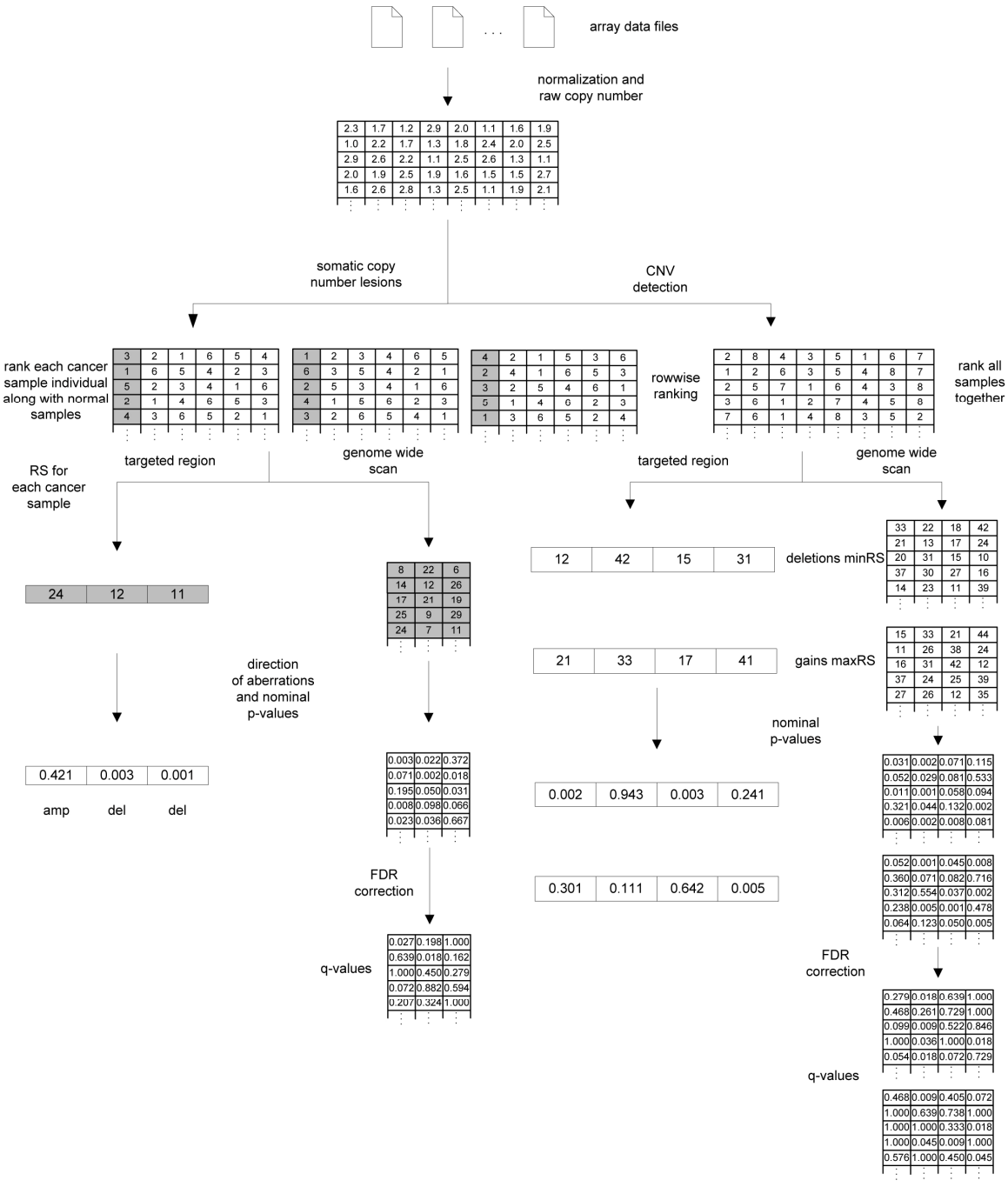
Supplementary Methods

General workflow for rank-based procedure: Beginning with the raw intensity files (top), normalization and raw copy number extraction are performed using platform-appropriate methods (dChip [1], PLASQ [2], ITALICS [3], CRMA [4], Bead Studio [5], CGHpro [6], etc.). The raw copy numbers are then ranked across samples, marker by marker. In the somatic setting (left side), each test sample (gray columns) is ranked separately along with the reference ("normal") samples, and each test rank sum is computed. A genome-wide scan for copy number lesions will produce a large matrix of rank sum statistics *RS* from the sliding window, whose *P*-values need to be corrected for multiple tests. If a specific region is targeted (lower left), nominal *P*-values may be used. Significantly large values for *RS* imply amplification, while small values indicate deletion. In the CNV detection setting (right side) all samples are typically test samples, and are therefore ranked together. The min *RS* and max *RS* statistics are computed recursively (see text), and nominal (for a targeted genomic region) or FDR-corrected *P*-values (for a genome-wide sliding window) are used to measure statistical significance of each putative deletion (using min *RS*) and gain (using max *RS*).



Derivation of Null Distributions

Here we derive the large-sample null distribution of the statistics RS, $\min RS = \min_n R_n$ and $\max RS = \max_n R_n$.

Proposition: For large N, the approximate null distributions of RS, $\min_n R_n$ and $\max_n R_n$ are given by

$$\begin{aligned}
P(RS < x) &= \frac{1}{2(K-1)!}\Big(\sum_{j=0}^{\lfloor y \rfloor}((-1)^j \binom{K}{j} \frac{(K-j)^K - (y-j)^K}{K}) \\
&\quad + \sum_{j=0}^{\lfloor y \rfloor+1}((-1)^j \binom{K}{j} \frac{(K-j)^K + (y-j)^K}{K})\Big) \\
P(\min_i R_i < x) &= 1 - \Big[\frac{1}{2(K-1)!}\Big(\sum_{j=0}^{\lfloor y \rfloor}((-1)^j \binom{K}{j} \frac{(K-j)^K - (y-j)^K}{K}) \\
&\quad + \sum_{j=0}^{\lfloor y \rfloor+1}((-1)^j \binom{K}{j} \frac{(K-j)^K + (y-j)^K}{K})\Big)\Big]^N \\
P(\max_i R_i < x) &= \Big[1 - \frac{1}{2(K-1)!}\Big(\sum_{j=0}^{\lfloor y \rfloor}((-1)^j \binom{K}{j} \frac{(K-j)^K - (y-j)^K}{K}) \\
&\quad + \sum_{j=0}^{\lfloor y \rfloor+1}((-1)^j \binom{K}{j} \frac{(K-j)^K + (y-j)^K}{K})\Big)\Big]^N
\end{aligned}$$

where $y = \frac{(x-K)}{N-1}$ and $\lfloor \rfloor$ denotes the "floor" (largest lower integer) function.

Proof: First, note that

$$P(\min_n R_n < x) = 1 - P(\min_n R_n \geq x) = 1 - P(R_n \geq x \text{ for all } n).$$

It is straightforward to show that the rank sums R_n are asymptotically independent under the null hypothesis. Therefore,

$$P(R_n \geq x \text{ for all } n) \approx [P(R_n \geq x)]^N. \tag{1}$$

Under the null hypothesis, the ranks r_{kn} are randomly generated integer between 1 and N for each n. For large N, therefore, the $\frac{r_{kn}-1}{N-1}$ are distributed approximately as uniform $U(0,1)$ random variables. Summing over the K loci, it follows that the transformed rank sum $Y_n = \frac{R_n - K}{N-1}$ is asymptotically distributed as a sum of K independent $U(0,1)$ random variables. This probability function of this is the well-known Irwin-Hall distribution [7, 8]:

$$P_Y(y) = \frac{1}{2(K-1)!}\sum_{j=0}^{K}(-1)^j \binom{K}{j}(y-j)^{K-1}\text{sgn}(y-j).$$

We may integrate to find

$$\begin{aligned}
P_Y(y \geq y_0) &= \int_{y_0}^{K} \frac{1}{2(K-1)!} \sum_{j=0}^{K} (-1)^j \binom{K}{j} (y-j)^{K-1} \mathrm{sgn}(y-j)\, dy \\
&= \frac{1}{2(K-1)!} \sum_{j=0}^{K} (-1)^j \binom{K}{j} \int_{y_0}^{K} (y-j)^{K-1} \mathrm{sgn}(y-j)\, dy \\
&= \frac{1}{2(K-1)!} \sum_{j=0}^{\lfloor y_0 \rfloor} (-1)^j \binom{K}{j} \int_{y_0}^{K} (y-j)^{K-1} \mathrm{sgn}(y-j)\, dy \\
&\quad + \frac{1}{2(K-1)!} \sum_{j=\lfloor y_0 \rfloor + 1}^{K} (-1)^j \binom{K}{j} \int_{y_0}^{K} (y-j)^{K-1} \mathrm{sgn}(y-j)\, dy \\
&= \frac{1}{2(K-1)!} \sum_{j=0}^{\lfloor y_0 \rfloor} (-1)^j \binom{K}{j} \frac{(y-j)^K}{K} \Big|_{y_0}^{K} \\
&\quad + \frac{1}{2(K-1)!} \sum_{j=\lfloor y_0 \rfloor + 1}^{K} (-1)^j \binom{K}{j} \left(\frac{(y-j)^K}{K} \Big|_{j}^{K} - \frac{(y-j)^K}{K} \Big|_{y_0}^{j} \right) \\
&= \frac{1}{2(K-1)!} \Big(\sum_{j=0}^{\lfloor y_0 \rfloor} \left((-1)^j \binom{K}{j} \frac{(K-j)^K - (y_0-j)^K}{K} \right) \\
&\quad + \sum_{j=0}^{\lfloor y_0 \rfloor + 1} \left((-1)^j \binom{K}{j} \frac{(K-j)^K + (y_0-j)^K}{K} \right) \Big).
\end{aligned}$$

Making the substitution $Y_n = \frac{R_n - K}{N-1}$ yields the result for RS. Plugging the resulting distribution function into expression (1) yields the result for $\min RS$. The $\max RS$ case is completely analogous. This completes the proof.

References

[1] Lin,M. et al. (2004) dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics, 20, 1233–40.

[2] LaFramboise,T. et al. (2007) PLASQ: a generalized linear model-based procedure to determine allelic dosage in cancer cells from SNP array data. Biostatistics, 8, 323–36.

[3] Rigaill,G. et al. (2008) ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays. Bioinformatics, 24, 768–74.

[4] Bengtsson,H. et al. (2008) Estimation and assessment of raw copy numbers at the single locus level. Bioinformatics, 24, 759–67.

[5] Illumina (2006) Sentrix HumanHap550 genotyping BeadChip data sheet. San Diego (California).

[6] Chen,W. et al. (2005) CGHPRO – a comprehensive data analysis tool for array CGH. BMC Bioinformatics, 6, 85.

[7] Hall, P. (1927) The Distribution of Means for Samples of Size N Drawn from a Population in which the Variate Takes Values Between 0 and 1, All Such Values Being Equally Probable. Biometrika 19(3/4):240-245.

[8] Irwin, J.O. (1927) On the Frequency Distribution of the Means of Samples from a Population Having any Law of Frequency with Finite Moments, with Special Reference to Pearson's Type II. Biometrika 19(3/4):225-239.