SCIENTIFIC REPORTS

OPEN

Evaluating information content of SNPs for sample-tagging in re-sequencing projects

Received: 16 September 2014
Accepted: 07 April 2015
Published: 15 May 2015

Hao Hu[1,*], Xiang Liu[2,*], Wenfei Jin[3], H Hilger Ropers[1] & Thomas F Wienker[1]

Sample-tagging is designed for identification of accidental sample mix-up, which is a major issue in re-sequencing studies. In this work, we develop a model to measure the information content of SNPs, so that we can optimize a panel of SNPs that approach the maximal information for discrimination. The analysis shows that as low as 60 optimized SNPs can differentiate the individuals in a population as large as the present world, and only 30 optimized SNPs are in practice sufficient in labeling up to 100 thousand individuals. In the simulated populations of 100 thousand individuals, the average Hamming distances, generated by the optimized set of 30 SNPs are larger than 18, and the duality frequency, is lower than 1 in 10 thousand. This strategy of sample discrimination is proved robust in large sample size and different datasets. The optimized sets of SNPs are designed for Whole Exome Sequencing, and a program is provided for SNP selection, allowing for customized SNP numbers and interested genes. The sample-tagging plan based on this framework will improve re-sequencing projects in terms of reliability and cost-effectiveness.

Next-generation-sequencing (NGS) has gained its ground in medical research in recent years, and NGS-based re-sequencing has become a prevalent procedure in revealing causative variants[1,2]. In the foreseeable future, re-sequencing-based diagnosis will be conducted on a routine basis, not only in research labs, but also in clinical facilities. This heralds an increasing number of samples, and unavoidably exacerbates one of the existing problems in re-sequencing projects, namely, sample mix-up, which will lead to wrong diagnosis. Although there has been hitherto no general investigation on its incidence, our in-house records and the anecdotal sources show the sample mix-up at a rate between 0.1% and 1%. This problem can originate in each step of a re-sequencing project, thus entailing step-wise quality control. The intuitive solution to sample mix-up is sample-tagging and sample-matching, pre and post re-sequencing procedure, respectively, which was simply explored and discussed in recent studies[3–5]. On the other hand, there are various studies showed that SNPs could provide information to distinguish different populations, cell lines and individuals[6–12], and approaches for optimizing SNP selection have been proposed[7,8,11,13–15]. These approaches consider haplotypes and linkage disequilibrium, however, development of these approaches is usually earlier than the NGS era and the application in the NGS data has not been fully considered[7,8,11,13,14]. Also, most of these approaches are based on empirical procedures, thus lacking expansibility and robustness with the increasing haplotype data in diversified populations[7,8,11,15]. Therefore, a sophisticated SNP selection approach with a strict theoretical framework for sample-tagging in large-scale re-sequencing is still in need.

Any sample-tagging plan usually utilizes polymorphic markers, especially SNPs. The human genome is a hive of SNPs which accumulate in the long history of human evolution, diversifying individual genomes. There are more than 10 million SNPs in the human genome[16], and the latest version of 1000 Genomes Project database (phase 3) even includes more than 78 million SNPs, but not all of them

[1]Department of human molecular genetics, Max-Planck Institute for Molecular Genetics, Berlin, 14195, Germany. [2]BlackBerry Deutschland GmbH, Bochum, 44799, Germany. [3]Systems Biology Center, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD, 20892, USA. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to H.H. (email: hu@molgen.mpg.de)

can serve as useful discriminative markers. On one hand, only SNPs at a certain level of minor allele frequency (MAF) can be used, e.g., SNPs with a very rare minor allele should not be engaged. On the other hand, varying SNP frequencies in different ethnic groups render it hard to define the versatile markers for all populations. Even not to mention that SNPs with alleles in linkage disequilibrium lead to the slowed ascent of cumulative discriminative capability. Because the cost of sample-tagging plan is approximately proportionate to SNP number, it is of significant value to illustrate the least number of SNPs required to distinguish different samples considering all the aforementioned scenarios.

Suppose the present world population is 7 billion and the two-state markers are used to label each individual, then there are at least 33 markers needed ($2^{33} = 8.5$ billion). In the same way, suppose the markers with three genotypes, e.g., homozygous wild-type, homozygous mutant, and heterozygous mutant, are used to label, the least number of markers is 21 ($3^{21} = 10.4$ billion). It is obvious that these plans can work only if the markers have the two characteristics: 1. each genotype of a marker should be equally frequent; 2. the genotypes of one marker should be statistically independent of all other markers. Since we know that these conditions do not hold in reality, there are definitely more markers needed. According to information theory, the information contained in the aforementioned 33 markers and 21 markers are identical, equal to 33 bits of Shannon-entropy[17]. In other words, any marker plan of labeling the world population has to provide the information content no less than 33 bits. Therefore, in this work, we construct a model to evaluate the information content of SNPs, so that the optimized SNP sets can differentiate individual genomes in a reliable and efficient way.

Results

Sample collection.

Evaluation of SNPs' information content depends on their MAF, so that it is favorable to obtain genotypes of individual genomes. Three large samples with genotypes are included in this work, i.e., 1,092 individuals of the 1000 Genomes Project[18], 1,115 individuals of the International HapMap Project[16], and 1,043 individuals of the Human Genome Diversity Project[19]. The allele frequencies of ~ 2 million variants in ~ 6,500 individuals from the Exome Sequencing Project are involved in part of the modeling[20]. The genotypes drawn from the in-house whole exome sequencing (WES) of 265 individuals are used in testing the model. In the ensuing sections of this paper, these notations are used, i.e., HapMap for the International HapMap Project, HGDP for the Human Genome Diversity Project, ESP for the Exome Sequencing Project, MPIMG265 for the in-house 265 individuals.

MPIMG265 is comprised of individuals from multiple ethnic groups, focusing on the populations from the Middle East, which are presently under-represented in the polymorphism databases. MPIMG265 includes 176 individuals from Iran, 9 from Pakistan, 6 from Turkey, 1 from Saudi Arab, and 73 from Germany and other European countries. For every individual, the informed written consent was obtained from each family. MPIMG265 samples are processed under the supervision of GENCODYS consortium, the European Commission Framework Program 7 (Health-241995). On the other hand, among these 265 individuals, there are 14 trios (42 individuals) and 4 quartets (16 individuals). The genotypes of these 265 individuals are extracted from the WES results, and the procedures of variant calling and genotyping are described previously[2,21].

Only the coding SNPs on the autosomal chromosomes (RefSeq gene model, downloaded from UCSC Genome Browser in June of 2013) are considered in this work. There are 472,514, 1,093,534, 9,472 and 14,521 suitable SNPs collected from 1000 Genomes Project, ESP, HGDP, and HapMap, respectively.

In order to test the possible sample overlap between 1000 Genomes Project, HapMap, and HGDP, 2,236 coding SNPs shared by these three datasets are used to compare the individual genotypes mutually. Individuals with >95% identity in genotype are regarded as sample overlap. This test shows there is no sample overlap existing between 1000 Genomes Project, HapMap, and HGDP.

In order to test the model, the simulated genotypes of four major ethnic groups, i.e., African (AFR), European (EUR), East-Asian (ASN), and American (AMR), are generated according to the allele frequencies drawn from 1000 Genomes Project and HapMap. The simulation size for each ethnic group is 100 thousand individuals, and the simulated genotypes obey the Hardy-Weinberg equilibrium.

When evaluating the performance of the optimized SNP sets in differentiating individuals, random SNP sets with SNP numbers from 20 to 50 are each generated by 1,000 times. The random SNPs are all located on the autosomal chromosomes, with MAF between 0.4 and 0.5 according to 1000 Genomes Project and HapMap. These random SNP sets stand for an intuitive design which is expected to reach suboptimal performance, much better than random SNP sets without MAF restriction while inferior to the optimized SNP sets.

Information content of single SNP in different datasets.

The information content, or Shannon-entropy of single SNP in 1000 Genomes Project, ESP, HGDP, and HapMap is calculated according to the variability of its three genotypes (homozygous wild-type, homozygous mutant, and heterozygous mutant). The information content is ranging from almost zero to as high as 1.585, which is the expected maximal entropy for a single SNP (Fig. 1). Since 1000 Genomes Project and ESP are both based on WES which covers all the coding SNPs indiscriminatingly, the majority of the SNPs have a low MAF, with entropies near zero accordingly. On the other hand, SNPs used in HGDP and HapMap are known to be common in population, therefore, the bulk of their entropies are larger than 1 bit. There are 6,745 SNPs shared by the four datasets, namely, 1000 Genomes Project, ESP, HGDP, and HapMap. In order

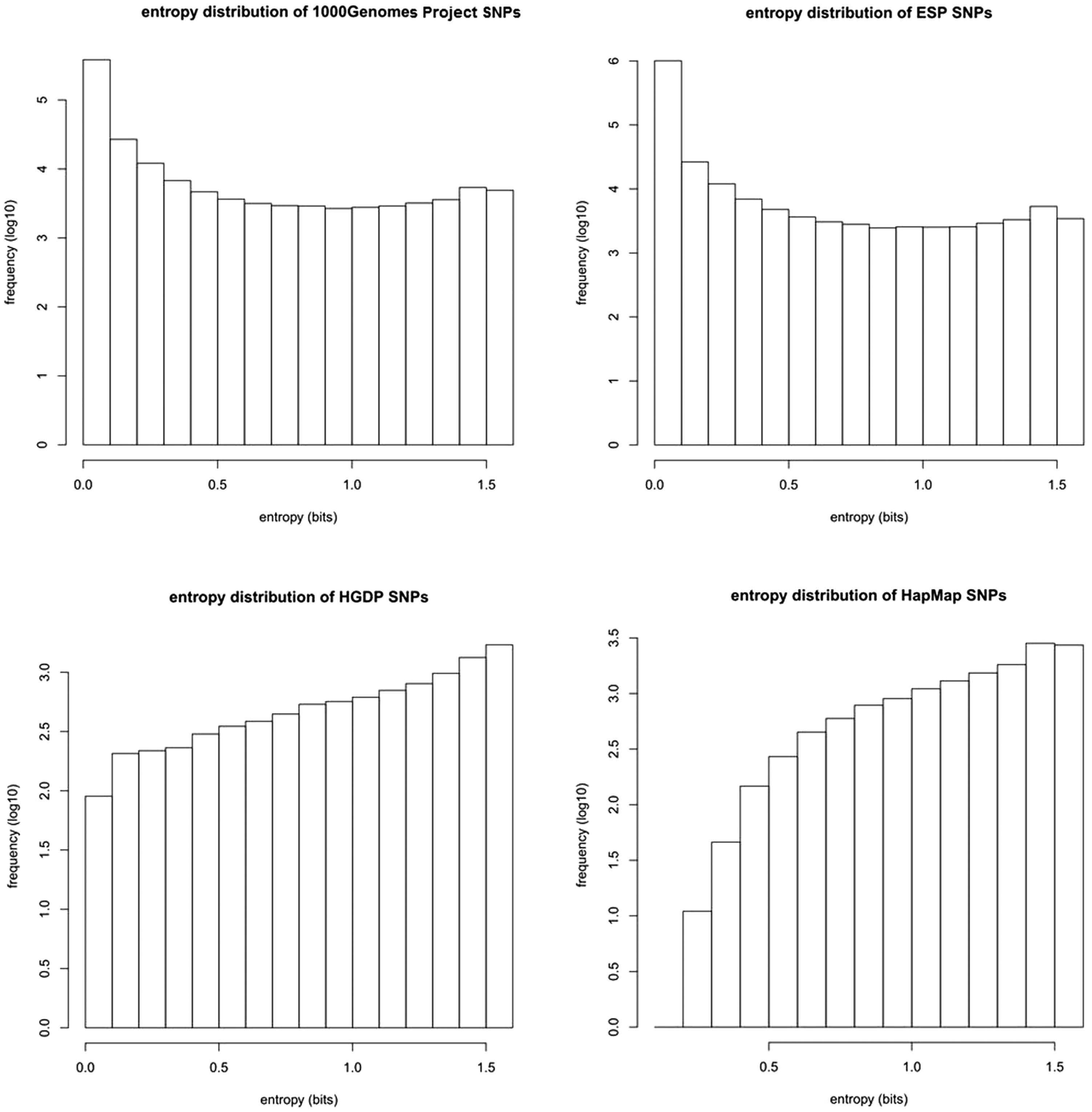

Figure 1. The entropy distribution of SNPs in 1000 Genomes Project, ESP, HGDP, and HapMap. The X-axis is the Shannon-entropy calculated for each SNP, the Y-axis is the frequency of entropy values in the specific database.

to evaluate the consistency of SNP entropy from different datasets, the pairwise Pearson correlation coefficients are calculated (Table 1). The shared SNPs in the four datasets show high correlation in terms of entropy, which indicates the potential of choosing SNPs consistent in entropy in different datasets.

To refine SNPs by entropy consistency, SNPs are regarded as qualified if their entropies do not fluctuate in different datasets. Pairwise bootstrap tests are conducted to identify the qualified SNPs among 1000 Genomes Project, HGDP and HapMap, for the three datasets have different composition of ethnic groups. Only SNPs shared by 1000 Genomes Project, HGDP, and HapMap, which are equal in entropy in the 95% bootstrap percentile confidence interval, are remained. And SNPs whose mappability are not 1[21] or GC-content beyond the expected interval of 0.35 - 0.55 are removed (the GC content scores are downloaded from UCSC Genome browser GC Percent (gc5Base) track). Consequently, there are 2,236 SNPs defined as the qualified SNPs for the subsequent procedures and their entropies are then re-calculated by the combined genotypes of 1000 Genomes Project, HGDP, and HapMap.

	ESP	HGDP	HapMap
1000 Genomes Project	0.893	0.965	0.957
ESP		0.84	0.874
HGDP			0.907
P-value $<10^{-16}$			

Table 1. The pairwise Pearson correlation coefficients among ESP, HGDP, HapMap, and 1000 Genomes Project, regarding the SNP entropy.

Mutual information of qualified SNPs. In order to measure the genotype dependency between SNP pairs, the mutual information of each pair of the qualified SNPs is computed, based on the combined genotypes of 1000 Genomes Project, HGDP, and HapMap, altogether 3,250 individuals (Fig. 2). Large values of mutual information correspond to high mutual dependency, and thus a strong pattern of linkage disequilibrium (LD) between SNP pairs. Not very surprising, the SNP pairs close to each other have relatively higher mutual information. It is notable that the SNP-pairs on the same chromosomes have reduced mutual information with increasing distance but not dropping to zero. Also, the mutual information of SNP-pairs on different chromosomes, although much lower than those on the same chromosomes, are still not low enough to be totally neglected.

Joint entropy of given-number SNPs. Information theory tells us that the total information harbored in the multiple variables is measured by the individual entropies substracting the mutual information. Since the entropy of each SNP has been obtained, as well as the mutual information of SNP pairs, from the genotypes of 3,250 individuals, the total information, or joint entropy of all possible SNP combinations, can be calculated according to the aforementioned algorithm. It has been known that the exact value of joint entropy of more than 3 ($N > 3$) SNPs is difficult to calculate, but the upper and lower bounds can be estimated as H_U and H_L. For the sets of SNPs with given number from 1 to 100 (N=1, ,100), the joint entropies are calculated for the maximal values, according to the aforementioned recursive algorithm. The gap between H_U and H_L grows as the number of SNPs increases (Fig. 3). Another interesting observation is that the joint entropy is 33 bits when the number of SNPs reaches 22. This indicates that at least 22 qualified SNPs are necessary for tagging the present world population.

Expected Hamming distance and duality between individuals. There are two major concerns in designing discriminative set of SNPs: the expected Hamming distance, or mismatches, between a pair of individuals and the probability of duality, or total match.

A population of L individuals is considered. There are S genotypes for each SNP. A set of N independent SNPs is used for differentiating all L individuals in the population, assuming that all S genotypes are equiprobable for each SNP, thus reaching the maximal joint entropy. If the Hamming distance between any two individuals is greater than t , these two individuals are regarded as different. The k-combination of n-set is denoted as

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} \tag{1}$$

The k-permutation of n-set is denoted as

$$P(n, k) = \frac{n!}{(n-k)!} \tag{2}$$

The total number of individuals which can be differentiated by the N-SNP set is S^N. The number of individual pairs which can be differentiated by the$N$-SNP set is S^{2N} . The number of individual pairs which has a Hamming distance of k in terms of the N-SNP set is

$$S^N \binom{N}{k} (S-1)^k \tag{3}$$

Hence the probability that two individuals having a Hamming distance of k in terms of the$N$-SNP set is

$$q_k = \frac{S^N \binom{N}{k} (S-1)^k}{S^N S^N} = \frac{\binom{N}{k} (S-1)^k}{S^N} \tag{4}$$

The probability that two individuals having a Hamming distance no more than t is

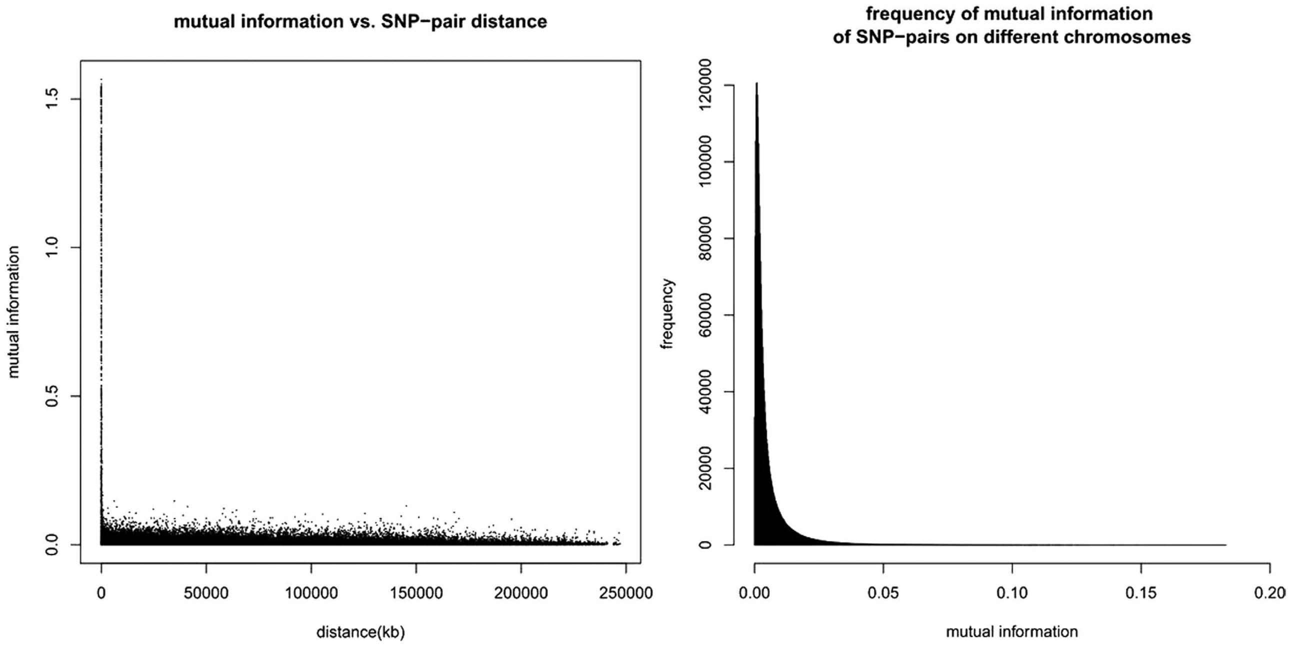

Figure 2. The mutual information distributions of SNP-pairs regarding distances and chromosomes.

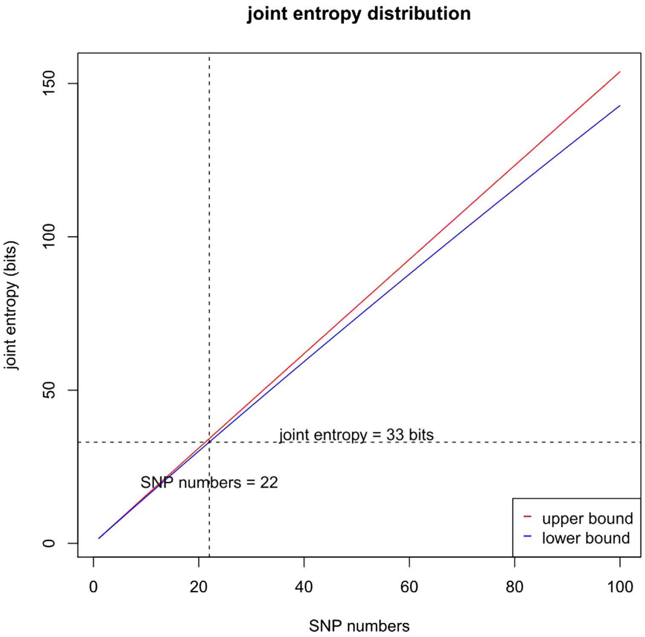

Figure 3. The lower bound and upper bound of joint entropy, i.e., H_L and H_U, increase following the increasing SNP number from 1 to 100, with widening gap between H_L and H_U. The joint entropy reaches 33 bits when the SNP number reaches 22.

$$q_{\leq t} = \sum_{k=0}^{t} q_k \tag{5}$$

The expected Hamming distance between two individuals:

$$d = \sum_{k=1}^{N} k q_k = \frac{1}{S^N} \sum_{k=1}^{N} k \binom{N}{k} (S-1)^k = \frac{1}{S^N} N(S-1) \sum_{k=0}^{N-1} \binom{N-1}{k-1} (S-1)^{k-1} = \frac{N(S-1)}{S} \tag{6}$$

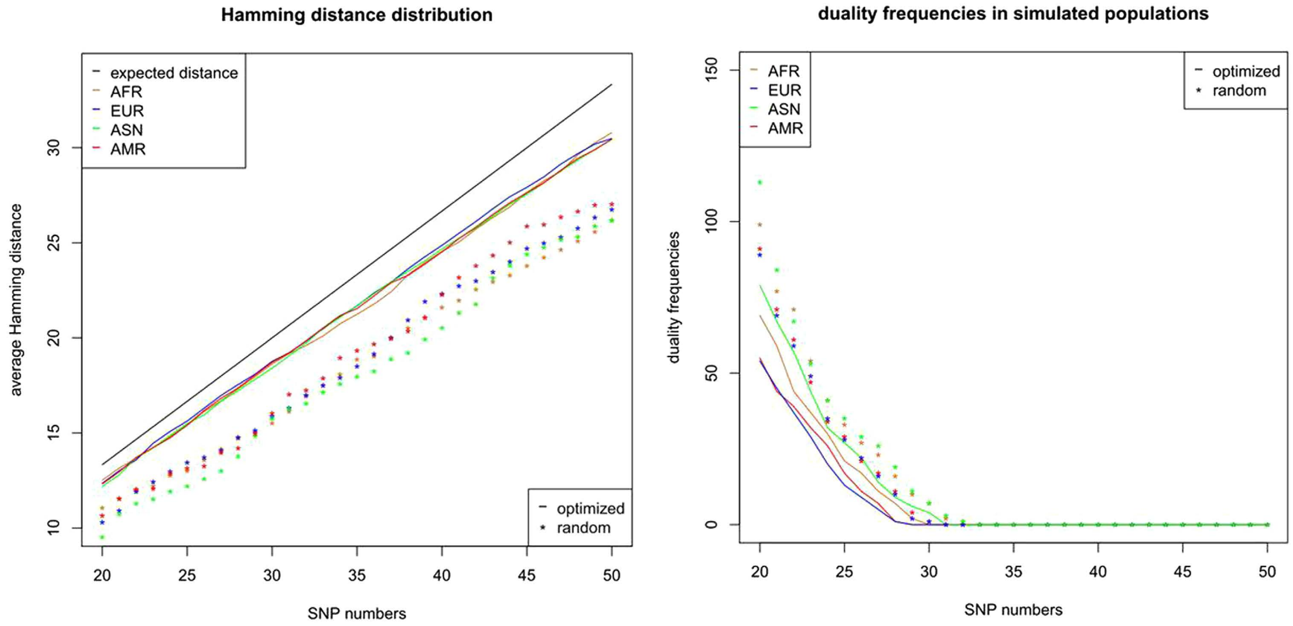

Figure 4. In the simulated populations (AFR, ASN, AMR, EUR), each of which has 100 thousand individuals, the Hamming distances and duality frequencies are generated by both the optimized sets of SNPs and the random sets of SNPs. The SNP numbers range from 20 to 50. The different populations are represented by different colors, and the lines stand for the optimized sets of SNPs and the asterisks stand for the average values of the random sets of SNPs.

Thus, we have the expected Hamming distance $2N/3$ in the case of$S = 3$ (Fig. 4).

The probability that any two individuals in a population of L individuals have a Hamming distance greater than t is approximated as

$$\overline{Q_{\le t}} \approx \left(1 - q_{\le t}\right)^{\binom{L}{2}} = \left(1 - \sum\nolimits_{k=0}^{t} q_k\right)^{\frac{L(L-1)}{2}} \quad (7),$$

assuming $L \le S^N$.

The probability that at least two individuals in a population of L individuals have a Hamming distance no more than t is $Q_{\le t} = 1 - \overline{Q_{\le t}}$, assuming $L \le S^N$. This is difficult to calculate in the case of a large population size of L. We may use an approximation assuming $L \le S^N$:

$$Q_{\le t} \approx 1 - \left(1 - q_{\le t}\right)^{\binom{L}{2}} \approx \frac{L(L-1)}{2} \sum_{k=0}^{t} q_k \quad (8)$$

Thus, we have the probability $Q_0 \approx$ 5.78e-10 when $L = 7e9$, $N = 60$, $S = 3$ and $t = 0$; and $Q_{\le 2} \approx$ 4.16e-6 when $L = 7e9$, $N = 60$, $S = 3$ and $t = 2$. Very interestingly, these results indicate that although the information content of an optimized set of 22 SNPs has the potential to label each individual in a population of 7 billion individuals, after assuming the independent distribution of genotypes, an optimized set of 60 SNPs can differentiate all 7 billion individuals with duality probability of 5.78e-10, even when tolerating 2 accidental SNP errors, the adjusted duality probability is still as low as 4.16e-6.

For example, if there is a population of 100 thousand individuals, and the optimized SNP sets are assumed independent in genotypes, according to our estimation of duality, the optimized set of 30 SNPs can give rise to a satisfactory duality probability of 2.4e-5.

The optimized sets of SNPs show high Hamming distance and low duality in the simulated populations. According to the allele frequencies in HapMap and following the Hardy-Weinberg equilibrium, there are 100 thousand individuals generated as the simulated populations for four major ethnic groups, i.e., African (AFR), European (EUR), East-Asian (ASN), and American (AMR). In order to compare with the optimized SNP sets regarding the performance, for each of SNP numbers from 20 to 50, there are 1,000 random sets of SNPs generated. These random SNPs are chosen when their MAFs in 1000 Genomes Project and HapMap are between 0.4 and 0.5 and they are all located on the autosomal chromosomes. This reflects an intuitive plan which is supposed to reach suboptimal performance. According to the calculated Hamming distances in these simulated populations, the optimized SNP sets

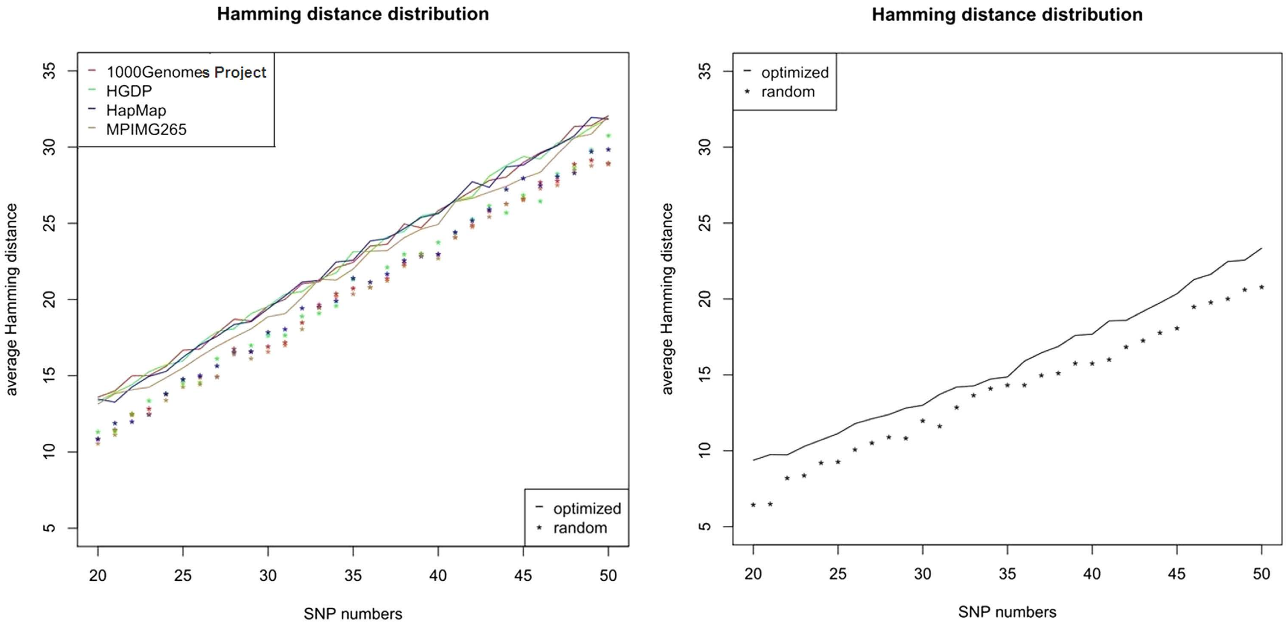

Figure 5. The Hamming distances are calculated based on the genotypes of 1000 Genomes Project, HGDP, HapMap, and MPIMG265 (left) and the relatives in MPIMG265 (right). The lines stand for the optimized SNP sets and the asterisks stand for the random SNP sets.

show superiority over the random SNP sets. For example, the average Hamming distance of an optimized set of given-number SNPs is larger than the random set by about 2. For example, an optimized set of 30 SNPs gives rise to the average Hamming distance of more than 18 in the simulation populations, while its random counterpart generates the average Hamming distance of around 16. And the duality frequencies of the optimized SNP sets are also much lower than the random ones, e.g., an optimized set of 30 SNPs can achieve very low duality frequency, which, even in the worst case of ASN, can be still less than 1 in 10 thousand (Fig. 4).

The optimized sets of SNPs show high Hamming distance in the real samples. According to both the optimized set of SNPs and the random sets of SNPs, the Hamming distances are generated for 1000 Genomes Project, HGDP, HapMap, and MPIMG265 (Fig. 5). The differentiating capability of the optimized set of SNPs are obviously better than the random ones, and even in the relatives (42 pairs of parents-offspring or siblings), the optimized set of SNPs can still give rise to satisfactory discrimination. As mentioned in the previous sections, MPIMG265 is mainly composed of samples from a population under-represented in the present databases. The optimized set of SNPs performs still well in MPIMG265, although MPIMG265 is not involved in constructing the entropy model. On the other hand, it is well known that parent-offspring and siblings share half of their genetic information, so that the reduction of discriminative capability of SNP sets will be unavoidable. However, even in these harsh conditions, the optimized set of SNPs seems still applicable (Fig. 5).

Software program for user-design SNP sets. A software program called SNP_Tagger has been developed for implementing the aforementioned algorithm. Users can download SNP_Tagger from https://sourceforge.net/projects/merap/files/SNP_Tagger20131220/. The software provides the 2,236 qualified SNPs as the starting menu, which have high and consistent information content in different datasets, and users can define the favorable SNP numbers and the gene list of interest. The default gene list includes all RefSeq genes on the autosomal chromosomes, assuming that most users will use WES. However, there are some users, especially clinicians, who may favor diagnostic panels including only part of the human gene repertoire, which thus can be defined as a curtailed gene list used by the program SNP_Tagger.

Discussion

Re-sequencing is entering the clinical research as a routine diagnostic method, while the large amount of samples processed will unavoidably bring about the problem of sample mix-up. The manual re-identification at the end of each re-sequencing project by Sanger sequencing is laborious and not cost-effective. Therefore, it is worthwhile to label samples before entering the diagnostic pipeline. The intuitive and cost-effective solution is tagging samples with a set of SNPs, which are abundant in the human genome and can distinguish individual genomes. However, it is the skewed allele frequency which makes the bulk of SNPs unsuitable as tagging markers, then how to define a set of competent SNPs in terms of differentiating capability, and consistence in different sample sources, is of major concern. Our motivation to design the algorithm is to create the set of SNPs with minimal numbers while with highest differentiating capacity. The information content, or entropy of SNP set, measures the variability of the SNP combinations, thus correlated with its competence in labeling individuals. To the limit of our knowledge, this is the first attempt by using information theory to select a set of qualified SNPs for sample-tagging in re-sequencing studies. We confirm the optimized SNP sets outperform the random ones, and also offer software by which the users can generate the SNP sets of their interest, so as to develop the specific diagnostic panels which harbor only a fraction of the human genes. In the same principle, this SNP-tagging plan also has potential application in forensic science, in order to match the SNP fingerprints of the concerned subjects.

In practice, the more SNPs are employed in tagging samples, the higher is the differentiating capability. The cost of SNP-tagging should be less than the re-sequencing cost times the sample mix-up rate, otherwise it will also make sense to re-identify the mixed samples by re-sequencing *per se*. This is the reason why we postulate that only 60 SNPs can label a population of 7 billion individuals, and we recommend using 30 SNPs to label the thousands of individuals in a typical re-sequencing project. The tests in our study also show that by using the optimized set of 30 SNPs, the Hamming distances in pairs are large enough to distinguish members of the same family. We hope that the implementation of this algorithm will result in an enhanced reliability and validity of re-sequencing projects, relieving the practitioners from the consequences of inadvertent sample mix-ups.

Methods

Calculation of information content of a single SNP. The Shannon-entropy, or information content, of a single SNP, is calculated in this work according to the variability of its three genotypes, i.e., homozygous wild-type, homozygous mutant, and heterozygous mutant. The information content of a single SNP X is given by:

$$H(X) = -\Sigma_{x\in\beth}\, p(x)\log_2 p(x) \qquad (9),$$

where $H(X)$ is the information contained in a single SNP X, and $\beth$ is the set of all three genotypes, i.e., $\beth =$ {homozygous wild-type, homozygous mutant, heterozygous mutant}, and $p(x)$ is the probability that the single SNP X has a specific genotype x, which belongs to $\beth$.

Obviously, the maximal information content of a single SNP is 1.585 bits, which is achieved when the three genotypes are equiprobable. For a SNP with MAF = 0.5, when it has reached the Hardy-Weinberg equilibrium in a population, its information content is 1.5 bits.

In order to evaluate the consistency of information content in different sample sources, the bootstrap, namely, random sampling with replacement from the original dataset, is employed[22]. Given the genotypes of a single SNP from two sources with sample sizes of n and m, respectively, the bootstrap procedures are:

1. Draw a re-sample of size n with replacement from the first source and a re-sample of size m with replacement from the second source. Compute the entropy difference (ΔH) between the first re-sample and the second re-sample, regarding the single SNP.
2. Repeat this re-sampling process 1,000 times.
3. Construct a bootstrap distribution of the 1,000 ΔH.
4. The expected value of ΔH is 0, assuming the genotype of a single SNP is identical in the two sources. A single SNP is regarded as consistent if the observed ΔH is located in the 95% bootstrap percentile confidence interval.

Furthermore, in order to prevent the unsatisfactory performance of SNP detection in a variety of platforms, two factors have been considered for refining SNPs, i.e., mappability and GC-content. The mappability of a SNP measures the possibility of short-read alignment ambiguity due to homologous sequences, and mappability of 1 indicates no alignment ambiguity[21]. In our application, only the SNPs with mappability of 1 are remained for further consideration. GC-content determines the bias of detection, thus, only the SNPs in the regions of GC-content from 0.35 to 0.55 are retained in our study (the GC content scores are downloaded from UCSC Genome browser GC Percent (gc5Base) track).

Calculation of mutual information of SNP pairs. Genotypes of any two SNPs in the human genome are in many cases not independent, because there might be linkage disequilibrium among their alleles, which thus reduces the cumulative discriminative capability of multiple SNPs. To measure the mutual dependency between SNP pairs regarding their genotypes, the mutual information (I) is used. The definition of mutual information between two SNPs X and Y is:

$$I(X;\, Y) = \Sigma_x \Sigma_y\, p(x,\, y)\log_2 \frac{p(x,\, y)}{p(x)p(y)} \qquad (10)$$

where $p\,(x,\,y)$ is the joint probability densities of X and Y regarding their three genotypes, and $p(x)$ $p(x)$ and $p(y)$ are the marginal probability densities of X and Y, respectively, regarding their three genotypes. The higher the mutual information between X and Y, the more redundance in the information sum of X and Y.

Estimation of combined information of multiple SNPs. The total information provided by N SNPs can be defined as the joint entropy $H\,(X_1,\,\cdots,\,X_N)$, where $X_k\,(k\,=\,1,\,\cdots,\,N)$ stands for the k-th SNP. Assuming that the joint entropy of the first $(N\,-\,1)$ SNPs is known, the joint entropy of the N SNPs should be:

$$H\,(X_1,\,\cdots,\,X_N)\,=\,H\,(X_1,\,\cdots,\,X_{N-1})\,+\,H\,(X_N)\,-\,I\,(X_1,\,\cdots,\,X_{N-1};\,X_N) \tag{11}$$

where $H\,(X_1,\,\cdots,\,X_N)$ is the joint entropy of the N SNPs, $H\,(X_1,\,\cdots,\,X_{N-1})$ is the joint entropy of the first $(N\,-\,1)$ SNPs, $H\,(X_N)$ is the individual entropy of the N-th SNP, $I\,(X_1,\,\cdots,\,X_{N-1};\,X_N)$ is the mutual information between the first $(N\,-\,1)$ SNPs $(X_1,\,\cdots,\,X_{N-1})$ and the$N$-th SNPX_N.

The exact value of $I\,(X_1,\,\cdots,\,X_{N-1};\,X_N)$ is difficult to calculate if N is large, but the upper bound and lower bound of the value can be estimated.

For the lower bound calculation, considering three SNPs, namely, X_1,$X_{2,}$$X_3$, and using the chain rule for mutual information[17], we have $I\,(X_1,\,X_2;\,X_3)\,=\,I\,(X_1;\,X_3)\,+\,I\,(X_2;\,X_3|X_1)\,\geq\,I\,(X_1;\,X_3)$, , since $I\,(X_2;\,X_3|X_1)\,\geq\,0$. Similarly we have $I\,(X_1;\,X_2;\,X_3)\,\geq\,I\,(X_2;\,X_3)$. Extending to the multivariate mutual information, we have $I\,(X_1,\,\cdots,\,X_{N-1};\,X_N)\,\geq\,\max_{i=1,\cdots N-1}\,I\,(X_i;\,X_N)$.

For the upper bound calculation, considering three SNPs, namely, X_1,$X_{2,}$$X_3$, and using the definition of mutual information, $I\,(X_1;\,X_2;\,X_3)\,=\,I\,(X_2;\,X_3)\,-\,I\,(X_2;\,X_3|X_1)$, since the mutual information $I\,(X_1;\,X_2;\,X_3)\,\geq\,0$, we have $I\,(X_2;\,X_3)\,\geq\,I\,(X_2;\,X_3|X_1)$. Substituting it into the chain rule for mutual information aforementioned, we have $I\,(X_1,\,X_2;\,X_3)\,\leq\,I\,(X_1;\,X_3)\,+\,I\,(X_2;\,X_3)$. Extending to multivariate mutual information, we have $I\,(X_1,\,\cdots,\,X_{N-1};\,X_N)\,\leq\,\sum_{i=1}^{N-1}I\,(X_i;\,X_N)$.

Therefore, the lower bound $(H_L)$and the upper bound $(H_U)$of the joint entropy of N SNPs can be defined as:

$$H_L(X_1,\,\cdots,\,X_N)\,=\,H\,(X_1,\,\cdots,\,X_{N-1})\,+\,H\,(X_N)\,-\,\sum_{i=1}^{N-1}I\,(X_i;\,X_N) \tag{12}$$

$$H_U(X_1,\,\cdots,\,X_N)\,=\,H\,(X_1,\,\cdots,\,X_{N-1})\,+\,H\,(X_N)\,-\,\max_{i=1,\cdots,N-1}\,I\,(X_i;\,X_N) \tag{13}$$

Optimal set of a given number of SNPs in terms of combined information. In order to select an optimized set of given-number SNPs for sample-tagging, we have to maximize the combined information. The lower bound of the joint entropy (H_L) of a set of SNPs is used to estimate the combined information in this work. Given the individual entropy of each SNP and the mutual information of all SNP pairs, the question then is, how to find a set of SNPs, which has the maximalH_L, or, has the highest discriminative capability. This problem is solved by an improved recursive algorithm. Suppose that there are N single SNPs as the candidates,

Step 1: when N= 1, take the 1,000 SNPs with the highest individual entropies;

Step 2: when$N$= 2, calculate the joint entropy $H\,(X_1,\,X_2)\,=\,H\,(X_1)\,+\,H\,(X_2)\,-\,I\,(X_1;\,X_2)$. Take the leading 1,000 SNP pairs which have the largest joint entropies.

Step 3: when$N$= 3, calculate the joint entropy

$$H\,(X_1,\,X_2,\,X_3)\,\approx\,H\,(X_1,\,X_2)\,+\,H\,(X_3)\,-\,\sum_{i=1}^{2}I\,(X_i;\,X_3) \tag{14}$$

Hereby, (X_1,X_2) are defined in the 1,000 SNP pairs from the previous step. Take the leading 1,000 3-SNP sets which have the largest joint entropies.

Step k: when N= k, calculate the joint entropy

$$H\,(X_1,\,\cdots,\,X_k)\,\approx\,H\,(X_1,\,\cdots,\,X_{k-1})\,+\,H\,(X_k)\,-\,\textstyle\sum_{i=1}^{k-1}I\,(X_i;\,X_k) \tag{15}$$

where $2\,\leq\,k\,\leq\,N$. Hereby, $(X_1,\,X_2,\,\cdots,\,X_{k-1})$ are defined in the 1,000 leading SNP sets from the previous step when N= k- 1. Take the leading 1,000k-SNP sets which have the largest joint entropies.

Evaluation of SNP sets by Hamming distance. The optimized SNP sets should differentiate individuals most efficiently. This can be evaluated by the mismatches between sample pairs regarding the SNP sets. In information theory, the Hamming distance between two strings of equal length is the number of positions where the corresponding symbols are different[23]. Therefore, the Hamming distance is used to evaluate the performance of SNP sets in differentiation, i.e., the larger the Hamming distance, the better the performance. If the Hamming distance between two individuals is zero, the two individuals are claimed duality and the SNP sets could not differentiate between them, which indicates that the SNP set fails incidentally.

References

1. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12,** 745–55 (2011).
2. Najmabadi, H. *et al.* Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature* **478,** 57–63 (2011).
3. Westra, H. J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* **27,** 2104–11 (2011).
4. Pengelly, R. J. *et al.* A SNP profiling panel for sample tracking in whole-exome sequencing studies. *Genome Med.* **5,** 89 (2013).
5. Xu, W. *et al.* Coding SNPs as intrinsic markers for sample tracking in large-scale transcriptome studies. *Biotechniques* **52,** 386–8 (2012).
6. Rosenberg, N. A., Li, L. M., Ward, R. & Pritchard, J. K. Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* **73,** 1402–22 (2003).
7. Stram, D. O. Tag S. N. P. selection for association studies. *Genet. Epidemiol.* **27,** 365–74 (2004).
8. Kidd, K. K. *et al.* Developing a SNP panel for forensic identification of individuals. *Forensic Sci. Int.* **164,** 20–32 (2006).
9. Pakstis, A. J., Speed, W. C., Kidd, J. R. & Kidd, K. K. Candidate SNPs for a universal individual identification panel. *Hum. Genet.* **121,** 305–17 (2007).
10. Qin, P. *et al.* A panel of ancestry informative markers to estimate and correct potential effects of population stratification in Han Chinese. *Eur. J. Hum. Genet.* **22,** 248–53 (2014).
11. Demichelis, F. *et al.* SNP panel identification assay (SPIA): a genetic-based assay for the identification of cell lines. *Nucleic Acids Res.* **36,** 2446–56 (2008).
12. Kim, J. J., Han, B. G., Lee, H. I., Yoo, H. W. & Lee, J. K. Development of SNP-based human identification system. *Int. J. Legal Med.* **124,** 125–31 (2010).
13. Meng, Z., Zaykin, D. V., Xu, C. F., Wagner, M. & Ehm, M. G. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am. J. Hum. Genet.* **73,** 115–30 (2003).
14. Zhang, K. *et al.* HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics* **21,** 131–4 (2005).
15. Sicotte, H., Rider, D. N., Poland, G. A., Dhiman, N. & Kocher, J. P. SNPPicker: high quality tag SNP selection across multiple populations. *BMC Bioinformatics* **12,** 129 (2011).
16. International HapMap 3 Consortium, *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–8 (2010).
17. Cover, T. M. & Thomas, J. A. *Elements of information theory*, xxiii, 748 p. (Wiley-Interscience, Hoboken, N.J., 2006).
18. 1000 Genomes Project Consortium, *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–73 (2010).
19. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298,** 2381–5 (2002).
20. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–9 (2012).
21. Hu, H. *et al.* Integrated sequence analysis pipeline provides one-stop solution for identifying disease-causing mutations. *Hum. Mutat.* **35,** 1427–35 (2014).
22. Moore, D. S., McCabe, G. P. & Craig, B. A. *Introduction to the practice of statistics*, (W.H. Freeman, New York, 2009).
23. MacKay, D. J. C. *Information theory, inference, and learning algorithms*, xii, 628 p. (Cambridge University Press, Cambridge, UK ; New York, 2003).

Acknowledgements

We are very grateful to Hossein Najmabadi and Kimia Kahrizi from Genetics Research Center, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran, for they provided the sample DNAs from the Middle East. This work was supported by the Max Planck Society and by the European Commission Framework Program 7 (FP7) project GENCODYS with grant number 241995.

Author Contributions

H.H., H.H.R. and T.F.W. initiated and coordinated this work. H.H. designed the scheme and collected the samples. L.X. and H.H. constructed and tested the mathematical models. J.W. provided the plan of assessing the model robustness in the different datasets. H.H., L.X. and T.F.W. wrote the manuscript.

Additional Information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Hu, H. et al. Evaluating information content of SNPs for sample tagging in re-sequencing projects. *Sci. Rep.* **5**, 10247; doi: 10.1038/srep10247 (2015).