European Journal of Human Genetics (2008) 16, 635–643

www.nature.com/ejhg

ARTICLE

Evaluation of coverage variation of SNP chips for genome-wide association studies

Mingyao Li*[,1,5], Chun Li[2,3,5] and Weihua Guan[4]

[1]*Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA, USA;* [2]*Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN, USA;* [3]*Center for Human Genetics Research, Vanderbilt University School of Medicine, Nashville, TN, USA;* [4]*Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA*

Genome-wide association (GWA) studies for complex human diseases are now feasible. Many GWA studies rely on commercial SNP chips, for which a common evaluation criterion is global coverage of the genome. Although providing an overall evaluation of an SNP chip, the global coverage does not tell us how the coverage varies across the genome, an important feature that should be taken into consideration, as coverage variation often results in power variation and potentially biased search in subsequent association analysis. To achieve a fuller understanding of SNP chip coverage, we conducted detailed evaluation of coverage, including (1) a map of local coverage – calculated over small consecutive genomic regions and (2) gene coverage – calculated for each known gene in the genome. These evaluations can reveal the degree of variation of each SNP chip in covering the genome and can facilitate SNP chip comparisons at a finer scale.

European Journal of Human Genetics (2008) **16,** 635–643; doi:10.1038/sj.ejhg.5202007; published online 6 February 2008

Keywords: genome-wide association; coverage; HapMap; linkage disequilibrium

Introduction

Genome-wide association (GWA) studies for complex human diseases have now become increasingly popular due to rapid decrease of genotyping costs and recent completion of the International HapMap Project.[1–4] With interrogation of hundreds of thousands of SNPs in a large collection of human subjects, GWA studies allow a comprehensive scan of the genome and have the potential to identify novel disease-related genes. The advent of GWA studies has led to the discovery of susceptibility genes for age-related macular degeneration,[5] cardiac repolarization,[6] obesity,[7] inflammatory bowel disease,[8] and type II diabetes.[9]

However, many issues in designing and analyzing GWA studies remain unclear. For example, when designing a GWA study, an investigator has to choose among several SNP chips. Ideally, one would wish to choose the SNP chip that provides the best genomic coverage for the studied population. However, given the increased cost of using a denser chip, one would also be interested in knowing how much power gain a denser chip has over a less dense chip. The decision is largely dependent on comparison of different SNP chips, thus making systematic and thorough evaluation inevitably important.

The most commonly used criterion for SNP chip evaluation is global coverage, defined as the fraction of common SNPs that are tagged by the SNPs on the chip.[10,11] The global coverage is clearly the most relevant criterion, as it represents the average level of coverage of all common SNPs. However, the HapMap data showed in great detail

*Correspondence: Dr M Li, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, 624 Blockley Hall, Philadelphia, PA 19104, USA.
Tel: +1 215 746 3916; Fax: +1 215 573 4865;
E-mail: mingyao@mail.med.upenn.edu
[5]These authors contributed equally to this work.
Received 19 April 2007; revised 30 November 2007; accepted 20 December 2007; published online 6 February 2008

the extent of local variation in linkage disequilibrium (LD) across the genome. Since coverage is calculated based on LD, one would expect variation in coverage as well. Although the global coverage provides an overall evaluation of an SNP chip, it does not tell us how the coverage varies across the genome, an important feature that should be taken into consideration because coverage variation often results in power variation in subsequent association analysis.

To achieve a fuller understanding of the coverage of SNP chips, we propose carrying out more detailed coverage evaluations, including a map of local coverage over small consecutive genomic regions, and gene coverage that is calculated for each known gene in the genome. These evaluations reveal the degree of variation of each SNP chip in covering the genome and can facilitate SNP chip comparisons at a finer scale. We evaluate both the local coverage and gene coverage for six currently available SNP chips, including Affymetrix SNP Array 5.0 and SNP Array 6.0, and Illumina HumanHap300, HumanHap550, HumanHap650Y, and Human1M. Since the power for regions or genes of low coverage is likely to be lower than that for regions or genes of high coverage, information on local coverage and gene coverage can help determine if supplementary genotyping is necessary for the success of a GWA study.

Methods

Data sets

We considered six most commonly used SNP chips in GWA studies: Affymetrix SNP Array 5.0 (500 568 SNPs) and SNP Array 6.0 (934 968 SNPs), and Illumina HumanHap300 (317 511 SNPs), HumanHap550 (555 352 SNPs), HumanHap650Y (660 917 SNPs), and Human1M (1 072 820 SNPs). The Illumina SNP chips include tag SNPs derived from over two million common SNPs (minor allele frequency MAF ≥ 0.05) in the HapMap data. The Affymetrix SNP Array 5.0 includes SNPs selected on the basis of sequence constraints when choosing the probes, and thus represents a set of quasi-random SNPs that ignores LD patterns.[10] The additional SNPs in the SNP Array 6.0 are mostly tag SNPs. Allele frequency and LD data for the four HapMap populations (CEU, CHB, JPT, and YRI) were obtained from HapMap release no. 21.

Local coverage

We estimated the coverage of the six SNP chips for chromosomal regions of sizes 1 Mb throughout the genome. We adapted the formula of Barrett and Cardon[10] to estimate local coverage rate for each of the four HapMap populations. Briefly, for each 1 Mb region, we obtained R – the number of common SNPs in the HapMap, T – the number of common SNPs on the SNP chip, and L – the number of common SNPs not on the SNP chip but are tagged at $r^2 \geq 0.8$ by at least one SNP in the chip within 250 kb. Let G denote the total number of common SNPs in the region under consideration, including those that have already been discovered and those that have yet to be discovered. Following Barrett and Cardon,[10] the local coverage rate is estimated by

$$[L/(R-T)\times(G-T)+T]/G. \quad (1)$$

Here $L/(R-T)$ computes the fraction of HapMap common SNPs tagged by SNPs on the chip but are not tags themselves. Multiplying this fraction by $G-T$ yields the number of common SNPs in the region that are not on the chip but can be tagged by SNPs on the chip. This number is then added by T to give an estimate of the total number of SNPs that are captured by either LD tagging or by inclusion on the chip. Compared to a naïve estimate of coverage, $(L+T)/R$, this formula corrects for overestimation of coverage.[10]

The value of G is unknown and needs to be estimated. For a 1 Mb region, the average number of common SNPs is estimated to be about 2631 based on the estimated numbers of common SNPs (7.5×10^6) and euchromatic base pairs (2.85×10^9) in the human genome.[10,11] We recognize that different estimates of G may lead to different values of local coverage rate. However, the above formula can be rewritten as $L/(R-T)+[1-L/(R-T)]\times T/G$, which indicates that the value of G has little effect on the final estimate as long as the fraction of common SNPs included in the SNP chip, T/G, is small, which is true for the six SNP chips we evaluated.

To calculate local coverage rate across the genome, we moved the 1 Mb window by 200 kb and repeated the calculation until the end of the chromosome. We did not calculate the values for a window if (1) the number of common SNPs in the HapMap is <20, (2) all common SNPs are located at the left or right half of the window, or (3) the common SNPs are clustered at the ends of the window with a big gap (≥ 500 kb) in between. As a result, coverage was not calculated for about 7% of the genome, most of which are in heterochromatic regions and have effectively no coverage from the current SNP chips.

Gene coverage

The local coverage calculation procedure can also be applied to calculate the coverage for each gene in the genome. To obtain the starting and ending positions of genes, we downloaded the known Gene table (contains positions of transcripts for known protein coding genes) and the kgXref table (contains cross reference between transcript IDs and gene symbols) from the UCSC human genome release hg17. A gene region is defined as the region from the transcriptional start to end positions, including both exons and introns. For a gene that has more than one transcript, the gene region is defined as the union of regions for all the transcripts. By merging the known Gene

and the kgXref tables and eliminating genes that map onto different chromosomes, we obtained 29 815 autosomal and X-linked gene regions. Gene regions vary greatly in size, and those containing very few HapMap common SNPs may have unreliable or inflated coverage results because the design of most current SNP chips relied on the HapMap data. Because of this, we considered gene regions containing only five or more HapMap common SNPs, resulting in 19 913 gene regions for the CEU sample in final analysis (19 299 for CHB, 19 211 for JPT, and 20 694 for YRI, respectively).

Coverage calculation for SNP Array 6.0 and Human1M

The local coverage and gene coverage were calculated based on the HapMap data. However, each of the latest two chips, SNP Array 6.0 and Human1M, has about 10% of the SNPs that are not in the HapMap. According to Affymetrix, the SNP Array 6.0 has 934 968 SNPs, but with 99 854 SNPs (10.7%) not in the HapMap, including 72 379 common SNPs for CEU, 76 016 for CHB, 70 356 for JPT, and 83 412 for YRI. According to Illumina, the Human1M has 1 072 820 SNPs, but with 125 688 SNPs (11.7%) not in the HapMap, including 70 995 common SNPs for CEU, 67 453 for CHB/JPT, and 77 729 for YRI. Because of this, their local coverage and gene coverage may be underestimated if only the HapMap SNPs were considered in coverage calculation. To address this problem, we calculated an alternative coverage estimate as follows, using the SNP Array 6.0 as an example. Suppose there is an 'updated HapMap data set' that consists of the current HapMap SNPs and the SNPs on the SNP Array 6.0. Based on this 'updated data', for each region, we could estimate the number of common SNPs, denoted as R_1, and the number of common SNPs on the chip, denoted as T_1. For example, if the region contains m non-HapMap common SNPs on the SNP Array 6.0, then $R_1 = R + m$ and $T_1 = T + m$. However, owing to the lack of LD information between the 'new' SNPs and the other HapMap SNPs, we do not know how many additional HapMap SNPs are tagged by these 'new' SNPs, therefore, L_1 cannot be directly estimated. However, if we assume that the number of tagged common SNPs that are not on the chip increases proportionally with the number of common SNPs on the chip, that is, $T_1/T = L_1/L$, then L_1 can be estimated as $(T_1/T) \times L$. Therefore, based on the 'updated HapMap data', we could calculate the local/gene coverage of the SNP Array 6.0 as

$$[L_1/(R_1 - T_1) \times (G - T_1) + T_1]/G \qquad (2)$$

The original estimate of genomic coverage in (1) ignored the SNPs that were on the SNP Array 6.0 but were not on the HapMap, and thus it can be viewed as a 'lower bound' of the coverage. On the other hand, the coverage in (2) might overestimate when $T_1 > T$ and T is small. In our analysis, we took the average of the coverage calculated using (1) and (2), which we believe may provide a more appropriate estimate for the coverage of the SNP Array 6.0. The coverage estimate for the Human1M was similarly calculated.

Results

A map of local coverage

We estimated the local coverage rate for Affymetrix SNP Array 5.0 and SNP Array 6.0, Illumina HumanHap300, HumanHap550, HumanHap650Y, and Human1M. As an example, Figure 1 displays the local coverage rate for chromosome 17 for the four HapMap populations. Detailed, high-resolution results for all chromosomes can be downloaded from http://biostat.mc.vanderbilt.edu/SNPChipCoverage. Not surprisingly, the Human1M has universally better coverage than the other five chips for all four populations. For the CEU sample, the coverage of the HumanHap550 is almost always better than the SNP Array 6.0, despite the fact that the latter chip has a significantly more number of SNPs; moreover, the HumanHap300 is almost always better than the SNP Array 5.0. As expected, the coverage of the HumanHap650Y is significantly improved for the YRI sample over the HumanHap550. For comparison's purpose, the global coverage of the six SNP chips is summarized in Table 1.

Figure 2 shows a wide range of local coverage across the genome, with some regions receiving low to moderate coverage. For Human1M, the percentage of the euchromatic genome that has $\geq 80\%$ local coverage rate is 98% for the CEU sample and 97% for the CHB + JPT samples. For HumanHap650Y, the corresponding percentages are 90 and 77%, respectively; for HumanHap550, the percentages are 88 and 73%; for HumanHap300, the percentages are only 41 and 11%. For Affymetrix chips, the percentages are 69 and 74% for SNP Array 6.0, and only 9 and 12% for SNP Array 5.0. All six SNP chips have low coverage rate for the YRI sample. Figure 2 indicates that evaluation of local coverage provides complementary information of an SNP chip in addition to global coverage.

We next evaluated the variation of coverage across chromosomes by calculating the average local coverage rates for all 1 Mb intervals on each chromosome. The coverage of different chromosomes is largely similar, except for chromosome 19, which appears to have lower coverage by all six SNP chips across all HapMap populations (Figure 3). For example, for the CEU sample and SNP Array 6.0, the coverage for chromosome 19 is 67%, whereas the coverage for the other chromosomes ranges from 75 to 86%. The lower coverage for chromosome 19 is presumably due to SNP ascertainment bias in the HapMap[12] or the unusually high density of repeat sequences and high prevalence of large segmental duplications on this chromosome.[13]

Gene coverage

Figure 4 displays the number of gene regions with coverage exceeding certain thresholds for all six SNP chips. For the

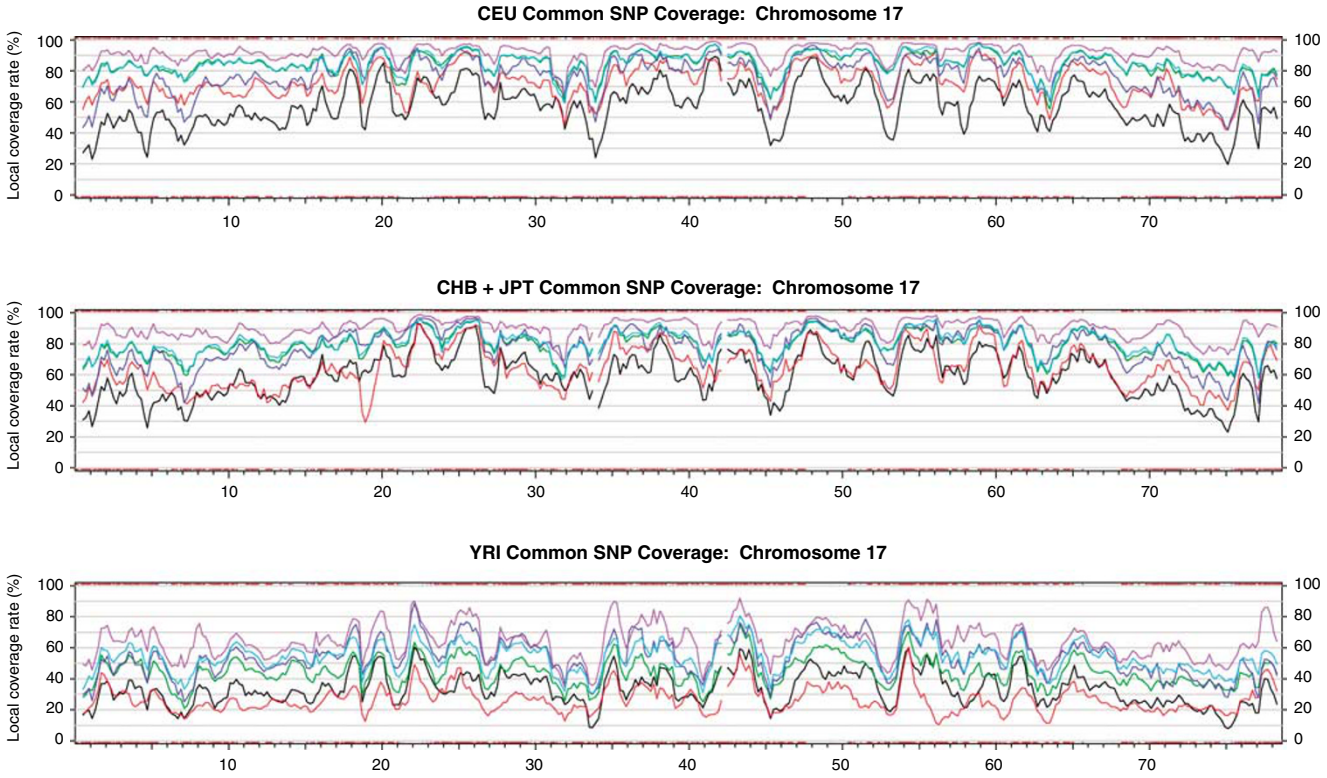

Figure 1 Local coverage map for each HapMap population for chromosome 17. The six SNP chips that were evaluated are SNP Array 5.0 (black), SNP Array 6.0 (blue), HumanHap300 (red), HumanHap550 (green), HumanHap650Y (cyan), and Human1M (purple). The red bars at the top and bottom indicate the transcription regions of known protein coding genes.

Table 1 Global coverage (%) by SNP chips

SNP chip	*CEU*	*CHB+JPT*	*YRI*
SNP Array 5.0	64	66	41
SNP Array 6.0	83	84	62
HumanHap300	77	66	29
HumanHap550	87	83	50
HumanHap650Y	87	84	60
Human1M	93	92	68

CEU sample, among the 19 913 genes with at least five common SNPs in the HapMap, 17 730 (89.1%) genes have $\geq$80% coverage by the Human1M, while the numbers are 16 210 (81.4%), 15 873 (79.7%), 11 207 (56.3%), 12 613 (63.3%), and 6820 (34.2%), respectively, for the HumanHap650Y, HumanHap 550, HumanHap300, SNP Array 6.0, and SNP Array 5.0. The numbers are slightly smaller for the CHB+JPT samples, but drop substantially for the YRI sample. We also note that there is a noticeable fraction of genes that are not well covered by all six SNP chips (Figure 5). For example, for the CEU sample, 1897 (9.5%) genes have coverage of <80% by all six SNP chips. The numbers of such genes are even greater for the CHB (2457, 12.7%), JPT (2295, 11.9%), and the YRI (10 722, 51.8%) samples. Moreover, for each SNP chip, there are some genes that have zero coverage at $r^2 = 0.8$, even though they contain five or more HapMap common SNPs (Table 2).

Similar to the analysis of local coverage, we also calculated the average coverage for genes on each chromosome (Figure 6). Again, we observed that the average coverage for genes on chromosome 19 is significantly lower than that for genes on other chromosomes. For example, for the CEU sample and SNP Array 6.0, the average coverage for genes on chromosome 19 is 61%, whereas the average coverage for genes on other chromosomes ranges from 73 to 85%. Since chromosome 19 has the highest density of genes among all human chromosomes, more than double the genome-wide average,[13] it is inevitably important to improve its coverage.

Table 3 lists genes that have <30% coverage for the CEU sample by all six SNP chips and that are known to be associated with pathways in the KEGG and BioCarta databases (lists for other samples can be obtained from http://biostat.mc.vanderbilt.edu/SNPChipCoverage). This list includes several genes that have been previously identified to be associated with human diseases. For example, Long *et al.*[14] noted that increased expression

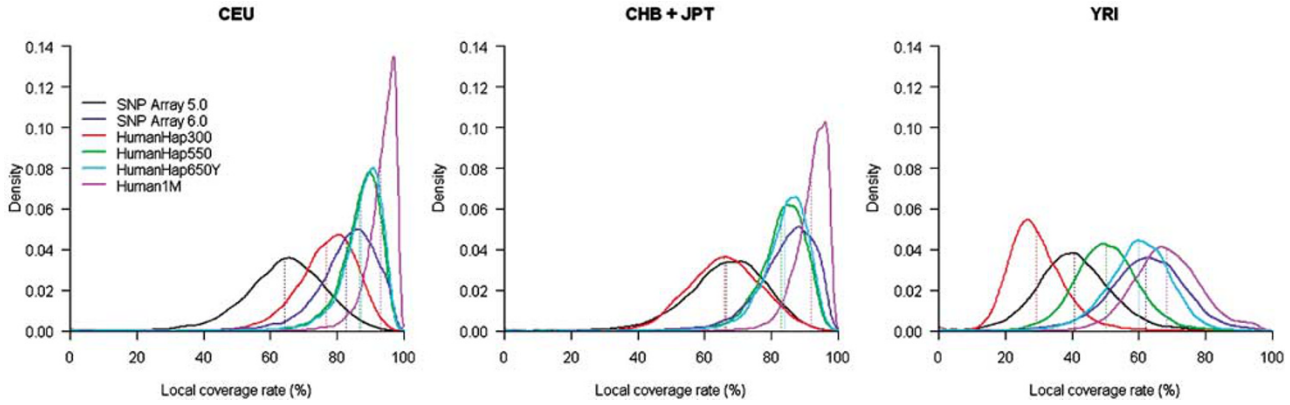

Figure 2 Distribution of local coverage. The vertical line is the global coverage rate.

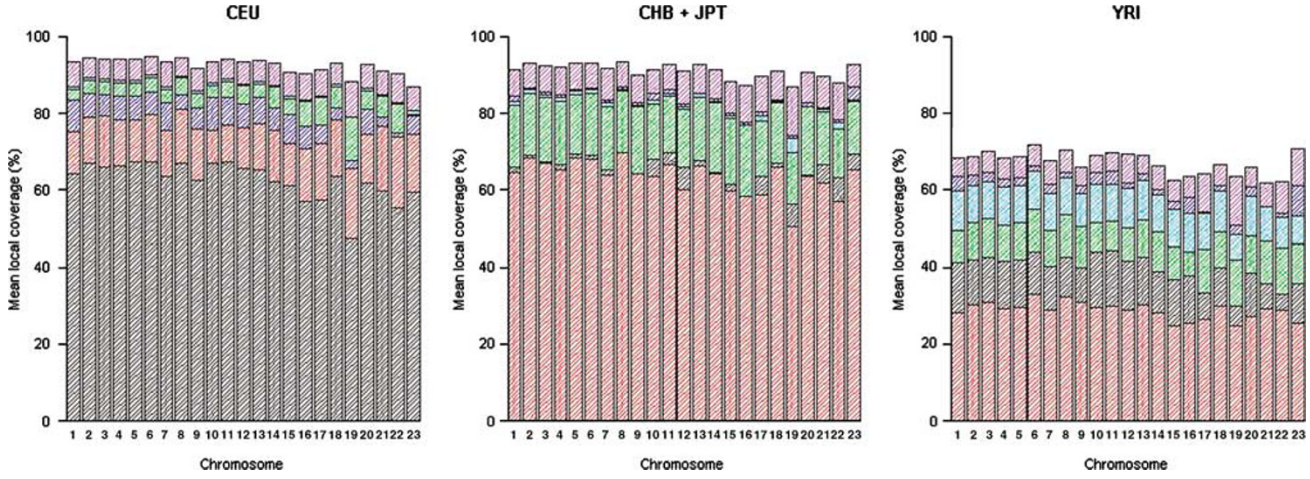

Figure 3 Mean local coverage by chromosome. The six SNP chips that were evaluated are SNP Array 5.0 (black), SNP Array 6.0 (blue), HumanHap300 (red), HumanHap550 (green), HumanHap650Y (cyan), and Human1M (purple).

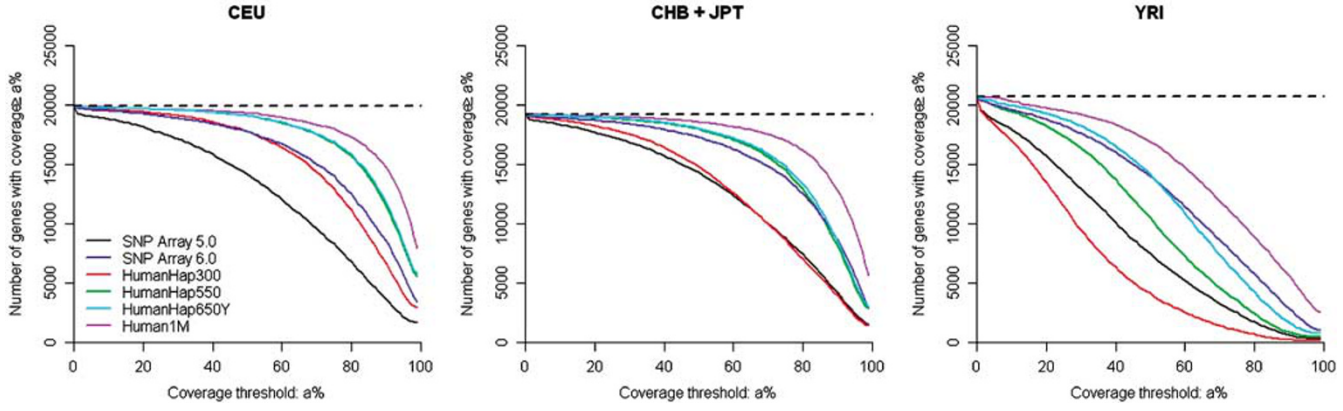

Figure 4 Number of genes covered at various coverage thresholds. Only gene regions containing with ≥5 HapMap common SNPs were considered, and coverage was evaluated at $r^2 \geq 0.8$.

and a polymorphism of *TGFB1* are associated with abdominal obesity and body mass index in humans. *TGFB1* has also been reported to play a role in many other diseases, including Duchenne muscular dystrophy,[15] kidney disease,[16] cancer,[17] scleroderma,[18] lung disease,[19] and herpes simplex virus-1 infection.[20] We recognize that

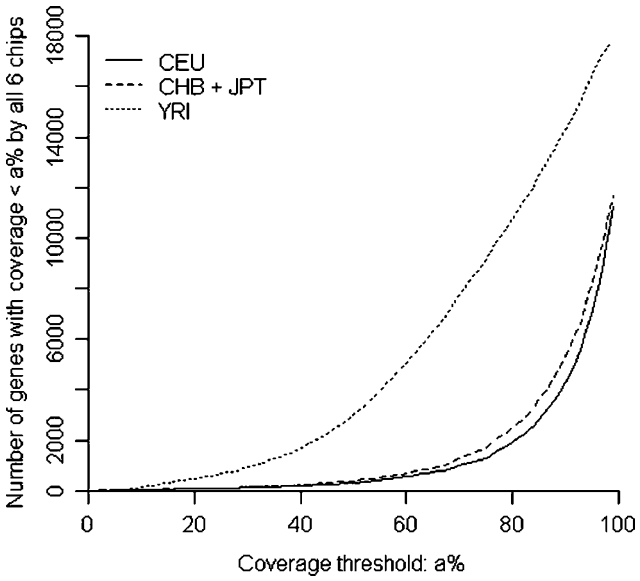

Figure 5 Number of genes with coverage less than a certain threshold by all six SNP chips. Only gene regions containing with ≥ 5 HapMap common SNPs were considered, and coverage was evaluated at $r^2 \geq 0.8$.

Table 2 Number of genes with 0% coverage by SNP chips

SNP chip	*CEU*	*CHB*	*JPT*	*YRI*
SNP Array 5.0	575	540	496	980
SNP Array 6.0	163	152	151	265
HumanHap300	106	209	236	1064
HumanHap550	46	50	56	225
HumanHap650Y	43	46	52	114
Human1M	8	8	9	16

Note: only gene regions containing with 5 HapMap common SNPs were considered, and coverage was evaluated at $r^2 0.8$.

these findings need to be replicated by future studies. However, despite the potential important role of *TGFB1* in many diseases, all six SNP chips we evaluated have poor coverage for this gene. If an investigator is mainly interested in studying these diseases, then it is likely that *TGFB1* will be missed in the initial scan. Understanding the coverage of known genes of different SNP chips will help investigators determine whether supplementary genotyping is needed for certain genes of high interest.

We next evaluated whether genes with poor coverage are more likely to be located in copy number variation (CNV) regions.[21,22] We obtained the CNV annotation file from Affymetrix, which assembled information of all known CNV regions. For a given coverage threshold, the genes were categorized into two groups, one with coverage higher than the threshold and the other lower than the threshold. Within each group, we calculated the fraction of genes that are located in known CNV regions. Not surprisingly, a higher fraction of low coverage genes fall into known CNV regions than high coverage genes, and the difference is greater for smaller coverage threshold values (Figure 7). This indicates that genes with poorer coverage are more likely to be located in known CNV regions. We also note that for the CEU sample, the fraction of low coverage genes in known CNV regions is slightly higher for the Illumina chips than the Affymetrix chips. This is presumably due to the fact that Illumina designed their products based on tag SNPs derived from the HapMap CEU sample, whereas Affymetrix designed their chips on the basis of sequence constraints when choosing the probes, which may result in a better coverage for CNV regions.

Another possible reason of poor coverage is due to weak LD, as such regions would require inclusion of the majority of SNPs in the region in order to achieve satisfactory coverage. For genes that are not located in known CNV

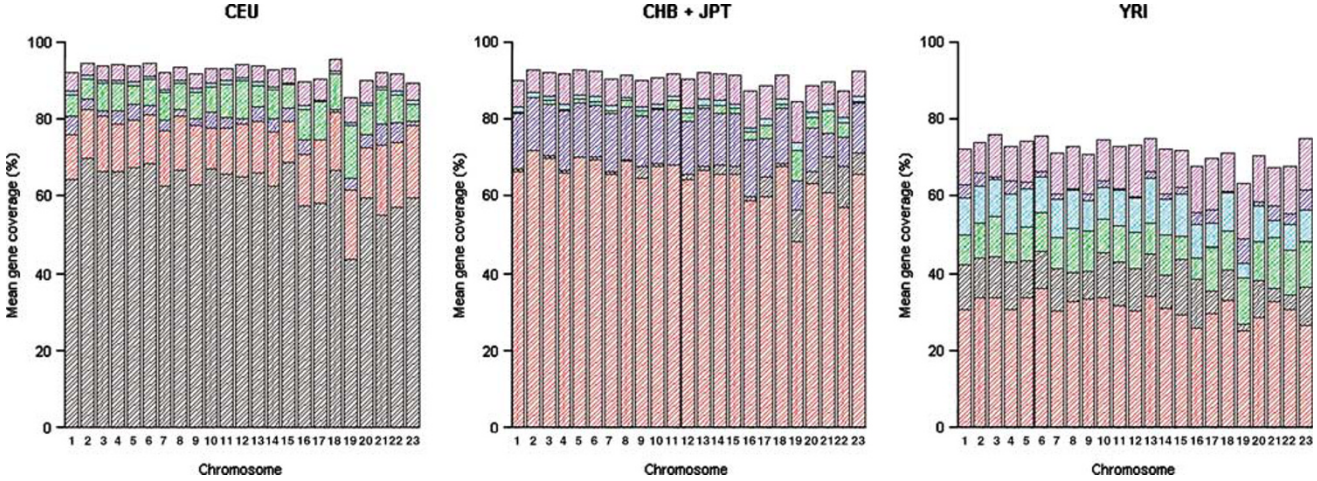

Figure 6 Mean gene coverage by chromosome. The six SNP chips that were evaluated are SNP Array 5.0 (black), SNP Array 6.0 (blue), HumanHap300 (red), HumanHap550 (green), HumanHap650Y (cyan), and Human1M (purple). Only gene regions containing with ≥ 5 HapMap common SNPs were considered, and coverage was evaluated at $r^2 \geq 0.8$.

Table 3 Genes with coverage less than 30% by all six SNP chips for the CEU sample

			Coverage (%)					
Gene	*Chromosome*	*Pathway*	*SNP Array 5.0*	*SNP Array 6.0*	*Human Hap300*	*Human Hap550*	*Human Hap650Y*	*Human1M*
APOBEC3C	22	Atrazine degradation	0.0	0.0	8.3	8.3	8.3	8.3
ADRA1D	20	Calcium-signaling pathway	14.7	21.6	23.4	23.4	23.4	23.4
MUC2	11	Cholera infection	0.0	11.2	10.3	10.3	10.3	10.3
TGFB1	19	Chronic myeloid leukemia	20.0	20.0	6.6	6.6	6.6	9.0
TNFRSF7	12	Cytokine–cytokine receptor interaction	29.2	29.2	29.2	29.2	29.2	29.2
GBGT1	9	Glycan structures-biosynthesis 2	0.0	0.0	13.8	13.8	13.8	13.8
PLA2G2F	1	GnRH-signaling pathway	3.3	3.3	10.0	10.0	10.0	15.0
TRIP10	19	Insulin-signaling pathway	3.1	3.1	9.4	9.4	9.4	9.4
ICAM1	19	Leukocyte transendothelial migration	14.3	14.3	9.8	12.2	12.2	12.2
MAPK8IP2	22	MAPK-signaling pathway	0.0	0.0	13.8	13.8	13.8	17.2
GSTM4	1	Metabolism of xenobiotics by cytochrome P450	0.0	4.0	4.0	8.0	8.0	8.0
UGT2B15	4	Metabolism of xenobiotics by cytochrome P450	0.0	0.0	0.0	0.0	0.0	0.9
FCGR3A	1	Natural killer cell-mediated cytotoxicity	4.6	28.3	25.0	28.0	28.0	28.0
KIR3DL1	19	Natural killer cell-mediated cytotoxicity	26.3	26.3	0.9	0.9	0.9	3.0
NCR1	19	Natural killer cell-mediated cytotoxicity	0.0	0.0	27.8	27.8	27.8	27.8
ADRBK2	22	Olfactory transduction	0.2	0.5	12.9	14.9	14.9	14.9
TRPM5	11	Taste transduction	20.1	20.1	6.1	8.2	8.2	14.3
NTRK1	1	Thyroid cancer	1.8	1.8	1.8	1.8	1.8	5.4
PDZK1	1	mta3 Pathway	0.0	20.8	26.6	26.6	26.6	26.6
TERT	5	Tel Pathway	0.9	2.7	4.5	4.5	4.5	5.4

Note: only gene regions containing with ≥ 5 HapMap common SNPs were considered, and coverage was evaluated at $r^2 \geq 0.8$.

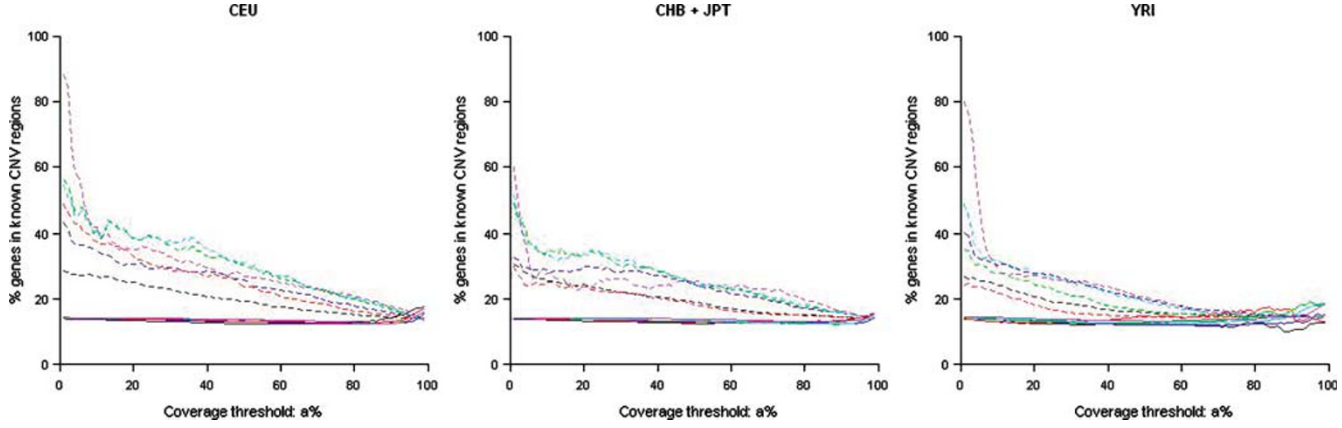

Figure 7 Percentage of genes in known CNV regions at various coverage thresholds. The six SNP chips that were evaluated are SNP Array 5.0 (black), SNP Array 6.0 (blue), HumanHap300 (red), HumanHap550 (green), HumanHap650Y (cyan), and Human1M (purple). Solid lines are for genes with coverage greater than the coverage threshold, and dashed lines are for genes with coverage less than the coverage threshold. Only gene regions containing with ≥ 5 HapMap common SNPs were considered, and coverage was evaluated at $r^2 \geq 0.8$.

regions, we calculated the average r^2 over all common SNP pairs that are 30 kb apart. As expected, genes with poor coverage tend to have significantly lower levels of LD than genes with high coverage (data not shown).

Discussion

For six currently available SNP chips, we calculated a map of local coverage across the genome as well as the coverage of all known genes. All six SNP chips have demonstrated variation in their coverage. As GWA studies are becoming a major approach toward disease gene discovery, such explicit evaluation of coverage variation will give a full picture of the genotyping products. We believe that our results can facilitate several aspects in GWA studies.

First, it will be of interest to investigators who have specific prior interest in certain regions in the genome (e.g. candidate genes, linkage peaks, conserved elements and so on). Knowing the extent of coverage for these regions or genes can help determine whether supplementary genotyping is needed in addition to the whole-genome SNP chip.

Second, evaluation of local coverage and gene coverage can ease interpretation and comparison of inconsistent

results from GWA studies using different SNP chips. Inconsistency of results in a region or gene across studies might be partly due to differences in coverage. Our results on local coverage (Supplementary Figure 1) and gene coverage (Supplementary Table 1) provide a clear visualization of coverage across the genome for several widely used SNP chips. With such information, an investigator can easily compare local coverage of different SNP chips, aiding interpretation of different results.

Third, knowledge on local and gene coverage can help design new SNP chips. We recognize that the selection of SNPs to be included in a chip will depend on practical constraints; for example, it may be difficult to improve coverage for certain regions due to structural variations such as CNVs or other segmental repeats.[21,22] However, our results indicate that many genes in the genome have low coverage simply due to weak LD. Previous studies have shown that some genes are preferentially located in such regions, for example, genes that are involved in immune response and sensory perception.[23] Low coverage of a gene will often result in low power to detect genetic association if the disease variant falls in the gene. Evaluation of local and gene coverage can provide guidance on which regions or genes should receive denser coverage in the new chip.

When calculating gene coverage, we used the transcriptional start and end positions to define gene regions. We recognize that functional variants may exist in the 5′ or 3′ UTRs. However, the UTR information is not available for all the known genes and there is no consensus on how large the UTRs should be. Indeed, we repeated our calculation by expanding each region by 5 kb on each end, and observed similar results (data not shown).

It is commonly believed that GWA studies offer an unbiased approach for identification of susceptibility variants for complex diseases. However, even if the investigator does not impose any prior information onto a GWA study, the analysis results still will be biased toward regions and genes that are better covered by the SNP chip that is used in the study. Thus, for current SNP chips, it is desirable to carry out supplementary genotyping if necessary and to employ more flexible data analysis approaches that can take prior information into account.

In summary, we have evaluated coverage variation of different SNP chips for GWA studies at a finer scale. Although we focused on six SNP chips in this paper, the procedures that we employed are general and are not restricted to a particular product. As whole-genome SNP chips continue to evolve, we believe that detailed coverage evaluation will be valuable for comparing different genotyping products and designing future GWA studies. All results presented in this paper can be downloaded from http://biostat.mc.vanderbilt.edu/SNPChipCoverage.

Acknowledgements

We thank Drs Goncalo Abecasis, Michael Boehnke, Vivian Cheung, and Richard Spielman for discussion and critical reading of an earlier version of the paper, and Dr Kai Wang for providing the KEGG and BioCarta pathway information. This work was supported by an internal grant from the Center for Human Genetics Research at Vanderbilt University (to CL), and by the University Research Foundation grant and the McCabe Pilot Award from the University of Pennsylvania (to ML).

References

1 Hirschhorn JN, Daly MJ: Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005; **6**: 95–108.
2 Wang WY, Barratt BJ, Clayton DG, Todd JA: Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005; **6**: 109–118.
3 Hinds DA, Stuve LL, Nilsen GB *et al*: Whole-genome patterns of common DNA variation in three human populations. *Science* 2005; **307**: 1072–1079.
4 International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
5 Klein RJ, Zeiss C, Chew EY *et al*: Complement factor H polymorphism in age-related macular degeneration. *Science* 2005; **308**: 385–389.
6 Arking DE, Pfeufer A, Post W *et al*: A common genetic variant in the *NOS1* regulator *NOS1AP* modulates cardiac repolarization. *Nat Genet* 2006; **38**: 644–651.
7 Herbert A, Gerry NP, McQueen MB *et al*: A common genetic variant is associated with adult and childhood obesity. *Science* 2006; **312**: 279–283.
8 Duerr RH, Taylor KD, Brant SR *et al*: A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 2006; **314**: 1461–1463.
9 Sladek R, Rocheleau G, Rung J *et al*: A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 2007; **445**: 881–885.
10 Barrett JC, Cardon LR: valuating coverage of genome-wide association studies. *Nat Genet* 2006; **38**: 659–662.
11 International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. *Nature* 2004; **431**: 931–945.
12 Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R: Ascertainment bias in studies human genome-wide polymorphism. *Genome Res* 2005; **15**: 1496–1502.
13 Grimwood J, Gordon LA, Olsen A *et al*: The DNA sequence and biology of human chromosome 19. *Nature* 2004; **428**: 529–535.
14 Long JR, Liu PY, Liu YJ *et al*: APOE and TGF-beta-1 genes are associated with obesity phenotypes. *J Med Genet* 2003; **40**: 918–924.
15 Bernasconi P, Torchiana E, Confalonieri P *et al*: Expression of transforming growth factor-beta-1 in dystrophic patient muscles correlates with fibrosis: pathogenetic role of a fibrogenic cytokine. *J Clin Invest* 1995; **96**: 1137–1144.
16 Ziyadeh FN, Hoffman BB, Han DC *et al*: Long-term prevention of renal insufficiency, excess matrix gene expression, and glomerular mesangial matrix expansion by treatment with monoclonal antitransforming growth factor-beta antibody in db/db diabetic mice. *Proc Nat Acad Sci* 2000; **97**: 8015–8020.
17 Derynck R, Rhee L, Chen EY, Van Tilburg A: Intron–exon structure of the human transforming growth factor-beta precursor gene. *Nucleic Acids Res* 1987; **15**: 3188–3189.
18 Dong C, Zhu S, Wang T *et al*: Deficient Smad7 expression: a putative molecular defect in scleroderma. *Proc Nat Acad Sci* 2002; **99**: 3908–3913.
19 Pittet JF, Griffiths MJ, Geiser T *et al*: TGF-beta is critical mediator of acute lung injury. *J Clin Invest* 2001; **107**: 1537–1544.

20 Gupta A, Gartner JJ, Sethupathy P, Hatzigeorgiou AG, Fraser NW: Anti-apoptotic function of a microRNA encoded by the HSV-1 latency-associated transcript. *Nature* 2006; **442**: 82–85.
21 Feuk L, Carson AR, Scherer SW: Structural variation in the human genome. *Nat Rev Genet* 2006; **7**: 85–97.
22 Redon R, Ishikawa S, Fitch KR *et al*: Global variation in copy number in the human genome. *Nature* 2006; **444**: 444–454.
23 Smith AV, Thomas DJ, Munro HM, Abecasis GR: Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res* 2005; **15**: 1519–1534.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)