Supplementary Information

Whole exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer

Jens G. Lohr[1,2,3,11]*, Viktor A. Adalsteinsson[1,4,11]*, Kristian Cibulskis[1,11]*, Atish D. Choudhury[1,2,3], Mara Rosenberg[1], Peter Cruz-Gordillo[1], Joshua Francis[1,2], Cheng-Zhong Zhang[1,2], Alex K. Shalek[5], Rahul Satija[1], John T. Trombetta[1], Diana Lu[1], Naren Tallapragada[4], Narmin Tahirova[4], Sora Kim[1], Brendan Blumenstiel[1], Carrie Sougnez[1], Alarice Lowe[6], Bang Wong[1], Daniel Auclair[1], Eliezer M. Van Allen[1,2,3], Mari Nakabayashi[2,3], Rosina T. Lis[2], Gwo-Shu M. Lee[2], Tiantian Li[2], Matthew S. Chabot[2], Amy Ly[7], Mary-Ellen Taplin[2,3], Thomas E. Clancy[2,3,6], Massimo Loda[1,2,3,6], Aviv Regev[1,8,9], Matthew Meyerson[1,2,3], William C. Hahn[1,2,3,6], Philip W. Kantoff[2,3], Todd R. Golub[1,2,3,9], Gad Getz[1,7]**, Jesse S. Boehm[1]**, J. Christopher Love[1,4,10]**

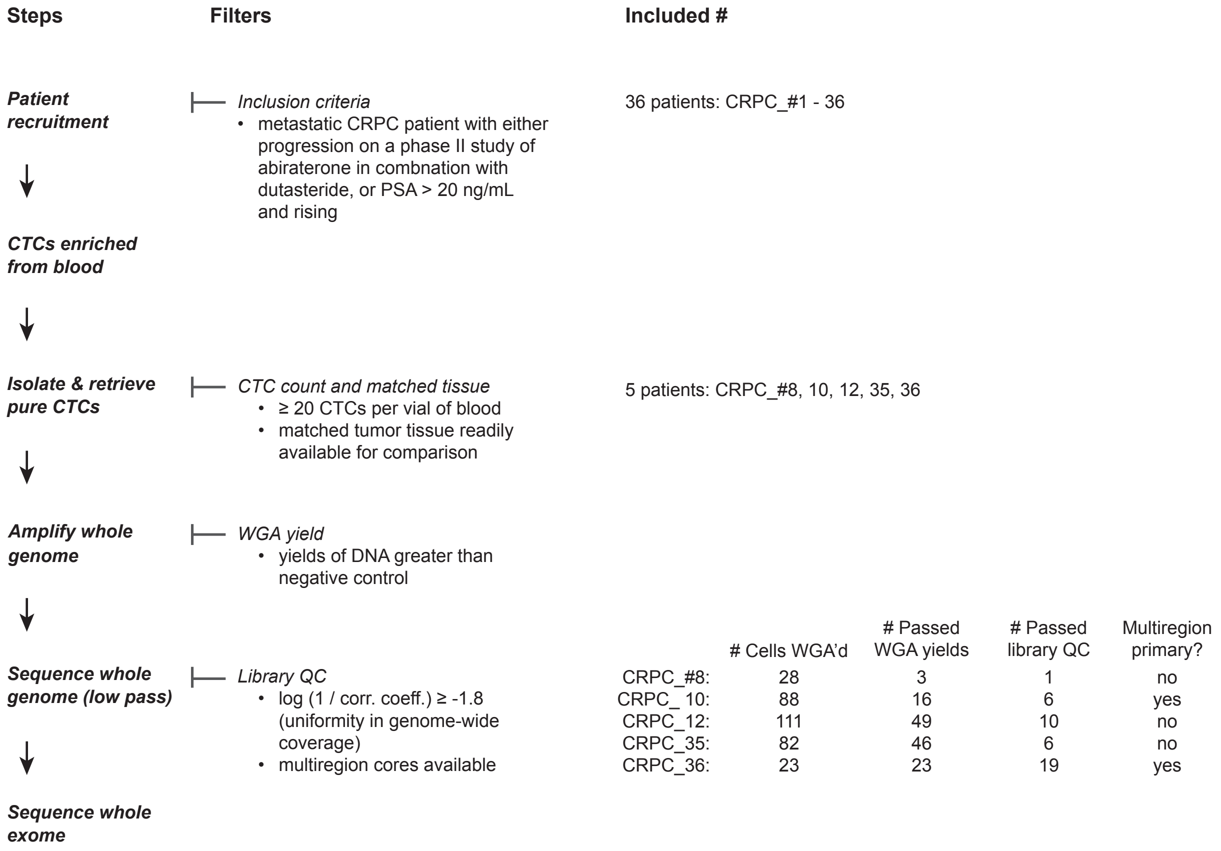
(1) The Eli and Edythe Broad Institute, Cambridge, Massachusetts 02412, USA
(2) Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA
(3) Harvard Medical School, Boston, Massachusetts 02115, USA
(4) Koch Institute for Integrative Cancer Research at MIT, Massachusetts Institute of Technology, 77 Massachusetts Ave., Bldg. 76-231, Cambridge, Massachusetts 02139, USA.
(5) Department of Chemistry and Chemical Biology and Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA
(6) Brigham and Women's Hospital, Boston, Massachusetts 02115, USA
(7) Massachusetts General Hospital, Boston, Massachusetts 02114, USA
(8) Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
(9) Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA
(10) Ragon Institute of MGH, MIT, and Harvard, Cambridge, Massachusetts 02139, USA
(11) *These authors contributed equally to this work.

****Corresponding authors**
J. Christopher Love
77 Massachusetts Ave., Bldg. 76-253
Cambridge, MA 02139
Phone: 617-324-2300
Email: clove@mit.edu

Jesse S. Boehm
7 Cambridge Center, 4021
Cambridge, MA 02142
Phone: 617-714-7494
Email: boehm@broadinstitute.org

Gad Getz
301 Binney Street
Cambridge, MA 02142
Phone: 617-714-7621
Fax: 617-714-8931
Email: gadgetz@broadinstitute.org

Nature Biotechnology: doi:10.1038/nbt.2892



Steps | **Filters** | **Included #**

Patient recruitment

Inclusion criteria
- metastatic CRPC patient with either progression on a phase II study of abiraterone in combnation with dutasteride, or PSA > 20 ng/mL and rising

36 patients: CRPC_#1 - 36

CTCs enriched from blood

Isolate & retrieve pure CTCs

CTC count and matched tissue
- ≥ 20 CTCs per vial of blood
- matched tumor tissue readily available for comparison

5 patients: CRPC_#8, 10, 12, 35, 36

Amplify whole genome

WGA yield
- yields of DNA greater than negative control

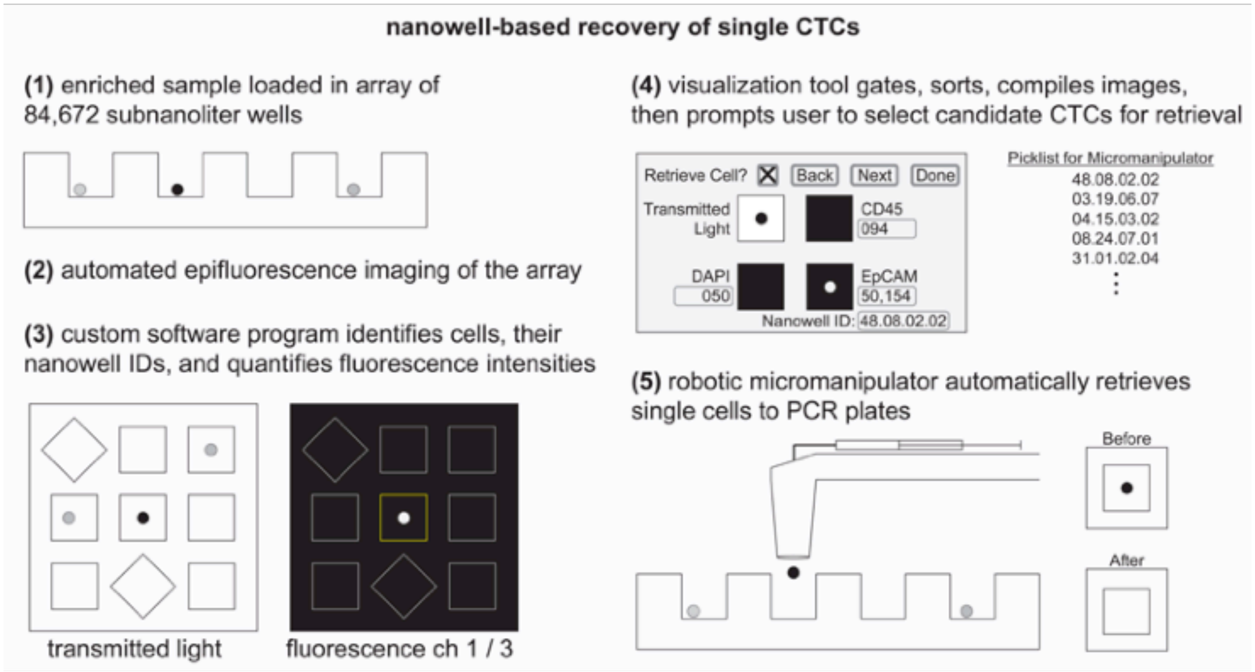
Sequence whole genome (low pass)

Library QC
- log (1 / corr. coeff.) ≥ -1.8 (uniformity in genome-wide coverage)
- multiregion cores available

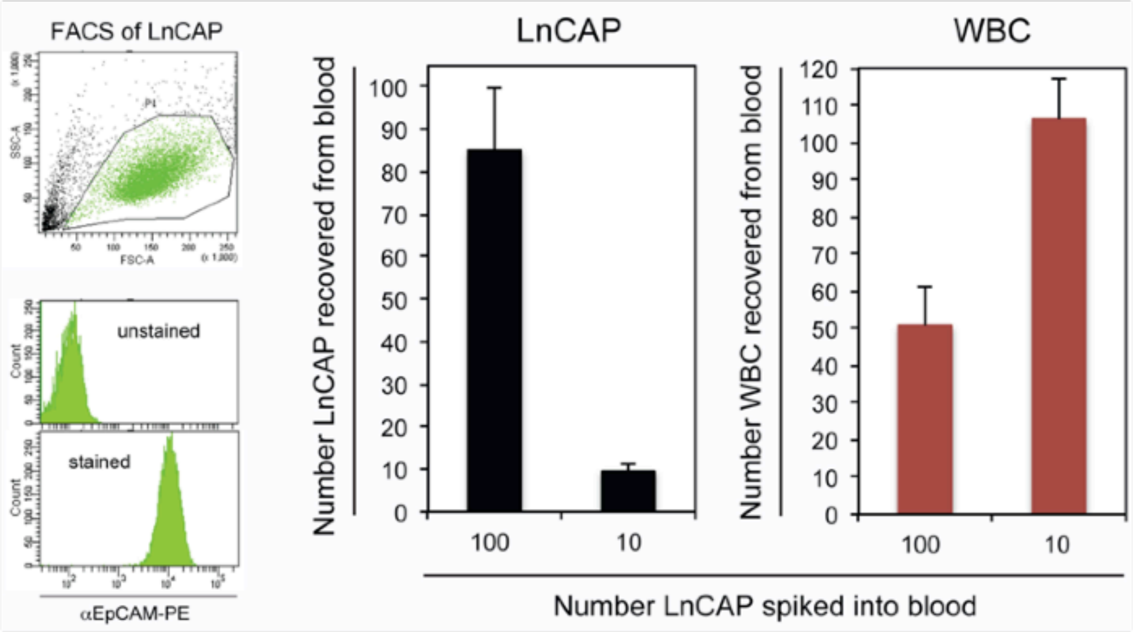
	# Cells WGA'd	# Passed WGA yields	# Passed library QC	Multiregion primary?
CRPC_#8:	28	3	1	no
CRPC_ 10:	88	16	6	yes
CRPC_12:	111	49	10	no
CRPC_35:	82	46	6	no
CRPC_36:	23	23	19	yes

Sequence whole exome

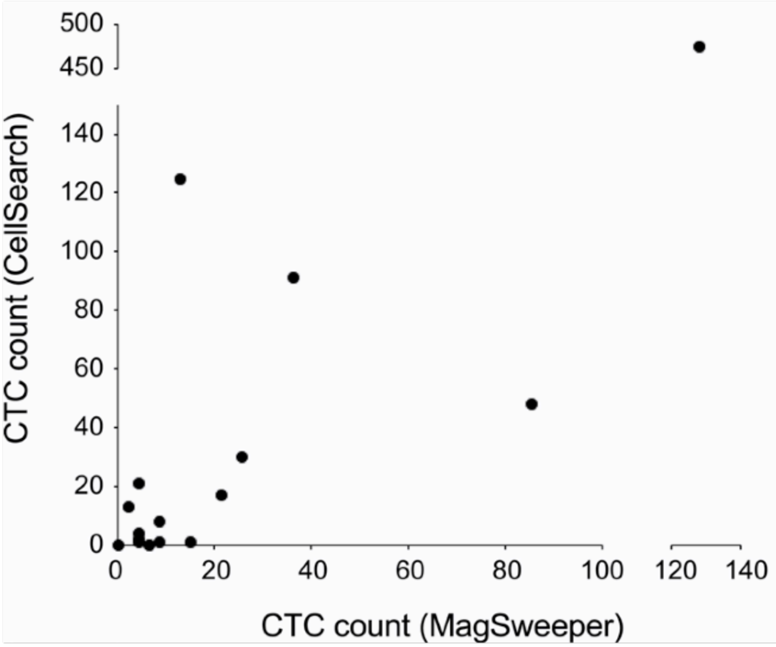
Supplementary Figure 1. Experimental process, selection criteria, and tabulation of data available for sequencing of CTCs. In this study, single vials of blood (3.75 mL) from patients with progressing metastatic prostate cancer were processed to retrieve CTCs. To account for variability in amplification from single cells and ensure sufficient numbers of libraries for variant calling, we set a ≥ 20 CTCs per vial of blood threshold, based on our empirical analysis of the samples processed. Additionally, to determine whether variants identified in CTCs were present in the tumor, we also required access to matched tumor tissue for comparison. Two subjects yielded sufficient numbers of CTCs and multiregion cores from the primary tissue were available.


nanowell-based recovery of single CTCs
(1) enriched sample loaded in array of 84,672 subnanoliter wells
(2) automated epifluorescence imaging of the array
(3) custom software program identifies cells, their nanowell IDs, and quantifies fluorescence intensities
transmitted light
fluorescence ch 1 / 3
(4) visualization tool gates, sorts, compiles images, then prompts user to select candidate CTCs for retrieval
Retrieve Cell? ☒ Back Next Done
Transmitted Light
CD45 094
DAPI 050
EpCAM 50,154
Nanowell ID: 48.08.02.02
Picklist for Micromanipulator
48.08.02.02
03.19.06.07
04.15.03.02
08.24.07.01
31.01.02.04
⋮
(5) robotic micromanipulator automatically retrieves single cells to PCR plates
Before
After

Supplementary Figure 2. Nanowell-based isolation of pure, single CTCs. **(a)** Following enrichment of patient blood via the Illumina MagSweeper or other platform, the enriched sample is loaded stochastically into an array of 84,672 subnanoliter (50 x 50 x 50 µm) wells. Automated epifluorescence microscopy is performed for the entire array, and a custom software program is used to segment all cells on the array and quantify fluorescence intensities. An image review program was written for this application to enable rapid review of candidate CTCs, including fluorescence intensities in each channel. Cells coated in beads and positive for EpCAM, while negative for CD45 and the nuclear stain DAPI, were selected within this program for retrieval. The program automatically generates a list of wells for retrieval that is compatible with an automated micromanipulator (ALS Aviso CellCelector) that retrieves single cells from individual wells and deposits into a PCR plate.

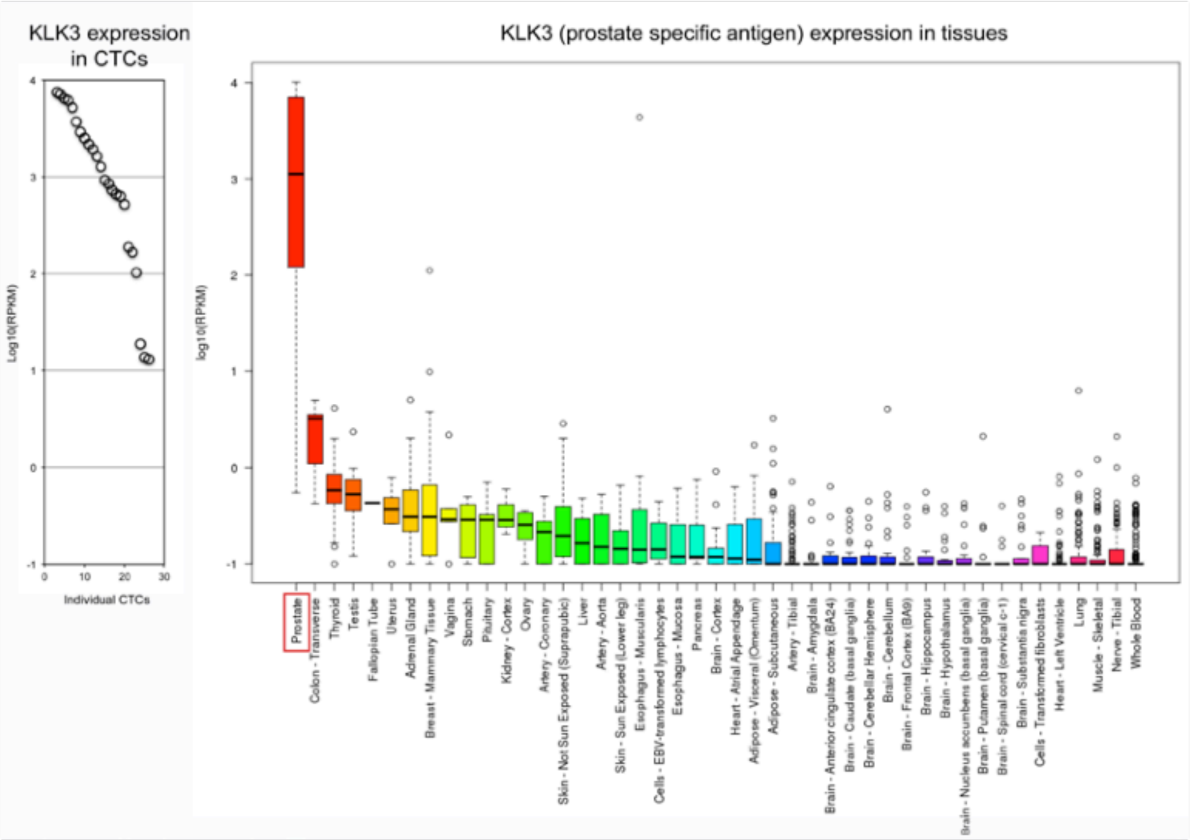


Supplementary Figure 3. Validation of prostate cancer cell recovery by the Illumina MagSweeper. Surface expression of the EpCAM epitope of LnCAP prostate cancer cells was determined by flow cytometry. Live cells, as determined by the indicated gate of forward scatter (FSC) and side scatter (SSC) in the upper left scatterplot, were stained with anti-EpCAM-PE or left unstained (histograms). To determine the sensitivity of the Illumina MagSweeper CTC isolation procedure, LnCAP cells were labeled with CFDA, a green fluorescent marker, and 10 or 100 CFDA-labeled LnCAP cells were spiked into 3.75 ml of blood from a healthy normal donor, and subjected to the MagSweeper isolation procedure, in triplicate. The numbers of LnCAP cells (EpCAM+, CFDA-labeled) and “contaminating” white blood cells (CD45+) recovered by the procedure were determined.

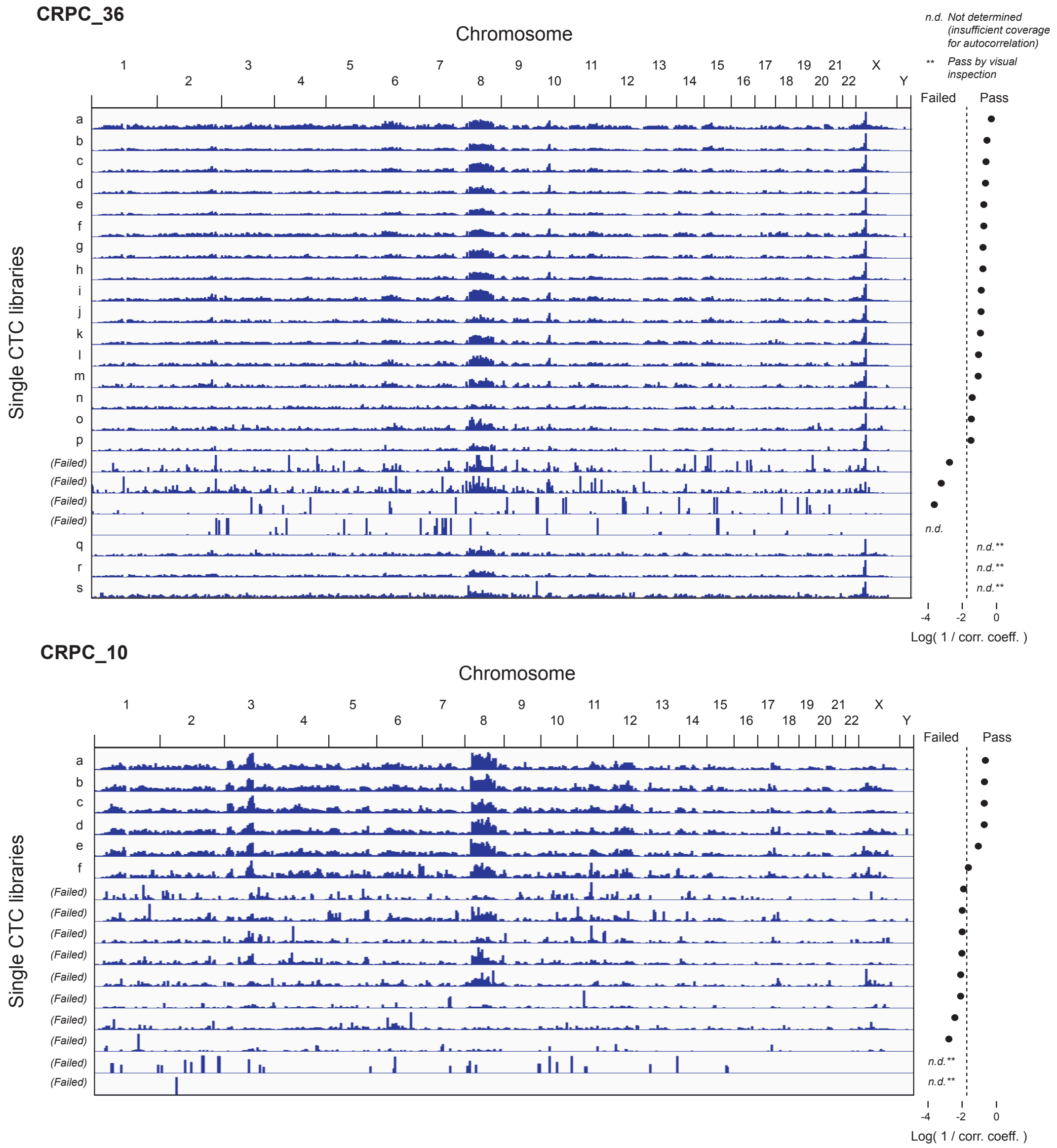


Supplementary Figure 4. Comparison of CTC counts from Illumina MagSweeper vs Veridex CellSearch. CTC counts per 7.5 mL of blood are compared for a subset of blood draws that were enumerated using both the MagSweeper and Veridex. A significant correlation between CTC counts obtained by the two platforms was observed (p = 0.006; Spearman, two-tailed).

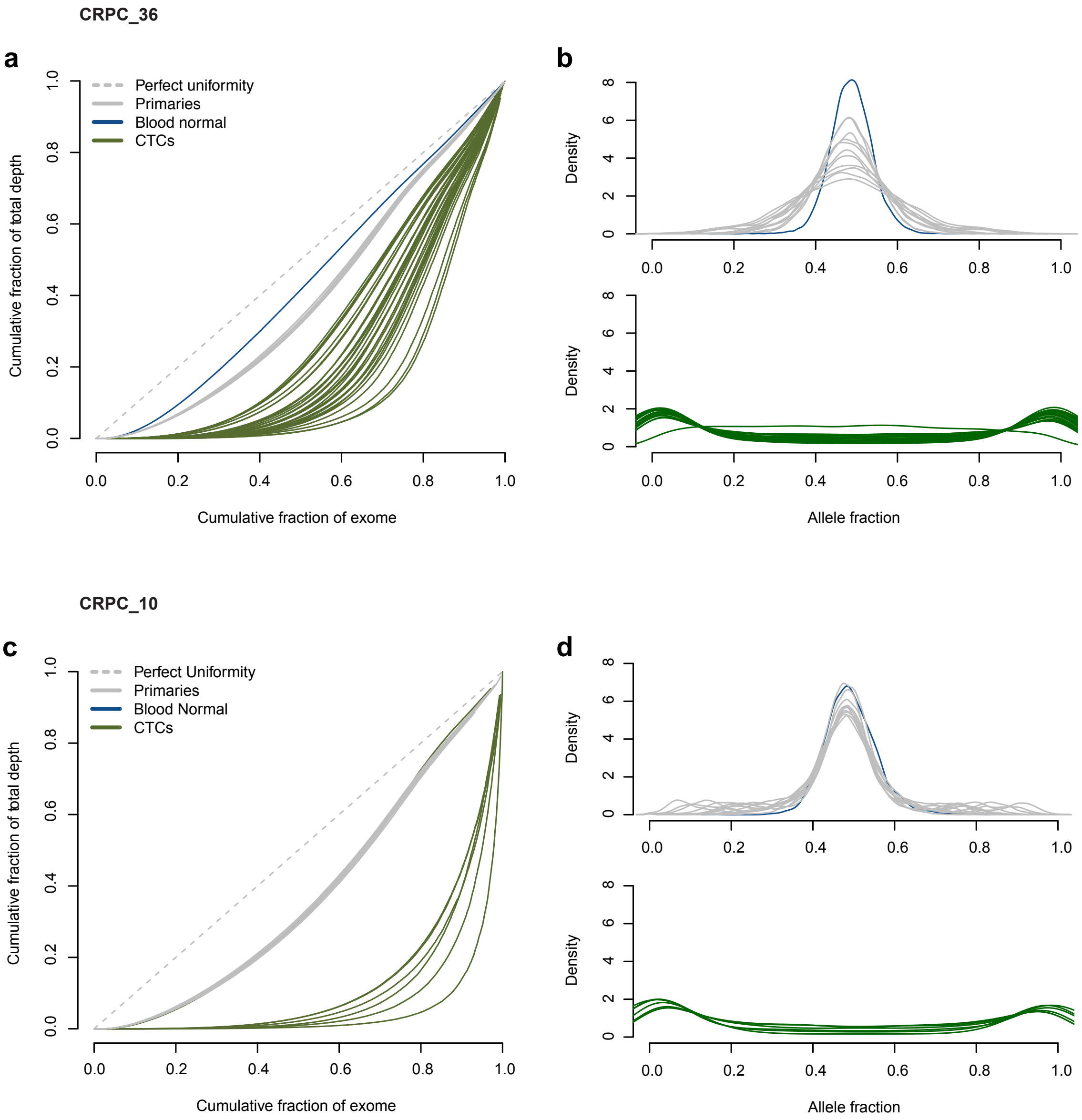
Nature Biotechnology: doi:10.1038/nbt.2892



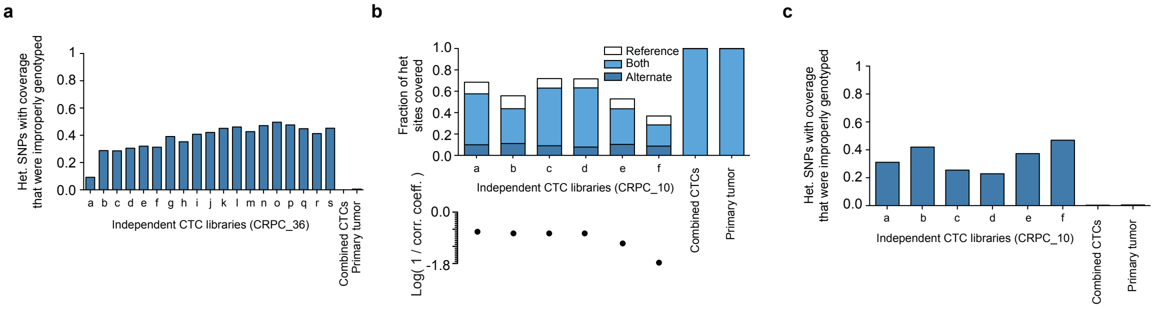
Supplementary Figure 5. *KLK3* (prostate specific antigen) mRNA is highly expressed in CTCs from patient with metastatic prostate cancer. CTCs were isolated from peripheral blood and picked as single cells as described in **Online Methods**. RNA sequencing of single cells was performed according to a recently published experimental procedure[32]. The left scatterplot represents the $\log_{10}$(RPKM) value of *KLK3* for 26 out of 48 individual single cells, for which *KLK3* expression was detected, and for which greater than 100 genes out of 8247 were detected. The barplot on the right represents the $\log_{10}$(RPKM) value of gene expression as measured by RNA sequencing, obtained from a publicly available dataset, in which RNA sequencing was performed to determine gene expression in the indicated human tissues (Genotype-Tissue Expression Portal, GTEx, http://www.broadinstitute.org/gtex). RPKM = reads per kilobase per million mapped reads.



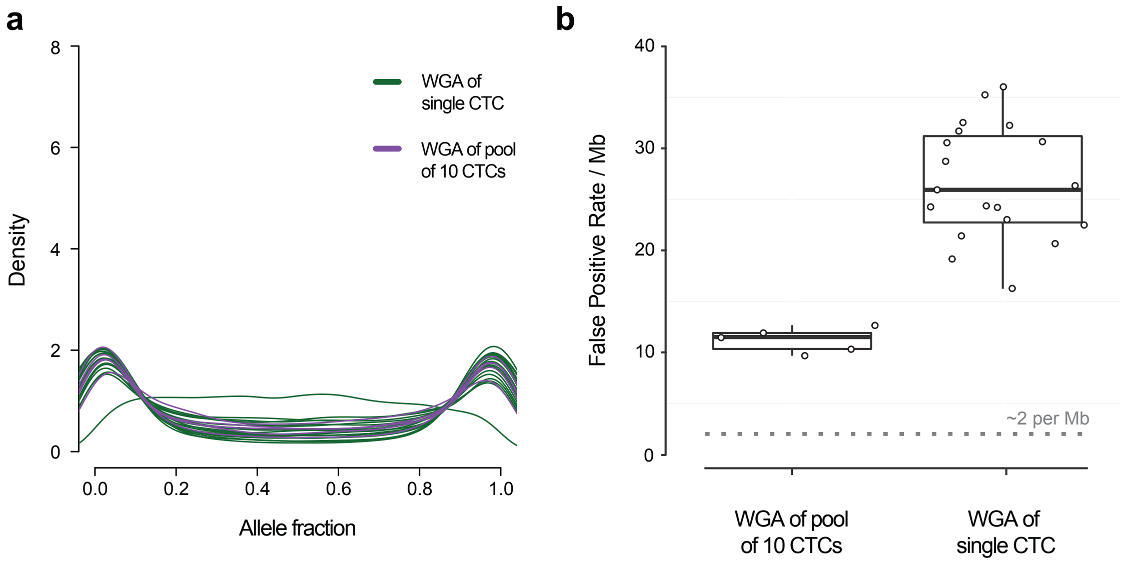
Supplementary Figure 6. Quality control on genome-wide uniformity for libraries from patients CRPC_36 and CRPC_10. Genome-wide read densities for 23 libraries from CRPC_36, and 16 libraries from CRPC_10. The calculated autocorrelation coefficients (**Online Methods**) for each library is shown (right side). The cut-off threshold for pass/fail is shown (dashed line). For a few libraries, there was insufficient coverage to calculate the autocorrelation coefficient (n.d., not determined); here, visual inspection of genome-wide read densities was used to assess pass (**) versus fail.



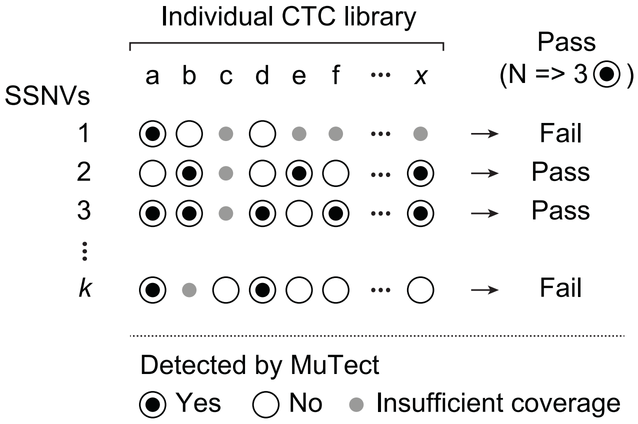
Supplementary Figure 7. Comparison of allelic coverage in bulk and WGA libraries. (a) Lorenz curves showing the cumulative fraction of reads as a function of the cumulative fraction of the exome for bulk blood normal DNA (blue), bulk primary tumors (grey) and single CTCs (green) for CRPC_36. Grey dashed line indicates perfectly uniform coverage. WES yielded 124 ± 12x mean target coverage for CTC libraries from CRPC_36 (**Supplementary Table 2**). **(b)** Allelic fraction distribution of 22,054 germline heterozygous SNPs as observed in exome sequencing data of bulk normal DNA (blue), primary bulk tumors (grey), and libraries derived from amplified single CTCs (green) for CRPC_36. One library derived from an amplified single CTC exhibited substantially less allelic distortion than the other single CTC libraries. **(c)** The same analysis as in **Supplementary 7a** is presented, but for CRPC_10. WES yielded 89 ± 8x mean target coverage for CTC libraries from CRPC_10 (**Supplementary Table 2**). **(d)** Same as **Supplementary 7b**, but for patient CRPC_10.



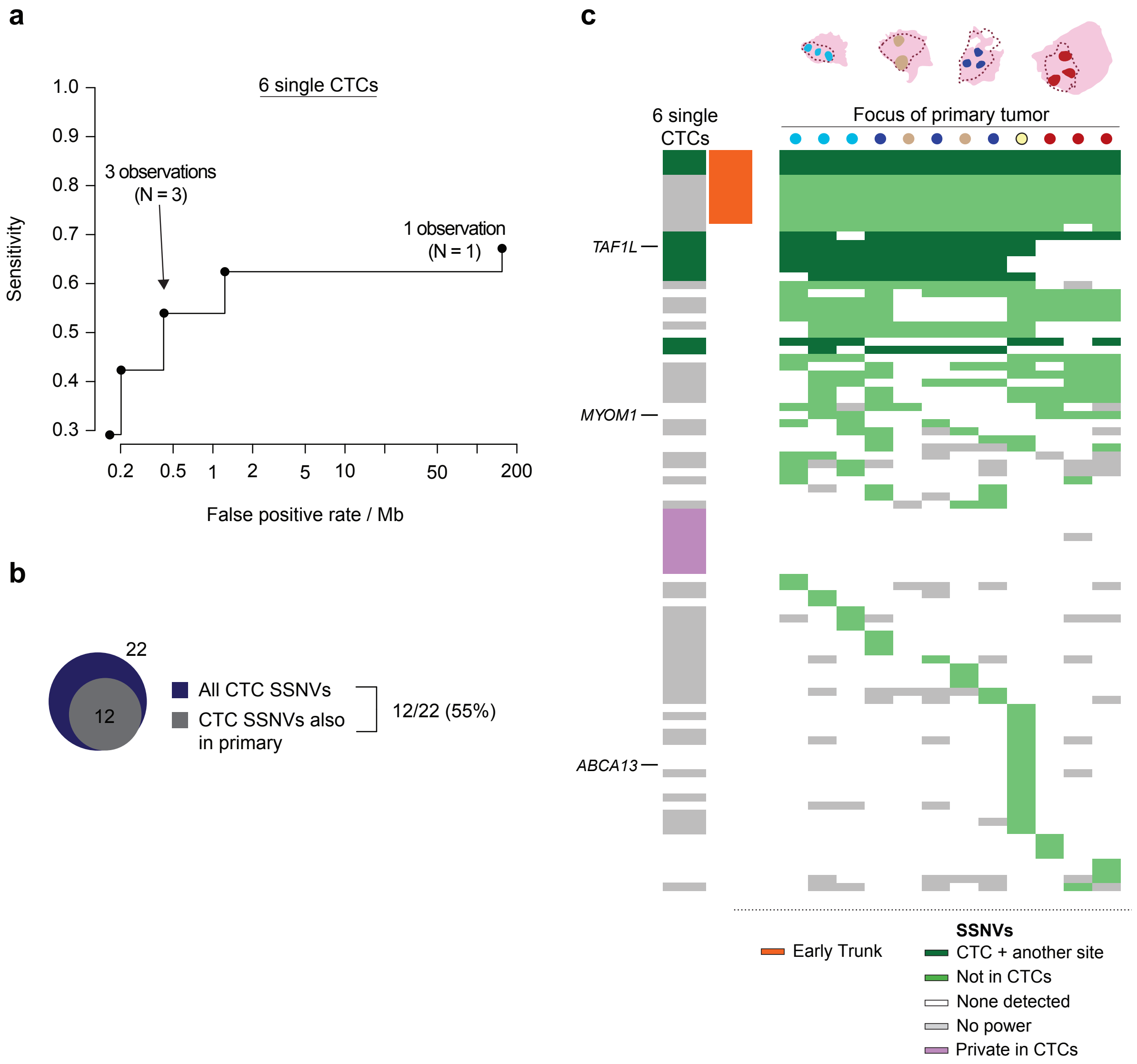
Supplementary Figure 8. Heterozygous SNP sites that were improperly genotyped. (a) The fraction of improperly genotyped sites is calculated from the total number with coverage from each of the CTC libraries from CRPC_36. This is calculated as the number of sites with loss of reference or alternate allele divided by the number of sites with coverage. **(b)** The same analysis as presented in **Fig. 2a** is presented for patient CRPC_10. Combining CTC libraries yielded coverage of both alleles at 99.892% of heterozygous SNP sites. **(c)** The same analysis as presented in **Supplementary Fig. 8a** is presented for patient CRPC_10.



Supplementary Figure 9. Sequencing a pool of CTCs does not provide comparable accuracy in calling of variants to census-based sequencing. (a) Distributions of allelic fractions for 22,054 germline heterozygous SNPs for exome sequencing data of 19 libraries derived from amplified products of single CTCs (green) and five libraries prepared from amplified pools of 10 CTCs (purple) from patient CRPC_36. The degree of allelic distortion is similar between both sets of libraries. **(b)** Estimation of false positive rate per megabase (Mb) for the 19 single-cell libraries and five libraries from pools of 10 CTCs. The estimated natural rate for treated prostate cancer is 2 mutations / Mb (grey dashed line)[19].


Individual CTC library
a b c d e f ··· x
Pass (N => 3 ◉)
SSNVs
1 ◉ ○ ● ○ ● ● ··· ● → Fail
2 ○ ◉ ● ○ ◉ ○ ··· ◉ → Pass
3 ◉ ◉ ● ◉ ○ ◉ ··· ◉ → Pass
⋮
k ◉ ● ○ ◉ ○ ○ ··· ○ → Fail
Detected by MuTect
◉ Yes ○ No ● Insufficient coverage

Supplementary Figure 10. Principle of census-based calling of SSNVs. SSNVs are identified in each library as described in **Online Methods** using MuTect[18]. SSNVs are compared one by one, across all libraries. If an SSNV is detected in N or more independent libraries, the SSNV is called in the CTCs. (N = 3 is shown here.)



Supplementary Figure 11. Comparison of mutation pattern across CTCs and primary cores from patient CRPC_10. (a) Sensitivity versus false positive rate / Mb as a function of the required number N of independent observations of the variant among 6 CTC libraries. **(b)** The number of SSNVs called in total among 6 CTC libraries (22) and those that were validated as being present in matched tumor tissue (12) are shown. **(c)** Hierarchical clustering using the Jaccard index for mutations called across twelve primary cores and CTCs (when observed in ≥ 3 out of 6 single CTCs). Only sites in the exomes that were considered to be powered for mutation calling, as described in **Online Methods**, were included in this analysis. Shading of green represents presence in CTCs and at least one other sample (dark green) or not present in CTCs (light green). The areas shaded in pink represent the pathology blocks from which cores of tissue were obtained (drawn to scale). The dotted lines represent the area with histological presence of tumor within each block. The sites from which the individual cores of tissue were obtained for sequencing are shown. Genes highlighted indicate non-synonymous mutations present in > 2 patients from a previous sequencing study in prostate cancer[31]. CTCs detected all 3 of the early trunk SSNVs for which they were powered (CTCs were not powered to call the remaining 6 in this patient, yet 2 of these were detected in less than 3 CTCs).

Patient ID	**Age**	**PSA**	**CTC / 3.75 mL**	**Matched tissue?**
CRPC_1	86	23	50	n
CRPC_2	77	1201	27	n
CRPC_3	76	91	3	n
CRPC_4	67	734	6	n
CRPC_5	55	4706	6	y
CRPC_6	65	2680	0	y
CRPC_7	63	1084	2	y
CRPC_8	63	1687	20	y
		1174	50	
CRPC_9	76	17	3	y
CRPC_10	56	322	90	y
	56	1072	200	
CRPC_11	65	61	0	y
		44	0	
CRPC_12	57	738	81	y
		738	30	
CRPC_13	56	193	0	y
	56	243	0	
CRPC_14	57	25	0	y
CRPC_15	69	191	15	n
CRPC_16	76	737	5	y

		17	0	
		15	0	
CRPC_17	54	817	60	n
CRPC_18	79	11	0	n
CRPC_19	67	25	2	y
CRPC_20	73	47	12	n
		61	40	
CRPC_21	64	10	2	n
CRPC_22	57	57	1	y
CRPC_23	69	8	2	n
CRPC_24	84	84	6	y
		24	10	
CRPC_25	66	9	0	n
CRPC_26	73	46	3	n
CRPC_27	65	5	3	y
		88	0	
CRPC_28	75	3	17	n
		5	7	
CRPC_29	61	44	7	y
		117	0	
CRPC_30	64	31	1	n
CRPC_31	53	28	0	n
CRPC_32	77	4	1	n
		8	0	
CRPC_33	73	8	1	y
CRPC_34	73	74	1	y
		149	1	
CRPC_35	65	382	100	y
		362	100	
CRPC_36	60	151.7	85	y

Supplementary Table 1. Patient information table. Blood draws were processed from 36 patients with castration-resistant prostate cancer (CRPC). In some cases, repeat blood draws were obtained on different dates. The present availability of matched tissue has been indicated and affected our decision on whether or not to perform WGA on the isolated CTCs.

Supplementary Table 2. Sequencing metrics. Standard exome sequencing metrics are presented for the samples sequenced from patients CRPC_10 and CRPC_36.

Supplementary Table 3. List of SSNVs called in patient CRPC_36. Somatic single nucleotide variants with their genomic positions and protein changes, which were identified in 9 individual cores of tissue from the primary prostate cancer, a lymph node metastasis, and in CTCs, are listed, including missense, silent, intron, nonsense, splice site, translation start site, and 5' and 3' untranslated region mutations. Any point mutation within the targeted exome territory was considered independent of its classification.

Supplementary Table 4. List of SSNVs called in patient CRPC_10. Somatic single nucleotide variants with their genomic positions and protein changes, which were identified in 12 individual cores of tissue from the primary prostate cancer and in CTCs, are listed, including missense, silent, intron, nonsense, splice site, translation start site, and 5' and 3' untranslated region mutation. Any point mutation within the targeted exome territory was considered independent of its classification.