Genomics 101 (2013) 20–23

Contents lists available at SciVerse ScienceDirect

Genomics

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

# Evaluating coverage of exons by HapMap SNPs

Xiao Dong [a,b,1], Tingyan Zhong [a,b,1], Tao Xu [c], Yunting Xia [d], Biqing Li [a,b], Chao Li [a,b], Liyun Yuan [a], Guohui Ding [a,e,*], Yixue Li [a,e,*]

[a] Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai, People's Republic of China
[b] Graduate University of Chinese Academy of Science, 19 Yuquan Road, Beijing, People's Republic of China
[c] Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Munich/Neuherberg, Germany
[d] Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York 11794, USA
[e] Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, Shanghai, People's Republic of China

ARTICLE INFO

*Article history:*
Received 22 November 2011
Accepted 7 September 2012
Available online 19 September 2012

*Keywords:*
Genome-wide association study
Coverage
HapMap
Single-nucleotide polymorphism

ABSTRACT

Genome-wide association (GWA) studies are currently one of the most powerful tools in identifying disease-associated genes or variants. In typical GWA studies, single-nucleotide polymorphisms (SNPs) are often used as genetic makers. Therefore, it is critical to estimate the percentage of genetic variations which can be covered by SNPs through linkage disequilibrium (LD). In this study, we use the concept of haplotype blocks to evaluate the coverage of five SNP sets including the HapMap and four commercial arrays, for every exon in the human genome. We show that although some Chips can reach similar coverage as the HapMap, only about 50% of exons are completely covered by haplotype blocks of HapMap SNPs. We suggest further high-resolution genotyping methods are required, to provide adequate genome-wide power for identifying variants.

## 1. Introduction

Single-nucleotide polymorphism is one of the most important kinds of markers in genetic and disease-genetic studies. By using hundreds of thousands of SNPs, genome-wide association (GWA) studies aim to identify disease-associated SNPs across the whole human genome, and lead to the discovery of novel susceptible genes [1]. Although a lot of GWA studies have been launched and many are finished [2–6], there are still disputable aspects in the study designs and analyses. One of them is that, an investigator has to select known SNPs for genotyping and most of the times choose from the commercially available alternatives [7,8]. Therefore, it is important to evaluate the coverage of a set of SNPs, such as commercial Chips SNPs and HapMap SNPs. [7,9–11].

In previous studies several different approaches were used to evaluate genomic coverage of SNP chips, but the common feature they share is that they all measured the coverage rate as the possibility or percentage of SNPs, un-genotyped but in linkage disequilibrium (LD) with genotyped SNPs (linked SNPs, LSNPs for short; and TSNPs for short of genotyped SNPs or tag SNPs) [7,9,10]. So we call it the percentage-of-SNP coverage. However, there are limitations when applying this concept. The main reason is that there are many variations with unknown LD information with TSNPs, so one has to assume the percentage of coverage are similar between these variations and LSNPs (see Materials and methods). Moreover, the problem becomes more serious when calculating local coverage (Fig. 1) for genes.

In this study, we used the concept of regional coverage instead of the previous percentage-of-SNP coverage, and to determine regional coverage we applied the concept of haplotype blocks [12]. A haplotype block is a genomic region with little historical recombination and has a few common haplotypes within it (see Materials and methods section) [12]. If a region is covered by one haplotype block, little historical recombination will occur in this region, leading to high correlation among the base pairs, and DNA variants in it will be more sufficiently captured by surrounding TSNPs. For a certain part of genomic region, we define its coverage as whether it's fully covered by one haplotype block. In this way, we not only solve the problem mentioned above, but also provide two other advantages. First, we can calculate the coverage of the whole set of HapMap SNPs, because our measurement needs only one set of TSNPs, while in previous studies [7,9,10], two are needed (LSNPs and TSNPs). Second, for the regions to be evaluated, a higher resolution can be reached. We can calculate the coverage for smaller region such as an exon rather than a whole gene. However, it also limits us to an estimation of small regions, because haplotype blocks are usually short (tens of kb). Since protein coding sequences are estimated to harbor 85% of disease related mutations in human [13], we applied our measurement to exons. The coverage of each exon for all HapMap SNPs and four currently available SNP Chips were estimated in this study. The results should be helpful to determine if it is necessary for supplementary genotyping for a GWA study [9].

[*] Corresponding authors. Fax: +86 21 54920369.
*E-mail addresses:* gwding@sibs.ac.cn (G. Ding), yxli@sibs.ac.cn (Y. Li).
[1] These authors contribute equally to this work.

0888-7543/$ – see front matter © 2012 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.ygeno.2012.09.003

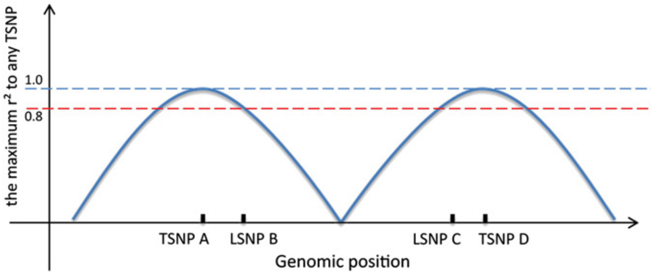

**Fig. 1.** A schematic diagram for the previously over-estimated situations. Given only four HapMap SNPs A to D located in this region, including two TSNPs (Tag SNPs) A and D, two LSNPs (Linked SNPs) B and C, and a distribution of maximum $r^2$ for every base pair to any TSNP, according to the previous definition of the percentage of SNPs [9], the coverage of this region is always 100%. However, there are weak-linked positions between B and C.

## 2. Results

### 2.1. Exon coverage by HapMap SNPs

We estimated the coverage of exons for four HapMap population(CEU, CHB, JPT and YRI) throughout the autosomal chromosomes by haplotype blocks (see Materials and methods). There are 49.9% exons covered by haplotype blocks in average among the four populations, and as shown in Fig. 2a, the coverage rates are similar for each individual chromosome. So a considerable large number of exons (~50%) are left uncovered (The word "uncovered" is defined in Materials and methods). With respect to population, CEU has the highest coverage rate (55.2%), the coverage rate for Asian populations is relatively the same (52.8%) and YRI has the lowest (41.6%). This can be expected since it was found that YRI has more genetic recombination and also smaller haplotype block size as mentioned above [12]. Detailed list of coverage for each exon can be accessed in Supplementary Table 1.

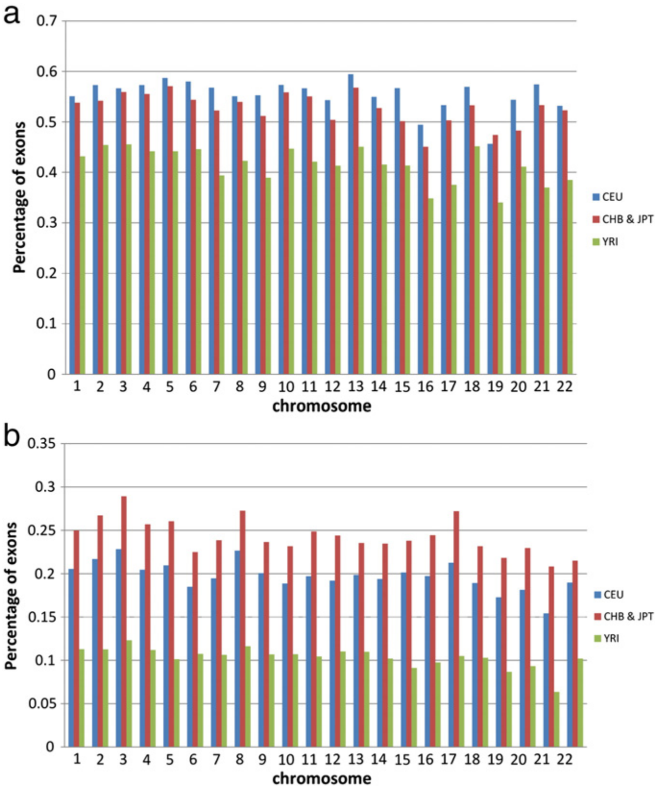

**Fig. 2.** Coverage by HapMap SNPs. a. The percentage of covered exons by HapMap SNPs for each chromosome in the four populations. CEU and Asian populations (55.2% and 52.8%, separately) have more exons covered than YRI populations (41.6%). b. The percentage of exons with flanking SNPs with $r^2 > 0.8$ [14]. The figure shows the percentage for each chromosome in four populations separately. In CEU and Asian populations, more exons have their flanking SNPs with high $r^2$ (20.1% and 24.8, separately), than those in YRI populations (10.6%).

The numbers of uncovered exons and genes for the four populations by all HapMap SNPs are summarized in Table 1. They are similar in CEU and Asian population, while in YRI populations, much more of them are uncovered. There are 44,250 exons uncovered in all the four populations, corresponding to nearly half of the protein coding genes (9720 out of 21,494).

As complementary information to the coverage capacity by haplotype blocks defined by Gabriel et al. [12], we calculate the correlation coefficient ($r^2$) between the two flanking SNPs of each exon (see Materials and methods). There are 20.1%, 24.5% and 10.6% of exons with high $r^2$ (>0.8 [14]) for CEU, CHB&JPT and YRI populations (Fig. 2b, and Supplementary Fig. 1 for the distribution of $r^2$ values). YRI is still obviously lower than other populations. Most of the exons with flanking SNPs, which have $r^2$ larger than 0.8, are covered by haplotype blocks (86.0%, 79.0% and 80.3% in CEU, CHB&JPT and YRI populations separately). The number of SNPs around each exon (±1 kb, ±2 kb and ±5 kb) is also calculated and provided in Supplementary Table 1. Although only a few number of exons are without a SNP in the surrounding region (for example 1.17%, 1.83% and 0.99% exons without a SNP in ±5 kb in CEU, CHB&JPT and YRI populations separately), if only considering this information, it may lead to the same problem as the previous definition of coverage [9] as we demonstrate in Fig. 1.

### 2.2. Exon coverage by commercial SNP Chips

We also estimated the block-coverage for each exon region for four widely used genome wide SNP Chips, the Affymetrix 5.0 and 6.0, and Illumina 660W and 1M. Fig. 3 shows the percentage of covered exons by the four Chips for different populations. Illumina 1M has the highest coverage rate in all populations. It almost reaches the coverage rate by all HapMap SNPs, which can explain the result from the previous study, using percentage of ungenotyped by genotyped SNPs [7,9,10]. The coverage rates of Affymetrix 6.0 and Illumina 660W are similarly lower than Illumina 1M but higher than Affymetrix 5.0. For the four populations, the CEU population is unsurprisingly the most well-covered and YRI has the lowest coverage rate.

### 2.3. Comparing the two definitions

To demonstrate the difference between the previous definition and the one that we proposed quantitatively, we calculated the local percentage-of-SNP coverage of each exon with its ±20 kb flanking sequence for Illumina 1M Chip CEU population, according to Ref. [9] as the following equation.

$$\frac{\left(\frac{L}{R-T}\right)(G-T)+T}{G} \tag{1}$$

where $T$ represents the number of TSNPs in a region, $L$ for LSNPs, and $R$ for a reference set of SNPs. $G$ is the total number of common SNPs, which are unknown and estimated as $(\mathrm{le}+40{,}000)\times n/e$, where le refers the length of an exon, $n$ is the number of common SNPs

**Table 1**
A summary of uncovered regions.

| | CEU | CHB & JPT | YRI | Total |
|---|---|---|---|---|
| Number of uncovered exons | 90,039 | 94,993 | 117,327 | 44,250 |
| Number of uncovered genes | 13,637 | 13,776 | 15,151 | 9,720 |

'Uncovered exons' means that the not fully covered exons by haplotype blocks and 'uncovered genes' refers to genes with at least one not fully covered exon. The last column "total" refers to the exons or genes that are not fully covered in any population.

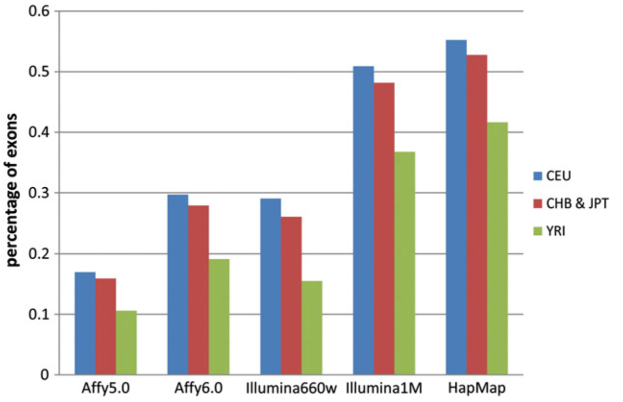

**Fig. 3.** The percentage of covered exons by HapMap SNPs and four commercial Chips for the four populations. Illumina 1M almost reaches a similar percentage of all HapMap SNPs. Their average percentages among four populations are 45.3% and 49.9%, separately. Illumina 660W and Affymatrix 6.0 are similar (23.5% and 25.6% separately) and Affymatrix 5.0 is the lowest (14.5%).

(estimated as $7.5\times10^6$) and $e$ is the number of euchrometic base pairs ($2.85\times10^9$) in human genome [7,15]. According to this definition, the coverage rate is a value between 0 and 1.

Unexpectedly, we found that, only 37.2% exons are with a high percentage-of-SNP coverage rate ($\geq 0.8$), and it was supposed to be at a similar level than the coverage rate of genes (93% in Ref. [9]). It is shown in Supplementary Fig. 3 that there is a weird distribution of coverage rate according to percentage-of-SNP definition. We found an extreme case that the percentage-of-SNP coverage rate of the exon ENSE00001784339 is larger than 1, because it has only 106 common SNPs as estimation ($G$) but actually has 110 TSNPs ($T$). This example suggests that the estimation of $G$ of a small region (such as an exon) may be quite different from the real one, and this would lead to inaccuracy and instability in estimations as the above. When comparing with the coverage rate defined by haplotype block, a similar trend was observed that the covered SNPs defined by regional coverage have larger percentage-of-SNP coverage rate than uncovered ($P<1.0\times10^{15}$, Wilcox rank sum test). However, the result of this comparison may be not informative, due to the low accuracy of percentage-of-SNP definition at the exon level.

## 3. Discussion

In this study, we evaluated the coverage by commercial Chips, and some of the Chips can almost reach the coverage by all HapMap SNPs. Therefore, with carefully selected set of SNPs, some of the current commercial Chips can capture most of the common variant, for example SNPs, in the genome.

However, the power by using even all HapMap SNPs to cover most of the variants in all exon regions is limited. Our result indicates about 50% of the exons are not fully covered by haplotype blocks. Variants within the uncovered exons may have low associations with surrounding genotyped SNPs than those within the covered ones. In consequence, many may failed to be captured, even if a researcher genotypes all HapMap SNPs.

To test possible difference in coverage capacity between coding region and the whole genome, we randomly picked 10,000 regions from the genome (finally 9988 were used, 22 were missed because of calculation with integers), with the same distribution of regional length as that of exon length (Supplementary Fig. 2). The coverage of these regions for the CEU population by HapMap SNPs was estimated in the same way as that of the exons. We find that 54.0% of the randomly picked regions are covered (Supplementary Table 2), compared to 55.2% covered for exons (p-value by $\chi2$ test is 0.2274). Although it is well known that the frequencies of SNPs are different between coding and non-coding regions [16,17] and DNA sequence polymorphism correlates with recombination rate [18,19], our result indicates similar rates of coverage for both exons and other non-coding regions. It suggests that recombination doesn't favor the exons or the other regions at the genome-wide level, and results from a previous study also lead to a similar suggestion, that coding bases are enriched in regions of both high and low LD (the top and bottom quartiles of the genome) [20].

Although the definition of regional coverage has the advantages, it also has some limitations. First, the accurate estimation of haplotype block depends on a carefully selected human population. In this article we use HapMap populations. Although in some GWA studies, no significant difference between the HapMap populations and GWA populations in LD structure was observed at the genome-wide level [2], it may exist at a local exon level. And for family-based studies and studies on diseases with strong hereditary background, the difference between our estimation and the study population may also exist. We suggest that one would estimate the specific coverage rate for the specific study population if possible, and it would correct the mistakes caused by inconsistency between ours and a specific study. Second, we currently cannot provide a parameter and criteria to indicate the confidence of our estimation, and it is also a problem when using the percentage-of-SNP definition. Third, according to the definition of regional coverage, we can only estimate local coverage of sequence with short length, because haplotype blocks are of limited size (usually tens of kb) [12].

In summary, we have evaluated the coverage for each exon by HapMap SNPs, and calculated $r^2$ of their flanking SNPs. The coverage by four commercial Chips was also evaluated. We believe that the low coverage by using SNPs as genetic markers is a problem for current GWA studies. With advances of next-generation sequencing technique, using exon capture and sequencing, one can avoid the problem of low coverage by SNP arrays, and directly identify variants in the exon regions [21,22]. For example, in one study on pancreatic cancer, novel variants in exon regions of PALB2 gene were discovered [23]. Currently this approach could be a compliment to SNP-based GWA studies, and it may be developed to be one of the major effective methods in the future.

## 4. Materials and methods

### 4.1. Data sets

We use SNPs and their information of linkage disequilibrium for four HapMap populations (CEU, CHB, JPT and YRI) from HapMap project release no. 27 and four SNP Chips in GWA studies, Affymetrix's Genome-Wide Human SNP Array 5.0, 6.0 and Illumina's Human 660W-Quad, Human 1M-Duo. The reference set of exons was downloaded from Ensemble Biomart, Ensemble Genes 66, Homo sapiens genes (GRCh37/hg19), with two filters, with an Ensemble exon ID and from a protein coding gene. The version of exon annotation was converted to NCBI36/hg18 for further analysis using utilities from UCSC Genome Browser [24].

### 4.2. Definition of coverage

In previous studies [7,9,10], the coverage defined in a similar way as in Ref. [7] follows. First, a naive estimate of coverage of all SNPs in the genome ($G$) is

$$\frac{L+T}{R} \tag{2}$$

where $T$ represents the number of TSNPs, $L$ for LSNPs, and $R$ for a reference set of SNPs, such as the Phase II HapMap. Second, they suggested that since $G>R$, it may cause overestimation using the above equation. So a corrected version is as in Eq. (1) and they set $G$ as 7.5 million in the genome-wide level. We call this definition as the percentage-of-SNP coverage. One can find that their correction was based on an assumption that,

$$\frac{L_R}{L_G} = \frac{R - T}{G - T} \tag{3}$$

where $L_R$ is the LSNPs from a reference SNP set, and $L_G$ is the variants from $G$ and also in LD with TSNPs. However, it is not certain whether the above assumption is reasonable, especially when discussing local coverage. In Fig. 1, we presented a situation when the percentage-of-SNP coverage definition fails.

To avoid the above problem, we use the concept of regional coverage. Rather than common SNP (with an allele frequency > 1% in a population), mutations may exist at any base pair in a sequence (such as an exon). So we wish to test if all sequence of an exon is in stable regions (such as haplotype blocks). Notice that when the two ends of an exon are in different haplotype blocks, an interval of exon sequence could be found, which is not within either of the blocks, we define an exon as "covered", when all of this sequence is within one and only one haplotype block, and others as "not fully covered". There are many methods to define haplotype blocks [12,25–31] (reviewed by Ref. [32]). We use the definition from Gabriel et al. [12]. They used the term "strong evidence for historical recombination" SNP pairs if the one-sided upper 95% confidence bound on D′, the normalized measure of allelic association [31,33], is less than 0.9. They called a genomic region as a haplotype block, when less than 5% of comparisons among informative SNP pairs in the region show strong evidence of historical recombination [12].

To estimate coverage of a certain set of SNPs for a target exon, we first selected a region including the target exon and ±20 kb sequences on both its sides. Second, the distribution of haplotype blocks over this region was estimated by Haploview [34], with a quality check for SNPs excluding those with a Hardy Weinberg p-value less than 0.001. Third, if all base pairs of the target exon are within one haplotype block, it is labeled as "covered", otherwise as "uncovered" (or not fully covered). The correlation coefficient ($r^2$) of two flanking SNPs of each exon is also provided as supplementary information to our analysis.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2012.09.003.

## Acknowledgments

We thank Drs Zhen Wang and Guangyong Zhen for their helpful comments and suggestions. This work was supported by grants from State key basic research program (973) (2011CB910204, 2011CBA00801), the Knowledge Innovation Program of the Chinese Academy of Sciences (KSCX2-YW-R-112, KSCX2-YW-R-190, 2011KIP204), and National Natural Science Foundation of China (90913009, 30900272).

Conflict of interest statement

The authors declare no conflict of interest.

## References

[1] J.N. Hirschhorn, M.J. Daly, Genome-wide association studies for common diseases and complex traits, Nat. Rev. Genet. 6 (2005) 95–108.
[2] P.R. Burton, D.G. Clayton, L.R. Cardon, et al., Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, Nature 447 (2007) 661–678.
[3] X. Chu, C.M. Pan, S.X. Zhao, et al., A genome-wide association study identifies two new risk loci for Graves' disease, Nat. Genet. 43 (2011) 897–901.
[4] T. Hirota, A. Takahashi, M. Kubo, et al., Genome-wide association study identifies three new susceptibility loci for adult asthma in the Japanese population, Nat. Genet. 43 (2011) 893–896.
[5] R. Saxena, B.F. Voight, V. Lyssenko, et al., Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels, Science 316 (2007) 1331–1336.
[6] R.H. Duerr, K.D. Taylor, S.R. Brant, et al., A genome-wide association study identifies IL23R as an inflammatory bowel disease gene, Science 314 (2006) 1461–1463.
[7] J.C. Barrett, L.R. Cardon, Evaluating coverage of genome-wide association studies, Nat. Genet. 38 (2006) 659–662.
[8] E. Zeggini, W. Rayner, A.P. Morris, et al., An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets, Nat. Genet. 37 (2005) 1320–1322.
[9] M.Y. Li, C. Li, W. Guan, Evaluation of coverage variation of SNP chips for genome-wide association studies, Eur. J. Hum. Genet. 16 (2008) 635–643.
[10] R. Magi, A. Pfeufer, M. Nelis, et al., Evaluating the performance of commercial whole-genome marker sets for capturing common genetic variation, Bmc Genomics 8 (2007).
[11] K.A. Frazer, D.G. Ballinger, D.R. Cox, et al., A second generation human haplotype map of over 3.1 million SNPs, Nature 449 (2007) 851–861.
[12] S.B. Gabriel, S.F. Schaffner, H. Nguyen, et al., The structure of haplotype blocks in the human genome, Science 296 (2002) 2225–2229.
[13] M. Choi, U.I. Scholl, W.Z. Ji, et al., Genetic diagnosis by whole exome capture and massively parallel DNA sequencing, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 19096–19101.
[14] C.S. Carlson, M.A. Eberle, M.J. Rieder, et al., Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium, Am. J. Hum. Genet. 74 (2004) 106–120.
[15] International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome, Nature 431 (2004) 931–945.
[16] B.A. Salisbury, M. Pungliya, J.Y. Choi, et al., SNP and haplotype variation in the human genome, Mutat. Res. Fundam. Mol. Mech. Mutagen. 526 (2003) 53–61.
[17] M. Cargill, D. Altshuler, J. Ireland, et al., Characterization of single-nucleotide polymorphisms in coding regions of human genes, Nat. Genet. 22 (1999) 231–238.
[18] M.W. Nachman, V.L. Bauer, S.L. Crowell, et al., DNA variability and recombination rates at X-linked loci in humans, Genetics 150 (1998) 1133–1141.
[19] D.J. Begun, C.F. Aquadro, Levels of naturally-occurring DNA polymorphism correlate with recombination rates in Drosophila-melanogaster, Nature 356 (1992) 519–520.
[20] A.V. Smith, D.J. Thomas, H.M. Munro, et al., Sequence features in regions of weak and strong linkage disequilibrium, Genome Res. 15 (2005) 1519–1534.
[21] J. Majewski, J. Schwartzentruber, E. Lalonde, et al., What can exome sequencing do for you? J. Med. Genet. 48 (2011) 580–589.
[22] E. Hodges, Z. Xuan, V. Balija, et al., Genome-wide in situ exon capture for selective resequencing, Nat. Genet. 39 (2007) 1522–1527.
[23] S. Jones, R.H. Hruban, M. Kamiyama, et al., Exomic sequencing identifies PALB2 as a pancreatic cancer susceptibility gene, Science 324 (2009) 217.
[24] D. Karolchik, R. Baertsch, M. Diekhans, et al., The UCSC genome browser database, Nucleic Acids Res. 31 (2003) 51–54.
[25] M.P.H. Stumpf, D.B. Goldstein, Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium, Curr. Biol. 13 (2003) 1–8.
[26] M.S. Phillips, R. Lawrence, R. Sachidanandam, et al., Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots, Nat. Genet. 33 (2003) 382–387.
[27] K. Zhang, M.H. Deng, T. Chen, et al., A dynamic programming algorithm for haplotype block partitioning, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 7335–7339.
[28] N. Wang, J.M. Akey, K. Zhang, et al., Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation, Am. J. Hum. Genet. 71 (2002) 1227–1234.
[29] E. Dawson, G.R. Abecasis, S. Bumpstead, et al., A first-generation linkage disequilibrium map of human chromosome 22, Nature 418 (2002) 544–548.
[30] N. Patil, A.J. Berno, D.A. Hinds, et al., Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21, Science 294 (2001) 1719–1723.
[31] M.J. Daly, J.D. Rioux, S.E. Schaffner, et al., High-resolution haplotype structure in the human genome, Nat. Genet. 29 (2001) 229–232.
[32] J.D. Wall, J.K. Pritchard, Haplotype blocks and linkage disequilibrium in the human genome, Nat. Rev. Genet. 4 (2003) 587–597.
[33] R.C. Lewontin, Interaction of selection and linkage .I. general considerations - heterotic models, Genetics 49 (1964) 49–67.
[34] J.C. Barrett, B. Fry, J. Maller, M.J. Daly, Haploview: analysis and visualization of LD and haplotype maps, Bioinformatics 21 (2005) 263–265.