SUPPLEMENTAL METHODS

Amplification of genomic DNA. Samples of DTCs or CD45+ normal cells were dried using a Speedvac and reconstituted in 4 µl of proteinase-K buffer (1 × Pharmacia One-Phor-All-Buffer-Plus (GE Healthcare, Piscataway, NJ), 0.67% Tween 20, 0.67% Igepal, and 0.67 mg/ml Proteinase K) and incubated for 10 h at 42ºC. The material was then amplified as described previously (1), and is referred to as rare cell genomic amplification (RCGA). For proof-of-principle tests, RCGA was applied to bulk LNCaP DNA (200 ng/reaction) collected by ethanol precipitation from 10^6 LNCaP tissue culture cells and four pools of 10-20 LNCaP cells collected by micromanipulation from the same passage. LCM-collected samples (tumor and normal) each containing 2000-4000 cells were incubated overnight in proteinase K buffer (1% Tween 20, 1 mg/ml Proteinase K, and 1 x TE). Genomic DNA was then isolated using the Qiagen QIAamp DNA Micro Kit (Qiagen Inc, Valencia, CA). The isolated genomic material was amplified using the WGA2 method from Sigma-Aldrich (Sigma-Aldrich, St. Louis, MO) following the manufacturer's instructions.

The reference DNA for all arrays in this work was obtained from the peripheral blood of a single female individual. This reference DNA was isolated from whole blood using the QIAamp DNA Blood Mini Kit (Qiagen Inc, Valencia, CA) and quantitated using the Quant-iT PicoGreen dsDNA Assay (Invitrogen, Carlsbad, CA). The amplification method (i.e., RCGA or WGA2) and a roughly equivalent amount of input DNA for all reference samples matched the test sample. All amplified DNA samples were cleaned using the QIAquick PCR Purification column (Qiagen). Five µls of each amplification was run on a 1% agarose gel by electrophoresis to confirm quality (i.e., a smear from approximately 2 kb – 100 bp) and quantity (relative to a concentration standard prepared from sonicated salmon sperm DNA).

Array CGH. The amplified test and reference material, 1 µg each, was labeled with Cy-5-dUTP and Cy-3-dUTP (GE Healthcare Bio-Sciences Corp., Piscataway, NJ), respectively, using the Bioprime Array CGH kit (Invitrogen). Labeled material was cleaned using the columns included with the labeling kit. Each labeled sample with 125 µg of human Cot-1 DNA (Invitrogen) and 30 µg of yeast tRNA (Invitrogen) were concentrated with a YM-30 centrifugal filter unit (Millipore, Billerica, MA). Concentrated test and reference samples were then combined and dried in a Speedvac.

The bacterial artificial chromosome (BAC) clones that make up the array, the hybridization procedures, and the scanning technology for this study has been described previously (2). In brief, the clones that make up this array are a subset of those described by the BAC Resource Consortium (3) with the addition of a set of clones containing tumor-related genes (2). A total of 4204 different BAC clones from known genomic locations were used in the final analysis. The analyses here use the BAC coordinates in the May 2004 sequence assembly (Build 35). The median spacing of the clones is 413 kbp, when pericentric heterochromatic regions and the short arms of acrocentric chromosomes are excluded.

Array analysis. The log$_2$-ratio data for each array were normalized with a block-level Loess algorithm to correct for intensity- and location-based dependencies (4). The values for the duplicate spots representing each BAC were averaged post-normalization. The loess-normalized averaged data for each array were processed by Circular Binary Segmentation (CBS,(5)), a method for organizing array CGH data into genome segments of approximately equal copy number. Thresholds for calling loss and gain were determined using the array results obtained from the normal-cell samples that we collected in parallel with the DTCs (i.e., the 9 samples of RCGA-amplified CD45-positive bone marrow cells) and primary tumors (i.e., the 5 samples of WGA2-amplified LCM-isolated normal prostate cells), respectively. Specifically, the threshold values were the 99th percentile, as calculated from the mean and standard deviation, of the segment values of each chromosome across all normal-cell arrays in each set. These chromosome-specific thresholds were used to define copy-number changes from the CBS-segmented data of the DTCs or primary-tumor cells.

LNCaP FISH. DNA from five BACs (RP11-108K14, RP11-138P20, RP11-346N8, RP11-520H16, and RP11-678D20) was used as FISH probes. Each BAC DNA sample was directly labeled with SpectrumRed-dUTP using a nick translation kit (Abbott Molecular Inc., Des Plaines, IL). Metaphase preparation and hybridization were performed as described previously (6).

Comparison of matched primary-tumor and LDC pairs. We counted the number of concordant regions of gain or loss in a matched pair of samples; a concordant region was one in which >30% of BACs in one sample's deviation were encompassed by the other sample's deviation, and vice versa.

To determine the likelihood of seeing the observed number of concordant sites of loss or gain between paired primary and DTC samples, we simulated datasets with the same number and sizes of loss and gain as the DTC dataset in question. The following method is described in greater detail in Young et al. (Young et al., in preparation)[1]. Because our concordance counts depend on the number of overlapping BACs, real and simulated region sizes were expressed as the number of array BACs they encompass,

[1] Young JM, Endicott RM, Parghi SS, Walker M, Kidd JM, Trask BJ. The functional olfactory receptor repertoire varies greatly in the human population. Manuscript in preparation 2008.

rather than in bp. We constructed an artificial genome in which each autosome was represented once with its size expressed as the number of BACs that correspond to that chromosome on the array. The X and Y chromosomes were excluded from this analysis. A large number of possible start co-ordinates were picked randomly within the artificial genome, using R's "runif" function to sample from a uniform distribution. Next, we generated the start and end coordinates of the simulated region based on one of the random start positions and the size of largest real deviation. We continued generating simulated intervals in this way, using real region sizes in decreasing order until all real deviations were represented in the simulation. During this process, if any simulated region overlapped with any region(s) previously simulated, alternative randomly chosen start positions were iteratively tested until a region was identified with no overlaps.

For 10,000 simulated DTC-like datasets, we determined how many sites of loss or gain were concordant with regions found in the real dataset from the primary tumor. The proportion of the 10,000 simulated sets that showed at least as many concordant sites as the real paired datasets gives an approximate p-value for how likely such overlap would occur by chance.

GO analyses. To determine which functional categories of genes might be enriched in sites of loss or gain in particular groups of samples, we examined regions of change seen in at least 20% of LDC, primary tumor, or ADC samples. First, we obtained a list of all genes in the genome, together with their genomic location and Entrez IDs, using Bioconductor's biomaRt package (7) to connect to the ensembl_mart_37 database, the most recent Ensembl database that uses co-ordinates from the May 2004 version of the genome assembly. Genes without chromosome location information or Entrez IDs were eliminated from further analysis. Genes without GO category information according to Bioconductor’s org.Hs.egGO were also eliminated. The resulting list of genes formed our "gene universe" in the GO analyses described below.

Second, we determined the subset of genes from the “universe” that were encompassed by regions of copy-number change in >20% of LDC, primary tumor, or ADC samples. Losses and gains were considered separately. We used the “changed gene” list, together with the "gene universe" list to test for enrichment of all Biological Process ("BP") GO terms using the hyperGTest function of Bioconductor's GOstats package (8), with additional parameters as follows: annotation = org.Hs.eg.db, hgCutoff = 0.001, conditional = FALSE. GOstats implements a hypergeometric test to determine the probability of finding the observed enrichment of each category by chance; it should be noted that we tested a large number of categories in our GOstats analyses, and the p-values it reports do not include any adjustment for multiple testing.

Third, due to our concern that genomic clustering of functionally related genes might invalidate the hyperGtest’s assumption of independence, we used simulations to determine the likelihood that a category would be enriched by chance in a dataset of alterations with the same characteristics as our real datasets. These simulations were conducted generally as described above for the concordance analysis of matched pairs of primary tumors and LDCs and as detailed in Young et al.[1]. However, unlike the concordance analysis, here the real and simulated data were expressed in bp and the artificial genome consisted of two copies of each autosome and one copy of each sex chromosome (all chromosome lengths reflected their length in the May 2004 genome assembly). For each of the six real datasets (i.e., losses for >20% of LDC, primary tumor, or ADC samples and gains for >20% of LDC, primary tumor, or ADC samples), we produced 1000 simulated datasets. For each simulated dataset, we repeated the process of finding genes and performed GO enrichment tests. For each category enriched in the real data set, we determined the proportion of simulations that also showed enrichment of that category. This proportion is our “SimPValue”, an estimate of how likely it is that a given category enriched by chance in a set of genomic regions with the same size distribution as the one analyzed.

SUPPLEMENTAL RESULTS

RCGA and WGA2 are comparable. We have compared arrays produced by amplification of normal cells by either RCGA (n=9) or WGA2 (n=5). Note that these arrays are excellent representations of the experimental bias that might affect our LocDCs or primary tumors, respectively, as the normal cells subjected to RCGA were small numbers of normal cells collected in parallel with our LocDC samples and the normal cells amplified by WGA2 were stromal cells collected by LCM from normal prostate. Our threshold method for calling loss and gain detects some site of change in both these of normal data sets. These observations are presumably a result of experimental noise. We compared the number and amount of deviant material identified in the RCGA- and WGA2-produced normal-cell arrays and found no statistical difference between the two amplification schemes (Student’s t-test p=0.3485 and 0.3438, respectively).

Next, we compared the array outputs from RCGA and WGA2 amplifications performed on two aliquots of the same isolate of LNCaP DNA. In this comparison, we first found no significant difference in the deviant segment values (as an indicator of dynamic range) between these two amplification schemes (p=0.4820). Second, we found significant concordance in the array results. Our tests of concordance were performed as defined and described above in the Supplemental Methods. The DNA amplified by RCGA showed 15 sites of change, and the WGA2 DNA showed 16 deviant sites. There were 12 concordant alterations, all previously observed for LNCaP. These 12 sites all surpassed our criteria for concordance. For six sites, 100% of the encompassed BACs registered as deviant using both methods. For the other six sites, over 60% of the encompassed BACs did so. Furthermore, we found that this pair of LNCaP arrays has significantly more concordant changes than the number expected if the deviant segments were randomly distributed ($p<0.0001$).

The differences between the amplification of LNCaP DNA by the two methods are worth noting. Three deviations in the RCGA-LNCaP array and four in the WGA2-LNCaP array were not observed for in the other array. The former have all been previously reported for LNCaP and the latter have not, implying that RCGA might be more sensitive to genomic change than WGA2. Thus, a total of 7 deviations were not consistent between RCGA and WGA2 for the LNCaP amplifications. These analyses indicate that a comparison of samples amplified by RCGA to those subjected to WGA2 is reasonable and will reveal the underlying relationship between two related samples, but might yield some inconsistencies. The two methods produce a comparable degree of experimental noise, show no significant difference in dynamic range, and detect highly concordant deviations in test DNA.

REFERENCES

1. Klein CA, Schmidt-Kittler O, Schardt JA, Pantel K, Speicher MR and Riethmuller G. Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. Proc Natl Acad Sci U S A 1999;96:4494-9.

2. Loo LW, Grove DI, Williams EM, et al. Array comparative genomic hybridization analysis of genomic alterations in breast cancer subtypes. Cancer Res 2004;64:8541-9.

3. Cheung VG, Nowak N, Jang W, et al. Integration of cytogenetic landmarks into the draft sequence of the human genome. Nature 2001;409:953-8.

4. Yang YH, Dudoit S, Luu P, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002;30:e15.

5. Olshen AB, Venkatraman ES, Lucito R and Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 2004;5:557-72.

6. Trask B and Pinkel D. Fluorescence in situ hybridization with DNA probes. Methods Cell Biol 1990;33:383-400.

7. Durinck S, Moreau Y, Kasprzyk A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 2005;21:3439-40.

8. Falcon S and Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics 2007;23:257-8.

Supplemental Table 1. Gleason grade and TNM stage for localized patients (A) and advanced patients (B). For the latter, chemical castration status is also indicated.

A.

Disseminated cell ID	Gleason grade	TNM stage
1586*	6	T2cN0
1685	6	T2b
1688	6	T2c
1695*	6	T2c
1728	6	T2b
1734	6	T2c
1762	6	T2a
1778	6	T2c
1785*	6	T2c
1815	6	T2cN0
1848	6	T2c
1888	6	T2cN0
1892	6	T2cN0
1918*	6	T2cN0
1923	6	T2cN0
1953	6	T2c
1967	6	T2c
2121	6	T2cN0
1508	7	T3a
1527	7	T2c
1585	7	T2cN0
1617	7	T2aN0
1636	7	T3a
1654*	7	T3a
1658*	7	T2a
1682*	7	T2c
1719	7	T2c
1720*	7	T3a
1788	7	T3a
1798	7	T3a
1833	7	T2c
1834*	7	T2a
1845	7	T2c
1847	7	T2c
1849	7	T2c
1851	7	T2c
1891	7	T2cN0
1922	7	T2cN0
1924	7	T2aN0
1937	7	T2cN0
1945	7	T3a
2120	7	T2cN0
1588	9	T2cN0
1767	9	T3b
1300	Unknown	Unknown
1710	Unknown	Unknown
1721	Unknown	Unknown
1894	Unknown	Unknown

*Matched primary tumor samples were collected for the indicated disseminated cell IDs.

B.

Disseminated cell ID	Gleason grade	TNM stage	Chemical castration
1696	7	T4N1	No
1773	7	T3aN1	No
1823	7	T2a	Yes
1856	7	T3aN1	No
1796	8	T2cM+	Yes
1865	9	M+	Yes
1877	9	T3aM+	No
1989	9	M+	Yes
1677		M+	Yes
1776		M+	Yes
1965		M+	Yes

Supplemental Table 2. Regions of copy-number loss (A) and gain (B) identified in 3 or more (≥27%) AdvDCs. The chromosome, genome start, and end positions of deviant regions are given. Overlapping and nested deviant regions are subdivided to indicate the regions with even higher frequency of copy-number change. The loci in bold are those sites that were also observed in >20% of the LocDC samples. Base pair positions have been rounded to the nearest one hundredth.

A.
Losses

Chr.	Start	End	% of ADCs
1	36361300	57320100	27
2	129446100	135823300	27
3	1859900	4016400	27
3	54007400	62456000	27
3	71172800	80133700	27
3	124525900	131810600	27
3	151508500	163022400	27
3	178755600	198504700	27
4	101800	8096400	27
	8096400	10705200	36
	10705200	11010400	27
4	17670300	40593600	27
4	182769200	183933500	27
5	97928000	103075900	27
6	75016100	93870300	36
	93870300	108616100	27
	108616100	120812100	36
	120812100	123463900	27
6	136180100	159454900	27
	159454900	161455600	36
	161455600	170881200	27
8	**304200**	**3899600**	**27**
	3899600	24892800	36
	24892800	30555600	27
9	77172300	78282000	27
10	214400	7937400	27
	7937400	12334300	36
10	21903100	22702300	27
10	33372800	42817200	27
	42817200	43432300	36
	43432300	43603000	45
	43603000	49230200	55
	49230200	50248100	45
	50248100	73256800	36
	73256800	78621100	27
	78621100	103808000	36
	103808000	105726300	27
	105726300	**112075300**	**36**
	112075300	113462400	27
	113462400	115937500	36
	115937500	**129427500**	**45**
	129427500	**135117700**	**55**

B.
Gains

Chr.	Start	End	% of ADCs
1	120000200	146522900	27
	146522900	161466500	36
	161466500	164824300	45
	164824300	166889000	36
	166889000	196846000	27
	196846000	198600100	36
	198600100	212033800	45
	212033800	242400400	36
2	44100	8639100	36
	8639100	9443700	27
	9443700	26184700	36
	26184700	28514900	45
	28514900	31327600	36
	31327600	42391900	27
	42391900	43985600	36
	43985600	49511100	27
	49511100	77192000	36
	77192000	86778900	27
	86778900	112461200	36
	112461200	124769800	27
2	168466100	224418400	27
2	224577900	242221600	27
3	231500	1859900	27
3	4016400	4330000	27
	4330000	12025500	36
	12025500	15780400	27
4	140977600	143001500	27
5	561600	4965700	36
	4965700	18874700	27
5	25999400	50107900	27
5	150272800	150300400	27
	150300400	151552300	36
	151552300	179467100	27
	179467100	180611400	36
6	180600	52696900	27
	52696900	53474200	36
	53474200	55544200	27
7	65973900	79491500	27
7	126695300	133346500	27
8	38183000	40762900	27
8	48143900	57031700	27
	57031700	99900700	36

	135117700	135315300	45
11	7687400	12608600	27
	12608600	14198500	36
	14198500	28748100	27
	28748100	32183800	36
	32183800	38672900	45
	38672900	43681700	36
	43681700	44879200	27
11	93214200	101149100	27
11	104171100	107883700	27
	107883700	127647400	36
	127647400	133686900	27
13	32090300	35802700	27
	35802700	49029900	36
	49029900	**57108100**	**45**
	57108100	**66378400**	**36**
	66378400	**73284100**	**45**
	73284100	**73620400**	**55**
	73620400	**76931000**	**45**
	76931000	**77899000**	**36**
	77899000	**80203300**	**45**
	80203300	**93718500**	**55**
	93718500	**101030200**	**36**
	101030200	**113351500**	**27**
14	19570800	30430800	27
14	34404800	69266000	27
	69266000	69652600	36
	69652600	75620300	27
15	30180000	41591900	27
	41591900	53535600	36
	53535600	61375300	45
	61375300	77888800	36
	77888800	88050500	27
16	**31443700**	**53822000**	**27**
	53822000	**54805500**	**36**
	54805500	**59251000**	**27**
	59251000	**60811600**	**36**
	60811600	**64421400**	**45**
	64421400	70010000	27
	70010000	88125800	36
	88125800	88612600	27
17	1335600	24318000	27
19	**241300**	**63560200**	**27**
21	25756900	26441500	27
	26441500	26777000	36
	26777000	26923000	27
21	40165200	43237400	27

	99900700	146081700	27
9	222300	4575600	36
	4575600	7302900	27
9	78282000	89644000	27
	89644000	98799000	36
	98799000	103308700	27
	103308700	111819300	36
	111819300	113124900	45
	113124900	127428500	36
	127428500	**130681600**	**45**
	130681600	138256300	36
11	44879200	57196000	27
*11	**63363100**	**76964700**	**27**
	76964700	77218000	36
	77218000	93214200	27
11	101149100	104171100	27
11	127647400	131037500	27
16	75800	2425400	36
	2425400	3292400	27
16	29550800	34476100	27
17	42087500	73516500	27
	73516500	78311500	36
22	20289400	22928700	27
X	7483600	28209600	36
	28209600	39019200	27
	39019200	66108300	36
	66108300	67961300	45
	67961300	69047600	36
	69047600	73682500	27
	73682500	100670300	36
	100670300	148896100	27
	148896100	153951900	36

Abbreviations: **Chr**, chromosome.

*These deviations are probably artifact as they were also observed in ~20% of arrays on 10-20 normal cells.

Supplemental Table 3. Regions of copy-number loss and gain in >20% of LocDCs (A), primary tumors (B), or AdvDCs (C) used for gene ontology analysis. The chromosome (Chr.), genome start, and end positions of deviant regions are given.

A.

LocDCs

Losses

Chr.	Start	End
4	28028903	30613014
5	4965703	9240476
5	18874717	26168302
8	304159	4386750
10	214399	1474060
10	107761235	111047380
10	128180749	135165811
11	40333699	41710577
12	124562362	129332501
13	52183617	113351493
16	49274146	54394461
16	57249741	64579967
16	81539093	81784852
18	56872541	57340436
19	33315121	37120074

Gains

Chr.	Start	End
3	42812666	53827245
3	123243312	131961221
3	193633588	198504724
9	125271843	133650036
11	56357374	77094745

B.

Primary tumors

Losses

Chr.	Start	End
1	154391405	163508931
1	235875218	237932584
2	12289583	19306523
2	121500476	156109000
4	4538586	5119073
5	66078534	128933634
5	150272828	171316611
6	64287483	137498692
6	159454947	170881221
7	35648493	42617046
8	304159	38453623
10	77226915	79157041
10	86468638	95935762
10	127832427	129796619
11	10492617	44271172
11	78034239	83842293
11	119158031	133686875
12	4485584	5783146
12	63300097	91514464
12	94705537	102604940
12	124562362	132310722
13	30131054	56820599

Gains

Chr.	Start	End
1	1036980	47620504
1	142916565	153594838
2	6608798	10952850
2	25347012	31515961
3	46857872	52743025
5	561584	1540913
6	26151436	44815782
7	106476	6396697
7	71922030	75326524
7	98937882	105570154
8	126655187	136154996
9	125698693	138256276
10	69643018	73406411
10	80532209	82331976
10	101635168	104959528
10	126193027	126665243
10	129686061	135315306
11	44098955	48247226
11	117662914	118992466
12	44524861	60652543
12	107617696	111206853
12	119040684	123499450

15	25673627	31975487
15	45134880	49448561
15	85413252	86472154
16	46203946	54988902
16	60811618	64579967
16	75563743	82429929
17	12002245	15511681
17	28794278	29841486
18	72100759	73593885
19	19747489	37612371
20	6090695	24282120
20	35432771	43172910
22	24443742	26243615

13	111758242	113351493
14	89130447	92239415
14	99137698	106175506
15	38241119	41750990
15	42546116	43812588
15	99149458	100021943
16	10159016	15681446
16	55369591	57653333
16	64952312	69066728
17	117304	8365794
17	34153267	38083571
17	68729134	78311473
18	75021265	76089909
19	241269	19877489
19	37482371	63560213
21	41630602	46912065
22	16233834	23142856
22	35686144	49441620
X	7483618	153951934

C.

AdvDCs

Losses

Chr.	Start	End
1	36361290	56762346
2	129446050	135593357
3	1859870	2675754
3	54007388	62347194
3	71172842	76611609
3	124525892	128983062
3	151508469	162207623
3	178755564	198504724
4	101785	10835224
4	17670296	40214695
4	182769217	183096645
5	97928034	102841325
6	75016060	123465414
6	136180051	170881221
8	304159	29507616
9	77172311	78142517
10	214399	11961014
10	21903086	22813984
10	33372796	135315306
11	7687390	44614067
11	93214215	100132253
11	104171070	133686875
13	32572850	113351493
14	19570807	29163547
14	34404822	74551240
15	30179961	87936225

Gains

Chr.	Start	End
1	142916565	240032059
2	44073	121679355
2	168466065	224444130
2	224577925	242221648
3	231485	397141
3	4016396	15387250
4	140977612	143144965
5	561584	16775431
5	25999396	43795937
5	150272828	180611420
6	180642	55416965
7	65973863	78862373
7	126695314	133018477
8	38183042	40742585
8	48143884	146081698
9	222268	6601726
9	78281981	138256276
11	44879165	57277679
11	63363052	93087588
11	101149074	103965030
11	127647353	130996287
16	75836	3158832
16	29550781	31443695
17	42087497	78311473
22	20289397	20719548
X	7483618	153951934

16	34476094	88612553
17	1335633	21191548
19	241269	63560213
21	25756854	26960676
21	40165171	40295171

Supplemental Table 4. GO categories enriched ($p < 0.001$ as determined by the hyperGTest package) in regions of loss observed in >20% of LocDC (A), primary tumor (B), and AdvDC (C) samples, respectively. The number of “changed genes” and the number of GO categories they represent is given under each sample heading. The “gene universe” included 12,639 genes for all tests. The first column for each sample type is the GO identifier (GO_ID) as designated by the GO consortium. Also given is the p-value determined by the hyperGTest algorithm (Pvalue), the number of genes for that GO category that were identified from the changed gene list (Count), the number of genes for that GO category in the total gene list (Size), and the term associated with that GO category (Term). We also give, from our simulations that take into account the physical distribution in the genome of the genes of particular gene categories, the number of times that the GO category was observed at a hyperGTest significance level of $p < 0.001$ in the simulations (NumSims), and the p-value as determined from our simulations (SimPValue). Those SimPValues in bold are significant with $p < 0.05$.

A.

LocDCs

157 genes in dataset and 716 GO IDs tested

GO_ID	Pvalue	Count	Size	Term	NumSims	SimPValue
GO:0016337	1.14×10^{-7}	14	223	cell-cell adhesion	1	0.051
GO:0007156	1.36×10^{-7}	10	112	homophilic cell adhesion	2	0.054
GO:0016109	9.05×10^{-4}	2	4	tetraterpenoid biosynthetic process	0	**0.005**
GO:0016114	9.05×10^{-4}	2	4	terpenoid biosynthetic process	0	**0.005**
GO:0016117	9.05×10^{-4}	2	4	carotenoid biosynthetic process	0	**0.005**

B.

Primary tumors

1372 genes in dataset and 2380 GO IDs tested

GO_ID	Pvalue	Count	Size	Term	NumSims	SimPValue
GO:0007268	4.12×10^{-6}	51	247	synaptic transmission	0	**0.011**
GO:0019226	1.05×10^{-5}	55	282	transmission of nerve impulse	0	**0.009**
GO:0007267	5.03×10^{-5}	94	584	cell-cell signaling	2	**0.014**
GO:0045639	1.66×10^{-4}	7	13	positive regulation of myeloid cell differentiation	0	**< 0.001**
GO:0051046	1.92×10^{-4}	14	45	regulation of secretion	0	**0.002**
GO:0042742	3.12×10^{-4}	19	75	defense response to bacterium	9	0.078
GO:0006813	3.38×10^{-4}	31	151	potassium ion transport	0	**< 0.001**
GO:0006874	3.58×10^{-4}	23	100	cellular calcium ion homeostasis	3	**0.015**
GO:0055074	3.58×10^{-4}	23	100	calcium ion homeostasis	3	**0.015**
GO:0009617	3.77×10^{-4}	20	82	response to bacterium	8	0.074
GO:0045648	6.32×10^{-4}	4	5	positive regulation of erythrocyte differentiation	0	**0.001**
GO:0007269	6.33×10^{-4}	10	29	neurotransmitter secretion	0	**0.001**
GO:0006875	9.81×10^{-4}	23	107	cellular metal ion homeostasis	3	**0.013**
GO:0055065	9.81×10^{-4}	23	107	metal ion homeostasis	3	**0.013**

C.

AdvDCs

3548 genes in dataset and 3225 GO IDs tested

GO_ID	Pvalue	Count	Size	Term	NumSims	SimPValue
GO:0006351	3.54×10^{-8}	668	2018	transcription, DNA-dependent	22	0.165
GO:0032774	3.76×10^{-8}	669	2022	RNA biosynthetic process	22	0.162
GO:0006355	3.94×10^{-8}	652	1966	regulation of transcription, DNA-dependent	22	0.167
GO:0016070	4.78×10^{-8}	809	2494	RNA metabolic process	22	0.149
GO:0045449	3.99×10^{-7}	680	2087	regulation of transcription	22	0.158
GO:0006350	6.10×10^{-7}	704	2173	transcription	22	0.157
GO:0031323	6.46×10^{-7}	748	2322	regulation of cellular metabolic process	20	0.145
GO:0019219	1.21×10^{-6}	693	2145	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	21	0.154
GO:0050794	1.44×10^{-6}	1132	3647	regulation of cellular process	13	0.101
GO:0019222	1.73×10^{-6}	770	2410	regulation of metabolic process	19	0.143
GO:0050789	1.13×10^{-5}	1207	3943	regulation of biological process	13	0.094
GO:0065007	1.41×10^{-5}	1299	4268	biological regulation	10	0.087
GO:0006118	9.36×10^{-4}	113	312	electron transport	0	**0.001**

Supplemental Table 5. GO categories enriched ($p < 0.001$ as determined by the hyperGTest package) in regions of gain observed in >20% of LocDC (A), primary tumor (B), and AdvDC (C) samples, respectively. The number of “changed genes” and the number of GO categories they represent is given under each sample heading. The “gene universe” included 12,639 genes for all tests. The first column for each sample type is the GO identifier (GO_ID) as designated by the GO consortium. Also given is the p-value determined by the hyperGTest algorithm (Pvalue), the number of genes for that GO category that were identified from the changed gene list (Count), the number of genes for that GO category in the total gene list (Size), and the term associated with that GO category (Term). We also give, from our simulations that take into account the physical distribution in the genome of the genes of particular gene categories, the number of times that the GO category was observed at a hyperGTest significance level of $p < 0.001$ in the simulations (NumSims), and the p-value as determined from our simulations (SimPValue). Those SimPValues in bold are significant with $p < 0.05$.

A.

LocDCs

604 genes in dataset and 1502 GO IDs tested

GO_ID	Pvalue	Count	Size	Term	NumSims	SimPValue
GO:0006071	3.40 x 10^{-4}	6	21	glycerol metabolic process	12	**0.012**
GO:0043666	4.19 x 10^{-4}	3	4	regulation of phosphoprotein phosphatase activity	0	**< 0.001**
GO:0019751	4.49 x 10^{-4}	6	22	polyol metabolic process	13	**0.013**
GO:0051180	4.93 x 10^{-4}	5	15	vitamin transport	5	**0.005**

B.

Primary tumors

3989 genes in dataset and 3233 GO IDs tested

GO_ID	Pvalue	Count	Size	Term	NumSims	SimPValue
GO:0031424	3.38 x 10^{-12}	31	35	keratinization	133	0.133
GO:0009913	6.54 x 10^{-10}	35	46	epidermal cell differentiation	133	0.133
GO:0006323	5.49 x 10^{-9}	137	287	DNA packaging	78	0.078
GO:0048730	2.59 x 10^{-8}	35	50	epidermis morphogenesis	133	0.133
GO:0008544	3.42 x 10^{-8}	69	125	epidermis development	133	0.133
GO:0006325	3.53 x 10^{-8}	132	281	establishment and/or maintenance of chromatin architecture	84	0.084
GO:0006139	4.15 x 10^{-8}	1154	3264	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	68	0.068
GO:0016070	4.20 x 10^{-8}	900	2494	RNA metabolic process	70	0.070
GO:0007398	4.38 x 10^{-8}	73	135	ectoderm development	133	0.133
GO:0032774	6.15 x 10^{-8}	741	2022	RNA biosynthetic process	77	0.077
GO:0006355	6.49 x 10^{-8}	722	1966	regulation of transcription, DNA-dependent	76	0.076
GO:0006351	7.39 x 10^{-8}	739	2018	transcription, DNA-dependent	77	0.077
GO:0031497	2.68 x 10^{-7}	51	88	chromatin assembly	129	0.129
GO:0006334	3.01 x 10^{-7}	46	77	nucleosome assembly	129	0.129
GO:0006333	3.08 x 10^{-7}	68	128	chromatin assembly or disassembly	126	0.126
GO:0007001	3.26 x 10^{-7}	154	348	chromosome organization and biogenesis (sensu Eukaryota)	59	0.059
GO:0006996	3.40 x 10^{-7}	381	982	organelle organization and biogenesis	16	**0.016**
GO:0006350	3.51 x 10^{-7}	785	2173	transcription	72	0.072
GO:0045449	4.93 x 10^{-7}	755	2087	regulation of transcription	72	0.072
GO:0051276	4.96 x 10^{-7}	157	358	chromosome organization and biogenesis	54	0.054
GO:0019219	8.89 x 10^{-7}	772	2145	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	77	0.077
GO:0065004	3.25 x 10^{-6}	69	137	protein-DNA complex assembly	121	0.121
GO:0019222	5.51 x 10^{-6}	852	2410	regulation of metabolic process	71	0.071
GO:0031323	6.83 x 10^{-6}	822	2322	regulation of cellular metabolic process	72	0.072
GO:0006259	1.68 x 10^{-5}	262	672	DNA metabolic process	22	**0.022**
GO:0002504	2.30 x 10^{-5}	14	17	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	124	0.124
GO:0019882	6.51 x 10^{-5}	32	56	antigen processing and presentation	126	0.126
GO:0043283	1.40 x 10^{-4}	1467	4360	biopolymer metabolic process	47	**0.047**
GO:0016568	1.53 x 10^{-4}	81	182	chromatin modification	1	**0.001**
GO:0048729	1.70 x 10^{-4}	37	70	tissue morphogenesis	132	0.132
GO:0016042	4.75 x 10^{-4}	49	103	lipid catabolic process	4	**0.004**
GO:0043170	7.44 x 10^{-4}	1910	5788	macromolecule metabolic process	43	**0.043**
GO:0048002	9.92 x 10^{-4}	14	21	antigen processing and presentation of peptide antigen	102	0.102

C.

AdvDCs

3904 genes in dataset and 3247 GO IDs tested

GO_ID	Pvalue	Count	Size	Term	NumSims	SimPValue
GO:0031424	2.92 x 10^{-10}	29	35	keratinization	272	0.272
GO:0006334	1.15 x 10^{-8}	48	77	nucleosome assembly	240	0.240
GO:0048730	7.05 x 10^{-8}	34	50	epidermis morphogenesis	270	0.270
GO:0006333	1.24 x 10^{-7}	68	128	chromatin assembly or disassembly	192	0.192
GO:0031497	1.26 x 10^{-7}	51	88	chromatin assembly	234	0.234
GO:0009913	3.66 x 10^{-7}	31	46	epidermal cell differentiation	271	0.271
GO:0048729	4.34 x 10^{-7}	42	70	tissue morphogenesis	266	0.266
GO:0002504	1.50 x 10^{-6}	15	17	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	267	0.267
GO:0065004	3.19 x 10^{-6}	68	137	protein-DNA complex assembly	172	0.172
GO:0019882	3.72 x 10^{-6}	34	56	antigen processing and presentation	256	0.256
GO:0007398	8.04 x 10^{-5}	63	135	ectoderm development	199	0.199
GO:0008544	8.99 x 10^{-5}	59	125	epidermis development	203	0.203
GO:0002252	9.52 x 10^{-5}	44	87	immune effector process	10	**0.010**
GO:0002526	2.54 x 10^{-4}	34	65	acute inflammatory response	17	**0.017**
GO:0002443	3.75 x 10^{-4}	36	71	leukocyte mediated immunity	2	**0.002**
GO:0002541	7.87 x 10^{-4}	19	32	activation of plasma proteins during acute inflammatory response	43	**0.043**
GO:0006956	7.87 x 10^{-4}	19	32	complement activation	43	**0.043**
GO:0002449	8.89 x 10^{-4}	33	66	lymphocyte mediated immunity	3	**0.003**