nature | **methods**

High-resolution mapping of copy-number alterations with massively parallel sequencing

Derek Y Chiang, Gad Getz, David B. Jaffe, Michael J T O'Kelly, Xiaojun Zhao, Scott L Carter, Carsten Russ, Chad Nusbaum, Matthew Meyerson & Eric S Lander

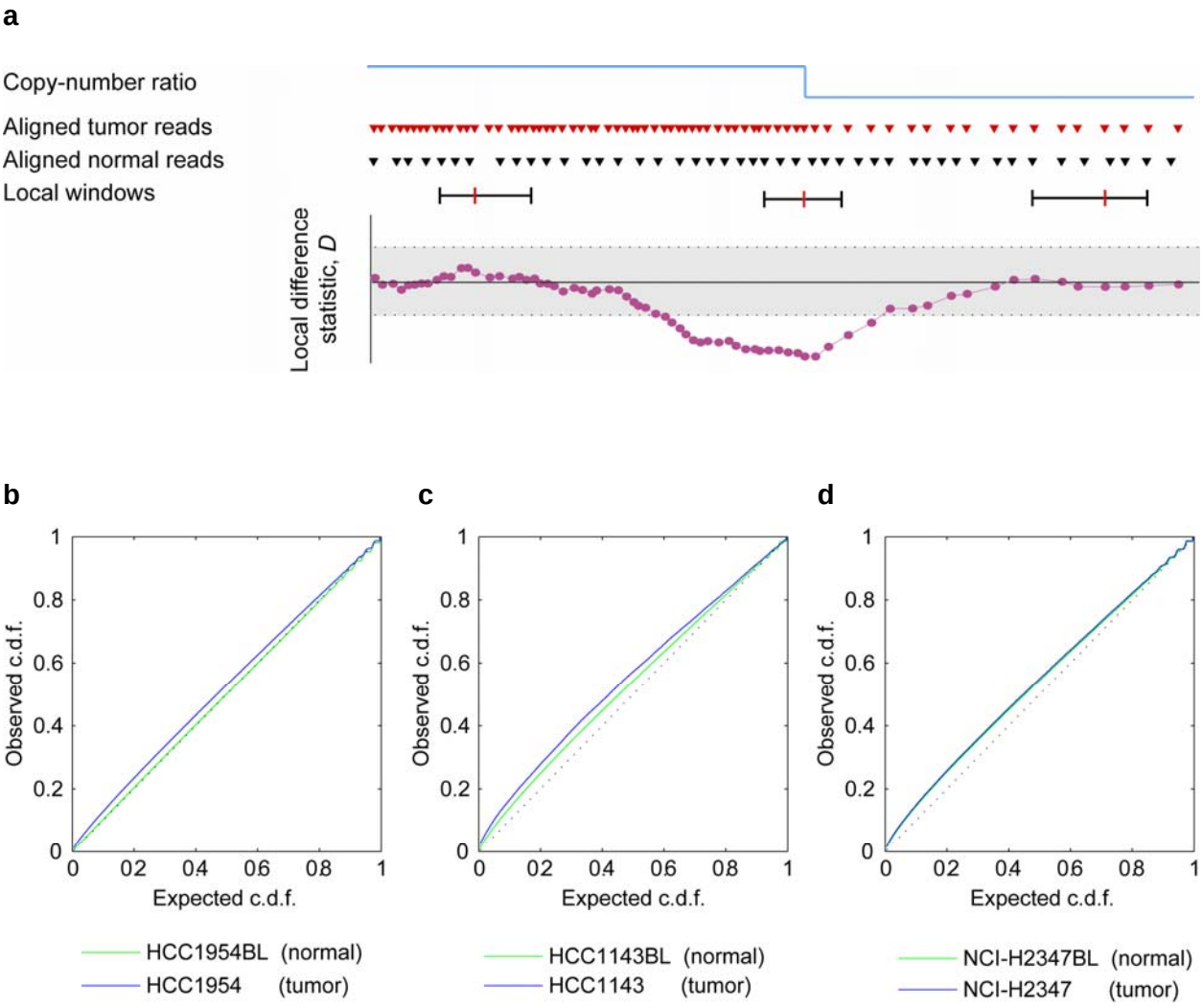
Supplementary figures and text:

Supplementary Fig. 1	Validity of *p*-values for the log-ratio difference statistic, *D*
Supplementary Fig. 2	Distribution of counts for aligned sequence reads in normal cell lines
Supplementary Fig. 3	G+C bias on counts of aligned sequence reads
Supplementary Fig. 4	Detailed view of segmentation results for HCC1954
Supplementary Fig. 5	Detailed view of segmentation results for HCC1143
Supplementary Fig. 6	Detailed view of segmentation results for NCI-H2347
Supplementary Fig. 7	Predicted homozygous deletions in HCC1143 and NCI-H2347 cell lines
Supplementary Fig. 8	Comparisons of copy-number alterations between sequencing and microarrays
Supplementary Fig. 9	*ERBB2* amplification in HCC1954
Supplementary Fig. 10	Mapped breakpoints of the *UTRN* homozygous deletion
Supplementary Fig. 11	Mapped breakpoints of the *PTPRD* homozygous deletion
Supplementary Fig. 12	Mapped breakpoints of the *HS3ST3A1* homozygous deletion
Supplementary Fig. 13	Approximations for the distribution of copy-number ratios, *R*
Supplementary Fig. 14	Parameter optimization for Circular Binary Segmentation of SNP 6.0 arrays
Supplementary Table 1	Statistics of whole genome shotgun sequencing
Supplementary Table 2	Effect of G+C bias on number of aligned reads
Supplementary Table 3	Candidate homozygous deletions detected in cell lines
Supplementary Methods	

Note: Supplementary Data is available on the Nature Methods website.

Supplementary Figure 1

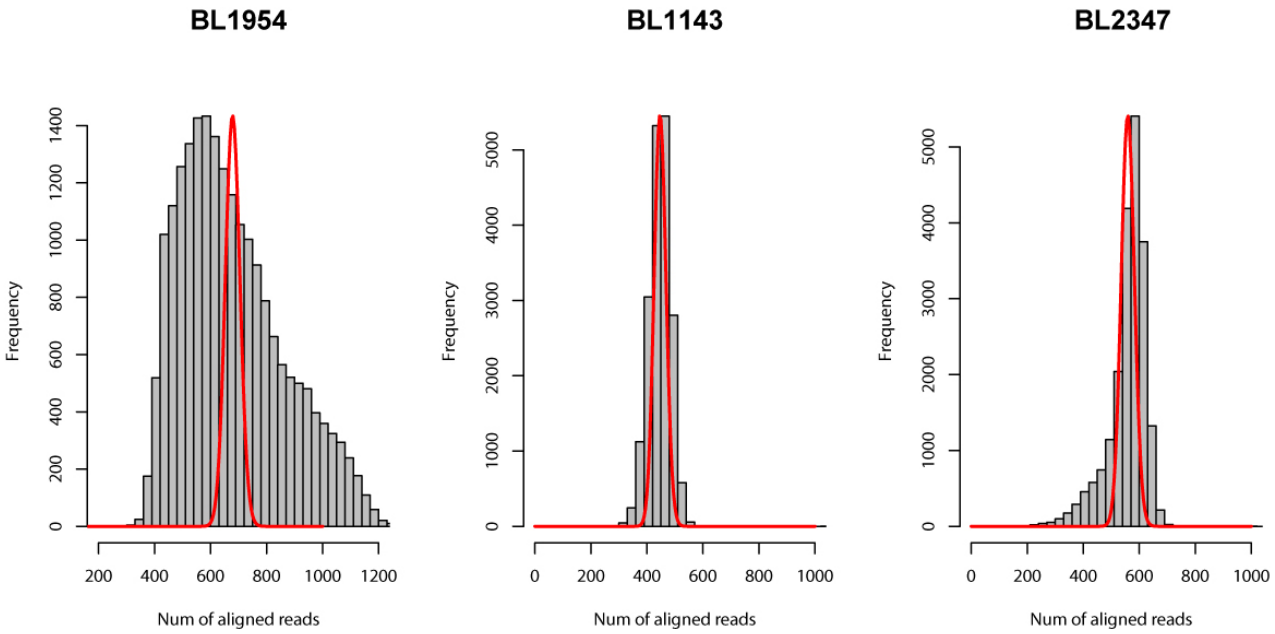
Validity of *p*-values for the log-ratio difference statistic, *D*



(**a**) Calculation of local difference statistic. The sizes of local windows flanking each position (vertical red line) are defined by a fixed number of uniquely aligning reads in the matched normal. Peaks in the local difference statistic that exceed the significance threshold (dotted horizontal lines) correspond to breakpoints of copy-number alterations. (**b-d**) Empirical null distribution for *D* for the three pairs of tumor cell lines and matched normals. We calculated the *p*-values for the empirical null distribution of *D* based on a log-normal approximation (see Supplementary Methods). The dashed black line indicates the cumulative fraction for a uniform distribution. Although our approximation inflates the significance of small *p*-values, similar skews are observed for the tumor sequencing library and the normal sequencing library that were processed at the same time.

Supplementary Figure 2

Distribution of counts for aligned sequence reads in normal cell lines

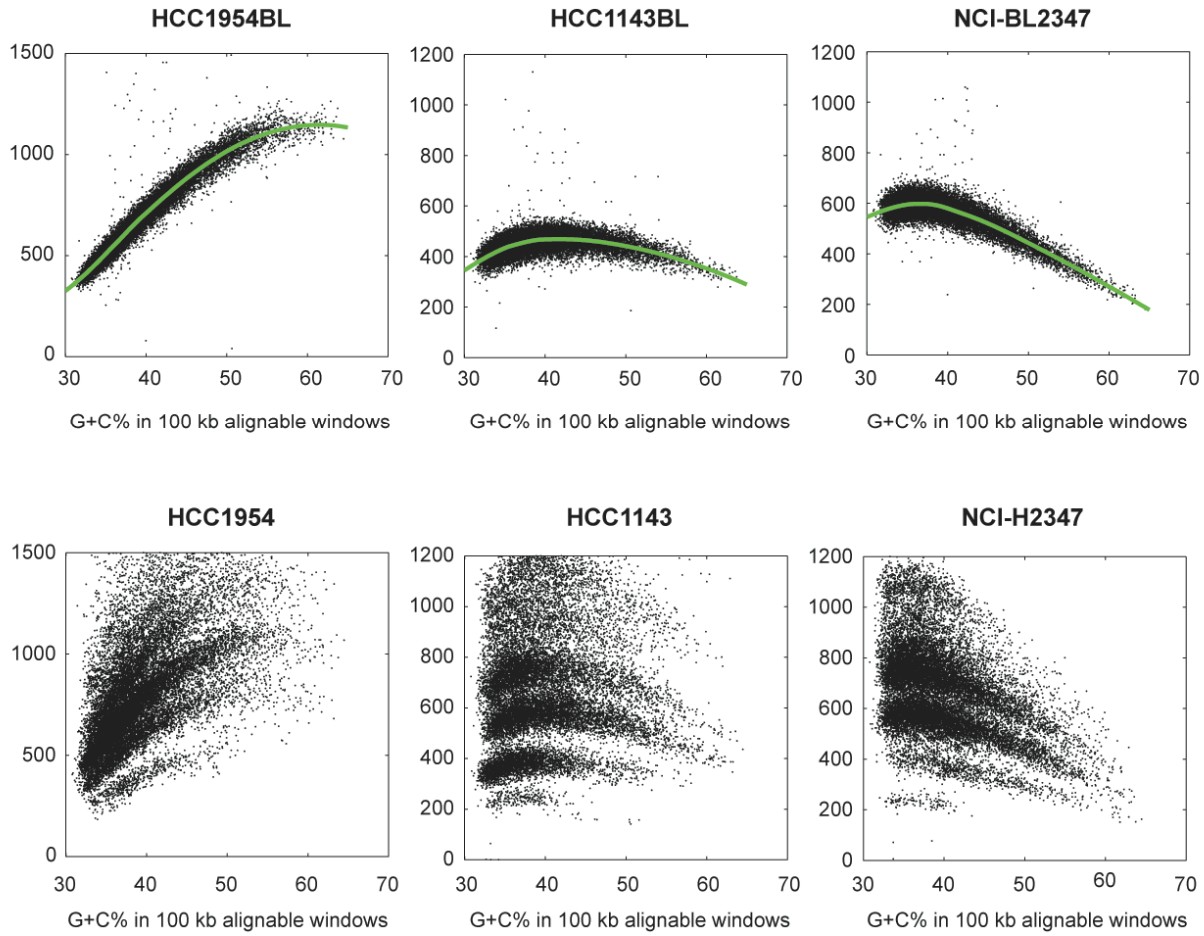


Each panel displays a histogram of the number of sequence reads aligning to 100 kb windows in the alignable portion of the human genome. Red lines indicate the expected Poisson distributions assuming a homogenous read density based on the average number of sequence reads in the lymphoblastoid normal cell lines HCC1954BL, HCC1143BL or NCI-BL2347. A single copy gain of chromosome 2 (whose presence was also confirmed by microarrays) was excluded from the HCC1143BL cell line.

Nature Methods: doi: 10.1038/nmeth.1276

Supplementary Figure 3

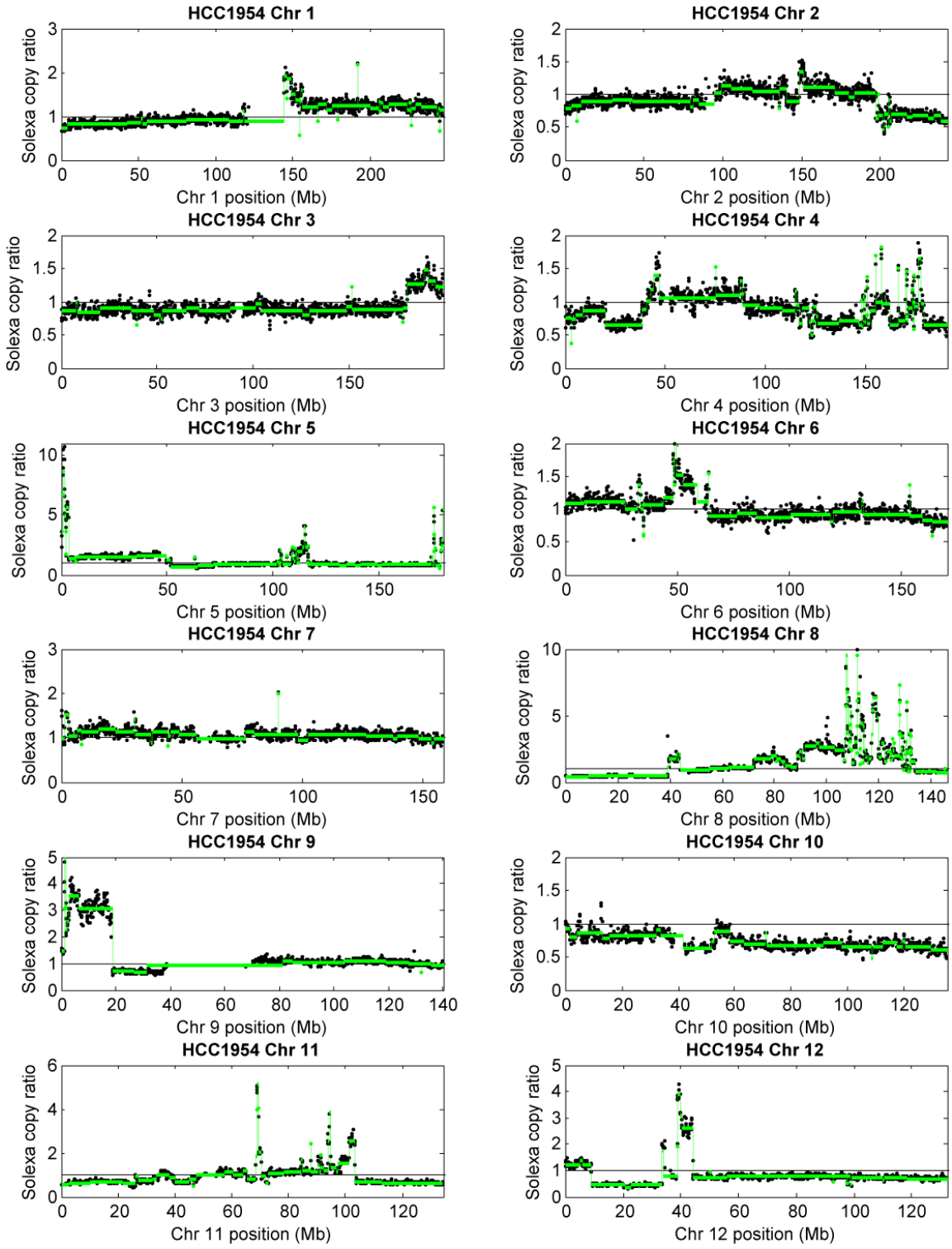
G+C bias on counts of aligned sequence reads



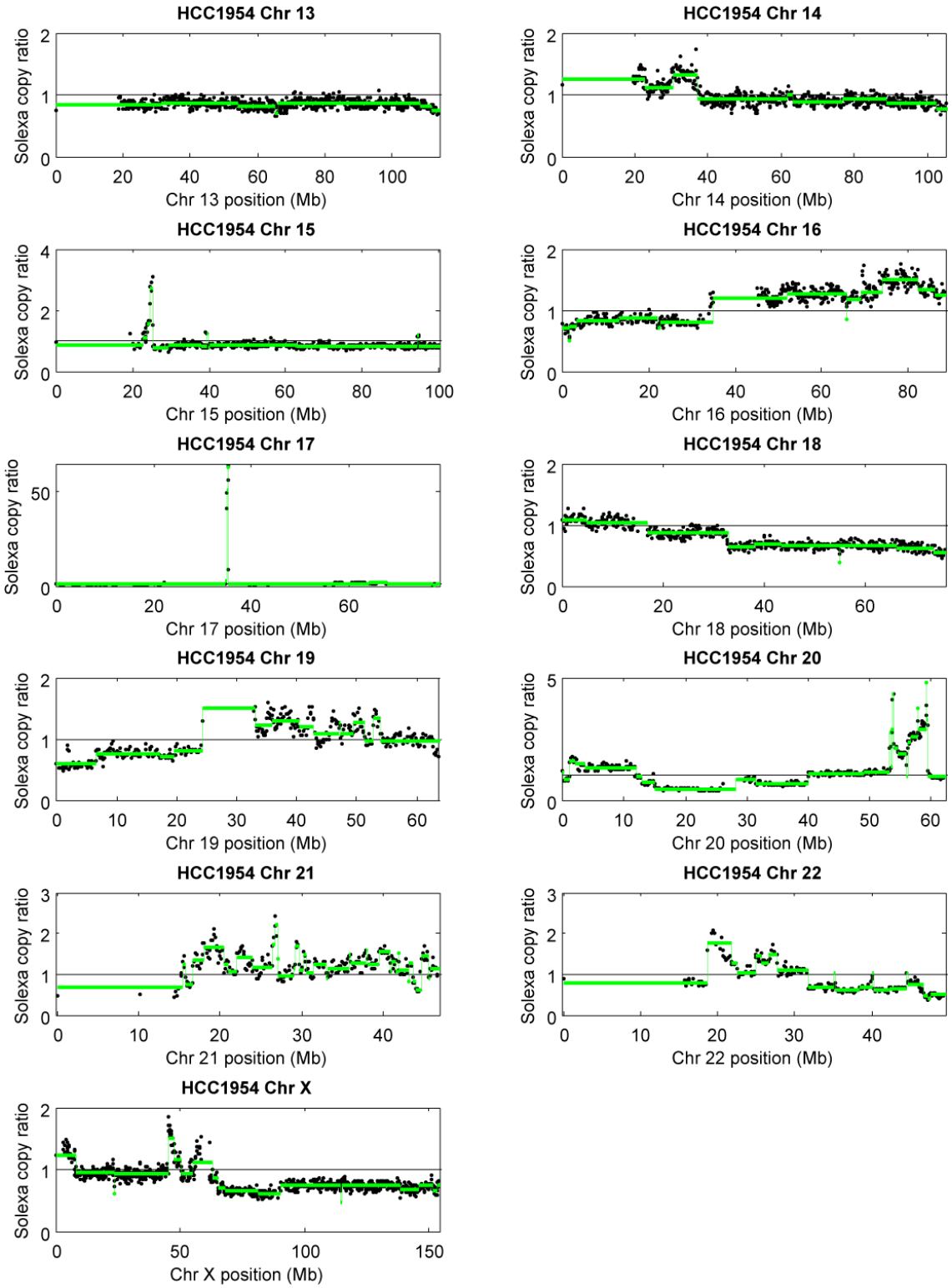
Each point represents the number of sequence reads aligned to a 100 kb window in the alignable portion of the reference human genome. The horizontal axis represents the G+C percentage of all possible 36 bp reads in each 100 kb alignable window. The vertical axis indicates the number of counts observed in each window. A single copy gain of chromosome 2 was excluded from the analysis of the HCC1143BL cell line, and a single copy loss of chromosome 6p was excluded from the analysis of the NCI-BL2347 cell line. The green lines indicate a loess local regression fit for each of the normal cell lines. The G+C content bias accounted for 92% of the variance in the HCC1954BL cell line, 26% of the variance in the HCC1143BL cell line, and 64% of the variance in the NCI-BL2347 cell line (Supplementary Table 2).

Supplementary Figure 4

Detailed view of segmentation results for HCC1954



Supplementary Figure 4 (cont.)

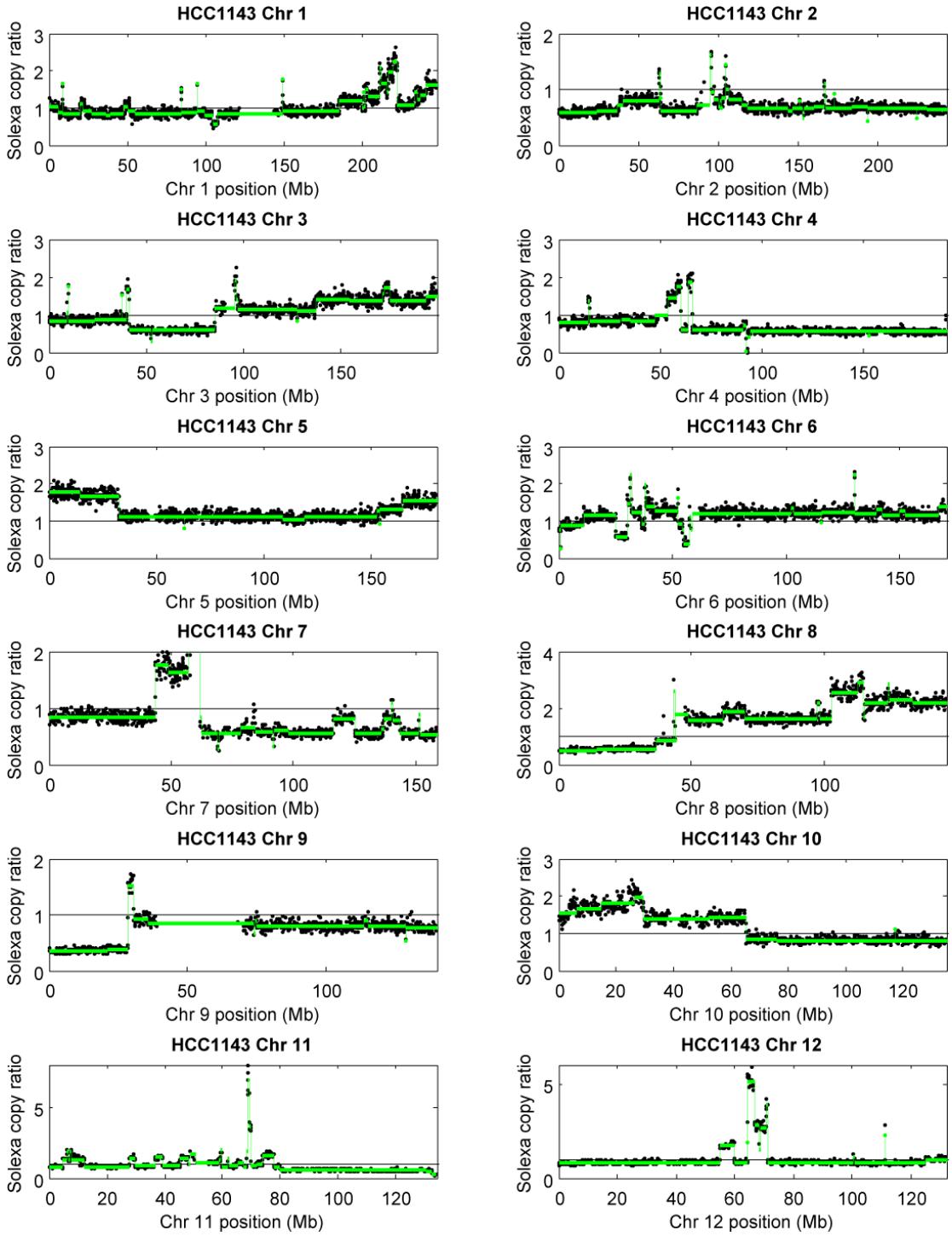


Each point represents a copy-number ratio between the number of tumor reads and normal reads in a 100 kb window in the alignable portion of the reference human genome. Horizontal green lines represent the copy-number levels determined by our segmentation algorithm on aligned sequence reads.

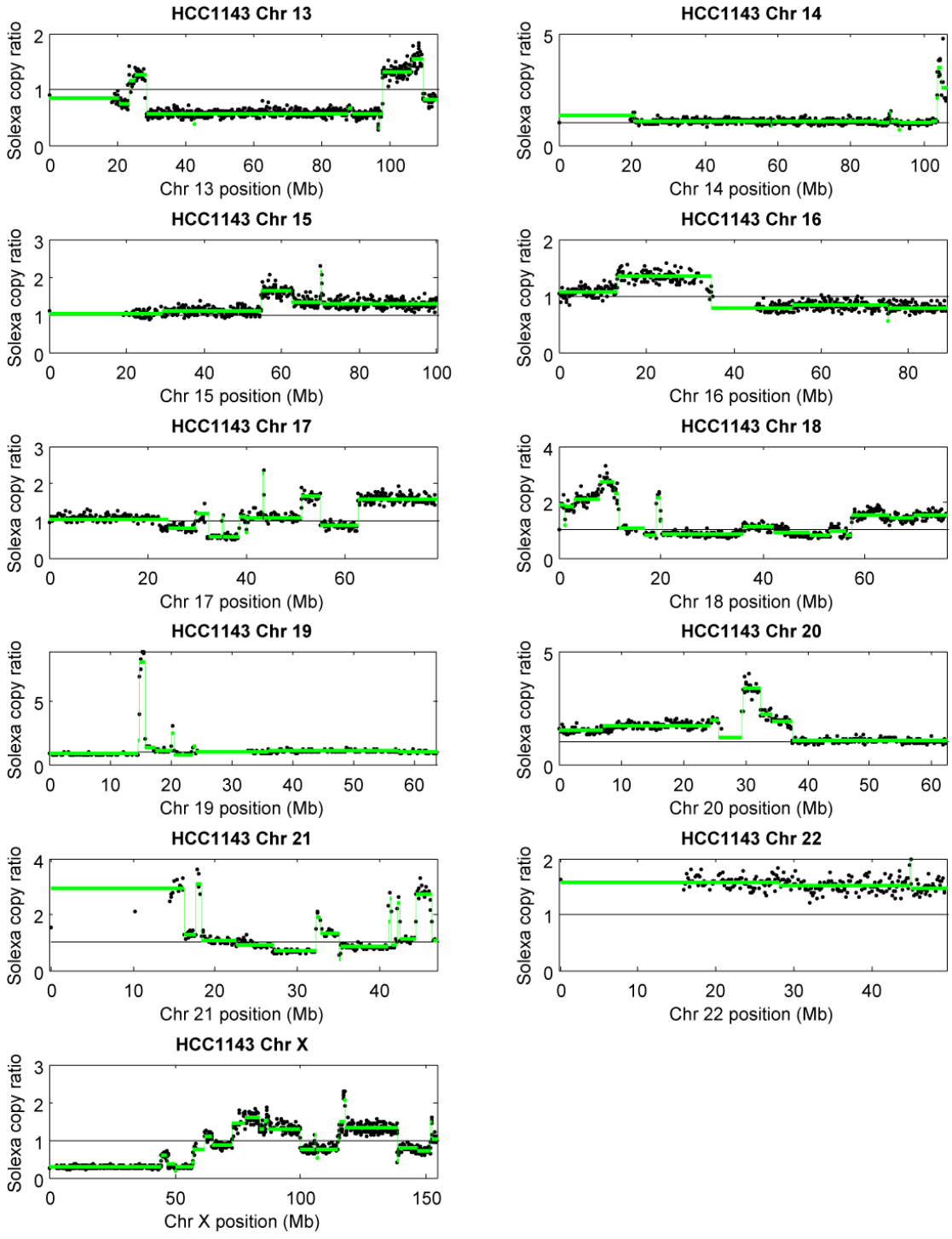
Nature Methods: doi: 10.1038/nmeth.1276

Supplementary Figure 5

Detailed view of segmentation results for HCC1143



Supplementary Figure 5 (cont.)

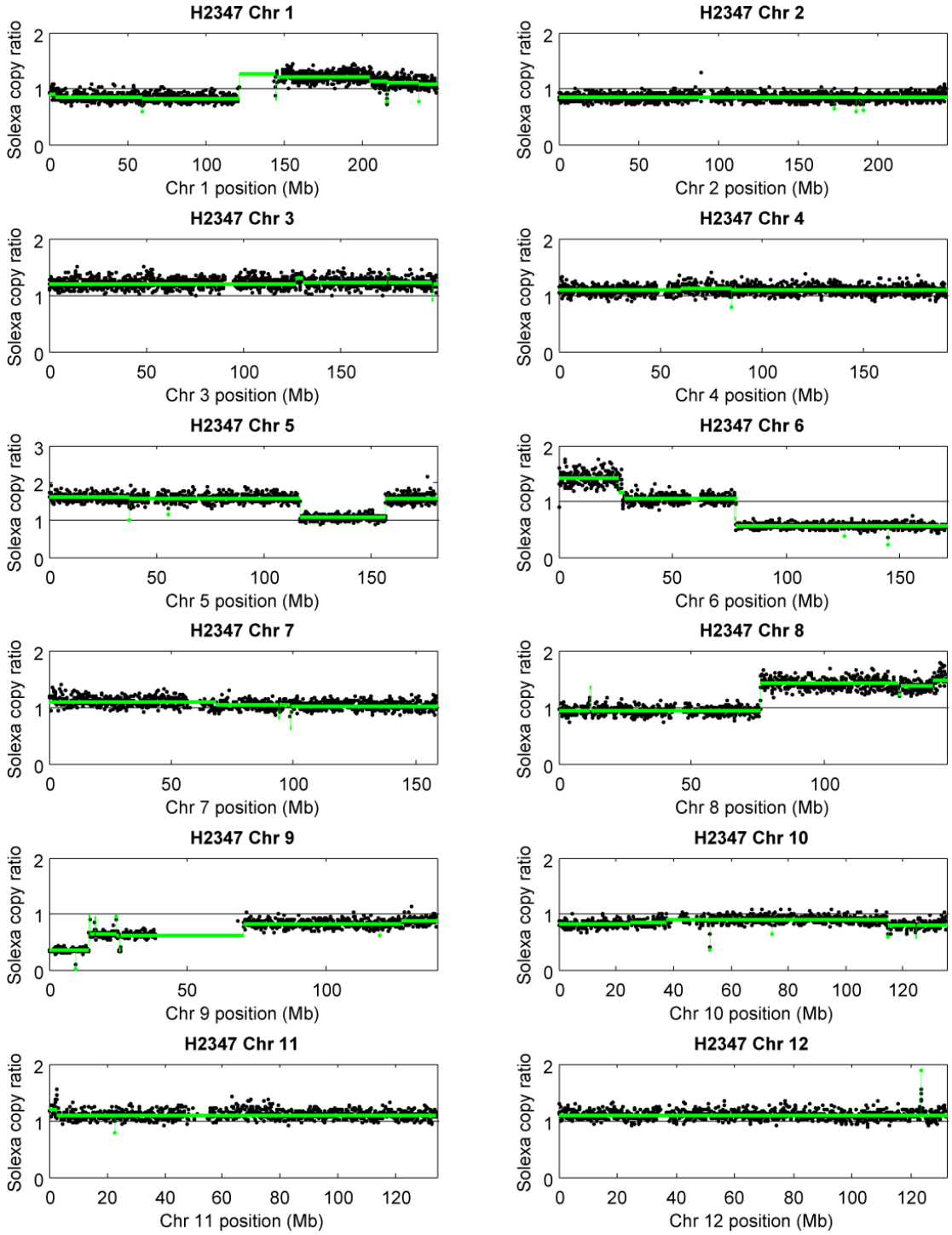


Each point represents a copy-number ratio between the number of tumor reads and normal reads in a 100 kb window in the alignable portion of the reference human genome. Horizontal green lines represent the copy-number levels determined by our segmentation algorithm on aligned sequence reads.

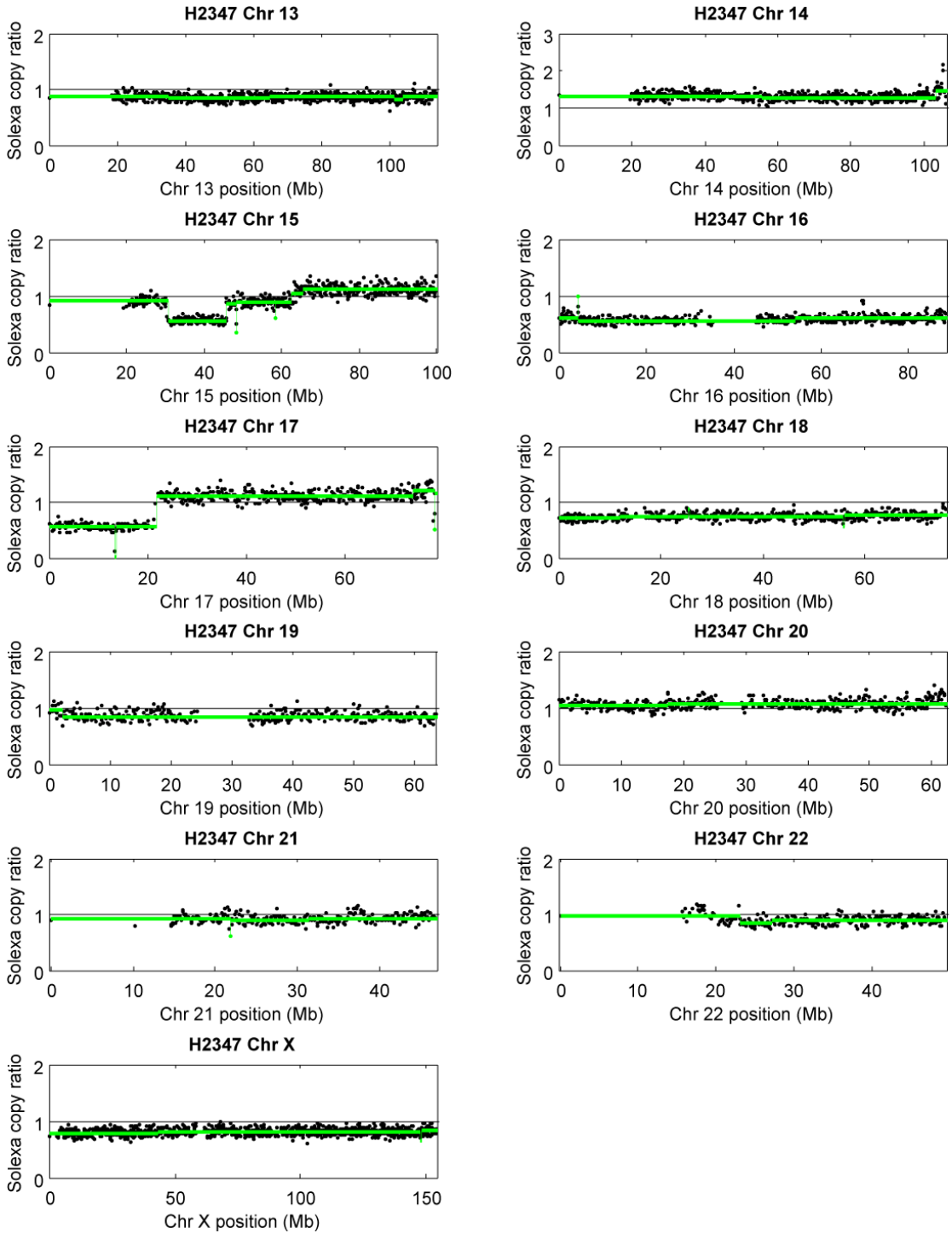
Nature Methods: doi: 10.1038/nmeth.1276

Supplementary Figure 6

Detailed view of segmentation results for NCI-H2347



Supplementary Figure 6 (cont.)

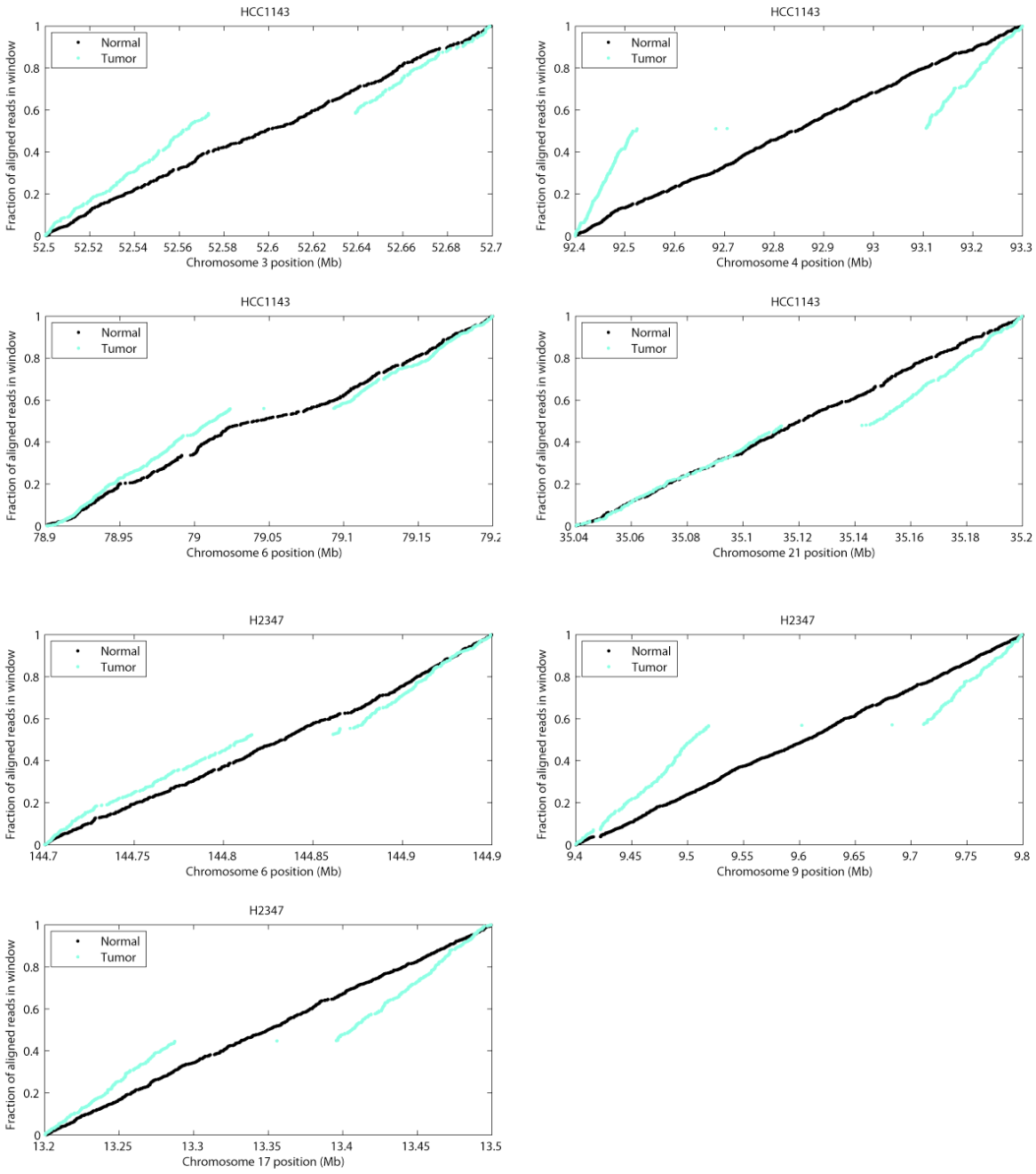


Each point represents a copy-number ratio between the number of tumor reads and normal reads in a 100 kb window in the alignable portion of the reference human genome. Horizontal green lines represent the copy-number levels determined by our segmentation algorithm on aligned sequence reads.

Nature Methods: doi: 10.1038/nmeth.1276

Supplementary Figure 7

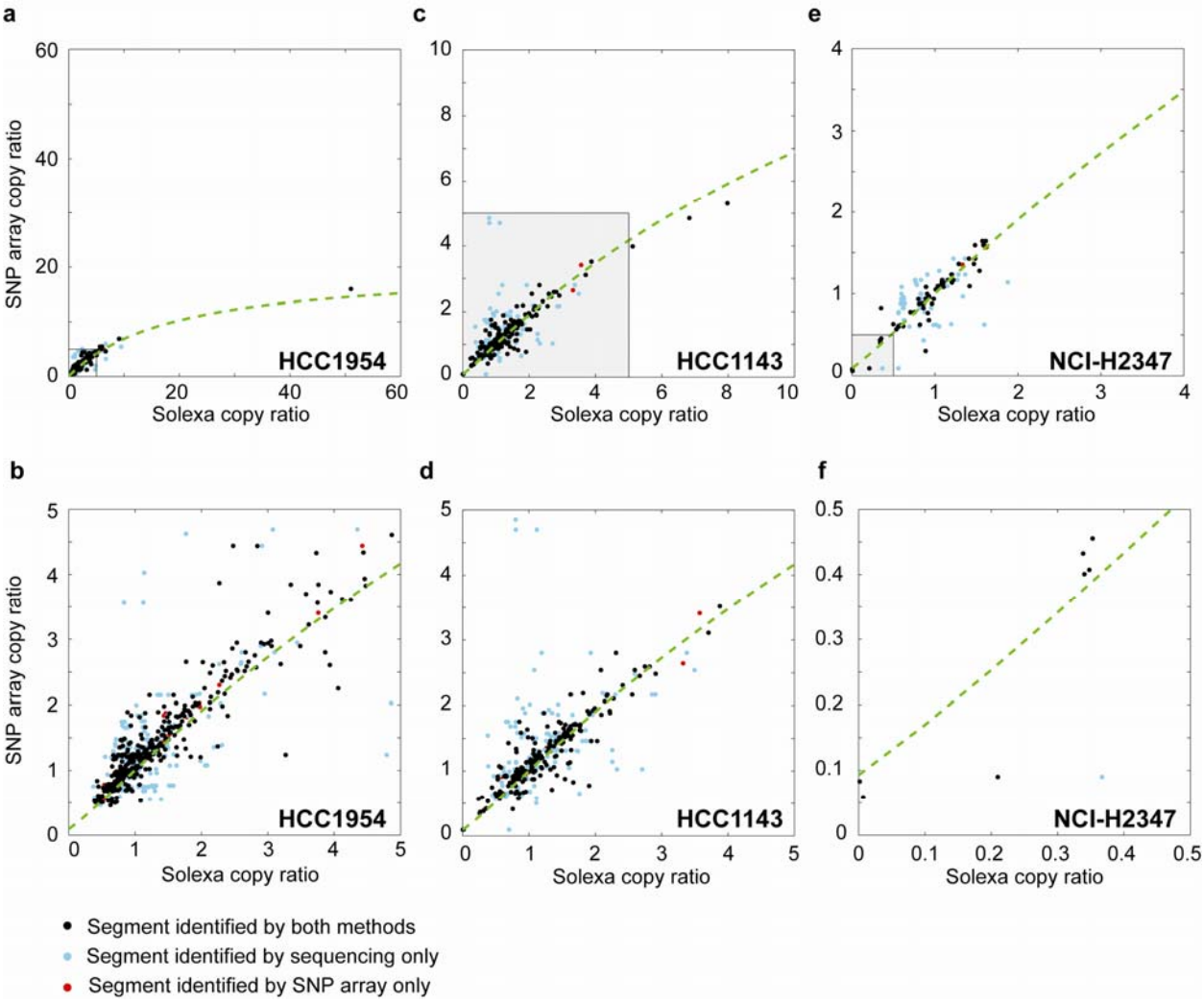
Predicted homozygous deletions in HCC1143 and NCI-H2347 cell lines



Each panel represents the cumulative distribution of sequence reads that aligned to the chromosomal window indicated on the horizontal axis. Each point represents the location of a sequence read aligning to the tumor cell line (blue) or its matched normal (black). Somatic homozygous deletions are predicted to occur in the chromosomal regions with a marked absence of aligned reads in the tumor cell line, yet are covered by aligned reads in the normal cell line.

Supplementary Figure 8

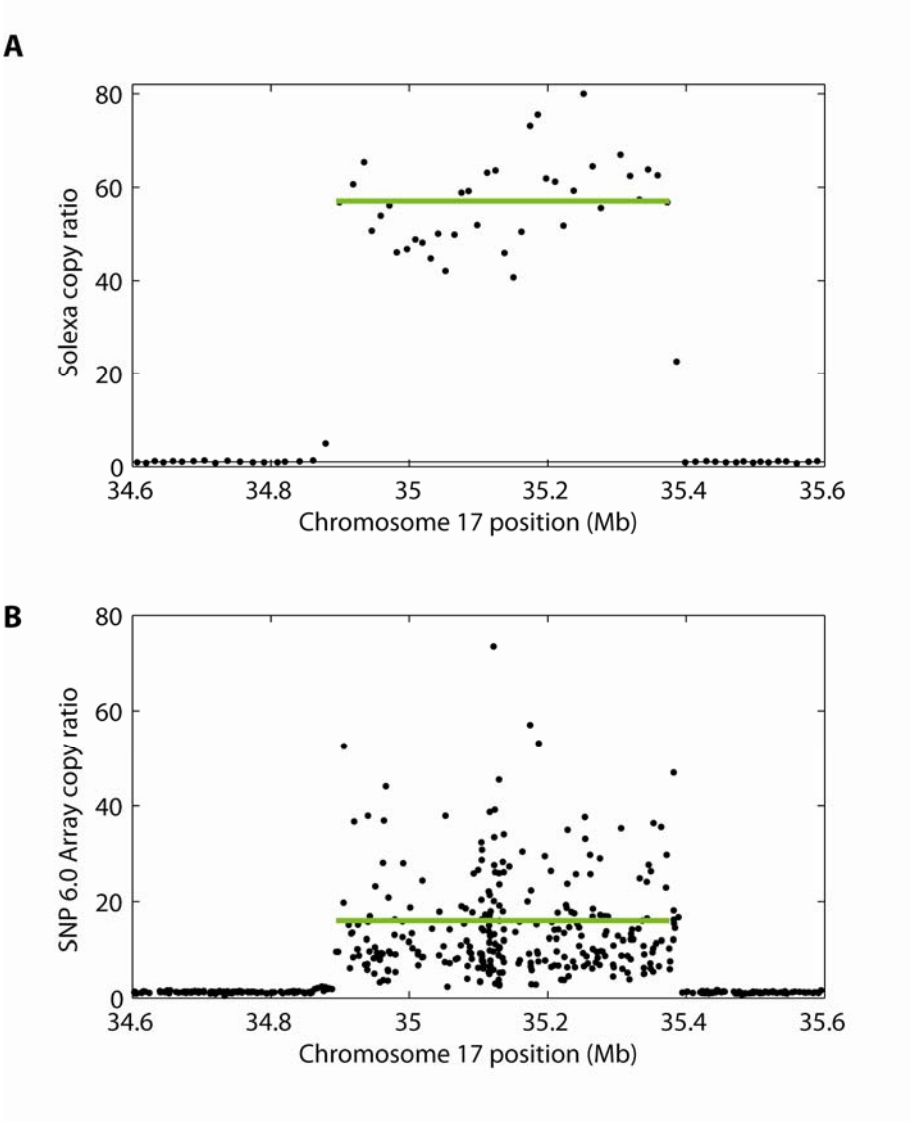
Comparisons of copy-number alterations between sequencing and microarrays



Copy-number alterations in the (**a**, **b**) HCC1954, (**c**, **d**) HCC1143 and (**e**, **f**) NCI-H2347 cell lines were inferred by massively parallel sequencing and SNP array-based hybridization. The shaded grey boxes in the top panels indicate the boundaries for the zoomed insets displayed in the bottom panels. Each point represents the copy-number ratio for a single chromosomal segment calculated from the tumor/normal ratio of aligned sequence reads (horizontal axis) or from GLAD segmentation of copy numbers on the Affymetrix Genome-Wide SNP 6.0 Array (vertical axis). Points in black indicate chromosomal segments with at least one concordant breakpoint between sequencing and microarrays. Points in red indicate segments with both breakpoints identified by microarrays, while points in blue indicate chromosomal segments that were uniquely found by our segmentation procedure on aligned sequence reads.

Supplementary Figure 9

ERBB2 amplification in HCC1954



Copy ratio estimates of the *ERBB2* high-level amplification in the HCC1954 cell line. (A) Each point represents the copy ratios estimated by massively parallel sequencing for a 10 kb window of the alignable portion of the reference genome. The green line indicates the average copy ratio within the *ERBB2* amplicon. (B) Each point represents the copy ratios estimated by a single probe set on the Affymetrix Genome-Wide SNP 6.0 Array. The green line indicates the mean copy ratio within the *ERBB2* amplicon.

Supplementary Figure 10

Mapped breakpoints of the *UTRN* homozygous deletion

```
Seq   000000310 ttggggtttcaccatgttggccagactggtctcaaactcctgacctctag 000000359
      >>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
hg18  144816139 ttggggtttcaccatgttggccagactggtctcaaactcctgacctctag 144816188

Seq   000000360 tgatccacccccctcagccttccaaagtgttgggattacaggtgtgagcc 000000409
      >>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
hg18  144816189 tgatccacccccctcagccttccaaagtgttgggattacaggtgtgagcc 144816238

Seq   000000410 actatgcccagcctatccttttttcttgttagttaatttttgttcaacaa 000000459
      >>>>>>>>> || |||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
hg18  144816239 accatgcccagcctatccttttttcttgttagttaatttttgttcaacaa 144816288

Seq   000000460 cttctgtcataacacattcaattctttagactcaaaccattatttttatg 000000509
      >>>>>>>>> ||||||||||||||||||||||||| ||||||||||||||||||||| >>>>>>>>>
hg18  144816289 cttctgtcataacacattcaattctttgactcaaaccattatttttatg 144816338

Seq   000000510 gcatggcgtact 000000521
      >>>>>>>>> |||||||||||| >>>>>>>>>
hg18  144816339 gcatggcgtact 144816350



Seq   000000522 cagcctggtcaacatggcaaaaacccgtctctactcaaaaaaaaaaaaaa 000000571
      >>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
hg18  144860307 cagcctggtcaacatggcaaaaacccgtctctactcaaaaaaaaaaaaaa 144860356

Seq   000000572 aaa 000000574
      >>>>>>>>> ||| >>>>>>>>>
hg18  144860357 aaa 144860359
```

Chromosomal breakpoints of the 44 kb interstitial homozygous deletion at the *UTRN* locus in the NCI-H2347 cell line were mapped by sequencing a PCR product spanning the deletion. The coordinates of the sequence read and reference genome are indicated above; the forward strand was sequenced.

Supplementary Figure 11

Mapped breakpoints of the *PTPRD* homozygous deletion

```
Sequence 000443 atatatttggtgggaatgaacatacattttataatttttaaatttaatga 0000492
       <<<<<<< ||||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<
hg18   9711291 atatatttggtgggaatgaacatacattttataatttttaaatttaatga 9711242

Sequence 000493 taaaatctgagatgtcttgaagatattccatgatcctctattgcaacttt 0000542
       <<<<<<< ||||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<
hg18   9711241 taaaatctgagatgtcttgaagatattccatgatcctctattgcaacttt 9711192

Sequence 000543 ctatactctttttaaaagatgtttatccacactttagtattttatttaat 0000592
       <<<<<<< ||||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<
hg18   9711191 ctatactctttttaaaagatgtttatccacactttagtattttatttaat 9711142

Sequence 000593 ttttactttagt 0000604
       <<<<<<< |||||||||||| <<<<<<<
hg18   9711141 ttttactttagt 9711130
```

```
Sequence 000605 ttcgagaccagtctggctaacatggtgaaaccccatctctatcaaagata 0000654
       <<<<<<< ||||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<
hg18   9519156 ttcgagaccagtctggctaacatggtgaaaccccatctctatcaaagata 9519107

Sequence 000655 caaaaaattagctaggcgtggtggcatgcacctgtaatcccagatactct 0000704
       <<<<<<< ||||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<
hg18   9519106 caaaaaattagctaggcgtggtggcatgcacctgtaatcccagatactct 9519057
```

Chromosomal breakpoints of the 139 kb interstitial homozygous deletion at the *PTPRD* locus in the NCI-H2347 cell line were mapped by sequencing a PCR product spanning the deletion. The coordinates of the sequence read and reference genome are indicated above; the reverse complement strand was sequenced.

Supplementary Figure 12

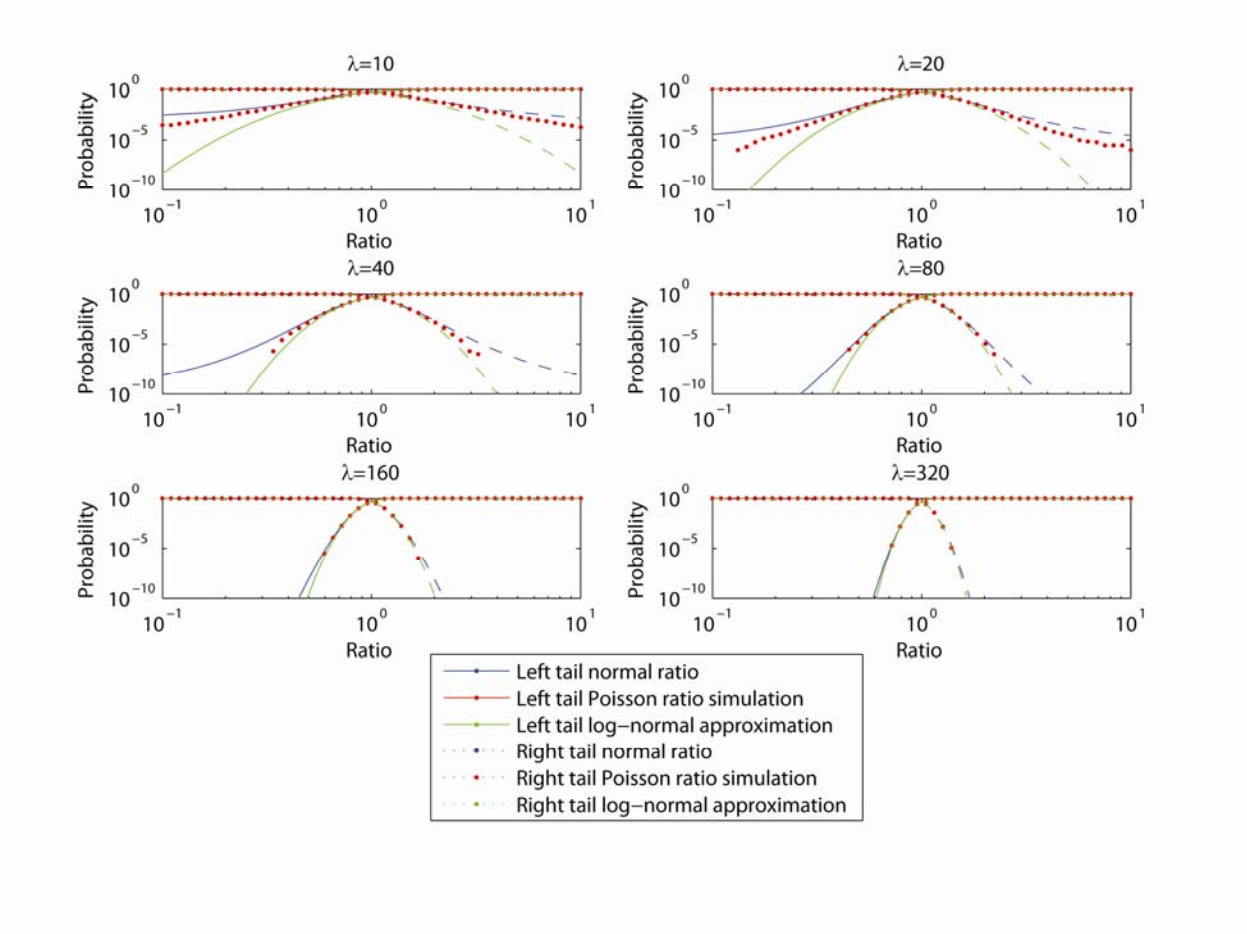
Mapped breakpoints of the *HS3ST3A1* homozygous deletion

```
Sequence 00000010 tcgc.gcacaaggcaggtagttttgtctgcaacagccattgtggctgtcg 00000058
         >>>>>>>> |||| |||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>
hg18     13287133 tcgcagcacaaggcaggtagttttgtctgcaacagccattgtggctgtcg 13287182

Sequence 00000059 gcaatccctattcccacatgcacaggaaatgaatttgtgtgtataaatat 00000108
         >>>>>>>> |||||||||||| ||| |       |  || |||||||||||||||||| >>>>>>>>
hg18     13287183 gcaatccctattgccagacatttgtgctataaatttgtgtgtataaatat 13287232

Sequence 00000109 gaatacatgcaactggaaggccatcattctgtg 00000141
         >>>>>>>> |||||||||||||||||||||||||||||||| >>>>>>>>
hg18     13287233 gaatacatgcaactggaaggccatcattctgtg 13287265


Sequence 00000157 aggcccggcgcagtggctcacgcctgtaatcccagcactttgggaggcca 00000206
         >>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>
hg18     13394828 aggcccggcgcagtggctcacgcctgtaatcccagcactttgggaggcca 13394877

Sequence 00000207 aggcgggcagatctcgatgtccgaagatcgagaccatcctggctagcaga 00000256
         >>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>
hg18     13394878 aggcgggcagatctcgatgtccgaagatcgagaccatcctggctagcaga 13394927

Sequence 00000257 gtgaaaccccgtctctacgaaaaatacaaaaattagccgggcatggtggc 00000306
         >>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>
hg18     13394928 gtgaaaccccgtctctacgaaaaatacaaaaattagccgggcatggtggc 13394977

Sequence 00000307 aggtgcctgtagccccagctactggggaggctgaggcaggagaatggcat 00000356
         >>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>
hg18     13394978 aggtgcctgtagccccagctactggggaggctgaggcaggagaatggcat 13395027

Sequence 00000357 gaacccgggaggcgaagcttgcagtgagccgagatcgcgtcactgcactc 00000406
         >>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>
hg18     13395028 gaacccgggaggcgaagcttgcagtgagccgagatcgcgtcactgcactc 13395077

Sequence 00000407 cagcctgggcgacagagcgagactctgtctcaaaaaaaaaaa 00000448
         >>>>>>>> |||||||||||||||||||||||||||||||||||||||| >>>>>>>>
hg18     13395078 cagcctgggcgacagagcgagactctgtctcaaaaaaaaaaa 13395119
```

Chromosomal breakpoints of the 108 kb homozygous deletion at the *HS3ST3A1* locus in the NCI-H2347 cell line were mapped by sequencing a PCR product spanning the deletion. The coordinates of the sequence read and reference genome are indicated above; the forward strand was sequenced.

Supplementary Figure 13

Approximations for the distribution of copy-number ratios, *R*

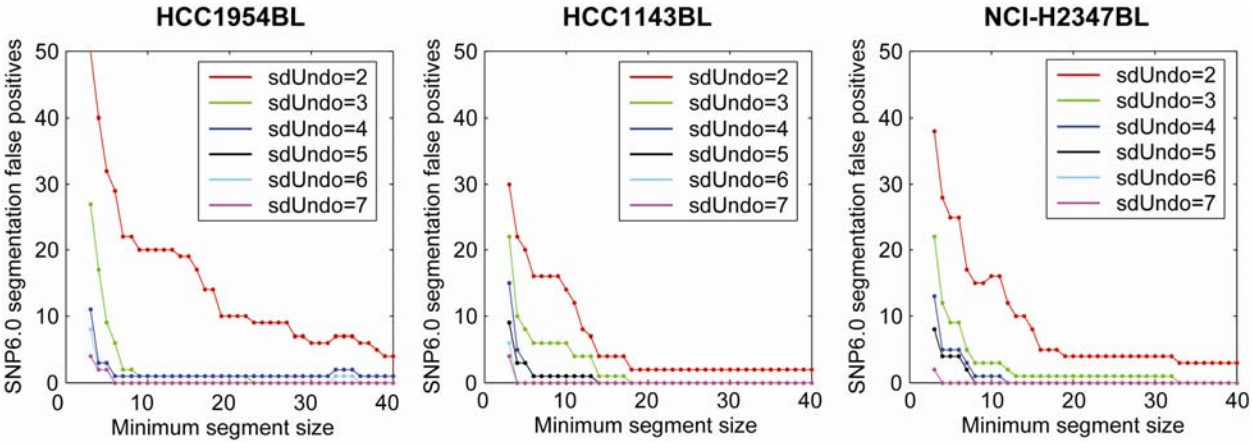


Each panel displays the left and right tail of the cumulative distribution function of the copy-number ratio, *R*. Three plots represent: (1) Exact distribution of ratios of normally distributed variables (blue) – used for the power analysis; (2) Ratios between simulated Poisson random variables (red dots); (3) Log-normal approximation for the distribution of normal ratios (green line) – used for estimating the distribution of *D* and assessing the significance of breakpoints.

Nature Methods: doi: 10.1038/nmeth.1276

Supplementary Figure 14

Parameter optimization for Circular Binary Segmentation of SNP 6.0 arrays



For each of the diploid cell lines HCC1954BL, HCC1143BL and NCI-H2347BL, one representative Affymetrix Genome-Wide SNP 6.0 Array was chosen with the median variance in copy numbers. Log_2 copy numbers were segmented with the Circular Binary Segmentation algorithm (Venkatraman & Olshen, 2007). We tested different stringencies for the sdUndo parameter, which sets the number of standard deviations between means to keep a segment split. We performed an additional merging step of segments with fewer than *k* consecutive probes on the SNP 6.0 Array, as indicated on the horizontal axis. Each line represents the number of false positive segments detected for the combination of an sdUndo parameter and minimum segment size. We found that the combination of sdUndo = 2 and a minimum segment size of 8 consecutive probes were the most lenient parameters that achieved the greatest reduction in the number of false positives.

Supplementary Table 1

Statistics of whole genome shotgun sequencing

Cell line	Type	Total Reads	Aligned reads	Aligned sequence	Effective coverage
HCC1954	Breast adenocarcinoma	40,805,074	18,927,988 (46.4%)	636.6 Mb	0.29×
HCC1954BL	EBV-transformed lymphoblastoid	31,210,602	14,634,320 (46.9%)	502.7 Mb	0.23×
HCC1143	Breast adenocarcinoma	44,762,968	15,038,736 (33.6%)	541.4 Mb	0.25×
HCC1143BL	EBV-transformed lymphoblastoid	34,293,547	10,012,495 (29.2%)	360.4 Mb	0.17×
NCI-H2347	Lung adenocarcinoma	40,116,322	13,983,159 (34.9%)	503.4 Mb	0.23×
NCI-BL2347	EBV-transformed lymphoblastoid	43,854,991	12,124,804 (27.6%)	436.5 Mb	0.20×

Effective coverage equals the amount of aligned sequence, divided by the 2,163,378,178 bp of the human genome (NCBI Build 36.1) that are accessible to aligned 36 bp reads using the stringent alignment criteria of Mikkelsen *et al.*(2007).

Supplementary Table 2

Effect of G+C bias on number of aligned reads

Normal cell line	Total variance, *TV*	Variance after G+C loess fit, *LV*	ΔR_{GC}^2	Poisson variance, *PV*	Residual variance, *LV – PV*	$\frac{PV}{LV}$
HCC1954BL	35761	2909.3	0.919	679.1	2230.2	0.233
HCC1143BL	1977.8	1460.9	0.261	447.2	1013.7	0.226
NCI-BL2347	5469.4	1978.5	0.638	559.7	1418.8	0.283

DEFINITIONS

Let X_1, X_2, …, X_W represent the number of sequence reads that align to W equivalently spaced windows in a normal cell line. These calculations used 100 kb windows, as measured in the alignable portion of the reference genome.

Total variance

$$TV = \frac{1}{W}\sum_w \left(X_w - \overline{X_w}\right)^2$$ where $\overline{X_w}$ represents the average number of aligned counts

Variance after G+C loess fit

The predicted number of counts for each window, L_w, was obtained via a loess local regression fit of the number of aligned counts against the G+C content, rounded to the nearest 0.5% increment. The residual variance after normalization to the loess fit was calculated as:

$$LV = \frac{1}{W}\sum_w \left(X_w - L_w\right)^2$$

Contribution of G+C bias to total variance

The proportion of total variance accounted for by the G+C bias was calculated as:

$$\Delta R_{GC}^2 = 1 - \frac{LV}{TV}$$

Poisson variance

Assuming a Poisson distribution, the expected variance in X_1, X_2, …, X_W equals:

$$PV = \overline{X_w}$$

Residual variance

The residual variance unexplained by either G+C bias or Poisson sampling was defined as:

$$LV - PV = \frac{1}{W}\sum_w \left(X_w - L_w\right)^2 - \overline{X_w}$$

Supplementary Table 3

Candidate homozygous deletions detected in cell lines

NCI-H2347

Locus	# of exons	Chr	Predicted Start (bp)	Predicted End (bp)	Mapped Start (bp)	Mapped End (bp)	Solexa ratio	SNP ratio	# SNP
UTRN	15	6	144815824	144861313	144816351	144860306	0	0.05	24
AK093114	3' UTR	6	164112429	164119214				0.02	14
PTPRD	5' UTR	9	9518931	9711181	9519157	9711129	0.002	0.04	137
HS3ST3A1	1	17	13287264	13395656	13287266	13394827	0	0.02	70

HCC1143

Locus	# of exons	Chr	Predicted Start (bp)	Predicted End (bp)	Mapped Start (bp)	Mapped End (bp)	Solexa ratio	SNP ratio	# SNP
PBRM1	12	3	52573049	52638899			0	0.07	35
KIAA1680	3' UTR	4	92524213	93106238			0.0008	0.05	332
Noncoding	0	6	79023817	79093433			0.0065	0.23	105
RUNX1	1	21	35113747	35142457			0	0.06	24

Sample preparation, sequencing and alignment

We performed DNA sequencing with the Illumina 1G Genome Analyzer according to the manufacturer's directions. Briefly, we sheared the DNA, ligated adaptors, performed gel purification to obtain fragments of ~150 bases in size, performed PCR and carried out sequencing. We obtained sequence reads of length of 32 - 36 bases from each cell line. These reads were compiled, post-processed and aligned to NCBI Build 36.1 of the human genome according to the procedure described by Mikkelsen *et al*[1]. In brief, a sequence read was considered uniquely mapped to the reference genome if its second best alignment had 2 or more additional mismatches when compared to its best alignment. This criterion allows slight sequence variants (such as single nucleotide polymorphisms) to be aligned, while discarding sequence reads that map to multiple genomic loci. Duplicate reads that aligned to a unique genome location were not omitted. With these alignment parameters, we determined that 72% of the human genome could be uniquely mapped by 36 bp reads.

[1] Mikkelsen, T.S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560 (2007).

Nature Methods: doi: 10.1038/nmeth.1276

Statistical analysis of tumor-normal copy-number ratios

Consider a genomic window of length L within the alignable portion of the reference genome with length A. Let a_N and a_T denote the total number of aligned sequence reads from the normal or tumor sample, respectively. In a given genomic window from x_L to x_R, let N and T denote the number of aligned sequence reads from the normal and tumor samples, respectively.

Log-normal approximation. Assuming that there are no copy-number alterations in a window of length L, N follows a Poisson distribution with parameter $\lambda_N = a_N \times L / A$. Similarly, T follows a Poisson distribution with parameter $\lambda_T = r \times a \times \lambda_N$, where $a = a_T / a_N$. The tumor-normal copy ratio, R, is defined as: $R(x_L, x_R) = \frac{T(x_L, x_R)/a_T}{N(x_L, x_R)/a_N}$ if $N>0$, else R is undefined. We approximate the log copy ratio, $\log(R)$, with a log-normal distribution and show that it conforms well for $\lambda_N > 80$ (see Supplementary Fig. 13).

Log-ratio difference statistic. We consider two adjacent genomic windows of arbitrary size: a left window from position x_L to x, and a right window from position x to x_R ($x_L < x < x_R$). We calculate the difference in log ratios between the right window and the left window:

$$D_x(x_L, x_R) = \log\left(R(x, x_R)\right) - \log\left(R(x_L, x)\right) \quad \textbf{(1)}$$

We obtain two-sided p-values for this log-ratio difference statistic, $p(\,|D_x(x_L, x_R)| > d\,)$ by convoluting the log-normal approximations to the individual distributions for R_x^R and R_x^L (see Supplementary Fig. 1, "Log-normal approximation for the distribution of log copy-number ratios" below).

Approximate distribution of copy-number ratios

First, we calculate the copy-number ratios between tumor and its matched normal in a local window of size $L = x_R - x_L + 1$:

$$R(x_L, x_R) = \frac{T(x_L, x_R)/a_T}{N(x_L, x_R)/a_N}$$

Our rationale for calculating copy-number ratios was to compensate for the non-uniform density of reads, even among adjacent genomic regions with the same copy-number. These variations can be mostly explained by local sequence properties such as repeat density and different G+C prevalence (Supplementary Table 2, Supplementary Fig. 2, 3).

Nature Methods: doi: 10.1038/nmeth.1276

We assumed the tumor and normal counts were Poisson distributed with local average counts of λ_N and $\lambda_T = r \times a \times \lambda_N$ respectively, where $a = a_T / a_N$ and r corresponds to the actual tumor normal ratio. When λ_N and λ_T are sufficiently large (≥80), one can approximate the Poisson distribution with a Gaussian distribution: Poisson($x;\lambda$) ≈ Gaussian(x; $\mu = \lambda$, $\sigma^2 = \lambda$);. Supplementary Fig. 13 compares two approximations for the distribution of Poisson ratios with simulated data, using different values for λ. From this empirical null distribution, we observed that the approximation based on ratio of normals is an upper bound and the log-normal approximation is a lower bound.

The distribution of ratios of two independent normal random variables has an exact formula[2], $f(x)$:

$$a(x) = \sqrt{\frac{x^2}{{\sigma_T}^2} + \frac{1}{{\sigma_N}^2}}$$

$$b(x) = \frac{\mu_T x}{{\sigma_T}^2} + \frac{\mu_N}{{\sigma_N}^2}$$

$$c(x) = e^{\frac{1}{2}\frac{b^2(x)}{a^2(x)} - \frac{1}{2}\left(\frac{{\mu_T}^2}{{\sigma_T}^2} + \frac{{\mu_N}^2}{{\sigma_N}^2}\right)}$$

$$f(x) = \frac{b(x)c(x)}{a^3(x)} \frac{1}{\sqrt{2\pi}\sigma_T\sigma_N}\left[2\Phi\left(\frac{b(x)}{a(x)}\right) - 1\right] + \frac{1}{a^2(x)\pi\sigma_T\sigma_N} e^{-\frac{1}{2}\left(\frac{{\mu_T}^2}{{\sigma_T}^2} + \frac{{\mu_N}^2}{{\sigma_N}^2}\right)}$$

where $\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\, du$.

There is also an exact formula for the cumulative distribution function $F(x)$:

$$F(x) = G\left(\frac{\mu_T - x\mu_N}{\sigma_T\sigma_N a(x)}, -\frac{\mu_N}{\sigma_N}; \frac{\sigma_N x}{\sigma_T\sigma_N a(x)}\right) + G\left(-\frac{\mu_T - x\mu_N}{\sigma_T\sigma_N a(x)}, \frac{\mu_N}{\sigma_N}; \frac{\sigma_N x}{\sigma_T\sigma_N a(x)}\right)$$

where $G(h,k;\gamma)$ is the standard cumulative distribution of a bivariate normal distribution;

$$G(h,k;\gamma) = \frac{1}{2\pi\sqrt{1-\gamma^2}} \int_{-\infty}^{h}\int_{-\infty}^{k} \exp\left(-\frac{1}{2}\begin{pmatrix} x & y \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix}\right) dxdy .$$

[2] Hinkley, D.V. On the ratio of two correlated normal random variables. *Biometrika* 56, 635-639 (1969).

Nature Methods: doi: 10.1038/nmeth.1276

Using the above formulas we calculate the significance level (*p*-values) and power for detecting copy-number alterations in non-overlapping windows of various sizes (see "Analytical power calculations" below, Fig. 1). The number of reads needed to obtain a desired power is estimated to within 10% when $\lambda \geq 80$.

Log-normal approximation for the distribution of log copy-number ratios

In our algorithm we use the local difference of log ratios statistic (from **Eq. 1** above), $D_x(x_L, x_R) = \log\left(R\left(x, x_R\right)\right) - \log\left(R\left(x_L, x\right)\right)$, to identify significant copy-number changes. In order to calculate two-sided *p*-values, we first approximate the distribution of *R* with a log-normal distribution. Under this approximation, the log copy-number ratio in each side of a potential breakpoint follows a normal distribution with parameters that depend on the number of normal and tumor reads in each of the intervals. It follows that the distribution of *D* is the distribution of the difference between two normally distributed random variables, which is also normally distributed. Finally, *p*-values are calculated using the normal cumulative distribution function.

In more detail:

(i) The approximation of $f(x; \lambda_N, \lambda_T)$ (see above) by a log-normal distribution, $v(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$ is performed in three steps: (a) Transforming the variables $x \to e^y$ such that *y* represents the log ratio (note that a Jacobian is needed) and log transforming, $h(y) = \log(e^y f(e^y; \lambda_N, \lambda_T))$; (b) Estimating μ by numerically solving for a vanishing first derivative, $\left.\frac{\partial h(y)}{\partial y}\right|_{\hat{\mu}} = 0$; and (c) Estimating σ^2 using the second derivative $\hat{\sigma}^2 = -\left(\frac{\partial^2 h(y)}{\partial y^2}\right)^{-1}$. Supplementary Fig. 13 shows the accuracy of the log-normal approximation for various values of λ_N.

(ii) Under the same approximation, the distribution of *D* is also Gaussian with mean $\mu = \mu_L - \mu_R$ and $\sigma^2 = \sigma_L^2 + \sigma_R^2$ since the left and right sides are independent.

Nature Methods: doi: 10.1038/nmeth.1276

(iii) To test for significance, the null distribution of D can be calculated by assuming that the two windows have the same copy-number ratio. We can approximate the null distribution by calculating the distribution of D after partitioning the total number of tumor reads to the two windows such that they have the same ratio as between the normal reads. In this case, the distribution is symmetric with mean $\mu = 0$; the variance σ^2 depends on the number of tumor and normal reads observed.

Supplementary Figure 1 demonstrates that the distribution of D in real data mimicked the approximation above. Notice that in the case of HCC1954, the normal versus normal comparison followed the expected identity line representing a uniform p-value distribution. In other cases, there was a small deviation from the expected line. However, it is striking that the same deviation was observed both in the matched tumor and normal that were run in the same day. This observation supported the use of normal versus normal analysis to set tumor-specific p-value cutoffs, which improved our control of genome-wide false positives.

Analytical power calculations

We used the approximation based on ratios of normals, $f(x)$, to model copy-number ratios in genomic windows of a fixed size. As seen in Fig. 1 and Supplementary Figure 13, this approximation yielded a conservative estimate of the power. For simplicity, we assumed that the same number of aligned reads were obtained from the tumor and normal ($a = 1$). The expected number of reads, $\lambda_N = a_N \times A / L$, were calculated for genomic windows varying in size from $L = 10{,}000$ to $L = 100{,}000$, as measured in coordinates of the alignable genome. We calculated the power to detect a single-copy gain in a diploid sample ($r = 1.5$), as well as a single-copy loss ($r = 0.5$). The critical values r_{low} and r_{high} were chosen to yield 1 false positive gain and 1 false positive loss across the A / L windows in the genome of size A. The power equals the percent of the distribution for R (in the alternative hypothesis) that exceeded r_{high} or below r_{low} for gains and losses, respectively.

Nature Methods: doi: 10.1038/nmeth.1276

Segmentation algorithm for the identification of copy-number alterations

Initialization. We consider every tumor read position and keep a short-list of candidate breakpoints with extreme values of the log-ratio difference statistic. For each tumor read position, b, we defined local windows that included exactly w consecutive reads in the normal sample to the left or right of position b. In other words, $N(x_L, b) = N(b, x_R) = w$. Next, we calculated D_x for all positions of aligned tumor reads on a particular chromosome. We then used the following procedure to initialize breakpoints: (i) the position $b = \min_x p\left(\left|D_x\left(x_L, x_R\right)\right|\right)$ was added to the list of candidate breakpoints; (ii) we removed from consideration all sequence reads on either side of the candidate breakpoint by setting $D_i(x_L, x_R) = 0$, where $b - w \le i \le b + w$. We chose additional candidate breakpoints by repeating the above two steps, until $p\left(\left|D_x\left(x_L, x_R\right)\right|\right) > p_{bkp}$. We performed the same procedure on the aligned sequence reads from the matched normal sample and set p_{bkp} such that there were exactly 1000 initial breakpoints.

Iterative merging of adjacent segments. Let $B^c = \{ b_1, b_2, \ldots, b_N \}$ denote the ordered list of candidate breakpoints for chromosome c. Each breakpoint k was associated with a left window from b_{k-1} to $b_k - 1$, and a right window from b_k to b_{k+1}. In an iterative procedure, the number of tumor and normal reads in each flanking window was compared to the null hypothesis of proportional distribution of reads across the breakpoint. The two segments on either side of the least significant breakpoint were merged, and the p-values for the flanking breakpoints b_{k-1} and b_{k+1} were updated. This merging procedure continues until $p\left(\left|D_{b_k}\left(b_{k-1}, b_{k+1}\right)\right|\right) < p_{merge}$. We chose the final p-value cutoff by merging the 1,000 initial breakpoints from the matched normal sample and setting p_{merge} such that there were exactly 10 false positive final segments.

Simulated copy-number alterations for parameter optimization

Assume that a_c sequence reads from a normal sample aligned to chromosome c. We created virtual reads at the midpoints between all pairs of adjacent reads and pooled these with the original reads. For each of 1,000 iterations, we randomly partitioned the sequence reads from this pool to a "reference" and an artificial "tumor". Next, we picked a random chromosomal position, m. We spiked in a copy-number alteration of size L by removing all sequence reads between positions m and $m + L$ in the artificial "tumor" sample, and replaced them with the corresponding reads in the actual tumor sample. For copy gains, we used chromosome 5 in the

Nature Methods: doi: 10.1038/nmeth.1276

NCI-H232347 cell line (copy ratio ~1.42); for copy losses, we used chromosome 16 (copy ratio ~0.59). We applied our segmentation procedure on this artificial tumor chromosome with local windows ranging in size from $w = 100$ to $w = 1{,}000$. We recorded a true positive if the segmentation algorithm correctly predicts breakpoints b_i and b_{i+1} that are both within w normal reads of m and $m + L$, respectively. Any additional predicted breakpoints were considered as false positives.

Copy-number inference for Affymetrix Genome-Wide Human SNP Array 6.0

Genotypes and hybridization intensities for over 906,600 single nucleotide polymorphisms and over 946,000 probes for copy number variants were measured with the Affymetrix Genome-Wide Human SNP Array 6.0. We inferred copy numbers from the raw intensity (CEL) files with a GenePattern pipeline that incorporated the following steps[3,4]. First, the SNPFileCreator module estimated a single value for each probeset representing a SNP allele or copy number probe via brightness correction, model-based expression and median polish. Second, the CopyNumberInference module estimated copy numbers for copy number probes from an X-chromosome dosage extrapolation[3] and for SNP probesets with allele-specific cluster centers from the Birdseed algorithm[5]. After removing outliers, copy numbers were divided by the 5 closest normal samples, as measured by minimum Euclidean distance in the $\log_2$ space.

For each genomic locus, we calculated the median copy number among replicate arrays of the HCC1954 ($n = 14$), HCC1954BL ($n = 13$), HCC1143 ($n = 22$) or HCC1143BL ($n = 23$) cell lines that were hybridized and scanned on different days. For the NCI-H2347 and NCI-H2347BL cell line, we used the copy numbers calculated for a single array. We optimized the parameter sdUndo = 2 by limiting the number of false positive segments to 10 per matched normal cell line, after merging segments less than 8 consecutive probe sets long (Supplementary Fig. 14). With these parameters, we found a total of 454 segments in the HCC1954 cell line, 300 segments in the HCC1143 cell line, and 70 segments in the NCI-H2347 cell line.

---

[3] Monti S., O'Kelly M.J.T., Stransky N., Gould J., Twomey D., Nadel M., Winckler W., Meyerson M., Getz G. Manuscript in preparation

[4] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068 (2008).

[5] Korn J.M. *et al*. Integrated genotype calling and association analysis of SNPS, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253-1260 (2008).

Nature Methods: doi: 10.1038/nmeth.1276

Comparison of copy-number alterations with single nucleotide polymorphism arrays

In order to compare the segmentation results from sequencing and SNP arrays, we first combined the set of chromosomal breakpoints predicted by each method. Predicted breakpoints from two different methods that were found within 50 kb were considered to be redundant, and the left-most breakpoint was retained. For each chromosomal segment defined by a pair of adjacent breakpoints, we calculated the copy-number ratios from sequencing data as $R = \frac{T / a_T}{N / a_N}$. For the microarray data, we identified the closest chromosomal segment predicted by the Circular Binary Segmentation algorithm[28]. The microarray copy-number ratio for the segment was taken as the geometric mean of the copy-number ratios for probe sets within the boundaries of the closest segment.

Parameter optimization for circular binary segmentation of SNP 6.0 arrays

We merged short segments in order to filter false positives that arise from probes with outlier values. A small subset of probes on any microarray may yield extremely high values (due to cross-hybridization) or a extremely low values (due to poor hybridization affinities). As negative controls, we considered microarray data from two diploid cell lines. First, we removed probes within known germline copy-number variants. Next, we used the Circular Binary Segmentation algorithm with different values of the "SD-undo" pruning parameter, from 1 to 10. Finally, we merged predicted copy-number segments that included fewer than *N* probes, where *N* ranged from 3 to 60.

We chose the most lenient segmentation parameters that yielded the fewest number of false positive segments predicted in these diploid cell lines (Supplementary Fig. 14). At all values of the "SD-undo" parameter, the minimum length of 8 consecutive probe sets led to the sharpest drop in false positives. In addition, we found that an SD-undo parameter of 2 leads to 22, 16 and 15 false positives in the HCC1954BL, HCC1143BL and NCI-H2347 cell lines, respectively.

Nature Methods: doi: 10.1038/nmeth.1276

Langmuir adsorption model for attenuation of oligonucleotide arrays

The Langmuir adsorption equation models the adsorption of molecules (*e.g.*, genomic DNA) on a solid surface at a constant temperature. Thus, the signal intensity for an oligonucleotide probe on a single-color array will depend on the copy-ratio, *r*:

$$I \propto \frac{\alpha r}{\alpha r + 1}$$

Note that probe-specific effects (such as G+C bias) will be common to the normal and the tumor sample. Since $r = 1$ in a normal sample, the copy-number ratio will be expressed as:

$$\begin{aligned} R &= \frac{I_T}{I_N} \\ &= \frac{\frac{\alpha r}{1+\alpha r}}{\frac{\alpha}{1+\alpha}} \\ &= \frac{1+\alpha}{1+\alpha r} r \end{aligned}$$

Note that $\lim_{r\to\infty} \frac{1+\alpha}{1+\alpha r} r = \frac{1+\alpha}{\alpha}$, which corresponds to the asymptotic saturation level for the array. Thus, we define the parameter $\alpha' = \frac{1+\alpha}{\alpha}$ and re-write the above equation as $R = \frac{\alpha' r}{r + \alpha' - 1}$

Nature Methods: doi: 10.1038/nmeth.1276