Wang *et al.* Supplemental Information

Supplemental Tables

base pairs	mode	maps/read	S/N	autocorrelation
36	SE	0.727	5.609	0.023
36	PE	0.703	5.606	0.021
36	PE as SE	0.798	5.647	0.020
50	SE	0.746	5.620	0.019
50	PE	0.710	5.621	0.018
50	PE as SE	0.807	5.618	0.020
76	SE	0.706	5.656	0.016
76	PE	0.632	5.596	0.019
76	PE as SE	0.751	5.640	0.018
101	SE	0.560	5.562	0.019
101	PE	0.480	5.438	0.020
101	PE as SE	0.603	5.601	0.018
101	SMASH	3.096	5.289	0.019

Supplemental Table S1. WGS mode optimization for CNV detection. To determine the optimal read length and mapping method for copy number by low-coverage WGS, we clipped 101 bp paired-end reads to lengths equivalent of common run types. Reads were mapped using Bowtie in the following modes: (1) “SE,” uses read 1 only; (2) “PE,” uses read 1 and read 2 mapped in paired-end mode keeping only concordant pairs; and (3) “PE as SE,” uses read 1 and read 2, selecting the map from read 2 when read 1 is unmapped. To determine the performance characteristics of each mapping method, we select 10 million maps across 100 thousand bins to compute signal to noise (“S/N”) and autocorrelation as described in Table 1. SMASH data are shown here for comparison with values as in Table 2 for 100,000 bins at 100 maps per bin.

sample	all read	ignore 3′ excess	ignore all excess
12596_fa	0.758	0.880	0.977
12596_mo	0.756	0.879	0.977
12596_p1	0.754	0.876	0.977
12596_s1	0.763	0.884	0.977
12605_fa	0.750	0.874	0.976
12605_mo	0.756	0.879	0.977
12605_p1	0.756	0.880	0.977
12605_s1	0.755	0.879	0.977
skbr3	0.737	0.850	0.926

Supplemental Table S2. Proportion of bases in read 1 that pass the 20:4 rule. For each sample, we compute the proportion of bases that are included in a valid 20:4 map ("all read"). We also consider just the proportion of bases in a valid map excluding the bases that occur at the end of the read ("ignore 3′ excess"). This is for two reasons: 1) the 3′ end of the read is typically of lower quality, and 2) the fragment at the end of the read may not have a sufficient number of bases to qualify as a match. If we exclude unmapped bases on both ends of the read ("ignore all excess"), then nearly every base is used (>97%).

type	# of bins	maps per bin	autosomal auto-correlation	autosomal median	X Chrom median	autosomal MAD	X Chrom MAD	signal to noise	S/N from Table 2
SMASH	50000	20	0.006	2.020	1.120	0.295	0.161	2.679	2.497
WGS	50000	20	0.002	2.021	1.079	0.296	0.160	2.800	2.659
SMASH	50000	50	0.009	2.005	1.074	0.193	0.106	4.234	3.933
WGS	50000	50	0.008	2.005	1.036	0.192	0.099	4.480	4.173
SMASH	50000	100	0.015	2.002	1.060	0.140	0.077	5.886	5.487
WGS	50000	100	0.013	2.002	1.026	0.140	0.074	6.149	5.861
SMASH	100000	20	0.003	2.020	1.116	0.298	0.159	2.681	2.481
WGS	100000	20	0.005	2.020	1.078	0.299	0.157	2.788	2.633
SMASH	100000	50	0.016	2.006	1.074	0.197	0.108	4.148	3.833
WGS	100000	50	0.008	2.005	1.039	0.197	0.102	4.368	4.102
SMASH	100000	100	0.024	2.002	1.061	0.146	0.078	5.697	5.289
WGS	100000	100	0.017	2.002	1.024	0.145	0.077	5.979	5.633
SMASH	500000	20	0.009	2.021	1.113	0.317	0.170	2.524	2.342
WGS	500000	20	0.008	2.020	1.078	0.315	0.163	2.657	2.492
SMASH	500000	50	0.018	2.006	1.075	0.225	0.116	3.677	3.410
WGS	500000	50	0.017	2.005	1.037	0.222	0.109	3.906	3.656
SMASH	500000	100	0.028	2.001	1.060	0.182	0.090	4.622	4.319
WGS	500000	100	0.026	2.001	1.025	0.179	0.084	4.939	4.649

Supplemental Table S3. Empirical bin boundaries. The computations of Table 2 are repeated using bins of uniform observation of a reference instead of bins of uniform expectation. In this approach, bin boundaries are defined empirically; bins are established with the same number of maps as determined empirically. The signal to noise is improved over the results in Table 2 (“S/N from Table 2”), with little change to autocorrelation.

type	# of bins	maps per bin	autosomal auto-correlation	autosomal median	X Chrom median	autosomal MAD	X Chrom MAD	signal to noise
SMASH	50000	20	0.002	2.027	1.112	0.416	0.272	1.840
SMASH	100000	20	0.004	2.028	1.122	0.415	0.279	1.812
SMASH	500000	20	0.003	2.025	1.112	0.417	0.279	1.819
SMASH	50000	50	0.004	2.019	1.075	0.270	0.176	2.928
SMASH	100000	50	0.005	2.018	1.077	0.269	0.180	2.907
SMASH	500000	50	0.011	2.019	1.076	0.271	0.178	2.908
SMASH	50000	100	0.005	2.003	1.058	0.190	0.124	4.160
SMASH	100000	100	0.009	2.005	1.061	0.190	0.126	4.130
SMASH	500000	100	0.021	2.005	1.059	0.194	0.128	4.073
WGS	50000	20	0.004	2.034	1.080	0.413	0.261	1.953
WGS	100000	20	0.001	2.032	1.075	0.414	0.261	1.955
WGS	500000	20	0.003	2.030	1.078	0.413	0.263	1.944
WGS	50000	50	0.003	2.017	1.036	0.267	0.164	3.132
WGS	100000	50	0.006	2.016	1.039	0.266	0.168	3.111
WGS	500000	50	0.012	2.016	1.040	0.267	0.168	3.095
WGS	50000	100	0.008	2.003	1.017	0.189	0.114	4.452
WGS	100000	100	0.011	2.003	1.022	0.189	0.118	4.400
WGS	500000	100	0.025	2.003	1.024	0.189	0.119	4.385

Supplemental Table S4. Matching reference and sample coverage. Performance statistics are computed as in Table 2, but here the same number of maps for both the sample and the reference for each choice of bin resolution (50K, 100K, 500K) and for each map coverage (20, 50 and 100 reads/bin) are used. When the number of maps is equalized between sample and reference, the signal to noise is largely insensitive to bin resolution and depends strongly on map coverage for both WGS and SMASH, indicating that resolution is only limited by the depth of coverage.

Supplemental Table S5. SMASH and WGS segmentation comparisons (.xlsx file). The spreadsheet contains four labeled worksheets, containing the segmentation data depicted in Figs. 3 and 4. Both the family and SKBR3 data are normalized and segmented using either the SMASH or the WGS data to determine segment boundaries. For those segments, we compute the mean and standard deviation of the bins contained in that segment for both WGS and SMASH providing a detailed comparison of segment averages. Overall, SMASH and WGS are strongly concordant regardless of the data used for segmentation.

run type	mode	maps /read	lane cost	cost/genome (20 kb res)	cost/genome (10 kb res)	cost/ library	total cost (20 kb)	total cost (10 kb)
1x36	SE	0.727	1162	47.93	95.87	16.95	64.88	112.82
2x50	PE as SE	0.807	1819	67.58	135.17	16.95	84.53	152.12
2x125	SMASH	4.470	2670	17.92	35.84	18.10	36.02	53.94

Supplemental Table S6. SMASH and WGS analysis cost comparisons. We calculated costs (USD) of sample analysis on a HiSeq 2500 instrument (Illumina) using 2 optimal run modes for WGS and SMASH. Costs are calculated at current (early 2016) levels using CSHL Genome Center pricing. SMASH and WGS library preparation is similar in cost and time, but the increased number of maps/read from SMASH makes it substantially more cost-effective than low coverage WGS for high-resolution copy number analysis. Average bin size is expressed in kb.

Supplemental Methods

SMASH2 Protocol

The revised SMASH protocol utilizes a single NEBNext dsDNA fragmentase (NEB, M0348L) digestion step in place of the sonication steps and restriction enzyme digestion from the original protocol. 200 ng of genomic DNA are digested for 10 min at 37°C in a 10 µl reaction with 1 µl of dsDNA fragmentase, 1 µl 10X dsDNA fragmentase reaction buffer and 0.5 µl 20mM $MgCl_2$. Following this step, end repair is performed for 30 min at 20°C in the same reaction tube by adding 2 µl 10 mM dNTPs, 1 µl (3U/µl) T4 DNA pol, 0.5 µl Klenow (5U/µl), 1 µl (10U/µl) T4 PNK and 4.5 µl H_20 to a total reaction volume of 20 µl. Following this step, the reaction mix is size-selected with 2.5 volumes of AMPure XP beads and purified via the Qiagen nucleotide removal kit. Subsequent steps are identical to the original protocol (Methods).

In addition to the changes noted above, all enzyme heat-killing steps have been eliminated from the SMASH2 protocol. The overall time required for the SMASH2 protocol has been reduced by ~1 hour and cost reduced by nearly 50% relative to the original protocol.

Source Code and Bin Boundaries

The 1) longMEM package, SMASH mapping heuristics, map counting and copy number profiles and 2) bin boundaries are available as respective .zip archives.

1) Wang et al._Supplemental_Information__code.zip
2) Wang et al._Supplemental_Information__bins.zip