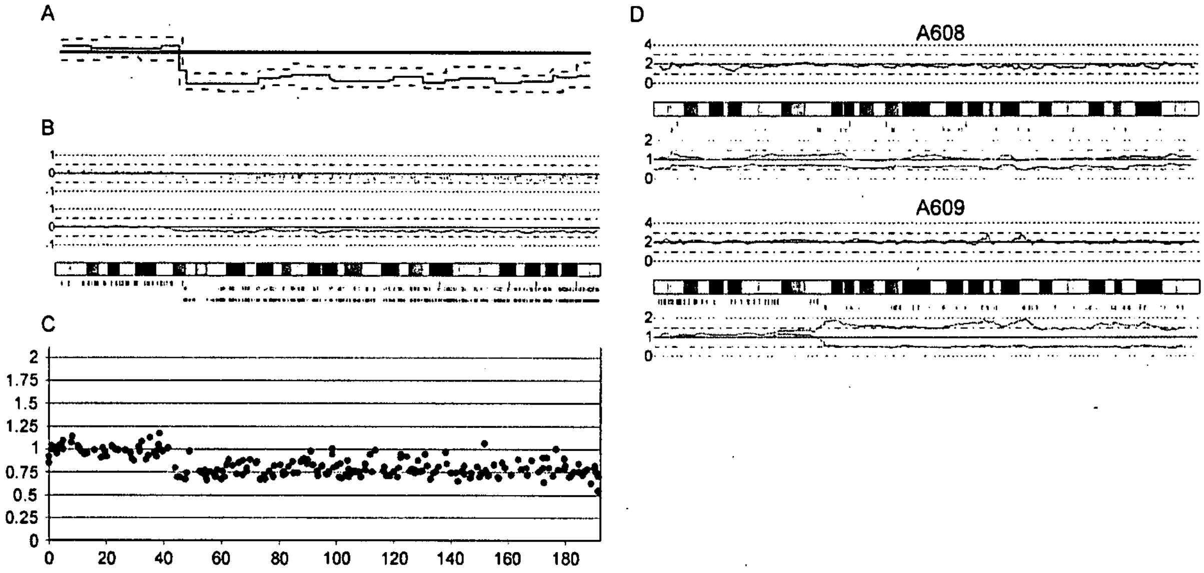

Figure 1. (A–C) Chromosome 4 for tumour A514. (A) Smoothed trend computed with dCHIP SNP data and quantile smoothing algorithm. The solid line represents the median, dashed lines represent 25% and 75% quantiles. (B) The upper panel shows raw copy number estimates by CNAG, while the lower panel shows the smoothed copy number. Stripes under the chromosome display heterozygous (green), homozygous SNPs (pink) and SNPs showing LOH (blue). (C) Array-CGH image for verification. (D) Allele-specific analysis with CNAG. Chromosome 5 is shown for tumour A608 (top) and A609 (bottom). The blue line indicates smoothed copy number, while the green and red lines indicate the two alleles. The pink stripes under the chromosome indicate LOH

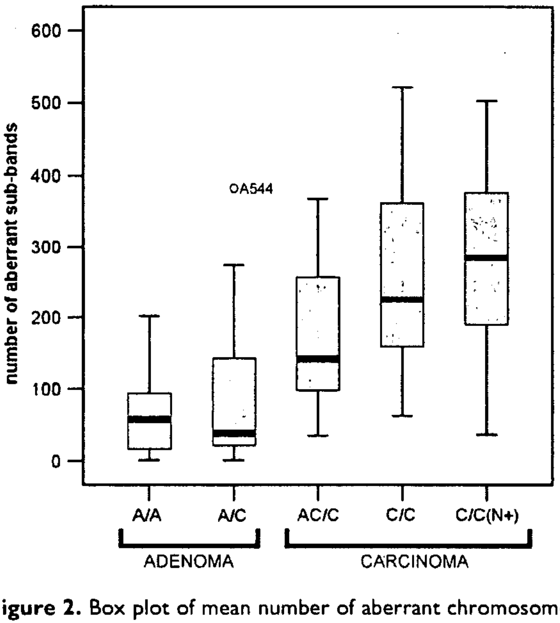

Figure 2. Box plot of mean number of aberrant chromosome sub-bands per group. Median, 25th and 75th percentiles, and range of expression levels are shown

Chromosomal aberrations in adenoma versus carcinoma tissue

For several aberrations, the minimal region of overlap was listed, although in most cases whole chromosome arms were involved. Figure 3 displays all the gains, physical losses and copy number neutral LOH events for each sample. Table 2 summarizes the most common events. Physical loss and copy number neutral LOH events are combined in this table. Loss of 1p36 (26%), 4q32-ter (26%), 5q (30%), and gain of 7p11–15 (26%), 12q13 (22%) were frequently observed aberrations in the adenomas. As the occurrence of these events had not significantly increased in carcinoma, they are indicated as early events (loss of 1p36, loss of 4q32-ter, loss of 5q, and gain of 12q13). There were no significant differences in frequency of these events between low-grade and high-grade dysplastic adenomas (data not shown). Gain of 7p11–15 was increased in carcinomas ($p = 0.005$). However, gain of 8q22–24 (50%), 13q (59%), 20q (86%) and loss of 17p (91%), 18q12–22 (86%) were the most frequent carcinoma aberrations. Percentages of these aberrations were all significantly higher in the carcinomas than in the adenomas ($p < 0.001$). Therefore, they were called carcinoma or malignant events. Gain of 1q23 was observed to be more common in carcinomas with lymph node metastases (62%) than in carcinomas without lymph node metastases (14%) ($p = 0.003$).

Frozen adenoma tissue from A/A cases was compared with frozen adenoma tissue from A/C cases in order to determine if carcinoma events were more common in the adenomas of cases with a carcinoma focus than in pure adenomas. Gain of 13q was such an event, which was only detected in one A/A case (out of 23), compared with four (out of 11) A/C cases ($p = 0.017$). A frequency table of all different combinations of chromosomal events in the different adenoma and carcinoma subgroups was created (see Supplementary Table 2, available at http://www.interscience.wiley.com/jpages/0022-3417/suppmat/path.2180.html). Gain of 13q in combination with 18q loss showed the highest frequency in the

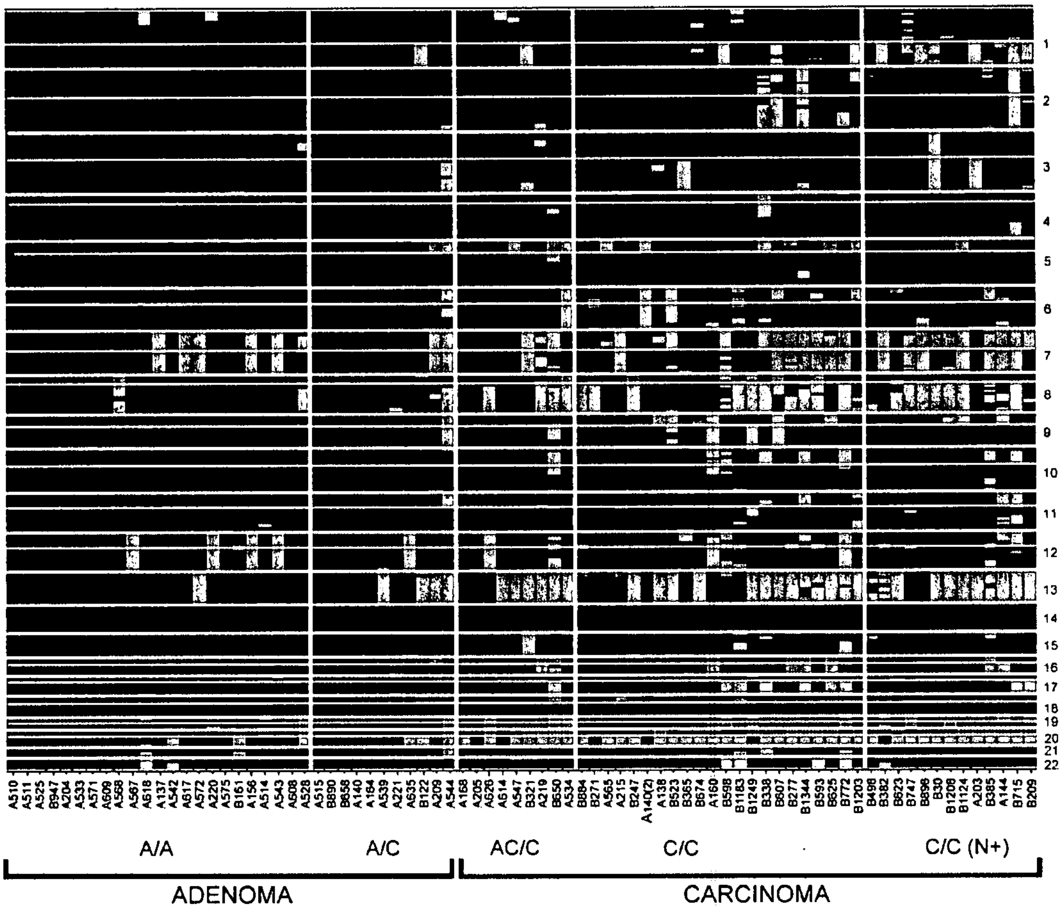

Figure 3. Overview of all gains (green), physical losses (red) and copy number neutral LOH (blue) in 78 rectal tumour samples. On the *x*-axis are all the different samples in the five different subgroups. On the *y*-axis are the chromosomes. The thick horizontal lines represent chromosomal borders; the thin lines represent centromere position in the meta-centric chromosomes

Table 2. The most common chromosomal aberrations

	Adenoma		Carcinoma			*p*-Value*		
	A/A ($n = 23$)	A/C ($n = 11$)	AC/C ($n = 9$)	C/C ($n = 22$)	C/C (*N*+) ($n = 13$)	Adenoma vs carcinoma	A/A vs A/C	C/C vs C/C (*N*+)
Early events								
Loss 1p36[†]	26	45	22	14	46	n.s.	n.s.	0.035
Loss 4q32-ter[†]	26	27	22	36	62	n.s.	n.s.	n.s.
Loss 5q[‡]	30	45	67	55	54	n.s.	n.s.	n.s.
Gain 7p11–15	26	18	33	50	77	0.005	n.s.	n.s.
Gain 12q13	22	9	22	32	23	n.s.	n.s.	n.s.
Carcinoma events								
Gain 8q22–24	9	18	44	50	62	<0.001	n.s.	n.s.
Gain 13q	4	36	67	59	85	<0.001	0.017	n.s.
Loss 17p[†]	17	18	44	91	62	<0.001	n.s.	0.038
Loss 18q12–22[†]	17	36	56	86	77	<0.001	n.s.	n.s.
Gain 20q	17	27	78	86	92	<0.001	n.s.	n.s.
Gain 13q combined with loss 18q12–22	0	27	56	59	62	<0.001	0.007	n.s.
Associated with lymph node metastasis								
Gain 1q23	0	9	11	14	62	0.002	n.s.	0.003

* *p*-Values were computed by χ^2 test.

[†] Most cases showed physical loss and LOH, while some showed copy number neutral LOH.

[‡] Most cases showed copy number neutral LOH, while some showed physical loss and LOH.

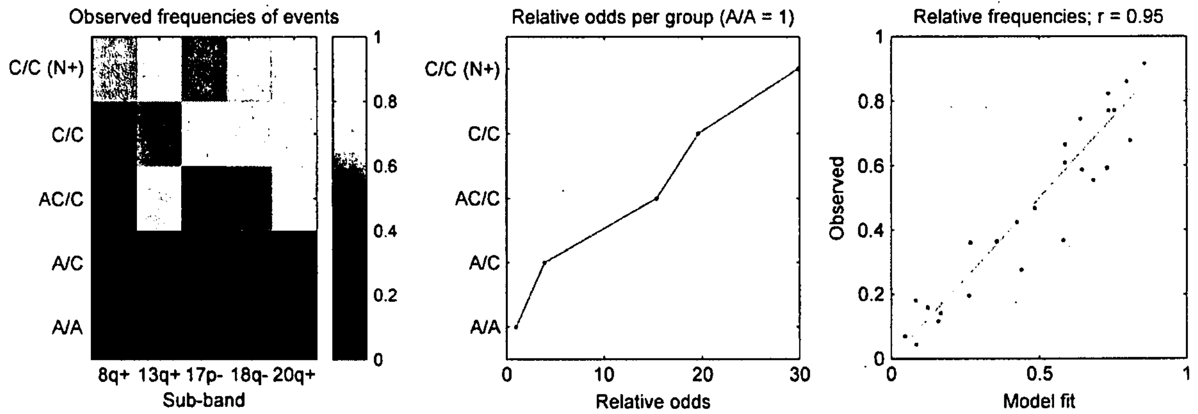

Figure 4. Logistic regression analysis. The observed frequency of the specific events is shown as a colour scale (A), the relative odds per group (B), and the model fit (C)

A/C group (27%), while it did not occur in the A/A cases ($p = 0.007$) (Table 2). All other combinations were more common in the carcinoma groups than in the adenoma groups, eg 18q loss in combination with 20q gain (4% in the A/A vs. 27% ranging to 80% in the different carcinoma groups).

Quantitative progression model

Loss of 17p and 18q12–22 and gain of 8q22–24, 13q and 20q were the most frequent events in carcinomas, in contrast to adenomas (all with a p-value <0.001 (Figure 3 and Table 2)). Therefore, these five events were used to build a quantitative progression model from adenoma to carcinoma. Figure 4A summarizes the observed frequencies of these five events in the different sample subgroups. The log-odds for each event and each sample group were calculated and fitted in a model through logistic regression analysis (Figures 4B and C). There was a strong correlation between the observed frequencies and the model fit ($r = 0.95$). The gradual increase of these events in the subgroups indicates that these five aberrations are highly correlated with the adenoma to carcinoma progression in rectal tumourigenesis.

Comparison of adenoma and carcinoma fractions of single cases

The frozen tissue contained sufficiently large adenoma and carcinoma fractions to be analysed separately in only a single case (A140). The remaining A/C cases showed large adenoma fractions with relatively small carcinoma foci that were not obtained during the procedure of snap freezing of the tumour material. For those cases, categorized as C/C and C/C (N+), macroscopy was quite different, as they contained no, or hardly any, precursor adenoma material. As we had recently demonstrated reliable LOH and copy number analysis in FFPE material using Illumina SNP Arrays [46,48], we analysed FFPE-derived tissue fractions from 16 cases (including case A140) and compared adenoma (either low- or high-grade dysplasia) and carcinoma tissue (Figure 5). The results were compared with those from the frozen tissue samples. With relatively minor differences the chromosomal aberrations in the corresponding tumour fractions were comparable. In 11 cases one or more 'malignant aberrations' (8q22–24 gain, 13q gain, 17p loss, 18q12–22 loss, 20q gain) were detected in the adenoma fractions. In sample A138 even four 'malignant aberrations' were detected in the adenoma fraction with low-grade dysplasia. The progression of adenoma to carcinoma was in 11 of 16 cases characterized by one to three extra 'malignant aberrations' (three cases 8q gain, six cases 13q gain, four cases 17p loss, six cases 18q loss, and four cases 20q gain). Five cases showed a comparable amount of aberrations in their adenoma and carcinoma fractions, respectively.

Discussion

After TEM treatment, 10–30% of presumed rectal adenomas prove to be carcinomas, infiltrating at least into the submucosa (tumour stage T1 and further). Since the introduction of preoperative endorectal ultrasound this figure has declined to 10% [10], but there is still a need for better preoperative parameters to discriminate rectal adenomas from carcinomas.

We used SNP array analysis and identified that gain of 8q22–24, 13q, 20q and loss of 17p and 18q12–22 are more prevalent in rectal carcinomas than in adenomas. Several other studies in colorectal cancer have also identified these regions [18,20–24,31]. Hermsen *et al* showed that two or more of seven specific chromosomal regions (loss on 8p, 15q, 17p, 18q, and gain on 8q, 13q, and 20q) were associated with colorectal cancer progression [20]. Leslie *et al* identified that chromosomal loss on 17p, and 18q and gain of 20 were related to the onset of high-grade dysplasia [31]. Diep *et al* have conducted a meta-analysis from 31 CGH studies and found that losses at 17p and 18 and gains of 8q, 13q, and 20 occur early in primary colorectal cancers [18]. Hence, the regions found in our