Supporting Information

Lee et al. 10.1073/pnas.1106233109

SI Text

Code for Lowess and Local Normalization. To remove variation in log intensity data caused by array spatial bias we perform local normalization independent for red and green channels using an algorithm described by pseudocode below.

```
for each channel
for each probe
subtract from the probe value the mean value of all autosomal probes that fall inside a square of size 17 centered at the probe.
end probe
end channel
```

To remove die bias we apply lowess (R package) subroutine as shown in pseudocode below.

```
for each channel
x is array of locally normalized log intensities of the channel (autosomal probes)
y is array of averages for locally normalized log intensities of both channels (autosomal probes)
z is array of locally normalized log intensities of the channel (sex chromosomes)
lowess returns new values for x
xnew = lowess(x,y)
for each w in z
if w > max(x)
wnew = max(x)
else if < min(x)
wnew = min(x)
else
[x1, x2] is smallest interval of autosomal log intensities covering value w
wnew = interpolate(w, x1,x2,x1new,x2new)
end z
end channel
```

MATLAB Code for Archive Generation, Principal Component Correction (PCC), Piecewise Principal Component Correction (PPCC), and G-C Correction (GCC).

```
%
% To simplify the code all procedures related to data IO are omitted.
% This file contains several matlab files concatenated together.
% These files are:
% archive_generation.m
% PCC_PPCC_drivers.m
% PCC.m
% PPCC.m
% GCC.m
%
% archive_generation.m file
%
% The following code assumes that SSH log ratio locally normalized data is
% loaded into array Yln; in the paper Yln is 2161679 x 132 array.
% auto is a vector of autosomal probe indices, in the paper it is 1:2038584
%
%%
% Creation and saving archive for PCC
%
[U S V] = svd(Yln, 'econ');
A = U(:, 1:14);
%%
save 'A.mat';
%%
% Creation and saving archive for PPCC
%
[dummy ia] = sort(abs(A(:,9)));
rperm = ia;
t = [0 6000:50000:2161679 2161679];
% The first and the last slices are smaller than 50000 probes,
% the rest are 50000 probes
slices = cell(length(t)-1, 1);
slicesa = cell(length(t)-1, 1);
for n = 1:(length(t)-1)
disp(n)
slices{n} = rperm((t(n)+1) : t(n+1))';
slicesa{n} = intersect(auto, slices{n})';
end
slices = slices';
slicesa = slicesa';
AA = cell(length(t)-1, 1);
AAa = cell(length(t)-1, 1);
for n = 1:(length(t)-1)
disp(n)
[U S V] = svd(Yln(slices{n}, :), 'econ');
AA{n} = U(:,1:14);
[dummy ia ib] = intersect(slicesa{n}, slices{n});
AAa{n} = U(ib,1:14);
end
AA = AA';
AAa = AAa';
%%
save 'slices.mat' slices;
save 'slicesa.mat' slicesa;
save 'AA.mat' AA;
save 'AAa.mat' AAa;
clear *
% end of archive_generation.m file
% PCC_PPCC_drivers.m file
%
%%
archives = {'AA.mat' 'AAa.mat'};
chunks = {'slices.mat' 'slicesa.mat'};
load('profile.mat')
result = PPCC(profile, archives, chunks);
%%
load('profile.mat');
load('archive.mat');
load('autosome_probes.mat')
result = PCC(profile, archive, autosome_probes);
% end of PCC_PPCC_drivers.m file
% PCC.m file
%
% PCC computes residual after projecting profile on the subspace spanned by
% archive column vectors
%
% profile - column vector of log of locally normalized hybridization ratio data.
% archive – matrix of principal components of SSH log ratio data.
% autosome_probes – vector of indices of autosome probes
%
%
function result = PCC(profile, archive, autosome_probes)
N = size(profile);
M = size(archive);
if N(1) ~= M(1)
disp ('number of rows in profile and archive should be equal')
return;
end
result = profile - archive*regress( profile(autosome_probes) , archive(autosome_probes, :));
%end of PCC.m file
% PPCC.m file
%
% PPCC computes residual after projecting profile on the subspace spanned by
% archive column vectors
%
% profile – column vector of log of locally normalized hybridization ratio data.
% archives – cell array containing names of archive matlab files (all and autosomal parts).
% chanks – cell array containing names of probe chunks (all and autosomal parts).
%
%
function result = PPCC(profile, archives, chunks)
load(chunks{1});
load(chunks{2});
load(archives{1});
load(archives{2});
N = size(profile);
result = zeros(N(1),1);
for m = 1:length(slices)
result(slices{m}) = profile(slices{m}) - AA{m}*regress( profile(slicesa{m}) , AAa{m});
end
% end of PPCC.m file
% GCC.m file
%
% GCC computes residual after adjusting for the GC content of the probes
%
%
% profile – column vector of log of locally normalized hybridization ratio data.
% gcsets – cell array containing arrays of probes of equal GC content.
% autosome_probes – vector of indices of autosome probes
%
%
function result = GCC(profile, gcsets, autosome_probes)
M = median(profile);
for n = 1:length(gcsets)
profile(gcsets{n}) = profile(gcsets{n}) - median(profile(intersect(gcsets{n}, autosome_probes)));
end
result = profile + M;
%end of GCC.m file
```

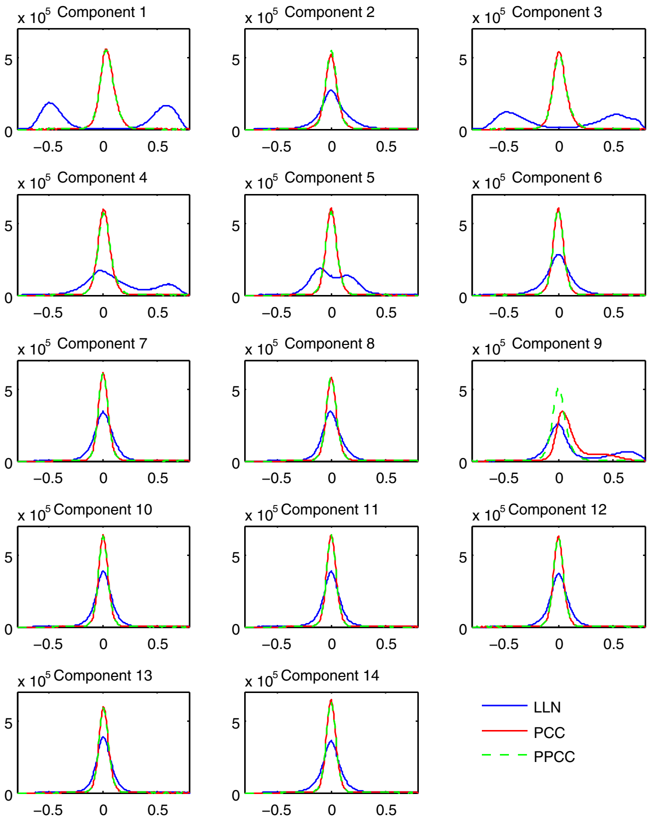

Fig. S1. Extent of probe correction following PCC and PPCC. For each of 14 components, a matrix of log ratios was created, consisting of 1,500 columns, one for each hybridization of parents, with about 4,200 rows, one for each probe with extreme loadings (most positive and negative 0.1% of values). Pearson correlations were computed between all pairs of rows. Histograms of these correlations are shown for all components before and after PCC or PPCC. The bin size for the histograms is 0.005.

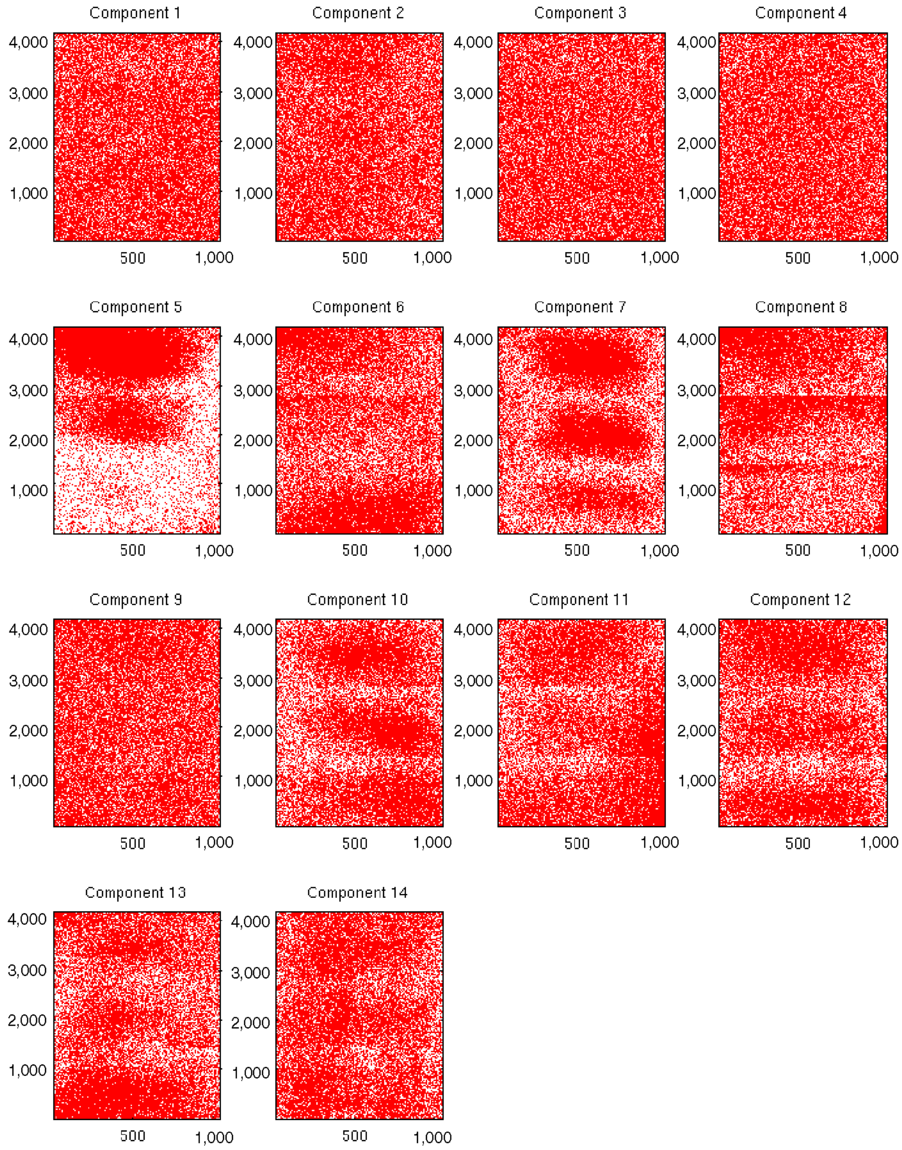

Fig. S2. Patterns of spatial distribution of extreme probes on CGH microarrays. Array coordinates for extreme probes (with topmost 1.5% loadings) in each of 14 principal components are displayed as the *X* and *Y* axes in each of 14 plots. Spatial clustering for components 5–8 and 10–14 reflects that the arrays are printed in three blocks, and each block is processed in a separate hybridization chamber. Components 1–4 and 9 do not show any dependence on array coordinates.

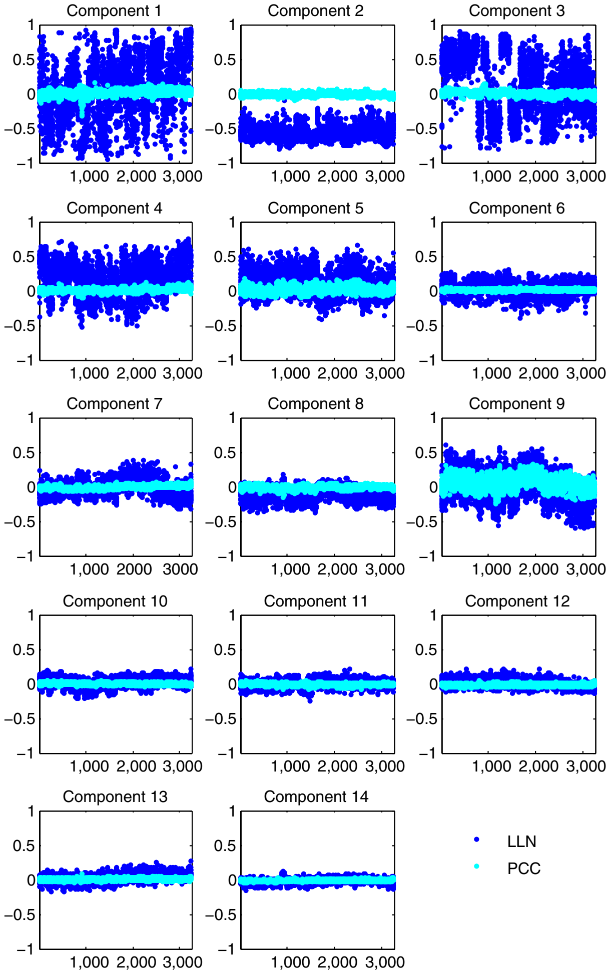

Fig. S3. Correlation of major principal components with hybridizations over time. The displayed plots indicate the correlation (*Y* axis) between the autosomal extreme (most positive and negative 1.5%) loadings of the indicated principal component with the log ratios of the probes with those extreme loadings for each 3,252 test-reference hybridizations, ordered by the "queue index" (*X* axis). Dark blue represents before (LLN) and light blue represents after PCC.

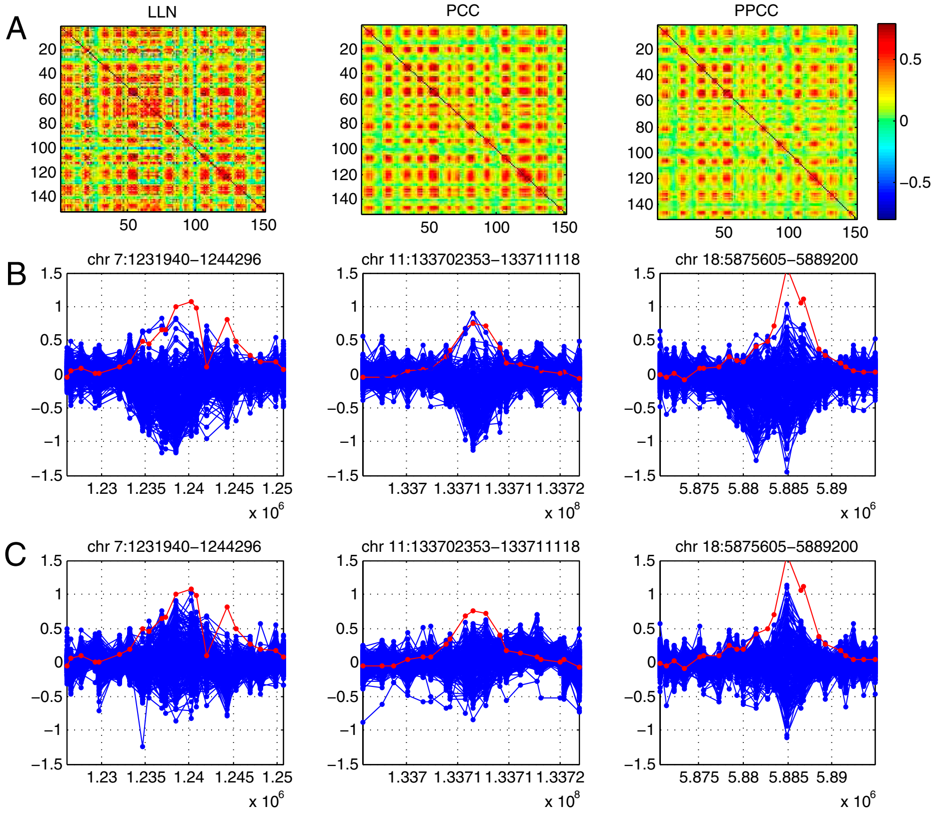

Fig. S4. Some false positive events are highly correlated with regions of extreme values of component 9. In panel (*A*), a heat map of correlations of log probe ratio for 151 probes covering 12 regions spread across genome, drawn from 313 array hybridizations of chronic lymphocytic leukemias (CLL). The correlation marices are: after LLN; after PCC; and after PPCC. In (*B*), LLN data from all CLL for three of twelve regions represented in panel (*A*) are displayed in blue together with scaled component 9 in red. (*C*) is similar, except the same regions as in *B* are from 313 random SSC hybridizations.

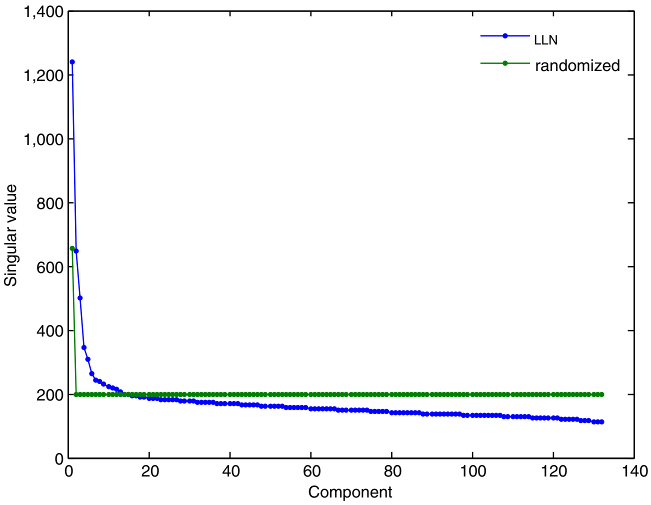

Fig. S5. Singular values for self-self hybridizations. Singular values for 132 self-self hybridizations were computed before (blue) and after (green) within-row-permutation (see text). To find the number of major principal components, we randomly permuted ratio values for each probe and recomputed the singular values of the resulting matrix. The intersection between the original and permuted singular values occurs at component 14.

Table S1. The impact of PCC on segmentation

	Mean	Median	Std
LLN	112.3	28.5	202.9
PCC	2.9	0	9.4

For each of 132 SSH, we created an archive of all other SSH removing the data from the respective hybridization. We then applied PCC using the first 14 components of the archive. After segmentation we computed mean, median, and standard deviation for the number.

Table S2. Compositional biases of principal components

Comp	A.neg	A.pos	C.neg	C.pos	G.neg	G.pos	T.neg	T.pos	Total Diff
1	**0.2981**	**0.2253**	**0.1998**	**0.2832**	**0.1829**	**0.2586**	**0.3192**	**0.2329**	**0.3183**
2	**0.3423**	**0.2451**	0.2103	0.2185	0.2020	0.2038	**0.2454**	**0.3326**	**0.1943**
3	**0.2416**	0.2882	**0.2635**	0.2121	**0.2226**	0.2057	**0.2723**	0.2941	**0.1368**
4	**0.2972**	**0.2306**	0.2191	**0.2575**	0.2024	**0.2281**	**0.2814**	**0.2838**	**0.1332**
5	**0.2624**	**0.3101**	**0.2385**	**0.1807**	**0.2230**	**0.1696**	**0.2761**	**0.3396**	**0.2222**
6	0.2896	0.2727	0.2233	0.2058	0.2061	0.1987	**0.2810**	**0.3228**	**0.0835**
7	0.2719	0.2898	0.2266	**0.1894**	0.1960	**0.1850**	0.3055	**0.3358**	**0.0964**
8	0.2869	0.2844	0.2172	0.2106	0.1901	**0.1872**	0.3058	**0.3177**	0.0238
9	**0.2475**	0.2841	0.2199	0.2258	0.1937	0.2086	**0.3390**	**0.2815**	**0.1149**
10	0.2826	0.2869	0.2145	**0.1982**	0.1984	**0.1842**	0.3045	**0.3307**	**0.0608**
11	**0.3011**	0.2694	**0.1831**	0.2175	**0.1841**	0.1962	**0.3316**	**0.3169**	**0.0929**
12	**0.2520**	**0.3186**	0.2161	**0.2000**	0.1912	**0.1886**	**0.3407**	0.2928	**0.1333**
13	0.2866	0.2860	0.2127	0.2050	**0.1823**	0.1982	**0.3184**	0.3108	0.0318
14	0.2718	0.2943	0.2181	0.2038	0.1932	0.1946	**0.3169**	0.3073	0.0476
Range in 1,000 random simulations	[0.2643, 0.2966]		[0.2038, 0.2324]		[0.1888, 0.2168]		[0.2850, 0.3155]		[0.0000, 0.0587]

For each of the 14 components, we computed the proportion of A, C, G, or T in those extreme probes with the 1.5% most negative or 1.5% most positive loadings, yielding "*.neg" and "*.pos," respectively. "Total Diff" is defined as the sum of the absolute values of the differences between *.pos and *.neg for each of the four bases. For each of 1,000 simulations, a random subset of 1.5% of probes (32,425 probes/simulation) was created and the range of the proportions of the four nucleotides was computed, creating the confidence intervals ($p = 10^{-3}$) shown in the bottom row of cells. The bottom right cell contains the confidence interval ($p = 10^{-3}$) of the total difference computed over 1,000 pairs of random subsets of 32,425 probes. Compositions outside the range ($p < 10^{-3}$) are in bold.

Table S3. Characterization of principal components

Comp	SV	Skew	Kurtosis	AC	Autocorr	Total ND	GC
1	1,238.0	−0.48	0.26	−	0.35	0.3183	−
2	648.6	−0.14	0.54	−	0.05	0.1943	−
3	500.6	−0.01	0.33	−	0.18	0.1368	−
4	346.1	0.23	0.62	−	0.09	0.1332	−
5	309.8	0.34	1.12	+	0.01	0.2222	−
6	263.7	−0.02	0.42	+	0.03	0.0835	−
7	244.9	0.22	0.81	+	0.04	0.0964	−
8	237.4	0.13	0.93	+	0.03	0.0238	−
9	232.2	3.69	46.52	−	0.33	0.1149	+
10	223.2	0.07	0.53	+	0.02	0.0608	−
11	219.7	0.29	1.36	+	0.01	0.0929	−
12	214.9	0.10	0.70	+	0.02	0.1333	−
13	206.5	0.05	0.70	+	0.01	0.0318	−
14	199.6	0.01	0.55	+	0.01	0.0476	−

Column guide: Comp = component number; SV = singular value, diagonal values of matrix D in formula (Eq. 2); Skew = skewness; Kurtosis = excessive kurtosis ; AC = array clustering observed in extreme loading values of component; Autocorr = autocorrelation, which is computed as correlation of the ratio vector shifted by one probe; Total ND = measure of nucleotide bias of probes with extreme loadings as described in Table S2, last column; GC = gene clustering defined as the overlap of extreme probes with regions of gene transcription starts near CpG islands. For this table, we regard probes with top or bottom 1.5% loadings in the specified component as "extreme."

Table S4. Probe cluster intervals overlapping with the 5′ ends of genes and/or CpG islands

Component	Cluster	Polarity	Gene & CpG	Gene Only	CpG Only
1	714	+	0.10	0.08	0.22
1	30	–	0.00	0.00	0.00
2	143	+	0.55	0.02	0.11
2	59	–	0.05	0.05	0.19
3	111	+	0.00	0.04	0.01
3	399	–	0.17	0.06	0.35
4	484	+	0.12	0.11	0.20
4	22	–	0.18	0.05	0.14
5	34	+	0.00	0.00	0.00
5	9	–	0.11	0.00	0.22
6	36	+	0.44	0.03	0.08
6	20	–	0.10	0.00	0.10
7	117	+	0.50	0.01	0.15
7	22	–	0.00	0.05	0.00
8	83	+	0.59	0.05	0.14
8	34	–	0.03	0.00	0.03
9	3,415	+	0.54	0.03	0.14
9	11	–	0.00	0.00	0.00
10	24	+	0.13	0.00	0.08
10	15	–	0.00	0.07	0.00
11	19	+	0.16	0.00	0.00
11	28	–	0.00	0.00	0.00
12	19	+	0.00	0.00	0.00
12	55	–	0.51	0.04	0.04
13	30	+	0.03	0.03	0.00
13	8	–	0.00	0.00	0.00
14	25	+	0.00	0.04	0.00
14	21	–	0.19	0.05	0.10

The column "Cluster" shows the total number of probe cluster intervals, defined as a maximally contiguous set of at least three probes from the top (+) and bottom (–) extreme 1.5% of loadings ("Polarity"), from each major component. The last three columns show the proportion of probe cluster intervals from the top and bottom extreme loadings for each principal component that overlap both 5′ ends of genes and CpG islands ("Gene & CpG"), 5′ end of genes only ("Gene Only"), and CpG islands only ("CpG Only"). While many components have clusters similar to component 9 in the proportion distributing to the 5′ ends of genes and CpG islands, none have these clusters in the abundance seen in the ninth.