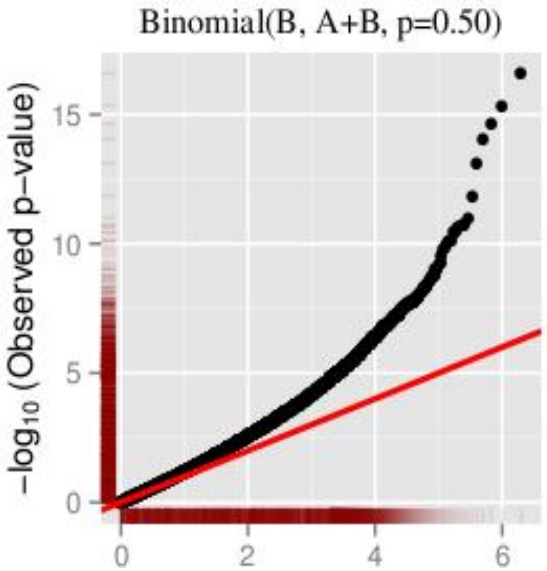
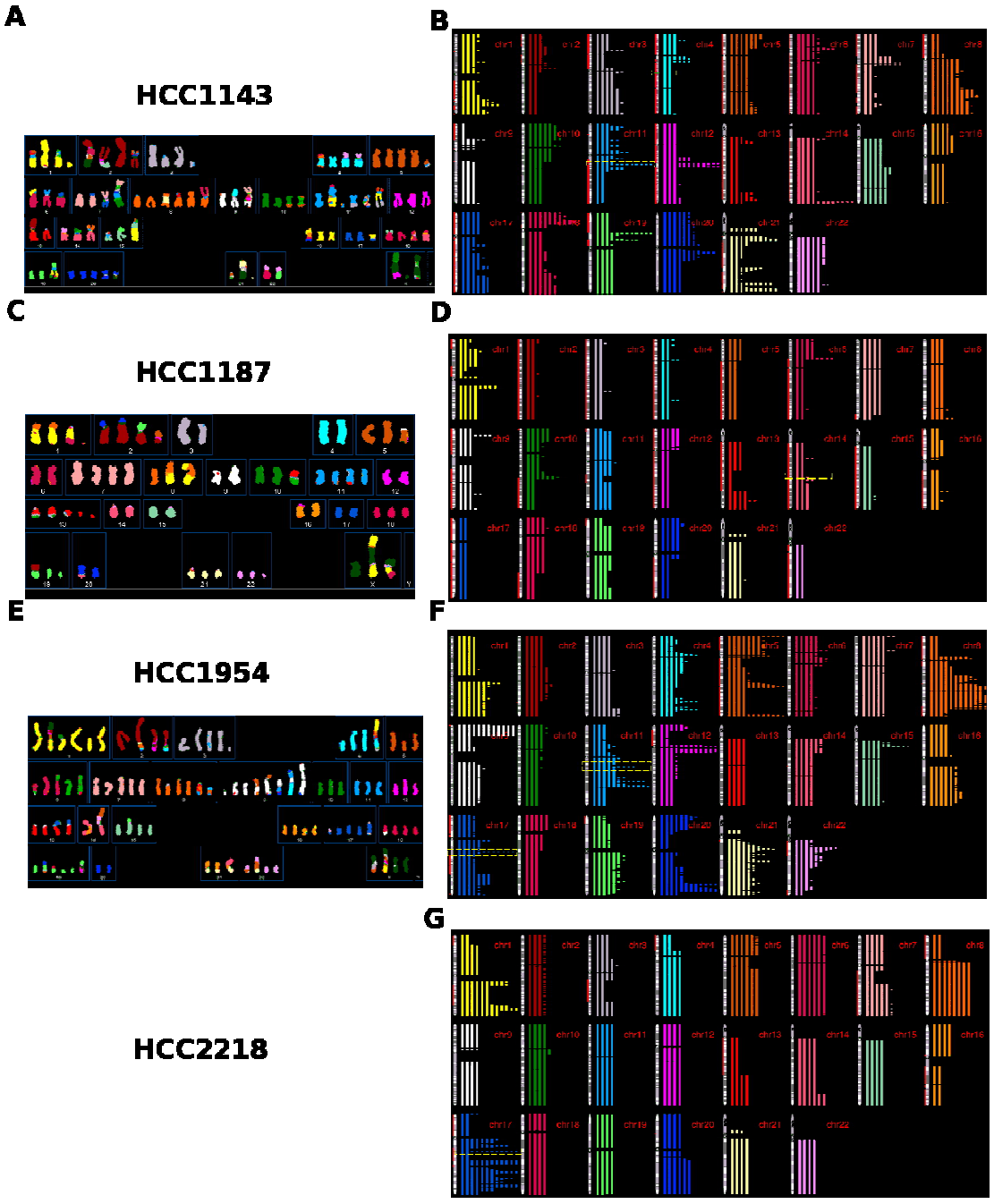
Supplementary Table 1.
A list of tools designed for copy number profile for impure tumor samples.

Tool	TC	Ploidy parameter	Paired Normal	GC-bias corr.	LOH	Data type	Segment based on	References
Absolute	Y	Y	Y	-	Y	Array NGS	HAPSEG	(Carter et al. 2011; Carter et al. 2012)
OncoSNP	Y	N	Y	Y	Y	Array	itself	(Yau et al. 2010)
PyLOH	Y	N	Y	Y	Y	NGS	BIC-seq	(Xi et al. 2011; Li and Xie 2014)
CLImAT	Y	N	N	Y	Y	NGS	itself	(Yu et al. 2014)
THetA2	Y	N	Y	Y	Y	NGS	BIC-seq	(Oesper et al. 2014)
OncoSNP-Seq	Y	N	Y	Y	Y	NGS	itself	(Yau 2013)
CNAnorm	Y	Y	Y	Y	N	NGS	itself	(Gusnanto et al. 2012)



Supplementary Figure 1. Inflation of test-statistic for detecting somatic copy number aberation for normal sample under null assumption of BAF = 0.5**.** Each of the black dots in the Q-Q plots corresponded to a P-value for a pair of allelic depths under different assumptions. P-values were calculated under the assumption that both alleles had equal copy numbers for a normal samples. The red line, y=x, is the expected fitness under null hypothesis.



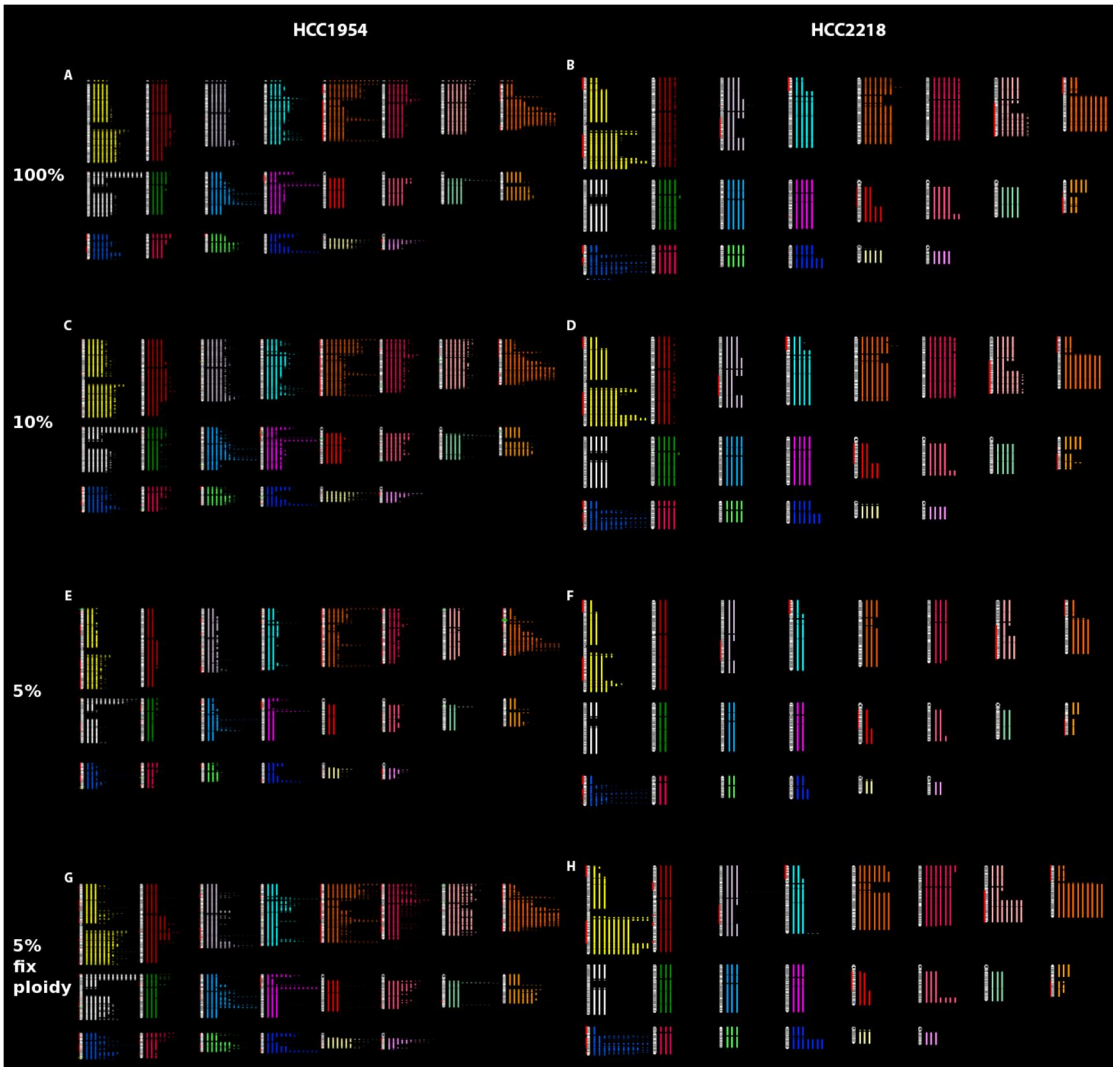
Supplementary Figure 2.
dSKY plots capture the same large scale copy number change present in existing SKY experiments. (A), (C) and (E) all show the results from a traditional SKY analysis of the cell-lines characterised in this study. Each chromosome is represented by a specific colour and are grouped together, making it possible to count the number of copies of each chromosome. This work was completed by Mira Grigorova & Paul A.W. Edwards (Department of Pathology, University of Cambridge) - see http://www.path.cam.ac.uk/~pawefish - and is reproduced with permission. We were unable to find publicly available results from a SKY analysis of HCC2218. To allow a direct comparison between the result produced by sCNAphase and the SKY experiments, each chromosome was painted in the same colour used in the traditional SKY analysis and are presented in a similar layout. (B),(D),(F) and (G) all show the corresponding dSKY plots. To show additional information, such as the location of each sCNA and the regions that have undergone an LOH event, a diagram of each chromosome is present in the dSKY plot. This diagram is coloured red to indicate an LOH event or green to highlight a homozygous

deletion. High resolution dSKY plots can be found from figshare at https://figshare.com/authors/_/1365237.

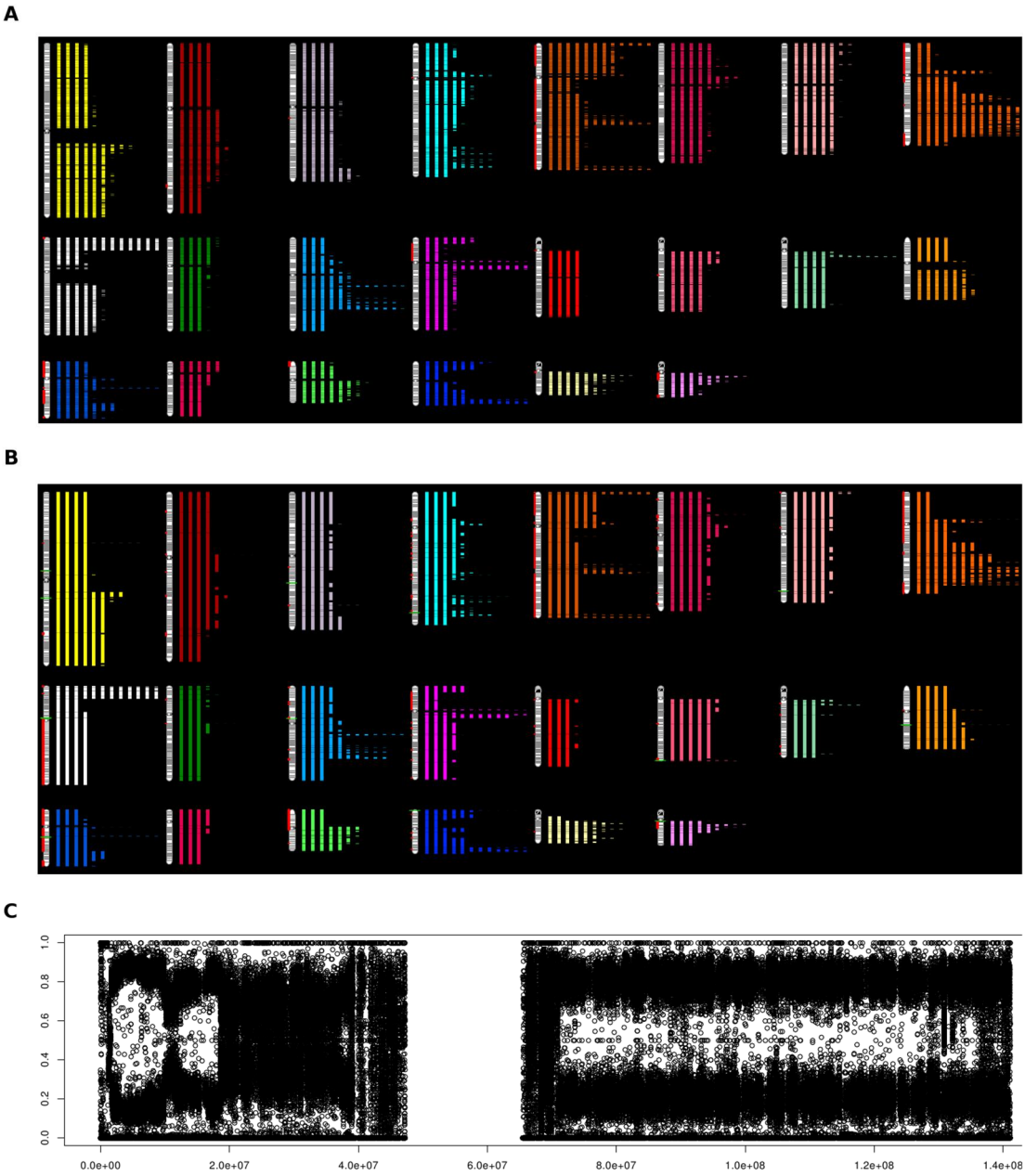
Supplementary Table 2
The estimated tumor purity, tumor cellularity, tumor ploidy for tumor-normal mixtures.

	HCC1143 (Hypo-tetraploid)					HCC1187 (Hypo-triploid)				
	sCNAphase			CLImAT		sCNAphase			CLImAT	
TP	*tc(%)*	*tc_d(%)*	*AP*	*tc(%)*	*AP*	*tc(%)*	*tc_d(%)*	*AP*	*tc(%)*	*AP*
100%	100	100	3.7	98	3.77	100	100	2.8	89	2.68
95%	91	95	3.7	82	3.80	93	95	2.7	85	2.70
80%	68	80	3.7	65	3.80	74	80	2.7	71	2.69
60%	45	60	3.8	43	3.91	52	60	2.7	52	2.70
40%	27	40	3.8	33	3.03	33	40	2.7	32	2.71
20%	12	21	3.8	19	3.19	19	23	2.7	18	2.75
10%	6	11	3.7	48	2.19	8	11	2.7	58	2.02
5%	4	7	3.7	71	2.16	4	5	2.8	62	2.02
0%	0.0	0.0	2	60	2.16	0.0	0.0	2	65	2.01
	HCC1954 (Hyper-tetraploid)					**HCC2218 (tetraploid)**				
	sCNAphase			**CLImAT**		**sCNAphase**			**CLImAT**	
TP	*tc(%)*	*tc_d(%)*	*AP*	*tc(%)*	*AP*	*tc(%)*	*tc_d(%)*	*AP*	*tc(%)*	*AP*
100%	100	1.00	4.5	49	2.67	100	100	4.2	48	2.3
95%	86	93	4.7	46	2.72	89	95	4.2	46	2.29
80%	60	78	4.7	41	2.62	64	79	4.2	76	2.14
60%	37	58	4.7	36	2.50	40	59	4.2	54	2.16
40%	22	39	4.7	20	2.98	23	39	4.4	34	2.12
20%	11	21	4.7	51	2.09	11	21	4.4	22	2.12
10%	6	12	4.9	50	2.05	5	11	4.4	41	2.02
5%	4	6	2.8	52	1.99	5	5	2.2	43	2.01
0%	0.0	0.0	2	64	2.16	0.0	0.0	2	45	1.98

In the table, *tc* denotes for tumor cellularity and tc_d denotes for tumor purity. acc stands for accuracy calculated as the average value of sensitivity and specificity. The numbers in acc column are in percentage. TP and AP stand for tumor purity and average ploidy. The average ploidy for HCC1143, HCC1954 and HCC1187 were derived from the SKY data (http://old-www.path.cam.ac.uk/~pawefish/BreastCellLineDescriptions/HCC1143.html). The ploidy of HCC2218 is found from COSMIC (http://grch37-cancer.sanger.ac.uk/cosmic/sample/overview?id=749716).



Supplementary Figure 3. Accurate copy number profiling relies on accurate determination of ploidy. Both sCNAphase and the independent SKY analysis of the pure HCC1954 and HCC2218 cell-lines classified them as tetraploids (A),(B), a result maintained across the majority of the mixtures, including those that only contained 20% of cell-line derived reads (C) and (D). However, in the 5% cell-line DNA mixtures, ploidy was miscalculated as a result, sCNAphase failed to produce an accurate copy number profile for these mixtures (E) and (F). When the incorrect plodiy value was manually replaced with the correct one, it was possible to identify the same broad copy number profile in the 5% mixture, that was present in the 100% cell-line mixture (G) and (H).



Supplementary Figure 4. Discrepancy about LOHs estimation for HCC1954 is presented in dSKY plots in (A) and (B) generated from sCNAphase and COSMIC respectively. The clear differences were found at chr9 q-arm, which is found to be LOHs only by COSMIC. The absolute copy number was predicted to be 4 from COSMIC than 5 from sCNAphase. However the B-allelic frequency at chr9q in (C) shows the allelic copy numbers at chr9 are very imbalance, allelic copy numbers of 4 to 1 from sCNAphase, but clearly the region remains heterozygous. Similarly the extra regions of LOHs from COSMIC at chr8, chr17 and chr19 are also likely to be heterozygous, as supported by the BAFs. These are also coincide with sight lower copy number estimation sCNAphase. COSMIC segmentation also under-estimated ploidy for chr7, chr15, which are 5 and 4 copies according to sCNAphase and the SKY plots (Supplementary Figure 2E) .

Supplementary Table 3.
The capacity of CLImAT to detect the focal sCNAs in COSMIC

Tumor purity	**HCC1143** **15 focal amplification from COSMIC**		**HCC1954** **94 focal amplification from COSMIC**		**HCC1187** **2 focal amplification from COSMIC**		**HCC2218** **18 focal amplification from COSMIC**	
	Sen	**Spe**	**Sen**	**Spe**	**Sen**	**Spe**	**Sen**	**Spe**
100%	33	17	44	55	50	25	6	33
80%	40	17	39	51	50	8	0	0
60%	47	13	14	36	100	13	0	0
40%	20	40	35	39	100	9	22	50
20%	53	31	0	0	100	6	22	25
10%	0	0	0	0	0	0	0	0

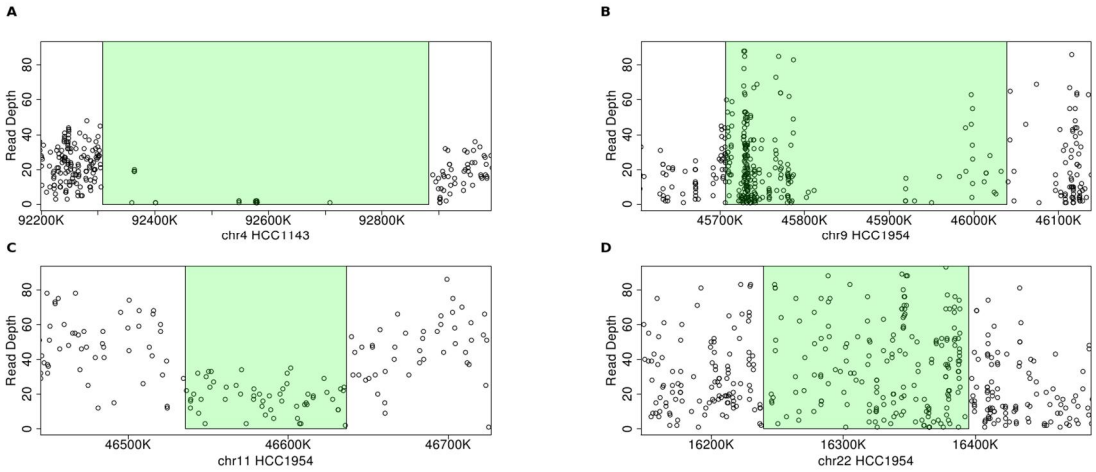
Sen for sensitivity; Spe for specificity.

Supplementary table 4
The detection of *ERBB2 and chr11q13* across the mixtures

	ERBB2		**chr11q13**	
TP	*HCC1954*	*HCC2218*	*HCC1143*	*HCC1954*
100%	>=12	>=12	>=12	>=12
95%	>=12	>=12	>=12	>=12
80%	>=12	>=12	>=12	>=12
60%	>=12	>=12	>=12	>=12
40%	>=12	>=12	>=12	>=12
20%	>=12	>=12	>=12	>=12
10%	>=12	>=12	>=12	>=12
5%	>=12	>=12	>=12	>=12

Supplementary Table 5
A list of the homozygous deletions greater than 100kb in the COSMIC data

Cell-Line	Location	Size	Detected by sCNAphase	Detected by CLImAT
HCC1143	Chr4:92,308,295-92,882,506	574 kb	10% Mixture	Not detected
HCC1187	Chr6:134,329,397-134,537,448	208 kb	Not detected	Not detected
HCC1187	Chr14:63,703,884-64,334,878	630 kb	10% Mixture	Not detected
HCC1187	Chr15:72,194,262-72,443,790	249 kb	10% Mixture	Not detected
HCC2218	Chr16:68,698,153-68,844,967	146 kb	Not detected	Not detected
HCC1954	Chr9:45,706,445-46,038,862	332 kb	Not detected	Not detected
HCC1954	Chr11:46,535,802-46,636,287	100 kb	Not detected	Not detected
HCC1954	Chr22:16,239,488-16,395,163	155 kb	Not detected	Not detected



Supplementary Figure 5. The **B-allelic frequency at regions with heterozygous deletions from COSMIC**.
The plots corresponds to the regions containing the homozygous deletions in HCC1143 and HCC1954 as shown in Supplementary Table 6. The green shading represents the location of the homozygous deletion listed in the COSMIC annotation of these cell-lines, while the black points present the read depth for the individual heterozygous SNPs located within these regions. For HCC1143 (A, C), there is a high level of consistency between deletion in the COSMIC annotation of this cell-line and the read depth of germ-line SNPs in the NGS data. In contrast, the results from HCC1954(B, D) do not show the same level of consistency. In the deletion on chromosome 9(B), the NGS data suggests that the deletion here is significantly smaller than the one listed by COSMIC. In contrast, the deletion on chromosome 22(D), is not consistent with the NGS data.

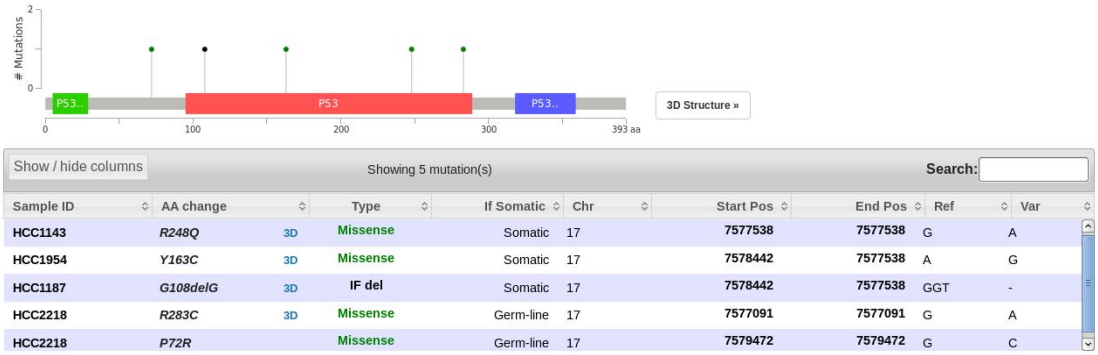
Supplementary Table 6.
Consensus detection of TP53 LOH at different levels of tumor purity.

	Location of TP3 (chr17:7,571,720-7,590,868)	
Cell-Line	Detected LOHs by sCNAphase	Detected LOHs by CLImAT
HCC1143	100%-20% Mixture (4 copies) 5% mixture (4 copies)	100%-20% Mixture
HCC1187	100%-5% Mixture (2 copies)	100%-20% Mixture
HCC2218	100%-10% Mixture (2 copies) 5% Mixture (1 copy)	80%-20% Mixture
HCC1954	100% Mixture (4 copies) 95% Mixture (4 copies) 80-40% Mixture (4 copies) 20%-10% Mixture (3 copies) 5% Mixture (1 copy)	80%-60% Mixture

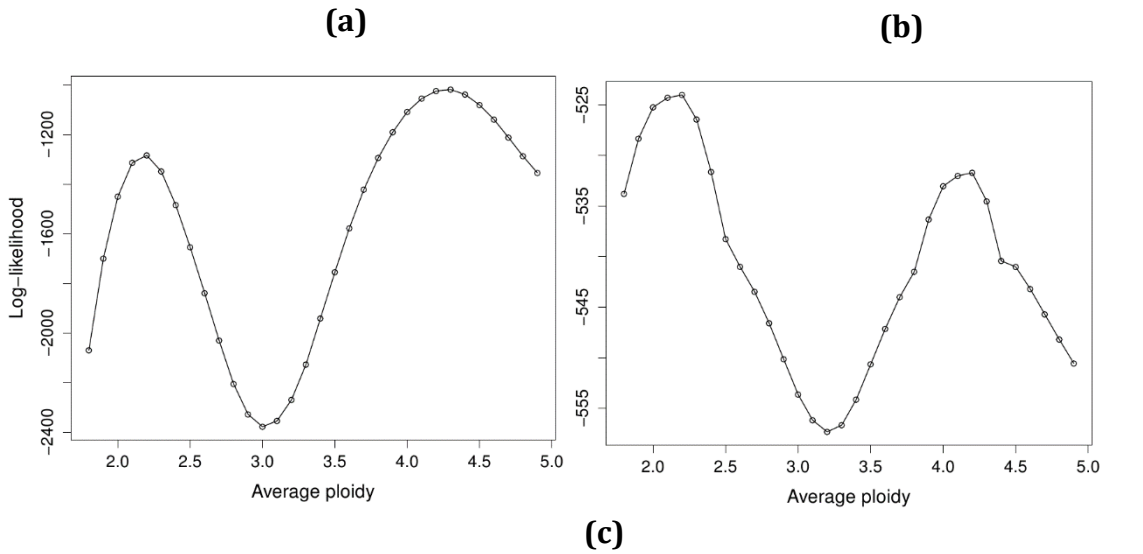
Supplementary Table 7.
Germline Heterozygous SNPs present in in the TP53 loci

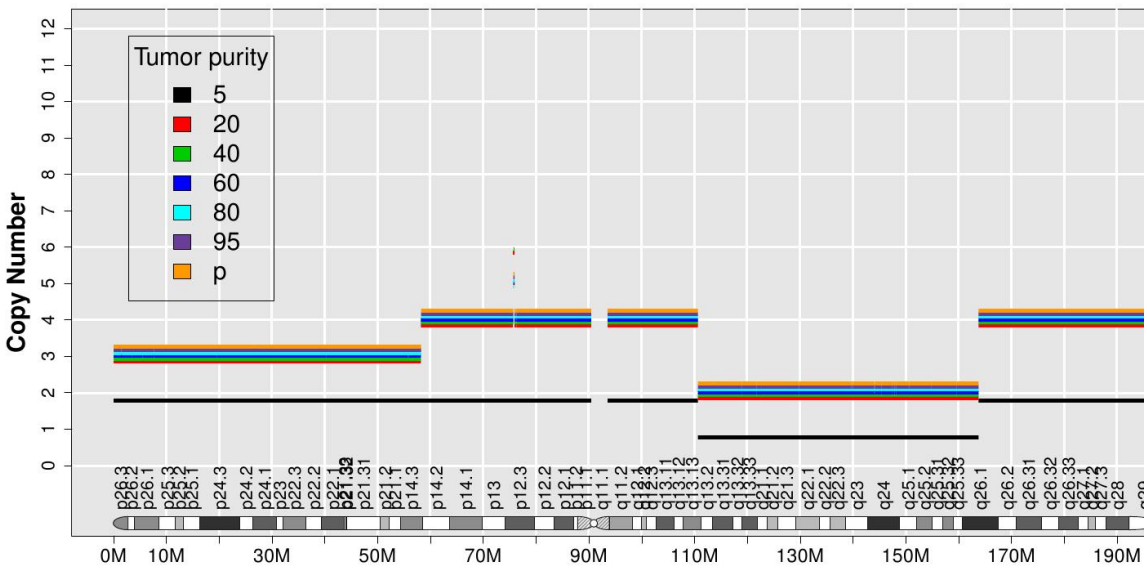
Position	**Ref**	**Alt**	**Depth**	**AF**	**Phase**	**Location**	**h0**	**h1**	**h0**	**Effect**
							Normal		Tumor	
chr17:7574775	C	T	31	0.54	01	Non-coding	C	T	C	
chr17:7574779	C	T	30	0.53	01	Non-coding	C	T	C	
chr17:7574936	C	T	41	0.41	01	Non-coding	C	T	C	
chr17:7577091	G	A	29	0.55	--	CDS	A	G	A	R283C*
chr17:7577644	C	G	36	0.47	10	Non-coding	G	C	G	
chr17:7578115	T	C	34	0.5	10	Non-coding	C	T	C	
chr17:7578645	C	T	29	0.55	10	Non-coding	T	C	T	
chr17:7578837	A	G	38	0.28	10	Non-coding	G	A	G	
chr17:7579472	G	C	37	0.59	10	CDS	C	G	C	P72R
chr17:7579801	G	C	31	0.35	10	Non-coding	C	G	C	

Ref and alt are short for reference and alternative allele. AF is for allelic frequency in the normal sample, HCC2218BL. H0 and H1 stand for the two parental chromosomes. The 01 in the phase column means that the reference allele is from H0, the alternative allele is from H1. The 10 means the opposite.

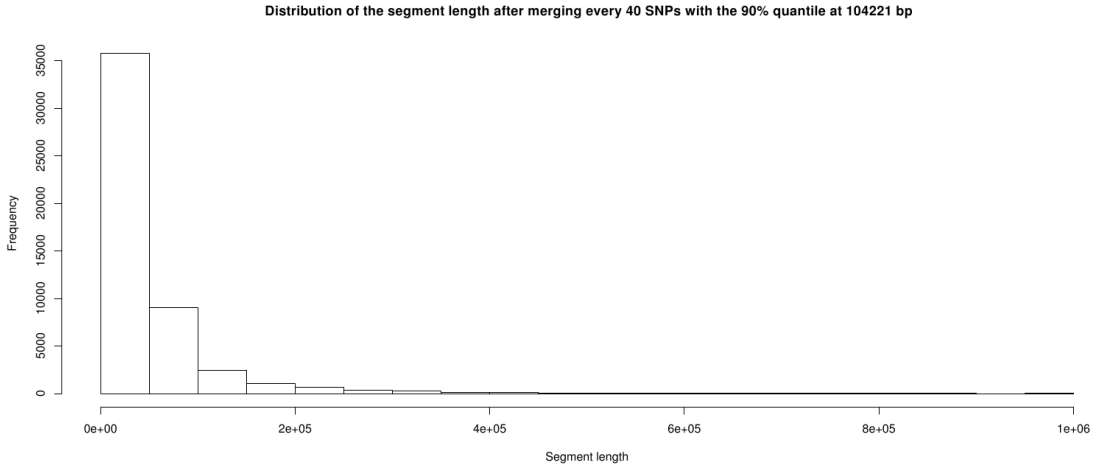


Supplementary Figure 6. *TP53* Germ-line and somatic mutations in in each cell line. Mutations annotated by COSMIC, shows the somatic mutations including 1 missense mutation for HCC1143, 1 missense mutation for HCC1954 and 1 inframe deletion for HCC1187. The two germ-line missense mutations for HCC2218 were from aforementioned analysis. The figure is generated from MutationMapper (Gao et al. *Sci. Signal.* 2013 & Cerami et al. *Cancer Discov.* 2012). Missense Mutations are showed in green and Inframe Mutations are showed in black. The top colored bars showed the location of the mutations in *TP53* as well as the domains of *TP53*. The table underneath shows the details about the mutation types.





Supplementary Figure 7. Ploidy estimation at different tumor purity for HCC2218. As sCNAphase searches possible average ploidy values, (a) shows the maximum likelihood calculated for the ploidy of 1.9 to 5 for 100% tumor purity sample. There are two peaks corresponding to two possible ploidy estimation roughly at 2.3 or 4.3. In **Supplementary Figure 2**, the dSKY of HCC2218 shows the number of the chromosome arms were mostly estimated to be even numbers, except for a region in chr3p estimated to be 3 copy. Because this region supported for average ploidy of ~4.3, the peak at 4.3 is clearly higher, a preferred ploidy estimation. However, at the 5% tumor purity, the lower tumor ploidy estimation became more likely in (b). As shown in the copy number profile of chr3 HCC2218 in (c), this copy number reduced by a half at 5% tumor purity. Even though the reduction from 3 to 2 copies is unfavorable in region from 0 to 50M, this region cannot produce as strong support for the higher ploidy as for the pure tumor sample. In (c) legend, p stands for pure tumor samples.



Supplementary Figure 8. The segment size distribution merging every 40 SNPs calculated from HCC1143.

References

Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir Ba et al. 2012. Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology* **30**(5): 413-421.

Carter SL, Meyerson M, Getz G. 2011. Accurate estimation of homologue-specific DNA concentration-ratios in cancer samples allows long-range haplotyping. *Scott L Carter*: 59.

Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, Berri S. 2012. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics* **28**(1): 40-47.

Li Y, Xie X. 2014. Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity. *Bioinformatics* **30**(15): 2121-2129.

Oesper L, Satas G, Raphael BJ. 2014. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics* **30**(24): 3532-3540.

Xi R, Hadjipanayis AG, Luquette LJ, Kim T-M, Lee E, Zhang J, Johnson MD, Muzny DM, Wheeler DA, Gibbs RA. 2011. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proceedings of the National Academy of Sciences* **108**(46): E1128-E1136.

Yau C. 2013. OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinformatics* **29**(19): 2482-2484.

Yau C, Mouradov D, Jorissen RN, Colella S, Mirza G, Steers G, Harris A, Ragoussis J, Sieber O, Holmes CC. 2010. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome biology* **11**(9): R92-R92.

Yu Z, Liu Y, Shen Y, Wang M, Li A. 2014. CLImAT : accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data. *Bioinformatics* **30**(5): 413-421.