Supplementary Information

Copy Number Analysis Indicates Monoclonal Origin of Lethal Metastatic Prostate Cancer

Wennuan Liu, Sari Laitinen, Sofia Khan, Mauno Vihinen, Jeanne Kowalski, Guoqiang Yu, Li Chen, Charles M. Ewing, Mario A. Eisenberger, Michael A. Carducci, William G. Nelson, Srinivasan Yegnasubramanian, Jun Luo, Yue Wang, Jianfeng Xu, William B. Isaacs, Tapio Visakorpi, and G. Steven Bova

Supplementary Table 1. Subject and Metastatic Prostate Cancer Sample Characteristics from 94 total samples studied from 30 men

Subject Race, Ethnicity	PELICAN Autopsy Study ("A" Study) Case Number	Sample Name (Met is abbreviation for Cancer Metastasis, CA is abbreviation for Carcinoma found at primary site, Xeno is abbreviation for Xenograft)	Anatomic Location Category (subdural met=1, liver met=2, adrenal met=3, pericardial met=4, lymph node met=5, bone met=6, ca found in prostate at autopsy=7, other=8)	Study Sample Identifier Used in Figure 1. Nine samples studied by Affymetrix 6 analysis only are asterisked*.	Affymetrix 6.0 Study specific Tissue Reagent ID Sample Identifier
White, Nonhispanic	1	A1 Subdural Met	1	1-1	-
White, Nonhispanic	2	A2 Liver Met C1	2	2-2a	-
	2	A2 JHU A2 Bone Met 2 Xeno	6	2-6	-
	2	A2 Liver Multiple Met pulverized	2	2-2b	-
African American, Nonhispanic	3	A3 Peritoneal Mass Met 3	5	3-5a	-
	3	A3 Pelvic Paraaortic LN Met	5	3-5b	15953
	3	A3 Subdural Pc B Met	1	3-1	-
	3	Pericardial Mass Met 1A	4	3-4*	16128
	3	Peritoneal Nodule pc1 Met	8	3-8*	15963
White, Nonhispanic	4	A4 Liver Met 17	2	4-2	-
White, Nonhispanic	5	A5 L Iliac LN Met	5	5-5	-
	5	A5 Soft Manubrium Mass Met	6	5-6	-
White, Nonhispanic	7	A7 R Post Subdural Met 1	1	7-1a	-
	7	A7 R Post Subdural Met 2	1	7-1b	-

White, Nonhispanic	8	A8 Multiple Liver Mets	2	8-2	-
	8	A8 R Inguinal LN Met	5	8-5	-
White, Nonhispanic	9	A9 Periportal LN Met	5	9-5	-
African American, Nonhispanic	10	A10 R Iliac LN Met	5	10-5a	-
	10	A10 Periportal LN Met	5	10-5b	-
	10	A10 Perigastric LN Met	5	10-5c	-
	10	A10 Prostate CA	7	10-7	-
White, Nonhispanic	11	A11 L Inguinal LN Met	5	11-5	-
African American, Nonhispanic	12	A12 Paraaortic LN Met	5	12-5a	15989
	12	A12 Mediastinal LN Met	5	12-5b	16053
	12	A12 R Pelvic LN Met	5	12-5c	16054
White, Nonhispanic	13	A13 S2 Vertebral Bone Met	6	13-6a	-
	13	A13 L4 Vertebral Bone Met	6	13-6b	-
White, Nonhispanic	14	A14 Liver Met	2	14-2	-
	14	A14 Thoracic Paraaortic LN Met	5	14-5	-
White, Nonhispanic	16	A16 R Adrenal Met	3	16-3	-
	16	A16 L Pulm Hilar LN Met	5	16-5	15979
	16	A16 R Temporal Subdural Met	1	16-1	15990
	16	A16 Pericardial Mets	4	16-4	15954
White, Nonhispanic	17	A17 Abd Paraaortic LN Met	5	17-5a	16060
	17	A17 R Iliac LN Met	5	17-5b	-
	17	A17 R Supraclavicular LN Met	5	17-5c	16061
	17	A17 R Femur marrow Met	6	17-6	15983
	17	A17 Subdural Met Fossa C	1	17-1a	15982
	17	A17 L Axillary LN #2 Met	5	17-5d	15986
	17	A17 R Subdural Tumor A Met	1	17-1b	-
	17	A17 Paraaortic LN Met	5	17-5e	-
White, Nonhispanic	18	A18 L Cervical LN Met 2	5	18-5a	-
	18	A18 L Cervical LN Met 4	5	18-5b	-
White, Nonhispanic	19	A19 Sternum Soft Met	6	19-6	15994
	19	A19 Paraaortic LN Met	5	19-5	16066
	19	A19 Subdural Met	1	19-1*	16065
White, Nonhispanic	21	A21 Single Liver Met #4	2	21-2a	16068
	21	A21 Single Liver Met #8	2	21-2b	16069
	21	A21 L Adrenal Met	3	21-3	15996
	21	A21 Single Liver Met #5	2	21-2c	15999
	21	A21 R Rib Nodular Met	6	21-6	15997
White, Hispanic	22	A22 L Humerus Bone Marrow Met	6	22-6	16002
	22	A22 Apical Prostate CA	7	22-7	16072
	22	A22 L Adrenal Met	3	22-3	16071
	22	A22 L Pelvic LN7 Met	5	22-5	16003
African American, Nonhispanic	23	A23 Liver Multiple Liver Mets	2	23-2a	-
	23	A23 Single Liver Met	2	23-2b	-
White, Nonhispanic	24	A24 R Diaphragmatic Met	8	24-8	16075

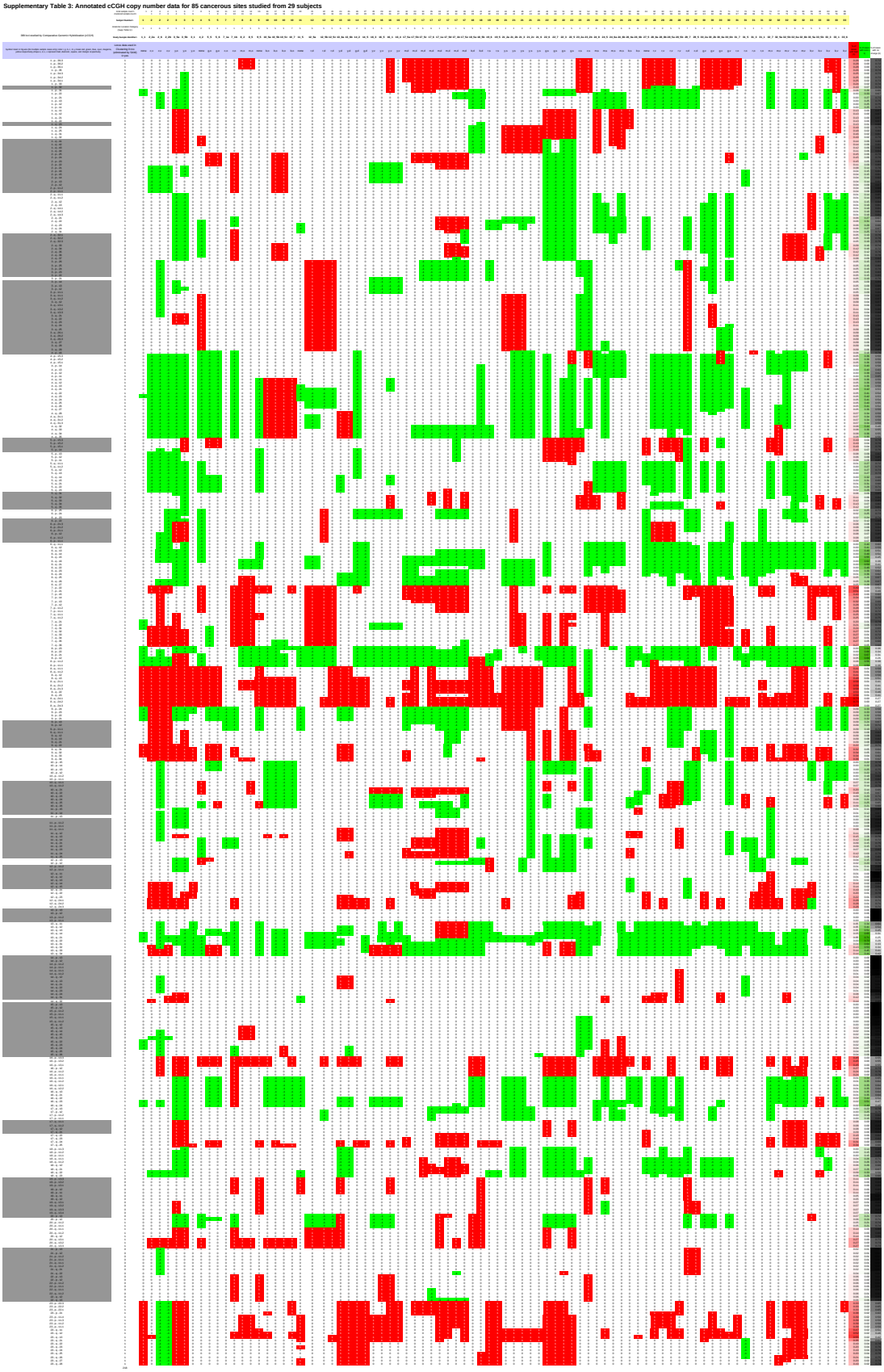
	24	A24 R Axillary LN Met	5	24-5	16008
	24	A24 R Rib7 Met	6	24-6a	16013
	24	A24 Xiphoid Met	6	24-6b	16032
White, Nonhispanic	26	A26 T7 Vertebral Bone Hemorrhagic Met 1-5	6	26-6a	-
	26	A26 L4 Vertebral Bone Hemorrhagic Met 1-9	6	26-6b	-
White, Nonhispanic	27	A27 R Axillary Lymph Node Met 2-5	5	27-5	-
White, Nonhispanic	28	A28 Posterior Bladder Polypoid Met A1	8	28-8a	16020
	28	A28 R Lower Lung Met A2	8	28-8b	16021
	28	A28 Anterior Mediastinal LN Met A8	5	28-5a	16079
	28	A28 L Superficial Ing LN Met A1	5	28-5b	16022
White, Nonhispanic	29	A29 Prostate CA	7	29-7	-
	29	A29 R Superficial Ing LN Met A1	5	29-5	-
White, Nonhispanic	30	A30 L Liver Single.Met 1-7	2	30-2a	16082
	30	A30 L Liver Single Met 2-5	2	30-2b	16083
	30	A30 R Femur Marrow Met 1	6	30-6a	16016
	30	A30 R Humerus Marrow Met 3	6	30-6b	16017
White, Nonhispanic	31	A31 Prostate 1-1-2 CA	7	31-7	16023
	31	A31 R Ing LN Met	5	31-5	16024
	31	A31 L Adrenal Met	3	31-3	16085
	31	A31 R Subdural Met	1	31-1	16086
	31	A31 R Rib 7 Met	6	31-6*	16025
White, Nonhispanic	32	A32 Prostate 10-1-3 CA	7	32-7	16026
	32	A32 L Cervical LN Met 1-2	5	32-5a	16027
	32	A32 L Subclavicular LN Met 1-5	5	32-5b	16033
	32	A32 R Rib 8 Met 1-11	6	32-6a	16034
	32	A32 R Humerus Met 1-12	6	32-6b	16028
White, Nonhispanic	33	A33 L Axillary LN Met	5	33-5a	16010
	33	A33 Paratracheal LN Met	5	33-5b	16029
	33	A33 L Adrenal Met	3	33-3	16035
	33	A33 L Subdural Met	1	33-1	16036
	33	A33 T12-1 Vertebral Met	6	33-6a	16031
	33	A33 R Rib 7 Met	6	33-6b*	16030
White, Nonhispanic	34	Liver Met 1	2	34-2a*	16109
	34	Liver Met 12	2	34-2b*	16110
	34	Liver Met 3	2	34-2c*	16111
	34	Spinal Cord Compressing Met 391T 11 yrs before death	8	34-6*	16090

Supplementary Table 2. Comparison Noncancerous Sample Characteristics for 14 Subjects studied with Affymetrix 6 technology.

PELICAN Autopsy Study ("A" Study) Case Number	Sample Name (NL is abbreviation for "Normal" noncancerous tissue)	Anatomic Location Category Blood=1, Kidney=2, Liver=3, Spleen=4	Affymetrix 6.0 Study Tissue Reagent ID Sample Identifier
3	Lymphs NL	1	16007
3	Kidney NL	2	16040
16	Liver NL	3	16059
17	Kidney NL	2	16062
19	Liver NL	3	16067
21	L Kidney NL	2	16070
22	Liver NL	3	16073
24	Spleen NL	4	16076
28	Spleen NL	4	16080
30	Spleen NL	4	16084
31	Spleen NL	4	16087
31	Liver NL	3	16088
32	Spleen NL	4	16037
33	Liver NL	3	16089
34	Blood 391B NL	1	16091
34	Spleen NL	4	16108

Supplementary Methods: Chromosomal metaphase-based comparative genomic hybridization (cCGH). Briefly, cancer DNA samples were labeled with FITC-dUTP (DuPont, Boston, MA) and normal reference male DNA with TexasRed-dUTP (DuPont) using nick translation. Labeled DNAs were hybridized to normal male lymphocyte metaphase slides (Vysis Inc., Downers Grove, USA) together with unlabelled Cot-1 DNA (10µg, Gibco-BRL). After hybridization, the slides were washed and counterstained with an antifade solution containing 4,6-diamidino 2-phenylindole (DAPI, Vector Laboratories, Burlingame, CA). Several metaphases from each hybridization were captured using a Photometrics ImagePoint CCD camera (Photometrics, Tucson, AZ, USA) mounted on an Olympus BX50 epifluorescence microscope (Tokyo, Japan) and IPLab Spectrum software program (Scanalytics Inc. Fairfax, VA, USA). Relative DNA sequence copy number changes were detected by analyzing the fluorescence intensities of green (tumor) and red (normal) signals along the length of all chromosomes in the metaphase spreads using Quips CGH analysis program (Vysis Inc.). CGH results were plotted as a series of green to red ratio profiles and the interpreted as previously published[1;2]. Hybridizations of FITC-labeled normal male DNA against Texas Red-labeled normal female DNA, in each hybridization batch, were used as negative controls. The mean green-to-red ratio and corresponding SD for all autosomes remained between 0.85 and 1.15. Based on these control hybridizations, chromosomal regions with a mean ratio of 0.85 or less were considered lost and those with a ratio 1.15 or more gained in the cancer samples studied. Chromosome Y was excluded from CGH analysis. MCF-7 breast cancer cell line was used as a positive control in each hybridization batch, and technical replicates performed in 7 samples revealed highly similar loss and gain patterns for each replicated pair based on visual interpretation of Vysis-generated CGH data plots. The complete cCGH dataset is shown in Supplementary Table 3.

Supplementary Table 3: Annotated cCGH copy number data for 85 cancerous sites studied from 29 subjects



Supplementary Methods: SAM analysis of cCGH data. SAM (Significance Analysis of Microarrays)[3] was used to calculate an estimate of the median false discovery rate (FDR) in the cCGH data. SAM uses repeated permutations of the data to determine if the expression of any genes is significantly related to the response. The cutoff for significance is determined by a tuning parameter delta, chosen by the user based on the false positive rate. By considering the CGH data as one class data and using 5000 permutations, a SAM delta value of 1.57 detected 218 significant loci with no false positives (Supplementary Table 3 contains all cCGH study data, and Supplementary Tables 4 and 5 contain SAM results for SAM-positive and SAM-negative loci). For hierarchical clustering Cluster/TreeView[4] software was used. To identify potentially clonally related metastases within and among the study subjects, we applied hierarchical clustering. In hierarchical clustering uncentered correlation was used. TreeView was used to visualize the results.

Supplementary Table 4: SAM analysis- Positive Loci (95)			
Row	Gene ID	Gene Name	Score(d)
192	locus	8 - q - 24.1	10.12405802
193	locus	8 - q - 24.2	10.12405802
194	locus	8 - q - 24.3	10.12405802
187	locus	8 - q - 21.1	7.03526069
188	locus	8 - q - 21.2	7.03526069
189	locus	8 - q - 21.3	7.03526069
191	locus	8 - q - 23	7.03526069
160	locus	7 - p - 21	6.603391584
159	locus	7 - p - 22	6.440874129
190	locus	8 - q - 22	6.377696807
382	locus	23 - q - 13	6.330319921
381	locus	23 - q - 12	5.874741692
186	locus	8 - q - 13	5.765116254
299	locus	16 - p - 13.3	5.66797971
300	locus	16 - p - 13.2	5.66797971
301	locus	16 - p - 13.1	5.66797971
161	locus	7 - p - 15	5.373814162
380	locus	23 - q - 11	5.296872324
183	locus	8 - q - 11.1	4.904984885
184	locus	8 - q - 11.2	4.904984885
185	locus	8 - q - 12	4.904984885
162	locus	7 - p - 14	4.664297685
210	locus	9 - q - 34	4.628997808
325	locus	17 - q - 25	4.525621691
375	locus	23 - p - 21	4.473925621
376	locus	23 - p - 11.4	4.473925621
377	locus	23 - p - 11.3	4.473925621
378	locus	23 - p - 11.2	4.473925621
379	locus	23 - p - 11.1	4.473925621
163	locus	7 - p - 13	4.387675737
324	locus	17 - q - 24	4.113374232
2	locus	1 - p - 36.3	3.976711061
3	locus	1 - p - 36.2	3.976711061
4	locus	1 - p - 36.1	3.976711061
164	locus	7 - p - 12	3.976711061
165	locus	7 - p - 11.2	3.976711061
166	locus	7 - p - 11.1	3.976711061
372	locus	23 - p - 22.3	3.94196485
373	locus	23 - p - 22.2	3.94196485
374	locus	23 - p - 22.1	3.94196485
353	locus	20 - q - 13.1	3.840159937
354	locus	20 - q - 13.2	3.840159937
355	locus	20 - q - 13.3	3.840159937
302	locus	16 - p - 12	3.840159937
383	locus	23 - q - 21	3.813471438
5	locus	1 - p - 35	3.703552785
209	locus	9 - q - 33	3.683780484
172	locus	7 - q - 32	3.682957548
173	locus	7 - q - 33	3.682957548
174	locus	7 - q - 34	3.682957548
252	locus	12 - q - 24.1	3.682957548
253	locus	12 - q - 24.2	3.682957548
254	locus	12 - q - 24.3	3.682957548
303	locus	16 - p - 11.2	3.566712626
304	locus	16 - p - 11.1	3.566712626
167	locus	7 - q - 11.1	3.566712626
168	locus	7 - q - 11.2	3.566712626
208	locus	9 - q - 32	3.443364829
207	locus	9 - q - 31	3.188120578
169	locus	7 - q - 21	3.152840293
6	locus	1 - p - 34.3	3.013328903
7	locus	1 - p - 34.2	3.013328903
8	locus	1 - p - 34.1	3.013328903
251	locus	12 - q - 23	3.004900613
175	locus	7 - q - 35	2.931192648
176	locus	7 - q - 36	2.931192648
9	locus	1 - p - 33	2.878367286
322	locus	17 - q - 22	2.729067872
323	locus	17 - q - 23	2.729067872
250	locus	12 - q - 22	2.7270844
249	locus	12 - q - 21	2.465564076
22	locus	1 - q - 24	2.465564076
350	locus	20 - q - 11.1	2.437102718
351	locus	20 - q - 11.2	2.437102718
352	locus	20 - q - 12	2.437102718
388	locus	23 - q - 26	2.362594627
171	locus	7 - q - 31	2.325557907
389	locus	23 - q - 27	2.324433087
23	locus	1 - q - 25	2.324433087
25	locus	1 - q - 32	2.324433087
17	locus	1 - q - 11	2.28709361
18	locus	1 - q - 12	2.28709361
19	locus	1 - q - 21	2.28709361
20	locus	1 - q - 22	2.28709361
282	locus	14 - q - 32	2.28709361
170	locus	7 - q - 22	2.272941219
387	locus	23 - q - 25	2.223988776
390	locus	23 - q - 28	2.181005567
24	locus	1 - q - 31	2.181005567
115	locus	5 - p - 15.3	2.083224106
116	locus	5 - p - 15.2	2.083224106
117	locus	5 - p - 15.1	2.083224106
384	locus	23 - q - 22	2.083224106
385	locus	23 - q - 23	2.083224106
386	locus	23 - q - 24	2.083224106

Supplementary Table 5: SAM analysis-Negative Loci (123)			
Row	Gene ID	Gene Name	Score(d)
264	locus	13 - q - 22	-9.855327412
178	locus	8 - p - 22	-8.883627765
177	locus	8 - p - 23	-8.448152799
151	locus	6 - q - 16	-8.240718279
152	locus	6 - q - 21	-7.843891394
179	locus	8 - p - 21	-7.680757499
263	locus	13 - q - 21	-7.57920995
180	locus	8 - p - 12	-7.231016379
150	locus	6 - q - 15	-7.110790625
265	locus	13 - q - 31	-7.098892905
262	locus	13 - q - 14	-6.89828392
148	locus	6 - q - 13	-6.281274922
149	locus	6 - q - 14	-6.281274922
153	locus	6 - q - 22	-6.281274922
312	locus	16 - q - 23	-5.874741692
313	locus	16 - q - 24	-5.727432309
94	locus	4 - p - 13	-5.373814162
261	locus	13 - q - 13	-5.267529789
181	locus	8 - p - 11.2	-5.265219466
182	locus	8 - p - 11.1	-5.265219466
154	locus	6 - q - 23	-5.229260929
147	locus	6 - q - 12	-5.229260929
311	locus	16 - q - 22	-5.156642618
95	locus	4 - p - 12	-5.086184086
146	locus	6 - q - 11.1	-5.086184086
96	locus	4 - p - 11	-4.803851592
93	locus	4 - p - 14	-4.743859597
155	locus	6 - q - 24	-4.628997808
306	locus	16 - q - 11.2	-4.492375271
307	locus	16 - q - 12.1	-4.492375271
308	locus	16 - q - 12.2	-4.492375271
309	locus	16 - q - 13	-4.492375271
310	locus	16 - q - 21	-4.492375271
12	locus	1 - p - 22	-4.387675737
305	locus	16 - q - 11.1	-4.356481468
260	locus	13 - q - 12	-4.209574463
100	locus	4 - q - 21	-4.209574463
105	locus	4 - q - 26	-4.209574463
126	locus	5 - q - 14	-4.113374232
101	locus	4 - q - 22	-4.08111305
102	locus	4 - q - 23	-4.08111305
103	locus	4 - q - 24	-4.08111305
104	locus	4 - q - 25	-4.08111305
127	locus	5 - q - 15	-3.976711061
13	locus	1 - p - 21	-3.840159937
90	locus	4 - p - 15.3	-3.825454785
91	locus	4 - p - 15.2	-3.825454785
92	locus	4 - p - 15.1	-3.825454785
266	locus	13 - q - 32	-3.802502372
111	locus	4 - q - 32	-3.773436616
315	locus	17 - p - 12	-3.703552785
97	locus	4 - q - 11	-3.698025361
98	locus	4 - q - 12	-3.698025361
99	locus	4 - q - 13	-3.698025361
106	locus	4 - q - 27	-3.698025361
128	locus	5 - q - 21	-3.682957548
14	locus	1 - p - 13	-3.566712626
15	locus	1 - p - 12	-3.566712626
16	locus	1 - p - 11	-3.566712626
123	locus	5 - q - 11.2	-3.566712626
124	locus	5 - q - 12	-3.566712626
125	locus	5 - q - 13	-3.566712626
213	locus	10 - p - 13	-3.566712626
129	locus	5 - q - 22	-3.548391011
259	locus	13 - q - 11	-3.443364829
211	locus	10 - p - 15	-3.429451327
212	locus	10 - p - 14	-3.429451327
314	locus	17 - p - 13	-3.429451327
214	locus	10 - p - 12	-3.291566911
130	locus	5 - q - 23	-3.280632315
122	locus	5 - q - 11.1	-3.278075189
156	locus	6 - q - 25	-3.059940407
157	locus	6 - q - 26	-3.059940407
158	locus	6 - q - 27	-3.059940407
112	locus	4 - q - 33	-3.048284954
113	locus	4 - q - 34	-3.048284954
107	locus	4 - q - 28	-3.006048125
108	locus	4 - q - 31.1	-3.006048125
109	locus	4 - q - 31.2	-3.006048125
110	locus	4 - q - 31.3	-3.006048125
114	locus	4 - q - 35	-2.927039316
329	locus	18 - q - 11.1	-2.901022803
330	locus	18 - q - 11.2	-2.901022803
136	locus	6 - p - 25	-2.878367286
137	locus	6 - p - 24	-2.878367286
316	locus	17 - p - 11.2	-2.871873429
241	locus	12 - p - 12	-2.866697973
50	locus	2 - q - 22	-2.840190883
267	locus	13 - q - 33	-2.837685366
138	locus	6 - p - 23	-2.742249488
215	locus	10 - p - 11.2	-2.729067872
216	locus	10 - p - 11.1	-2.729067872
317	locus	17 - p - 11.1	-2.729067872
268	locus	13 - q - 34	-2.725989212
334	locus	18 - q - 23	-2.712994252
45	locus	2 - q - 13	-2.585772646
46	locus	2 - q - 14.1	-2.585772646
47	locus	2 - q - 14.2	-2.585772646
48	locus	2 - q - 14.3	-2.585772646
49	locus	2 - q - 21	-2.58493812
327	locus	18 - p - 11.2	-2.584274811
328	locus	18 - p - 11.1	-2.584274811
53	locus	2 - q - 31	-2.499344813
333	locus	18 - q - 22	-2.494051326
51	locus	2 - q - 23	-2.455853656
331	locus	18 - q - 12	-2.455853656
326	locus	18 - p - 11.3	-2.437102718
195	locus	9 - p - 24	-2.372574655
196	locus	9 - p - 23	-2.372574655
52	locus	2 - q - 24	-2.325557907
44	locus	2 - q - 12	-2.296706748
332	locus	18 - q - 21	-2.252558798
197	locus	9 - p - 22	-2.250222652
198	locus	9 - p - 21	-2.250222652
347	locus	20 - p - 12	-2.193851225
42	locus	2 - q - 11.1	-2.148147595
43	locus	2 - q - 11.2	-2.148147595
348	locus	20 - p - 11.2	-2.136969772
349	locus	20 - p - 11.1	-2.136969772
227	locus	11 - p - 13	-2.133702786
240	locus	12 - p - 13	-2.034893037
67	locus	3 - p - 21	-1.998984722
11	locus	1 - p - 31	-1.931592539

Supplementary Statistical Analysis of cCGH Data

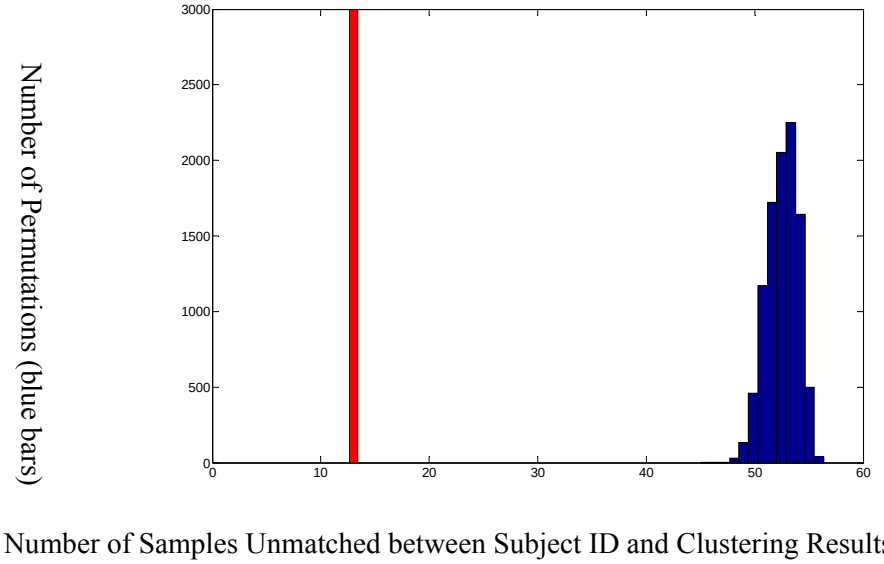
Assessing the statistical significance of observed "clonality"

We applied three different methods to jointly assess the statistical significance of observed "clonality", namely, unsupervised cluster-subject matching[5;6], supervised sample classification[7;8], and Fisher's distance statistics[9]. Based on a large number of random permutations, we generated the empirical distribution of the "summary statistics" under the null hypothesis that the observed "clonality" is a by-chance event. Accordingly, we used three summary statistics criteria to measure the degree of clonality[10], namely, cluster-subject matching error, predictive classification error, and Fisher's distance.

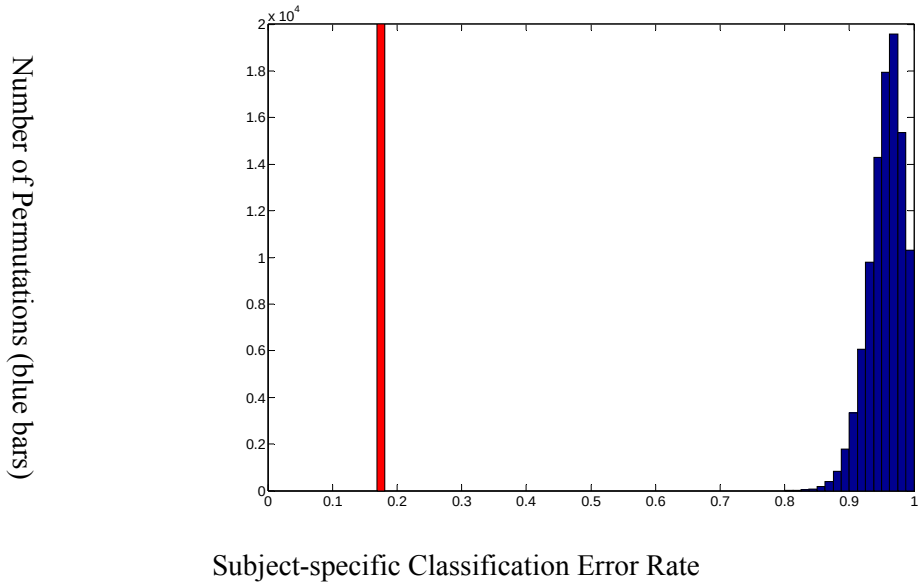
Specifically, in the unsupervised cluster-subject matching experiment, we used the matching error between subject ID assignments and cCGH data clusters (obtained via unsupervised hierarchical clustering) as the summary statistics. The underling null hypothesis is that the subject IDs are randomly assigned to tissue samples independent of samples' genomic signatures. We performed a large number of random permutations to assess the statistical significance of the observed label assignment. We searched exhaustively among different number of clusters, to find the minimum number of "unmatched" samples as the error rate of mismatching. We obtained the observed error of hierarchical clustering as 13/80, which means 13 samples are mismatched in total 80 samples. The histogram of the error rates obtained by 10,000 permutations is shown in Supplementary Figure 1. Error rate by permutation ranges from 45/80 to 56/80. The P value associated with the observed error rate of 13/80 is below 10^{-4}, upon which we can safely reject the null hypothesis and support the claim of "clonality".

We conducted similar permutation experiments using supervised sample classification and Fisher's distance statistics, and we reached the same conclusion. Detailed description for the statistical analyses using unsupervised cluster-subject matching, supervised sample classification, and Fisher's distance statistics are contained in the main manuscript, and below we provide detailed experimental results on the

statistical analyses using unsupervised cluster-subject matching, supervised sample classification, and Fisher's distance statistics.



Supplementary Figure 1. Histogram of error rates by Random Permutation Test: red line denotes the matching error of the observed label assignment; blue bar denotes the matching error of random label assignment.

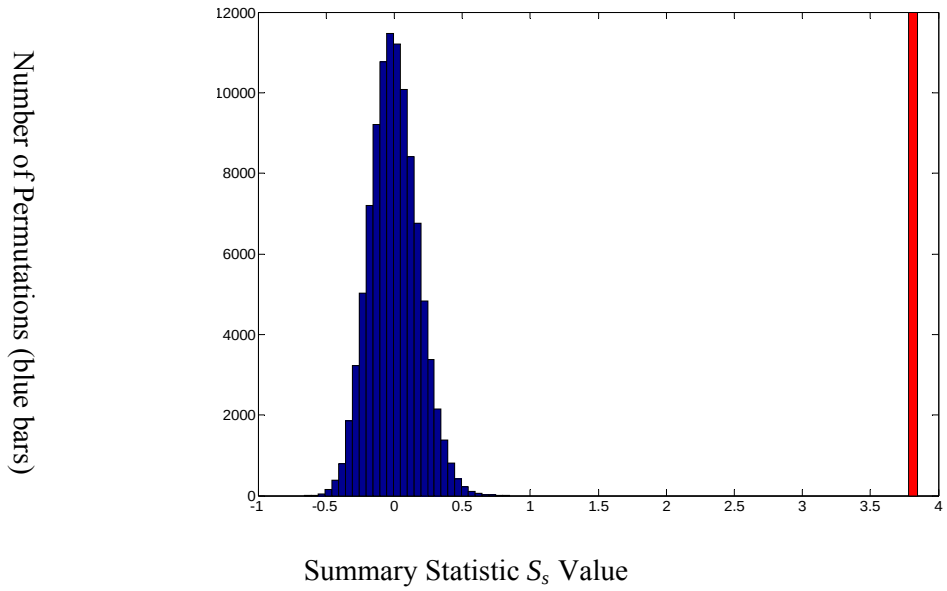


Supplementary Figure 2. The experimental result on the observed subject-specific supervised classification of metastatic prostate cancer samples using cCGH copy number data is given in Supplementary Figure 2. In the 80 samples from 24 subjects from whom 2 or more anatomically separate samples are available, subject-specific classification error rates estimated by 100,000 permutations of subject labels (blue bars), with numbers of permutations on the Y axis and error-rate on the X axis. The smallest error rate in all permutations is 0.800. The error rate based on the ground truth subject labels is 0.175 as indicated by the red bar whose associated P value is less than 10^{-5}.

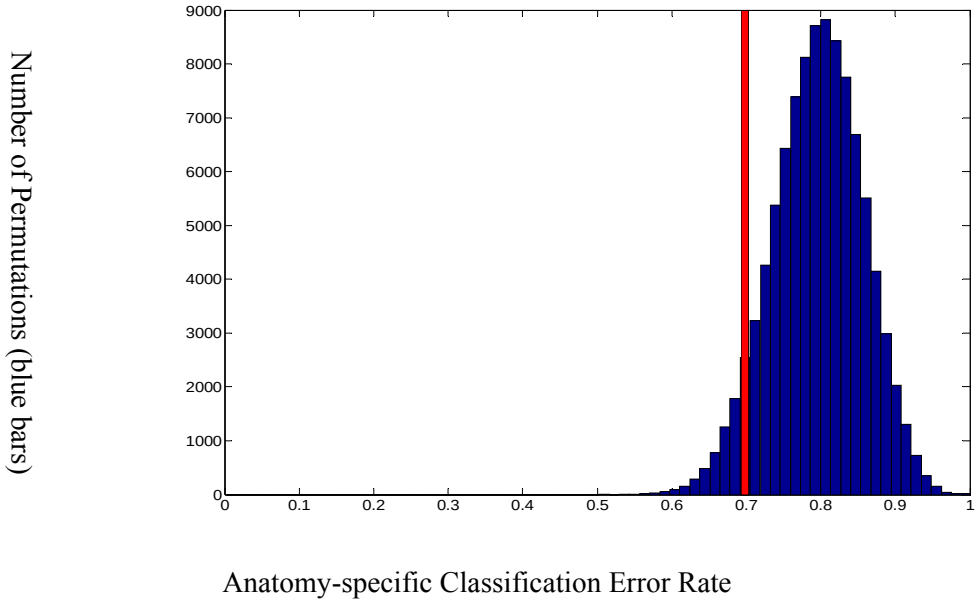
Subject Label	A2	A3	A5	A7	A8	A10	A12	A13	A14	A16	A17	A18
Num of Samples	3	3	2	2	2	4	3	2	2	4	8	2
Num of errors	0	1	0	2	0	0	0	0	1	0	0	0
Error Rate	**0**	**0.33**	**0**	**1**	**0**	**0**	**0**	**0**	**0.5**	**0**	**0**	**0**

Subject Label	A19	A21	A22	A23	A24	A26	A28	A29	A30	A31	A32	A33
Num of Samples	2	5	4	2	4	2	4	2	4	4	5	5
Num of Errors	0	1	0	2	0	0	0	2	0	1	1	3
Error Rate	**0**	**0.2**	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**0.25**	**0.2**	**0.6**

Supplementary Table 6. Distribution of classification errors among subjects studied: total 15 subjects consisting of 50 samples have been correctly classified without any misclassification; 3 subjects consisting of 6 samples have the misclassification error rate of 1; and there are 4 subjects consisting of 17 samples were imperfectly classified with small errors (error rate less than 0.33).



Supplementary Figure 3 Assessment of subject-specific similarity of metastatic prostate cancer using cCGH copy number data and Fisher's distance statistics, where we shown that the genomic similarity among the samples belonging to a specific subjects is significantly greater than the average/mixed similarity among all samples. In the 80 samples from 24 subjects from whom 2 or more anatomically separate samples are available, let D_{bs} represent the average "between-subject" Euclidian distance over all sample pairs belonging to different subjects and D_{ws} represent the average "within-subject" Euclidian distance over all sample pairs belonging to the same subjects, using the summary statistics (modified Fisher's distance) $S_s = D_{bs} - D_{ws}$, we compared experimentally the observed S_s (based on the ground truth subject labels) to the distribution of S_s under the null hypothesis calculated from 100,000 random permutations of subject labels. The maximum value of S_s in the 100,000 random permutations is 0.8467, while the value of experimentally observed S_s is 3.8159 (red bar) whose associated P value is less than 10^{-5}.



Supplementary Figure 4 Based on cCGH copy number changes, metastatic prostate cancers are not significantly related to anatomic location/category. Examining copy number data from all 85 samples from 29 subjects by anatomic location where cancer sample was isolated at autopsy, the observed error rate (0.6986) indicated by the red bar is reasonably within the distribution of anatomic-site-specific classification error rates under the null hypothesis with 100,000 permutations (blue bars). The p-value associated with the observed anatomic-site-specific classification error rate is 0.107.

Supplementary Methods: Affymetrix Genome-Wide Human SNP array 6.0 Analysis. 250 ng of genomic DNA were digested with either *Nsp* I or *Sty* I and then ligated to adapters that recognize cohesive four-basepair (bp) overhangs. A generic primer that recognizes the adapter sequence was used to amplify adapter ligated DNA fragments with PCR conditions optimized to preferentially amplify fragments in the 200 to 1,100 bp size range in a GeneAmp PCR System 9700 (Applied Biosystems, Foster City, CA). After purification with magnetic beads from Agencourt (Beverly, MA), the PCR product was fragmented using DNase I and a sample of the fragmented product was visualized on a 4% TBE agarose gel to confirm that the average size was smaller than 180 bp. The fragmented DNA was then labeled with biotin and hybridized to the Affy6 chip for 18 hrs. We washed and stained the arrays using an Affymetrix fluidics Station 450 and scanned the arrays using a GeneChip Scanner 3000 7G (Affymetrix, Inc., Santa Clara, CA). The Affymetrix GeneChip® Operating Software (GCOS) was used to collect and extract feature data from Affymetrix GeneChip® Scanners. We used Affymetrix® Genotyping Console™ Software 2.1 for genotype analysis. The average call rate for all samples was >97.7%.

Supplementary Methods: Affymetrix 6 chip-based Allele-specific copy number analysis. Allele-specific genomic analysis depicted in Figures 2, 3 and in Supplementary Table 8) was performed using the Partek Genomic Suite (PGS) verion 6.4 allele-specific analysis algorithm, which takes advantage of genotype information and allele-specific intensities from paired samples to estimate DNA copy number for each heterozygous SNP, and is further described in Supplementary Information. Allele-specific analysis can also help determine the effect of normal DNA contamination from nonmalignant cells in the tumor samples through comparison of allele ratios inside and outside regions of apparent hemizygous deletion. Please note that the currently released PGS allele-specific copy number algorithm for single sample analysis assigns one allele "Max" status and

colors its data red, and assigns the other allele "Min" status and colors its data blue based on the estimated copy number for the different alleles (max=red, min=blue). This labeling is meant to convey the structure of contiguous regions with differing allele prevalence, but by itself does not imply haplotype phase across regions of similar allele prevalence. Each allele specific display in Figures 2-4 is thus independently displayed and categorization of changes into omniclonal/subclonal/indeterminate groups are based on visual interpretation of overall pattern. Examples of homozygous deletion displayed in Figure 4 were identified by examination of the allele-specific copy number data using a combination of relative and absolute copy number (both alleles generally well below 0.5 copy number) and genomic length of affected segment containing more than approximately 20 probes (each dot in Figure 4 represents data from 10 probes).

Supplementary Statistical Analysis of Affymetrix 6 Data

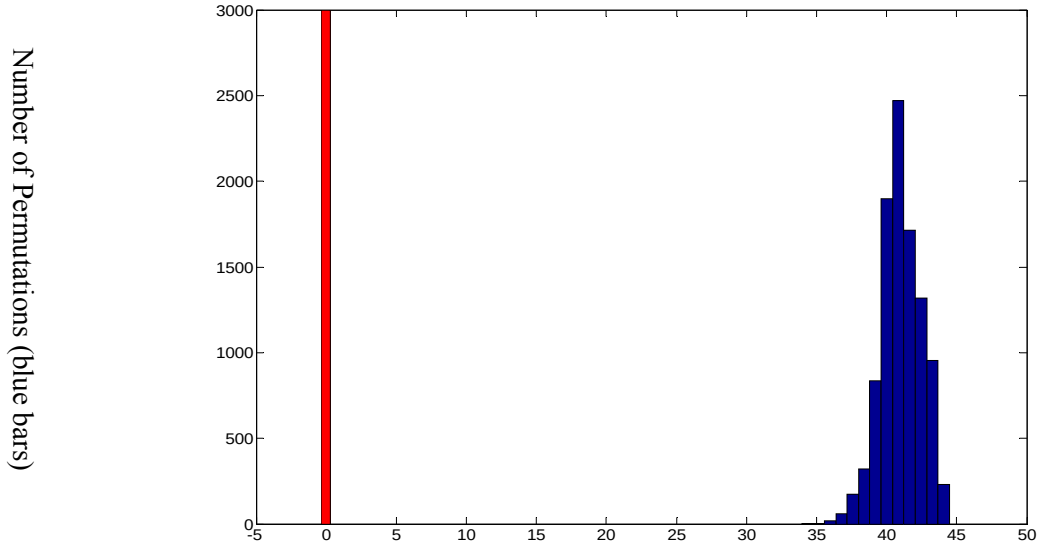
Assessing the statistical significance of observed "clonality"

We applied three different methods to jointly assess the statistical significance of observed "clonality", namely, unsupervised cluster-subject matching, supervised sample classification, and Fisher's distance statistics. Based on a large number of random permutations, we generated the empirical distribution of the "summary statistics" under the null hypothesis that the observed "clonality" is a by-chance event. Accordingly, we used three summary statistics criteria to measure the degree of clonality, namely, cluster-subject matching error, predictive classification error, and Fisher's distance.

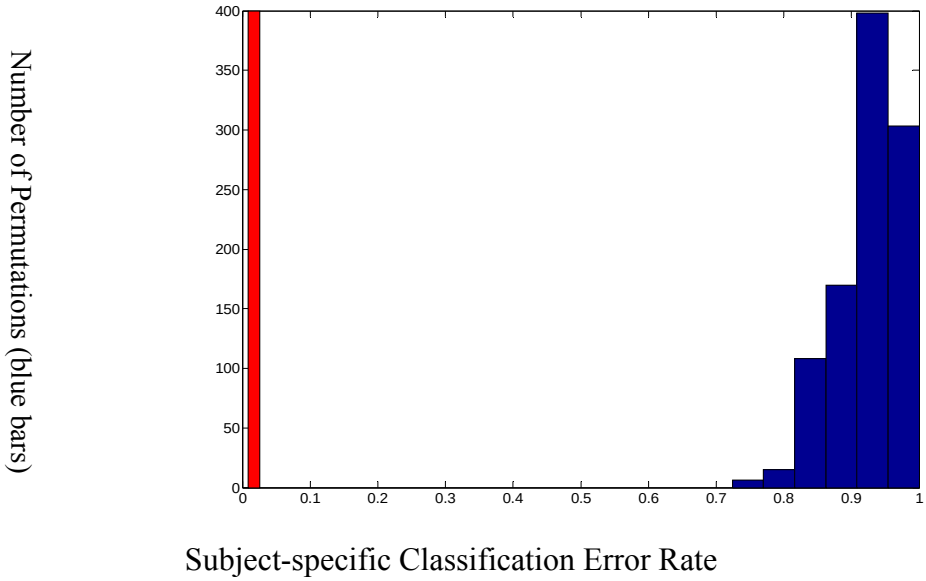
Specifically, in the unsupervised cluster-subject matching experiment, we used the matching error between subject ID assignments and Affymetrix 6 data clusters (obtained via unsupervised hierarchical clustering) as the summary statistics. The underling null hypothesis is that the subject IDs are randomly assigned to tissue samples independent of samples' genomic signatures. We performed a large number of random permutations to assess the statistical significance of the observed label assignment. We exhaustively search among different number of clusters, to find the minimum number of "unmatched" samples as the error rate of mismatching. We obtained the observed error of hierarchical clustering as 0/58, which means 0 samples are mismatched in total 58 samples. The histogram of the error rates obtained by 10,000 permutations is shown in Supplementary Figure 5. Error rate by permutation ranges from 34/58 to 44/58. The P value associated with the observed error rate of 0/58 is below 10^{-4}, upon which we can safely reject the null hypothesis and support the claim of "clonality".

We conducted similar permutation experiments using supervised sample classification and Fisher's distance statistics, and we reached the same conclusion. Detailed descriptions for the statistical analyses using unsupervised cluster-subject matching, supervised sample classification, and Fisher's distance statistics. are contained in the main manuscript, and below we provide detailed experimental results on

the statistical analyses using unsupervised cluster-subject matching, supervised sample classification, and Fisher's distance statistics.



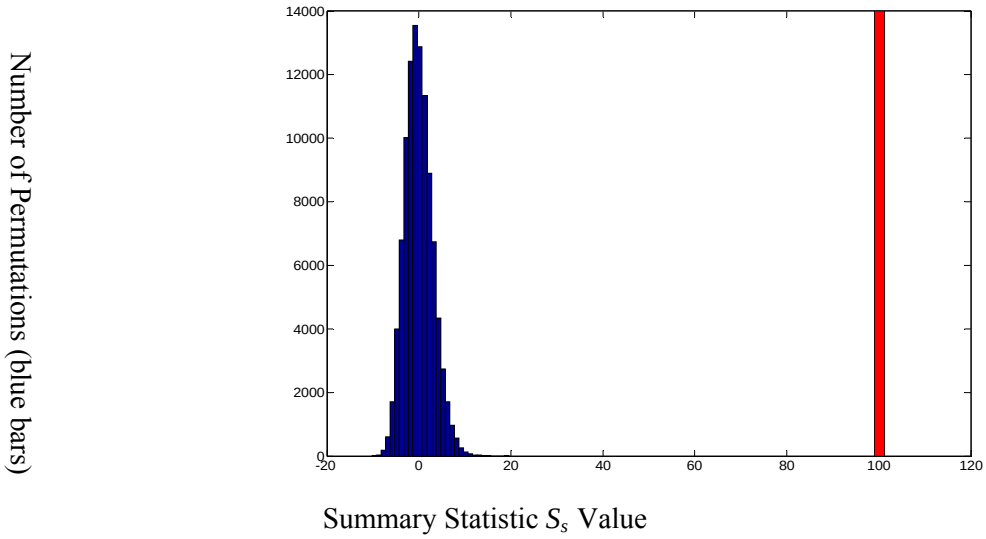
Supplementary Figure 5 Histogram of error rates by Random Permutation Test: red line denotes the matching error of the observed label assignment; blue bar denotes the matching error of random label assignment.



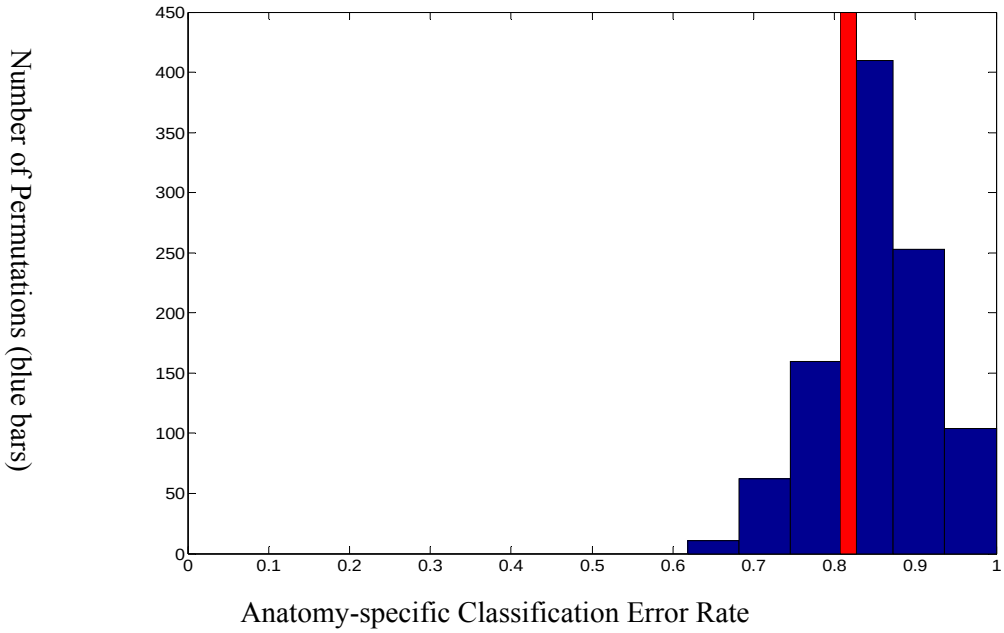
Supplementary Figure 6 The experimental result on the observed subject-specific supervised classification of metastatic prostate cancer samples using Affymetrix 6 data is given in Supplementary Figure 6. In the 58 samples from 14 subjects from whom 2 or more anatomically separate samples are available, subject-specific classification error rates estimated by 1,000 permutations of subject labels (blue bars), with numbers of permutations on the Y axis and error-rate on the X axis. The smallest error rate in all permutations is 0.7241. The error rate based on the ground truth subject labels is 0.0172 as indicated by the red bar whose associated P value is less than 10^{-3}.

Subject Label	A3	A12	A16	A17	A19	A21	A22	A24	A28	A30	A31	A32	A33	A34
Num of Samples	3	3	3	5	3	5	4	4	4	4	5	5	6	4
Num of errors	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Error Rate	**0.33**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Supplementary Table 7. The distribution of the classification errors among the subjects being studied: total 13 subjects consisting of 55 samples have been correctly classified without any misclassification; only one sample in one subject (A3) was misclassified.

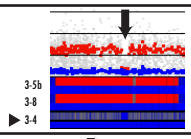
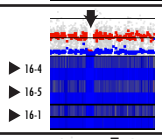
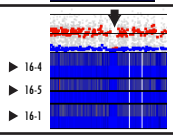
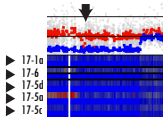
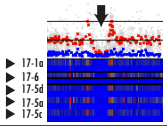
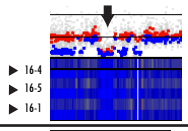
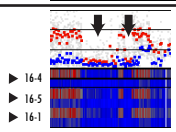
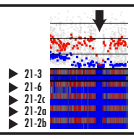


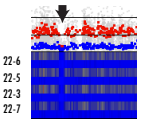
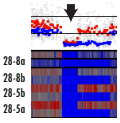
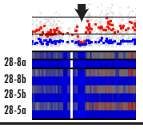
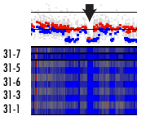
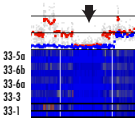
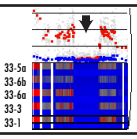
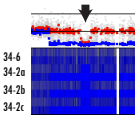
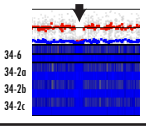
Supplementary Figure 7 Assessment of subject-specific similarity of metastatic prostate cancer using Affymetrix 6 copy number data and Fisher's distance statistics, where we shown that the genomic similarity among the samples belonging to a specific subjects is significantly greater than the average/mixed similarity among all samples. In the 58 samples from 14 subjects from whom 2 or more anatomically separate samples are available, let D_{bs} represent the average "between-subject" Euclidian distance over all sample pairs belonging to different subjects and D_{ws} represent the average "within-subject" Euclidian distance over all sample pairs belonging to the same subjects, using the summary statistics (modified Fisher's distance) $S_s = D_{bs} - D_{ws}$, we compared experimentally the observed S_s (based on the ground truth subject labels) to the distribution of S_s under the null hypothesis calculated from 100,000 random permutations of subject labels. The maximum value of S_s in the 100,000 random permutations is 19.62, while the value of experimentally observed S_s is 100.24 (red bar) whose associated P value is less than 10^{-5}.

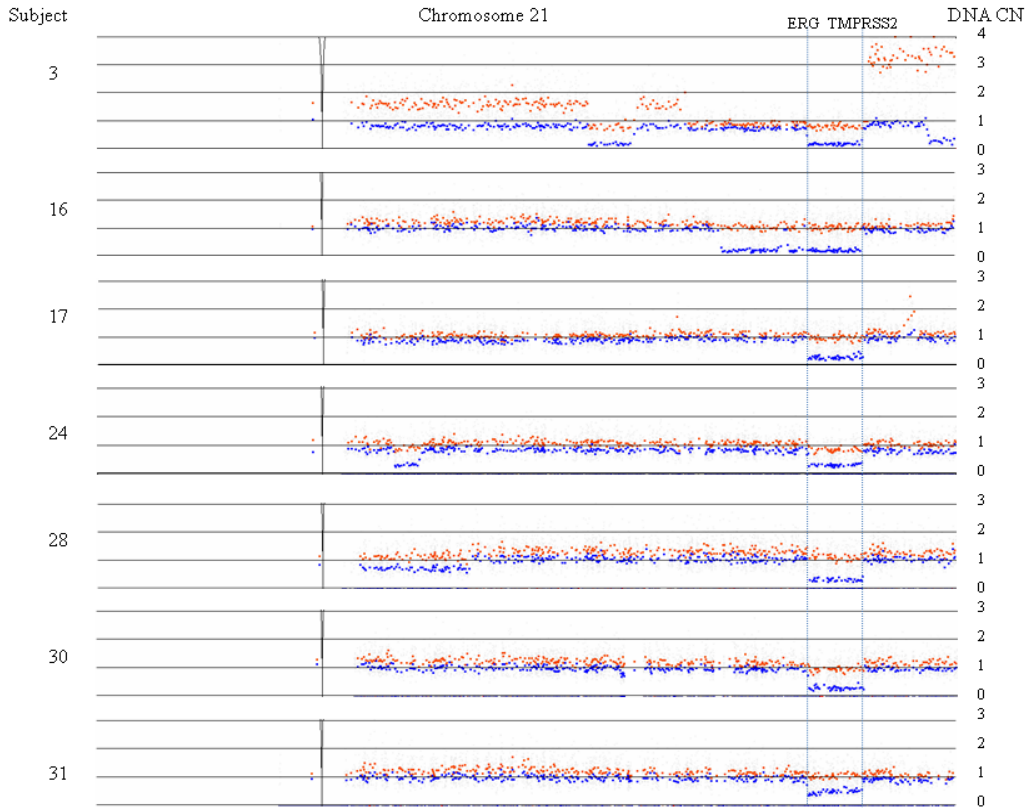


Supplementary Figure 8 Based on Affymetrix 6 copy number changes, metastatic prostate cancers are not significantly related to anatomic location/category. Examining copy number data from all 58 samples from 14 subjects by anatomic location where cancer sample was isolated at autopsy, the observed error rate (0.8182) indicated by the red bar is reasonably within the distribution of anatomic-site-specific classification error rates under the null hypothesis with 1,000 permutations (blue bars). The p-value associated with the observed anatomic-site-specific classification error rate is 0.326.

Supplementary Table 8: Homozygous Deletions detected in samples studied by Affy6

Subject	Chr.	Pos. (kb)	Omniclonal	Copy Number
3	8	25,093-25,978	No	
16	3	60,570-62,105	Yes	
16	8	26,062-27,078	Yes	
17	3	20,454-20,776	Yes	
17	8	25,372-27,399	Yes	
19	3	30,519-32,846	Yes	
19	9	23,613-25,313 26,114-26,911	Yes	
21	11	100,200-101,083	Yes	

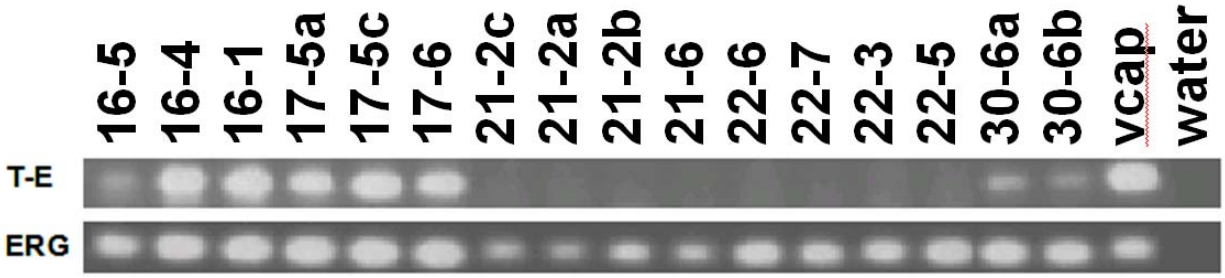
Subject	Chr.	Pos. (kb)	Omniclonal	Copy Number
22	1	8,436-9,593	Yes	
28	10	89,719-91,178	Yes	
28	12	49,732-50,670	Yes	
31	10	89,659-90,473	Yes	
33	6	112,578-117,101	Yes	
33	12	125,201-128,505	Yes	
34	1	65,104-67,340	No	
34	13	31,851-32,775	Yes	



Supplementary Figure 9: Sample Affy6-based Chromosome 21 copy number data with reference to position of ERG and TMPRSS2.

Supplementary Methods: Analysis of TMPRSS2-ERG fusion transcript and ERG transcript

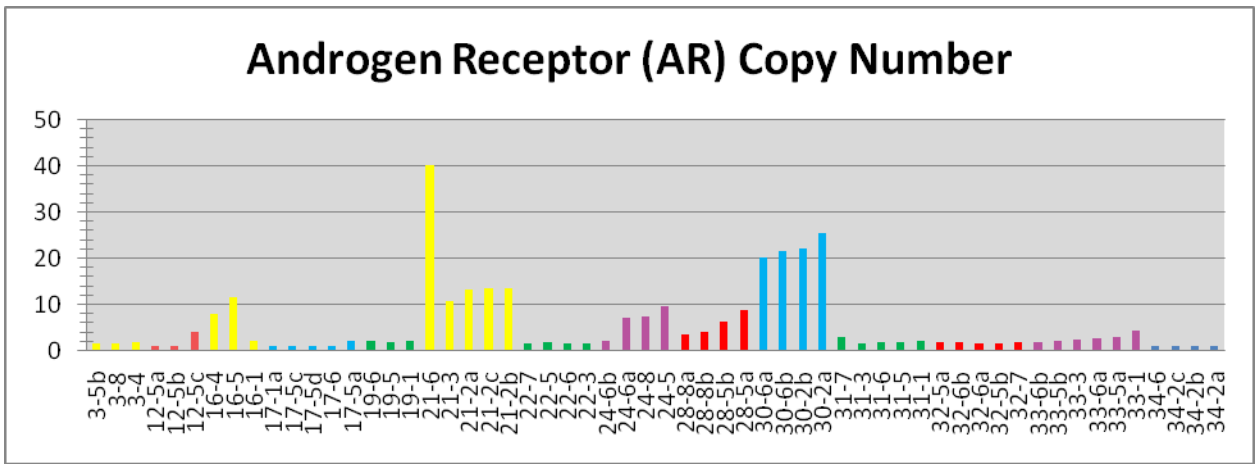
RNA Isolation and cDNA synthesis: Metastatic prostate cancer tissue sections were cryostat dissected as described previously[11] and total RNA was isolated as described previously[12]. The quality and concentration of the isolated RNA was determined using the Agilent 2100 Bioanalyzer Total RNA Nano Series II assay (Agilent, Santa Clara, CA). First strand cDNA synthesis was performed using 500 ng total RNA, 0.5 µg oligo (dT), and 200 units of SuperScript II reverse transcriptase (Invitrogen, Carlsbad, CA) in a volume of 20µL. Primers for TMPRSS2 and ERG real-time PCR reactions were obtained from Refseq sequence id numbers NM_005656 (TMPRSS2) and NM_004449 (ERG). Forward and reverse TMPRSS2-ERG fusion primers are (TMPRSS2 12-28) 5'-caggaggcggaggcgga-3' and (ERG 762-742) 5'-ggcgttgtagctgggggtgag-3'. Another primer set (ERG 992-1316 F 5'-ggcgttgtagctgggggtgag-3' and R 5'-ccgtggaagtcgaacttgt-3') was used to amplify the 3' end of ERG transcripts originating from both fused and non-fused (wt) transcripts. PCR was carried out with 5 µl of a 1 to 6 dilution of cDNA in a total reaction volume of 50 µl. Cycling conditions were 95C 2min, 95C 30 sec, 58C 30 sec, 72C 1 min, 36 cycles for the wild type erg amplification, and 39 cycles for the TMPRSS2-ERG fusion, 72C for 10min. Amplified products were resolved in 1% agarose and stained with Ethidium Bromide.



Supplementary Figure 10: TMPRSS2-ERG (T-E) Fusion Transcript and ERG Transcript in Metastatic Prostate Cancers, representative data. Lanes identified by Case Number and Sample Identifier contained in Supplementary Table 1. VCaP is included as a T-E fusion positive control. T-E transcript is uniformly present in all metastases studied in subjects with ERG deletion in genomic DNA, and uniformly absent in all samples in subjects without ERG deletion. 3' ends of ERG transcripts are present in all samples studied.

Autopsy Subject	Affy6 ERG Deletion Status	TMPRSS2-ERG Fusion Status (# anatomically separate cancer samples
3	positive	not done
16	positive	positive (3)
17	positive	positive (3)
19	negative	negative (4)
22	negative	negative (4)
28	positive	not done
30	positive	positive (2)
31	positive	positive (1)
32	negative	not done
33	negative	negative (1)
34	negative	not done

Supplementary Table 9: Summary of TMPRSS2-ERG (T-E) Fusion Transcript, ERG Transcript, and ERG genomic status in 18 anatomically separate prostate cancer metastases from 14 subjects studied by Affy6.



Supplementary Figure 11: Androgen Receptor Copy Number in 58 samples studied by Affy6. Note that standards for interpretation for very high copy number values using Affy6 do not yet exist, so copy number above 2 should be interpreted with caution. Sample Identifiers are detailed in Supplementary Table 1.

Supplementary Methods: Analysis of Subject-Specific Clonal and Nonclonal Genomic Change Frequencies

For each of 14 subjects whose samples were studied by Affy6, we classified each of the 52221 channels of segmented Affy6 data into one of the following four categories:

C_1: All samples have value 'loss' ("All Loss")

C_2: All samples have value 'gain' ("All Gain")

C_3: All samples have value 'no gain or loss' ("All No Change")

C_4: Samples have at least 2 of the 3 values above ("All Mixed").

For each subject, we count the number of segments belonging to the 4 categories respectively as $count(C_j),\ j = 1,2,3,4$, and take the empirical probabilities as a measure of genomic instability of this subject:

$$\hat{p}_j = \frac{count(C_j)}{d},\ j = 1,2,3,4$$

, where d is the number of segments.

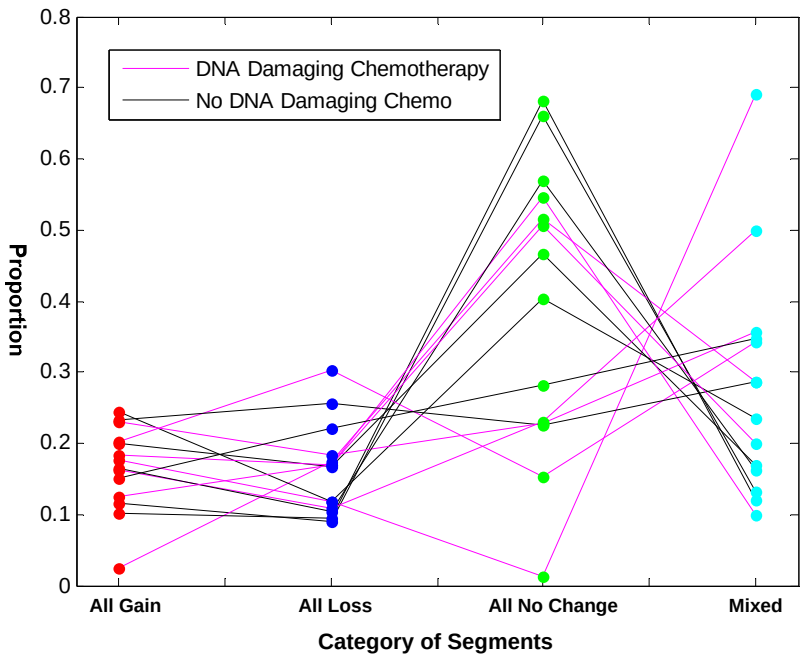
Since variable numbers of anatomically separate metastatic DNA samples were studied per subject (varies from 3-6), we made further adjustments to the proposed measure, in order to do fair comparison between subjects with different number of samples. Subjects with 3 samples studied use the formula above. For subjects where 4-6 samples were studied, we chose all possible 3-sample subsets, calculated the empirical probabilities of each subset, and averaged the empirical probabilities to obtain the adjusted measure for these subjects. The results of this analysis are contained in Supplementary Table 10.

Subject Number	3	12	16	17	19	21	22	24	28	30	31	32	33	34
All Gain	0.1775	0.1165	0.0246	0.1994	0.1025	0.2456	0.2027	0.1833	0.2319	0.2309	0.1626	0.1506	0.1659	0.1242
All Loss	0.1180	0.0893	0.1736	0.1661	0.0943	0.1171	0.3022	0.1696	0.2551	0.1841	0.1087	0.2200	0.1031	0.1693
All No Change	0.0131	0.6614	0.5157	0.4661	0.6820	0.4029	0.1521	0.5471	0.2259	0.2273	0.2293	0.2816	0.5689	0.5068
All Mixed	0.6914	0.1329	0.2861	0.1685	0.1212	0.2343	0.3430	0.1001	0.2871	0.3577	0.4994	0.3478	0.1621	0.1997

Supplementary Table 10: Analysis of Subject-Specific Clonal and Nonclonal Genomic Change Frequencies

Subject Number	3	12	16	17	19	21	22	24	28	30	31	32	33	34
DNA Damaging Chemo	Yes	No	Yes	No	No	No	Yes	Yes	No	Yes	Yes	No	No	Yes
Specific Chemo: C:Cyclophosphamide T:Topotecan E:Etoposide CP:Carboplatin	C		C				T	T		T	T			T,E,CP

Supplementary Table 11: DNA Damaging Agents Received by 14 subjects studied by Affy6. Subjects' treatment with DNA damaging drugs (alkylating agents, platinum compounds, topoisomerase poisons) are recorded below. Exposure to DNA damaging chemotherapy was analyzed because they are judged most likely to have an effect on DNA copy number, as compared to Microtubule disrupting drugs (vinca alkyloids, taxanes, others), or Other chemotherapy (phenylbutyrate, atrasentan, marimostat, suramin) which some of the subjects received. Small subject group size precluded analysis of frequency patterns in relation to specific agents received beyond the general DNA-damaging category.



Supplementary Figure 12: Plot of Subject-Specific Clonal and Nonclonal Genomic Change Frequencies by Treatment type. Subjects denoted by the magenta line received DNA-damaging chemotherapy, those marked by a black line received no DNA-damaging chemotherapy

Statistical Analysis of Subject-Specific Clonal and Nonclonal Genomic Change Frequencies by DNA-damaging chemotherapy status

We consider component "All gain", "All loss", "Mixed" in the proportion vector and define the summary statistic as the standardized distance between average proportion vector of subjects in the two treatment groups.

$$M = (\overline{\mathbf{p}_1} - \overline{\mathbf{p}_2})^T \Sigma^{-1} \left(\overline{\mathbf{p}_1} - \overline{\mathbf{p}_2}\right).$$

Then we set the null hypothesis as "there is no association between the treatment type and proportion vectors" and did the Random Permutation Test (RPT). The label (treatment type) assignment of subjects is random permuted, to calculate the summary statistic. We did 10000 permutations, and the estimated P value is about 0.2584, accepting the null hypothesis that there is no difference in combined "All gain", "All loss", and "Mixed" segment frequencies among Subjects according to DNA-damaging chemotherapy

status. We also did RPT based on any 3 components of the 4-D proportion vector, and the P value is very similar (around 0.25~0.26).

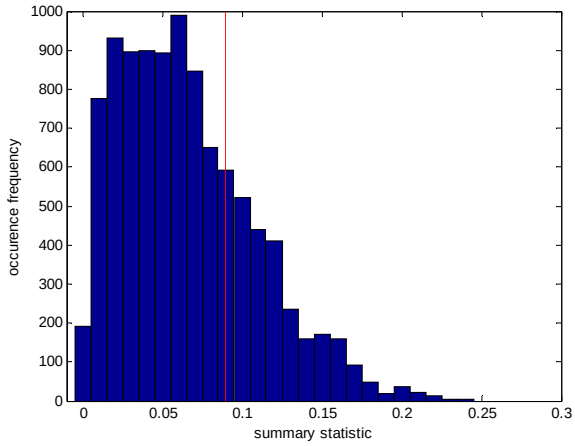


Figure. 13. Histogram of the Random Permutation Test, red line denotes the Mahalanobis distance calculated from ground truth label assignment.

Genomic segments whose copy number status among all samples for a given subject fall into the "mixed" category contain changes that are less likely to be clonal than those of the three other groups, and are more likely to have arisen after an initial genomic damage event leading to clonal changes shared among all samples. DNA-damaging chemotherapy was received by each subject long after the genomic damage event leading to the clonal changes (ie, metastasis had already occurred at the time chemotherapy had received). We separately analyzed the "Mixed" category of changes using techniques similar to those used for the analysis of all four categories of change discussed above. We calculated the mean of the "mixed" proportions of subjects in each treatment group, and used the difference between the mean of the 2 groups as a summary statistic. We set the null hypothesis as ""there is no association between the treatment type and the proportion of the Mixed category", and did a Random Permutation Test (RPT). The label (treatment type) assignment of subjects was randomly permuted to calculate the summary statistic. We did 10000 permutations, and the estimated P value is about 0.0893. Results are illustrated in Figure 13. The null hypothesis is accepted. We detected no difference in any of the Genomic Change

Frequencies calculated based on DNA-damaging chemotherapy status. These results are based on only 14 subjects' multiple metastatic samples studied. Because genomic change patterns do vary greatly among subjects, additional analysis in larger numbers of well-characterized subjects appears warranted.

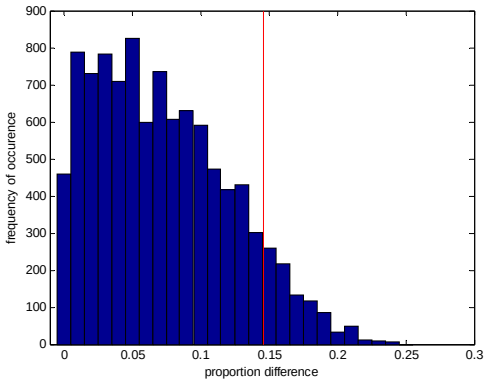


Fig. 14 Histogram of Random Permutation Test of Mixed Change Proportions among subjects with and without DNA Damaging Chemotherapy. There are 10000 permutations in total. Red line is the proportion difference based on ground truth treatment type. P value is 0.0863.

Supplementary References

(1) Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 1992; 258:818-821.

(2) Visakorpi T, Kallioniemi AH, Syvanen AC, Hyytinen ER, Karhu R, Tammela T, Isola JJ, Kallioniemi OP. Genetic changes in primary and recurrent prostate cancer by comparative genomic hybridization. Cancer Research 1995; 55(2):342-347.

(3) Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001; 98(9):5116-5121.

(4) Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998; 95(25):14863-14868.

(5) Zhu Y, Wang Z, Miller DJ, Clarke R, Xuan J, Hoffman EP, Wang Y. A ground truth based comparative study on clustering of gene expression data. Front Biosci 2008; 13:3839-3849.

(6) Zhu Y, Li H, Miller DJ, Wang Z, Xuan J, Clarke R, Hoffman EP, Wang Y. caBIG VISDA: modeling, visualization, and discovery for cluster analysis of genomic data. BMC Bioinformatics 2008; 9:383.

(7) Jain AK, Duin RPW, Mao J. Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000; 22:4-37.

(8) Hastie TR, Tibshirani TR, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 1st ed. New York: Springer, 2001.

(9) Loog M, Duin R, Haeb-Umbach R. Multiclass linear dimension reduction by weighted pairwise fisher criteria. IEEE Transactions on Pattern Analysis and Machine Intelligence 2001; 23:762-766.

(10) Zar JH. Biostatistical Analysis. 4th ed. Upper Saddle River: Prentice-Hall, Inc., 1999.

(11) Suzuki H, Freije D, Nusskern DR, Okami K, Cairns P, Sidransky D, Isaacs WB, Bova GS. Interfocal heterogeneity of PTEN/MMAC1 gene alterations in multiple metastatic prostate cancer tissues. Cancer Res 1998; 58(2):204-209.

(12) Luo J, Duggan DJ, Chen Y, Sauvageot J, Ewing CM, Bittner ML, Trent JM, Isaacs WB. Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. Cancer Res 2001; 61(12):4683-4688.