Supplementary Methods

Data Analysis

Intensity Pre-Processing

AB-ratio adjustment Affinity differences between allele ‘A’ and allele ‘B’ are estimated for each probe to allow a direct comparison of two DNA samples with different genotypes for the given SNP. The PM and MM intensities obtained from each probe are denoted as $P_{i,k}^{\mathrm{A}_l}, P_{i,k}^{\mathrm{B}_l}, M_{i,k}^{\mathrm{A}_l}$ and $M_{i,k}^{\mathrm{B}_l}, i = 1, ..., N_i, k = 1, ..., N_k,$ and $l = 1, ..., N_l$ with i representing the different samples, k representing different probe set numbers, l representing the probe pair number, and $A(B)$ representing the A(B) allele. A Gaussian mixture clustering algorithm was used to extract samples with exactly one ‘A’ allele and one ‘B’ allele for each SNP in the 270 HapMap samples.

$P_{i,k}^{'A_l}$ and $P_{i,k}^{'B_l}$ are normalized PM signals for allele A and B, respectively:

$$P_{i,k}^{'A_l} = \frac{P_{i,k}^{A_l}}{P_i^{\mathrm{autosome}}} \qquad (1)$$

$$P_{i,k}^{'B_l} = \frac{P_{i,k}^{B_l}}{P_i^{\mathrm{autosome}}} \qquad (2)$$

$$P_i^{\mathrm{autosome}} = \sum_{l=1}^{N_l} \sum_{k \in S_{autosome}} (P_{i,k}^{A_l} + P_{i,k}^{B_l}) \qquad (3)$$

P_i^{autosome} is the sum of PM signals in autosomal regions for i-th sample used to adjust the PM signals among different samples.

$x_{i,k}$ is a two dimensional vector defined as follows:

$$x_{i,k} = \left(\frac{1}{N_l}\sum_{l=1}^{N_l} \ln P_{i,k}^{'A_l}, \frac{1}{N_l}\sum_{l=1}^{N_l} \ln P_{i,k}^{'B_l}\right) \qquad (4)$$

The normalized PM signals of such samples are assumed to follow a Gaussian distribution.

$$p(X_k) = \sum_{m=1}^{N_{k,m}} w_{k,m} N(\mu_{k,m}, \sum\nolimits_{k,m}) \qquad (5)$$

$$\sum_{m=1}^{N_{k,m}} w_{k,m} = 1,\ \ w_{k,m} > 0 \qquad (6)$$

where *k* and *m* represent the different probe sets and the different clusters which have different numbers of 'A' alleles and 'B' alleles, respectively. The parameters of the Gaussian mixture model ($w_{k,m}$, $\mu_{k,m}$, $\Sigma_{k,m}$) were estimated with a fixed number of components using the EM algorithm (Dempster et al. 1977) while the Bayesian Information Criterion (BIC) (Schwarz 1978) was used to determine the number of clusters. BIC is used in several different stages of the algorithm. It is applicable to problems where the fitting is achieved by maximization of a log-likelihood. A heavy penalty proportional to the sample size is imposed with increased complexity of the model, encouraging a simple model that efficiently captures the intensity variation pattern defined by the sequence factors. The K-means algorithm was used to initialize the center of each cluster, and the cluster membership was determined based on the class posterior probabilities using the Bayes rule. After parameter estimation, the "AB" cluster was defined as the one with the minimum average signals that satisfies

$$m_{AB} = \arg\min_{m} \| \mu_{k,m} \| \qquad (7)$$

Finally, the signal ratio of allele A to B of probe *k* is

$$r_k^l = \operatorname*{median}_{i \in S_{\mathrm{AB}}} \frac{P_{i,k}^{A_l}}{P_{i,k}^{B_l}} \qquad (8)$$

where S_{AB} is the set of samples which belong to the cluster m_{AB}. The ratios are used to modify the intensity difference between allele 'A' and allele 'B'. When the number of samples with 'AB' genotype was too small (< 20) to identify the clusters correctly for a given SNP, the median of $P_{i,k}^{A_l}/P_{i,k}^{B_l}$ was used for all samples as r_k^l instead.

The modified signal ratio between two samples was calculated by using the affinity ratios. Let $g_{i,k} \in \{A, B, AB\}$ denote the genotype of probe k in sample i. Genotype A contains at least one A allele but does not contain a B allele, genotype AB contains at least one A allele and at least one B allele, and Genotype *B* contains at least one B allele but does not contain an A allele. Signal ratio between sample i and j for probe set k of the l-th quartet is calculated as follows:

$$r_{i,j,k}^{l} = \begin{cases} \dfrac{P_{i,k}^{\mathrm{A}_l}}{P_{j,k}^{\mathrm{A}_l}} & (g_{i,k}, g_{j,k}) = (A, A) \\ \dfrac{P_{i,k}^{\mathrm{A}_l}}{\gamma_k^l P_{j,k}^{\mathrm{B}_l}} & (g_{i,k}, g_{j,k}) = (A, B) \\ \dfrac{P_{i,k}^{\mathrm{A}_l}}{(P_{j,k}^{\mathrm{A}_l} + \gamma_k^l P_{j,k}^{\mathrm{B}_l})} & (g_{i,k}, g_{j,k}) = (A, AB) \\ \dfrac{(P_{i,k}^{\mathrm{A}_l} + \gamma_k^l P_{i,k}^{\mathrm{B}_l})}{P_{j,k}^{\mathrm{A}_l}} & (g_{i,k}, g_{j,k}) = (AB, A) \\ \dfrac{(P_{i,k}^{\mathrm{A}_l} + \gamma_k^l P_{i,k}^{\mathrm{B}_l})}{\gamma_k^l P_{j,k}^{\mathrm{B}_l}} & (g_{i,k}, g_{j,k}) = (AB, B) \\ \dfrac{(P_{i,k}^{\mathrm{A}_l} + \gamma_k^l P_{i,k}^{\mathrm{B}_l})}{(P_{j,k}^{\mathrm{A}_l} + \gamma_k^l P_{j,k}^{\mathrm{B}_l})} & (g_{i,k}, g_{j,k}) = (AB, AB) \\ \dfrac{\gamma_k^l P_{i,k}^{\mathrm{B}_l}}{P_{j,k}^{\mathrm{A}_l}} & (g_{i,k}, g_{j,k}) = (B, A) \\ \dfrac{P_{i,k}^{\mathrm{B}_l}}{P_{j,k}^{\mathrm{B}_l}} & (g_{i,k}, g_{j,k}) = (B, B) \\ \dfrac{\gamma_k^l P_{i,k}^{\mathrm{B}_l}}{(P_{j,k}^{\mathrm{A}_l} + \gamma_k^l P_{j,k}^{\mathrm{B}_l})} & (g_{i,k}, g_{j,k}) = (B, AB) \end{cases} \qquad (9)$$

Noise reduction and normalization Signal variation due to properties of the probe and restriction fragment sequences was estimated and removed by the GIM algorithm which has been described in detail previously (Komura et al. 2006). For this study, GIM was improved in several aspects to be more suitable for CNV detection. First, we have found that signal intensity variation often correlates with the long range GC content surrounding each SNP. Thus, the model now takes into account the GC percentage of 40kbp of sequence surrounding each SNP. Second, robust regression is now applied and robust Bayesian Information Criterion (robust BIC) (Qian and Kunsch 1996) is used to determine the optimal degree of polynomials in place of least-square regression and BIC (Schwarz 1978) which were used in the original algorithm. To reduce computational time, we estimated the optimal degree of polynomials for 100 randomly selected sample pairs beforehand and used the mode of each parameter for the analysis. Bi-square estimators were selected for the robust regression, which minimizes the objective function

$$\sum_{i=1}^{n} \rho(y_i - x_i b) \tag{10}$$

$$\rho(e) = \begin{cases} \frac{k^2}{6}\left\{1-[1-(\frac{e}{k})2]^3\right\} & |e| \leq k \\ \frac{k^2}{6} & |e| > k \end{cases} \tag{11}$$

We set k to 4.685σ (where σ is the standard deviation of the errors), which produces 95% efficiency when the errors are normal, and offers protection against outliers. In this analysis, the first round of crude copy number estimation used in the original version of GIM is omitted since the CNV regions are usually small and have little effect on the whole distribution of signal ratios; in addition, the robust regression that has been adopted protects against any such random perturbation. The algorithm was applied separately to each sample pair, each array (Nsp I and Sty I), each genotype combination and each restriction enzyme recognition site.

Scaling was carried out to make different experiments more comparable and to remove undesired bias derived from large copy number changes in some chromosomes. The median ratio was scaled to unity by dividing all the ratios with the median ratio of all autosomal probes, but leaving out the probes from the three chromosomes with the highest MAD (median absolute deviation). Here let r_i be a signal ratio of i-th probe, C_j be a set of probes on j-th chromosome and C_{autosome} be a set of probes on all autosomes. MAD for j-th autosome was calculated as follows:

$$\text{MAD}_j = \underset{i \in C_j}{\text{median}} \left| \log_2 r_i - \underset{k \in C_{\text{autosome}}}{\text{median}} \log_2 r_k \right|$$

Since chromosomes with high MAD may have large CNV regions, they were removed for scaling. Scaling was carried out as follows:

$$\log_2 r_i' = \log_2 r_i - \underset{k \in C_{\text{lowMAD}}}{\text{median}} \log_2 r_k$$

where C_{lowMAD} is the set of autosomal probes excluding the three chromosomes with the highest MAD and r_i' is a scaled signal ratio of i-th probe.

CNV Detection

Pair-wise comparisons. After signal ratios from Nsp I and Sty I arrays from the same sample were merged, SW-ARRAY was used to detect copy number changes in pair-wise comparisons. As an initial step, probes with signal ratios >1.4 (or <1/1.4) were reset to 1.4 (or 1/1.4) in order to reduce the effect of any outliers with extreme values. Signal ratios were converted back to their original values for subsequent analysis. Next, the background distribution was calculated. For this step, in order to avoid the possible effect of large CNVs in a chromosome reducing the overall detection sensitivity, probes were selected from the chromosomes with the lowest MAD. These probes were added iteratively until the number of probes was sufficiently large (> 39,189 in this case). The permutated signal ratios using this set of probes was then used to estimate the background distribution.

Detection of Homozygous Deletions Genotype calls in homozygous deletion regions are often 'no calls' which cannot be used by the algorithm proposed here. Since this would lead to significantly reduced resolution, homozygous deletions were detected separately using another algorithm that relies on the discrimination score. The discrimination score for each allele is defined as

$$D_{i,k}^{\mathrm{A}} = \sum_{l=1}^{N_l} \frac{P_{i,k}^{\mathrm{A}_l} - \frac{1}{N_l}(\sum_{l=1}^{N_l} M_{i,k}^{\mathrm{A}_l} + M_{i,k}^{\mathrm{B}_l})}{P_{i,k}^{\mathrm{A}_l} + \frac{1}{N_l}(\sum_{l=1}^{N_l} M_{i,k}^{\mathrm{A}_l} + M_{i,k}^{\mathrm{B}_l})} \qquad (12)$$

$$D_{i,k}^{\mathrm{B}} = \sum_{l=1}^{N_l} \frac{P_{i,k}^{\mathrm{B}_l} - \frac{1}{N_l}(\sum_{l=1}^{N_l} M_{i,k}^{\mathrm{A}_l} + M_{i,k}^{\mathrm{B}_l})}{P_{i,k}^{\mathrm{B}_l} + \frac{1}{N_l}(\sum_{l=1}^{N_l} M_{i,k}^{\mathrm{A}_l} + M_{i,k}^{\mathrm{B}_l})}, \qquad (13)$$

where *i* and *k* represent the different samples and the different probe sets, respectively. Consecutive probes with low discrimination scores are detected by applying SW-ARRAY to a series of $\max(D_{i,k}^{\mathrm{A}}, D_{i,k}^{\mathrm{B}})$ with a cut-off threshold of 0.1929 and 100,000 permutations. All probes are used irrespective of their genotypes. The signal distribution of homozygous deletions was simulated by generating artificial deletions through the digestion of sample DNA with the restriction enzyme Xba I prior to using the standard

500K EA assay. 2ug of genomic DNA was digested for 16 hours with Xba I (NEB) and then de-phosphorylated with shrimp alkaline phosphatase (Sigma Aldrich) to prevent re-ligation at later steps. 250ng of the purified genomic DNA was then used for the standard Nsp I WGSA protocol. Intensity signals from probes predicted to reside on Nsp I restriction fragments that contain internal Xba I restriction sites were used to train the algorithm for the detection of homozygous deletions.

Preliminary CNV extraction from multiple samples CNV regions are inferred based on summarization of all pair-wise comparisons. The ‘CNV density’ is defined as the fraction of pair-wise comparisons that show the target region as significant between the test sample and the reference set:

$$d_{i,k} = \frac{\sum_{j \in N} H_0(p_{\mathrm{th}} - p_{i,j,k})}{N} \quad (14)$$

where N is the total number of reference samples (269 in the HapMap analysis), p_{th} is set to 0.01, $p_{i,j,k}$ is a p-value of probe k calculated by SW-ARRAY between sample i and j, and $H_0(\cdot)$ is a Heaviside step function:

$$H_0(x) = \begin{cases} 0 & (x < 0) \\ 1 & (x \geq 0) \end{cases} \quad (15)$$

The CNV confidence score of probe k is defined as:

$$s_k = \max_{1 \leq n \leq \left\lfloor \frac{N}{2} \right\rfloor} \left(\sum_{i \in N} H_0 \left(d_{i,k} - \alpha \frac{N - n + 1}{N} \right) - n \right) \quad (16)$$

Here n is the size of the subset that shared the same altered copy number and is tested between 1 to $\frac{N}{2}$; α is the cut-off rate of CNVs successfully detected in all single pair-wise comparisons and accounts for the occasional false-positives and false-negatives. $s_k > 0$ indicates probe k resides inside a CNV region in any of the samples. CNV regions are extracted based on the CNV confidence score. The m-th CNV region is represented by $T_m = \{k_m \mid l_m \leq k_m \leq r_m, k_m \in N_k\}$ which satisfies

$$s_{k_m} \geq 0, \quad (17)$$

$$s_{l_m-1} < 0, \quad (18)$$

$$s_{r_m+1} < 0, \quad (19)$$

$$l_1 < r_1 <, \cdots, l_m < r_m, \cdots, < l_M < r_M \quad (20)$$

$$\bigcup_m T_m = \{k \mid s_k \geq 0\} \quad (21)$$

where M denotes the total number of candidate CNV regions. CNV regions that span centromeres were divided into two regions. Subsequent analysis is done separately for each CNV region.

Copy Number Inference

Identification of diploid samples The diploid group at any given region is initially defined under the assumption that it is the largest group with the same copy number. A graph-theory-based method is applied to find these groups. For a given CNV region that contains probes from l to r, an undirected graph $G=(V,E)$ is constructed where each node $v \in V$ represents each sample and each edge $e_{i,j} \in E$ between nodes v_i and v_j indicates that the copy number between the two samples is the same throughout the candidate CNV region. In other words, an edge connects node v_i to v_j if

$$\max_{l \leq k \leq r} H_0(p_{\mathrm{th}} - p_{i,j,k}) = 0. \quad (22)$$

This transforms the problem of finding the diploid group to finding the maximum clique in the graph, which is a well-known NP-complete problem and requires unrealistic computation time. As an alternative, a heuristic approach was developed that considers each node to be a clique of size one, and then merges cliques into larger cliques until there are no more possible merges. This requires only linear computational time and finds at least one maximal clique not contained in any larger clique; this local optimization satisfies the goal of finding as many diploid samples as possible.

The selection of diploid samples becomes more challenging when the candidate CNV region is present in high frequency in the reference population and has a complex genomic structure. In such cases, the maximum clique may be the one copy group, which should show loss of heterozygosity, or the three copy group, which should show A/B ratio's significantly different from one in the probes with an 'AB' genotype. Therefore, we use the genotype information to redefine the diploid groups in such complex regions to ensure that true diploid samples, which may not be the most frequent group anymore, are accurately identified. Therefore if the hetero SNP rate is < 0.05 for the assigned diploid group and the number of the samples classified as > 3 copies is more than 10%, or when the absolute log2 AB ratio of the assigned diploid group is > log2(1.197) and the number of the samples classified as deletions was more than 10%, we re-selected the diploid samples. Re-selection was carried by max clique algorithm after removing the diploid set in the previous iteration.

Boundary assignment After the diploid group is defined, diploid densities were calculated again as:

$$d_{i,k}^{\mathrm{dip}} = \frac{\sum_{j \in S_{\mathrm{dip}}} H_0(p_{\mathrm{th}} - p_{i,j,k})}{N_{\mathrm{dip}}} \qquad (23)$$

where S_{dip} and N_{dip} are the diploid group and the size of the group respectively. P_{th} is set to 0.01. As in the previous section, CNV regions $T_{i,m}$ of the sample i are extracted:

$$d_{i,k_m}^{\mathrm{dip}} \geq \alpha, \qquad (24)$$

$$d_{i,l_m-1}^{\mathrm{dip}} < \alpha, \qquad (25)$$

$$d_{i,r_m+1}^{\mathrm{dip}} < \alpha, \qquad (26)$$

$$l_{i,1} < r_{i,1} <, \cdots, < l_{i,m} < r_{i,m} <, \cdots, < l_{i,M} < r_{i,M} \qquad (27)$$

$$\bigcup_m T_{i,m} = \{k_m \mid d_{i,k_m}^{\mathrm{dip}} \geq \alpha\} \qquad (28)$$

The density is incorporated to infer the boundary of each CNV region for each sample. In addition, we assigned β% leftmost and β% rightmost boundary for each CNV region. The center of m-th CNV region for sample i is defined as

$$c_{i,m} = \operatorname*{argmax}_{k_m} d_{i,k_m}^{\mathrm{dip}}. \qquad (29)$$

The maximum value of the diploid density $\max d_{i,k_m}^{\mathrm{dip}}$ reflects the confidence of the CNV call.

The $\beta\%$ leftmost (l^{β}_{m}) and rightmost (r^{β}_{m}) boundary of m-th CNV region is defined as satisfying the following condition:

$$d^{\text{dip}}_{i,p^{\beta}_{m}} \ge \frac{\beta}{100} d^{\text{dip}}_{i,c_{i,m}} \tag{30}$$

$$d^{\text{dip}}_{i,l^{\beta}_{m}-1} < \frac{\beta}{100} d^{\text{dip}}_{i,c_{i,m}} \tag{31}$$

$$d^{\text{dip}}_{i,r^{\beta}_{m}-1} < \frac{\beta}{100} d^{\text{dip}}_{i,c_{i,m}} \tag{32}$$

$$p^{\beta}_{m} = l^{\beta}_{m},\cdots,r^{\beta}_{m} \tag{33}$$

For cell line artifacts that involved sub-chromosomal changes, the boundary was determined by maximizing the difference of the average log2 signal ratio between the test sample region and the corresponding reference set region.

Copy number estimation Absolute copy number of a CNV in a sample is inferred based on the representative signal ratio for each CNV region. Here the median of the ratios to the diploid samples are taken as the representative ratio:

$$R_{i,m} = \operatorname*{median}_{k\in p^{90}_{m}}\left(\operatorname*{median}_{j\in\left(S^{\text{dip}}_{m} \cap S^{\text{sg}}_{i,m,k}\right)} r_{i,j,k} \right) \tag{34}$$

where $r_{i,j,k}$ is a summarized signal ratio of probe *k* between sample *i* and *j* as calculated in (9), S^{dip}_{m} and $S^{sg}_{i,m,k}$ are the samples classified as diploid by the maximum clique algorithm and the samples with the same genotype as the i-th sample. Here the 90% boundaries were used as a conservative delineation of the region to minimize the effect of boundary estimation errors. The signal ratio between the probes with only the same genotype was used to avoid the effect of AB ratio estimation errors. Copy number of the region $C_{i,m}$ is defined as:

$$C_{i,m} = \begin{cases} 0 & R_{i,m} < 0.476 \\ 0.5 & 0.476 \le R_{i,m} < 0.558 \\ 1 & 0.558 \le R_{i,m} < 0.754 \\ 1.5 & 0.754 \le R_{i,m} < 0.868 \\ 2 & 0.868 \le R_{i,m} < 1.094 \\ 2.5 & 1.094 \le R_{i,m} < 1.197 \\ 3 & 1.197 \le R_{i,m} < 1.360 \\ 3.5 & 1.360 \le R_{i,m} < 1.424 \\ \ge 4 & 1.424 \le R_{i,m} \end{cases} \qquad (35)$$

These thresholds were determined based on the ratio of 2X DNA samples to 1X through 5X DNA samples (data not shown).

References

Dempster, A.P., N.M. Laird, and D.B. Rublin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. of Royal Statistical Society* **39:** 1-38.

Komura, D., K. Nishimura, S. Ishikawa, B. Panda, J. huang, H. Nakamura, S. Ihara, M. Hirose, K.W. Jones, and H. Aburatani. 2006. Noise reduction from genotyping microarrays using probe level information. *In Silico Biology* **6**.

Qian, G. and R.H. Kunsch. 1996. On model selection in robust linear regression. *Technical Report 80, Seminar Fur Statistik, Eidgenossische Technische Hochschule (ETH), Zurich, Switzerland*.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* **6:** 461-464.