nature biotechnology

LETTERS

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

Jens G Lohr[1–3,11], Viktor A Adalsteinsson[1,4,11], Kristian Cibulskis[1,11], Atish D Choudhury[1–3], Mara Rosenberg[1], Peter Cruz-Gordillo[1], Joshua M Francis[1,2], Cheng-Zhong Zhang[1,2], Alex K Shalek[5], Rahul Satija[1], John J 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[1–3], 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[1–3,6], Aviv Regev[1,8,9], Matthew Meyerson[1–3], William C Hahn[1–3,6], Philip W Kantoff[2,3], Todd R Golub[1–3,9], Gad Getz[1,7], Jesse S Boehm[1] & J Christopher Love[1,4,10]

Comprehensive analyses of cancer genomes promise to inform prognoses and precise cancer treatments. A major barrier, however, is inaccessibility of metastatic tissue. A potential solution is to characterize circulating tumor cells (CTCs), but this requires overcoming the challenges of isolating rare cells and sequencing low-input material. Here we report an integrated process to isolate, qualify and sequence whole exomes of CTCs with high fidelity using a census-based sequencing strategy. Power calculations suggest that mapping of >99.995% of the standard exome is possible in CTCs. We validated our process in two patients with prostate cancer, including one for whom we sequenced CTCs, a lymph node metastasis and nine cores of the primary tumor. Fifty-one of 73 CTC mutations (70%) were present in matched tissue. Moreover, we identified 10 early trunk and 56 metastatic trunk mutations in the non-CTC tumor samples and found 90% and 73% of these mutations, respectively, in CTC exomes. This study establishes a foundation for CTC genomics in the clinic.

Enabling precision medicine for each patient with cancer depends on the ability to access samples that accurately represent the genomic features of his or her tumors[1]. Two critical bottlenecks, however, are that metastatic tissue is often inaccessible and the purity and yield of biopsy samples are low. So far, genomic characterization of cancer has emphasized large-scale sequencing of primary tumors and, in a few cases, metastatic lesions[2]. Both circulating tumor DNA[3] and CTCs[4] are alternative sources that may overcome these sampling challenges. Comprehensive sequencing and confident determination of genomic variants in CTCs could provide routine monitoring of transiting cells with potential for metastatic colonization to complement the static sampling of resected or biopsied lesions[5].

Technologies for enriching and enumerating CTCs have provided prognostic value[4,6,7], and characterizing specific regions, genes or patterns of gene expression in CTCs is both possible and useful. PCR-based methods, array comparative genomic hybridization and high-throughput sequencing have revealed somatic single-nucleotide variants (SSNVs) and copy number alterations[8–10], and RNA sequencing has shown pathways that are implicated in metastasis[11]. For example, exome sequencing of lung cancer CTCs can uncover mutations shared with metastases[10]. Without comprehensive power statistics, however, it remains difficult to assess the fraction of CTC exomes that are being robustly and accurately sequenced and whether such approaches apply to other cancers such as prostate cancer.

Robust and accurate detection of SSNVs from CTCs is challenging. CTCs in a vial of blood are sparse[12], and whole-genome amplification is necessary to construct sequencing libraries. Yields of amplified DNA vary among CTCs[13], and whole-genome amplification introduces amplification bias and polymerase errors[14,15]. Census-based sequencing of multiple libraries from the same sample (requiring a variant to be present in more than one library), therefore, has helped distinguish private mutations from polymerase errors with some fidelity[14,15]. Despite the technical capabilities demonstrated for sequencing CTCs, no generalizable framework exists for confidently calling SSNVs, and design optimization of the experimental processes could provide a critical foundation for future comprehensive surveys of the genomics of CTCs across large numbers of samples.

On the basis of these considerations, we developed a modular set of experimental and analytical protocols for census-based whole-exome sequencing (WES) and confident calling of SSNVs from prostate CTCs.

[1]The Eli and Edythe L. Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [3]Harvard Medical School, Boston, Massachusetts, USA. [4]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts, USA. [6]Brigham and Women's Hospital, Boston, Massachusetts, USA. [7]Massachusetts General Hospital, Boston, Massachusetts, USA. [8]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [9]Howard Hughes Medical Institute, Chevy Chase, Maryland, USA. [10]Ragon Institute of MGH, MIT and Harvard, Cambridge, Massachusetts, USA. [11]These authors contributed equally to this work. Correspondence should be addressed to J.C.L. (clove@mit.edu), J.S.B. (boehm@broadinstitute.org) or G.G. (gadgetz@broadinstitute.org).

Received 25 October 2013; accepted 30 March 2014; published online 20 April 2014; doi:10.1038/nbt.2892

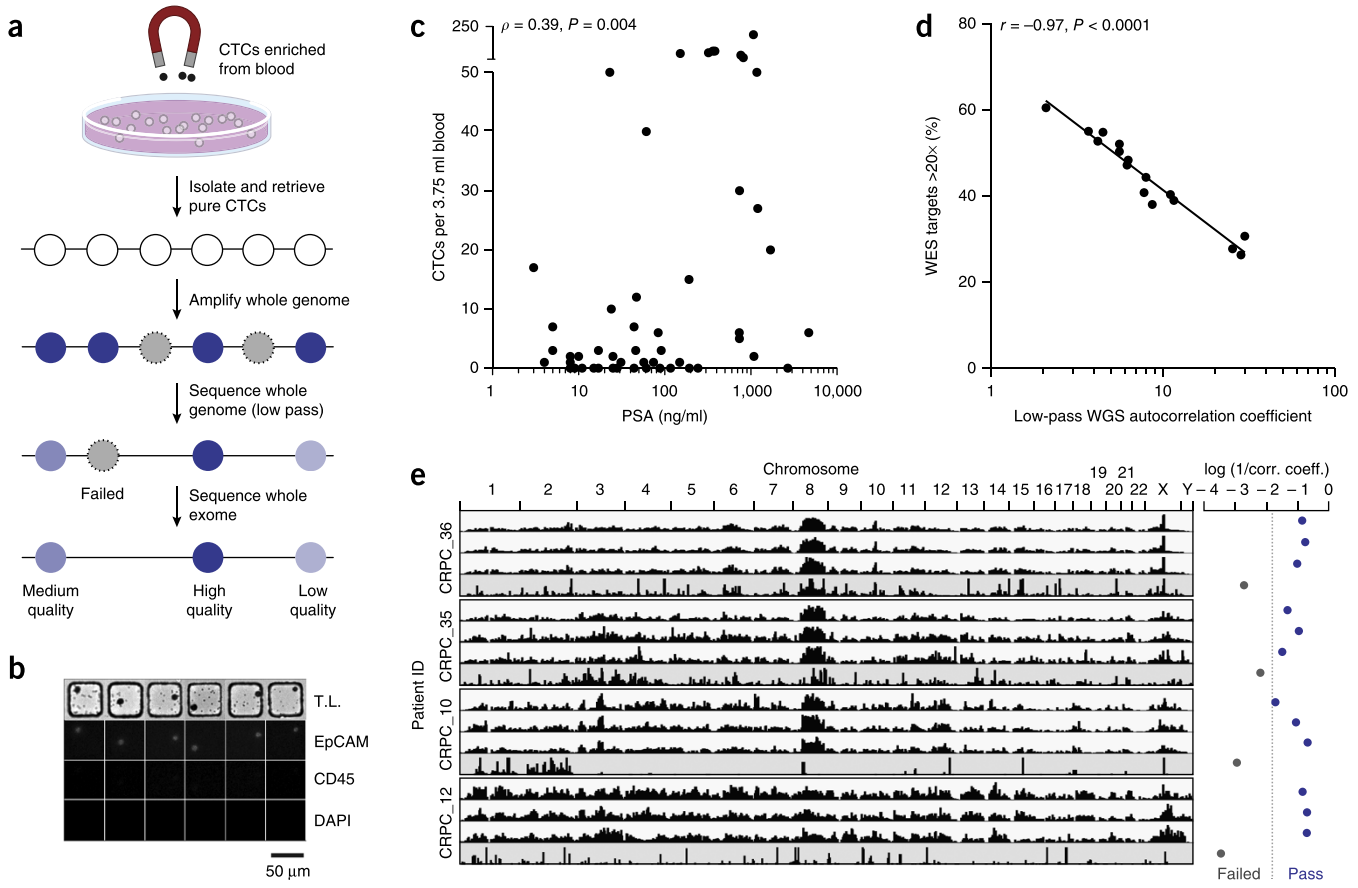

Figure 1 Experimental process for sequencing of CTCs. (**a**) Schematic of workflow for the enrichment, isolation and sequencing of CTCs. (**b**) Sample micrographs of CTCs isolated in nanowells with matched transmitted light (T.L.) and immunophenotyping for EpCAM, CD45 and 4,6-diamidino-2-phenylindole (DAPI) by epifluorescence. (**c**) Scatter plot of the number of CTCs enumerated versus the levels of PSA from 51 blood samples from 36 patients with prostate cancer (**Supplementary Table 1**) screened using the MagSweeper for enrichment. CTC numbers in the blood correlated with PSA levels ($P = 0.004$, Spearman, two tailed). (**d**) Scatter plot of the percentage of target bases covered >20× from WES versus the autocorrelation coefficient (Online Methods) calculated from low-pass whole-genome sequencing over chromosome 1 for patient CRPC_36 ($P < 0.0001$, Pearson, two tailed). WES yielded 124 ± 12× (mean ± s.d.) mean target coverage (**Supplementary Table 2**). Whole-genome sequencing yielded a mean coverage over chromosome 1 of between 0.0003× and 0.03×, with a median of 0.017×. (**e**) Genome-wide read densities (1-Mb bins) from low-pass whole-genome sequencing of CTC libraries from four different patients (CRPC_10, CRPC_12, CRPC_35 and CRPC_36). Examples of three quality libraries and one poor library are shown per patient. The log of the inverse correlation coefficient (corr. coeff.) was used to select high-quality libraries with a cutoff of –1.8.

We show that these techniques can provide a window into the genetics of metastatic prostate cancer in a manner that is potentially useful in the clinic.

We first created a standardized process to generate and qualify multiple independent libraries for WES from CTCs recovered from one vial of blood. The process involves cell enrichment and isolation, genomic amplification, library qualification and census-based sequencing (**Fig. 1a** and **Supplementary Fig. 1**). We used the Illumina MagSweeper to enrich epithelial cell adhesion molecule (EpCAM)-expressing CTCs[16]. The recovered cells, enriched with CTCs, were deposited into dense arrays of subnanoliter wells and imaged by automated epifluorescence imaging (**Fig. 1b**). Individual $EpCAM^+CD45^-$ CTCs were recovered by robotic micromanipulation for whole-genome amplification using multiple displacement amplification (MDA). This combined process reliably isolated single CTCs in a highly automated fashion (**Supplementary Fig. 2**).

We next validated our method for isolating CTCs. The yield of tumor cells spiked into whole blood was ≥85% (**Supplementary Fig. 3**) and concurred with an independent method for enrichment (Veridex CellSearch) (**Supplementary Fig. 4**). We also performed low-coverage single-cell RNA sequencing on cells recovered through our process from patients with prostate cancer to confirm that our isolated $EpCAM^+$ cells expressed prostate-specific antigen (PSA), confirming their prostate origins (**Supplementary Fig. 5**). We then enriched and enumerated CTCs from 51 blood samples of 36 patients with metastatic castration-resistant prostate cancer (CRPC) (**Supplementary Table 1**). We used the automated process for 45 of the 51 samples and performed manual picking for the remaining 6 samples. These samples yielded 0–200 CTCs per 3.75 ml of blood (median of 7 CTCs for samples with ≥1 CTC, with 27% having no detectable CTCs); 45% of samples had ≥5 CTCs, which is consistent with volume-adjusted counts previously reported in metastatic prostate cancer[12]. The number of CTCs also correlated with serum levels of PSA ($P = 0.004$, Spearman, two tailed) (**Fig. 1c**).

We reasoned that establishing methods to assess the quality and uniformity of genome-wide coverage of CTC-derived sequencing libraries before in-depth WES or whole-genome sequencing would help make census-based genomic sequencing of CTCs cost efficient and facilitate subsequent analysis of SSNVs. To address this challenge, we first performed whole-genome amplification on all single CTCs isolated from five patients with 20 or more CTCs and for whom matched tumor tissue was available in tumor banks for comparison. As expected, the

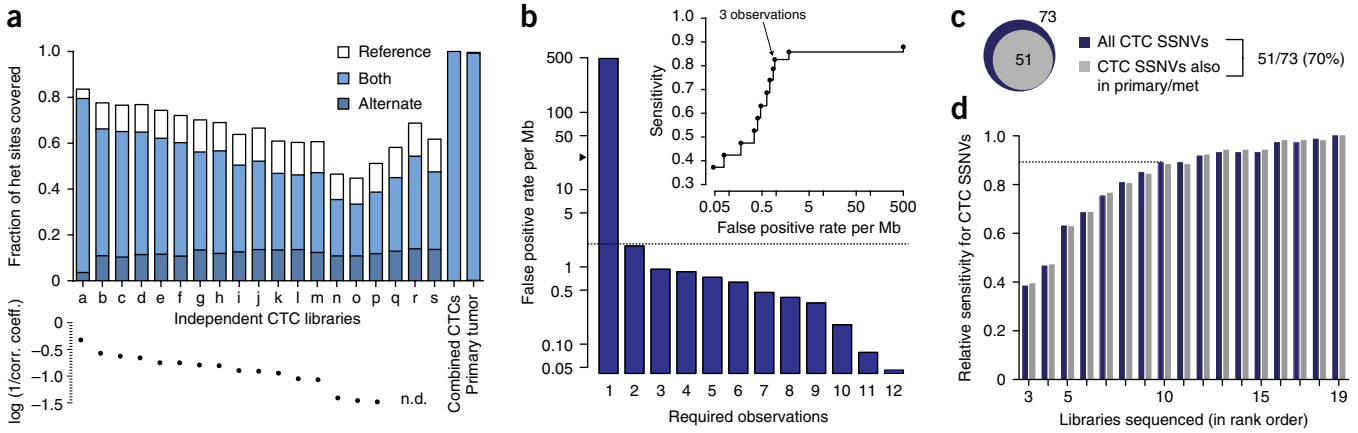

Figure 2 Census-based variant calling from WES of CTCs from patient CRPC_36. (**a**) Characterization of allelic coverage in each CTC sequencing library from the same patient compared to those libraries combined and the primary tumor as determined by 22,054 germline heterozygous (het) SNP sites; an allele was scored as covered if there were three total reads of the particular allele. For reference, the autocorrelation coefficient is plotted below all CTC libraries except for three CTC libraries (n.d., not determined) that had insufficient low-pass whole-genome sequencing coverage but passed quality control before exome sequencing on the basis of visual inspection of genome-wide read densities (**Supplementary Fig. 6**). Coverage of the alternate allele (either the alternate alone or both alleles) at germline heterozygous SNPs was correlated with the autocorrelation metric for individual CTC libraries ($P < 0.0001$, Spearman, two tailed). When the individual CTC libraries were combined (combined CTCs), 99.995% of sites were covered by both alleles, which is similar to bulk sequencing of the primary tumor. (**b**) Estimation of false-positive rate per Mb among 19 independent CTC libraries after requiring the variant to be observed in at least *n* independent CTC libraries (**Supplementary Fig. 10**). The gray dashed line indicates the reported mutation rate in bulk tumor sequencing of treated prostate cancer (~2 per Mb)[19]; the black arrowhead indicates the false-positive rate per Mb observed for a single CTC library. The inset shows sensitivity versus false-positive rate per Mb as a function of the required number of independent observations of the variant. (**c**) The number of SSNVs called in total among 19 CTC libraries (73) and those that were validated as being present in matched tumor tissue (51). Primary/met indicates primary tumor and/or matched metastasized tumor tissue. (**d**) Relative sensitivity to call CTC SSNVs (fraction of the total number called using 19 CTC libraries) as a function of the number of libraries sequenced ranked in order by the autocorrelation coefficient (blue bars). A sustained improvement in sensitivity was observed. Additionally, considering only the 51 CTC SSNVs also observed in bulk WES of matched tumor tissue, we observed a very similar increase in sensitivity for each additional library sequenced (gray bars).

rates of success in amplification of single prostate CTCs varied widely (11–100%) (**Supplementary Fig. 1**), which is consistent with the variability reported for amplifying lung cancer CTCs[13]. To assess the level of amplification bias in the recovered products, we developed a rapid and cost-effective method using low-pass whole-genome sequencing (<0.05×) and autocorrelation analysis of single-base coverage to qualify libraries before WES (Online Methods). This metric accurately predicted the fraction of well-covered targets in subsequent WES ($P < 0.0001$, Pearson, two tailed) (**Fig. 1d**). It also corresponded well to low-resolution views (1 Mb) of genome-wide read densities, highlighting prominent hallmarks of copy-number variants in prostate cancer, including chromosome 8q amplification, 8p deletion and amplification of chromosome X (**Fig. 1e**)[17]. When a library generated insufficient coverage (~0.0001×) to calculate an autocorrelation coefficient accurately, visual inspection of genome-wide read density could provide a qualitative means for qualification (**Supplementary Fig. 6**). Together these examples demonstrate that our integrated experimental approach generates independent, high-quality libraries from single CTCs for WES and that low-pass whole-genome sequencing can reliably predict which single CTC sequencing libraries are likely to yield high-quality data.

We then implemented a sensitive method to detect SSNVs from CTC libraries. We selected WES to generate high-coverage sequencing data of maximally informative genomic regions to enable cost-effective discovery of SSNVs. We sequenced 19 single CTC libraries of patient CRPC_36 to 124 ± 12× (mean ± s.d.) mean target coverage (**Supplementary Table 2**). As expected, individual libraries exhibited nonuniform coverage with only a fraction of the exome present (**Supplementary Fig. 7a**)[15] and bimodal allelic distortion at sites of germline heterozygous SNPs compared to normal distributions from bulk sequencing (**Supplementary Fig. 7b**). This effect caused improper genotyping at 38.3 ± 9.8% (mean ± s.d.) of such covered sites (**Supplementary Fig. 8a**). Coverage of either the alternate allele or both alleles ranged from 33.5% to 79.6% for individual CTCs and correlated strongly with the autocorrelation metric used for quality control ($P < 0.0001$, Spearman, two tailed) (**Fig. 2a**).

We hypothesized that combining data from independent CTC libraries would improve sensitivity. Indeed, the total coverage of both alleles (99.995%) at 22,054 SNP sites among 19 independent CTC libraries compared well to a representative bulk library from the primary tumor—only 0.005% of sites were improperly genotyped (**Fig. 2a** and **Supplementary Fig. 8a**). Analysis of whole exomes of CTCs from a second patient (CRPC_10) revealed similar extents of amplification bias that were also overcome by sequencing multiple independent libraries (**Supplementary Figs. 7c**,**d** and **8b**,**c**). (We found that amplifying a single pool of CTCs was sensitive to the same allelic distortion as any other individual MDA-derived library (**Supplementary Fig. 9a**).) Together these observations are consistent with stochastic loss of DNA from single cells, random preferential amplification of alleles and lack of systematic coverage biases in MDA products[14]; they also confirm that sequencing multiple independent libraries of CTCs for a patient can enable robust, highly sensitive determination of variants.

We next sought to assess the specificity of this approach. We estimated an upper bound for the rate of false positives by assuming that all variants, identified using MuTect[18], not present in bulk tumor samples from the same patient were false positives (Online Methods). This assumption is conservative because contemporary CTCs may have diverged biologically from previously resected samples. Although amplifying and sequencing a pool of CTCs exhibited a false-positive rate (~10 per Mb) less than that for a single CTC library (~25 per Mb), this rate was still insufficient for accurate calling of mutations on its

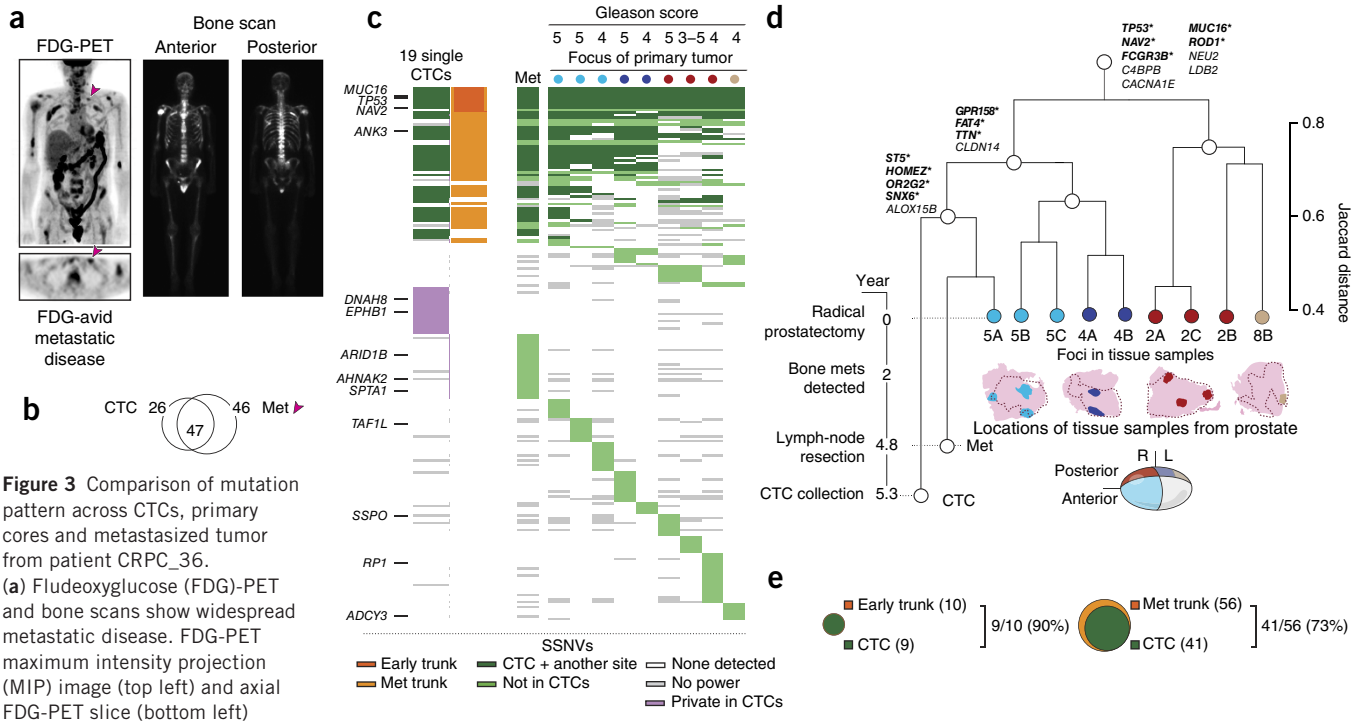

Figure 3 Comparison of mutation pattern across CTCs, primary cores and metastasized tumor from patient CRPC_36. (**a**) Fludeoxyglucose (FDG)-PET and bone scans show widespread metastatic disease. FDG-PET maximum intensity projection (MIP) image (top left) and axial FDG-PET slice (bottom left) demonstrate multifocal FDG-avid skeletal metastases throughout the axial and appendicular skeleton, as well as bilateral cervical, left supraclavicular (arrowheads), retroperitoneal and bilateral common iliac metastatic lymphadenopathy. The bone scan demonstrates widespread bone metastasis. (**b**) Venn diagram representing mutations called in the CTCs and metastasis (Met). Of note, 51% of mutations in the metastasis were called in CTCs. (**c**) Hierarchical clustering using the Jaccard index for mutations called across the nine primary cores, the metastasized tumor and CTCs (when observed in ≥3 out of 19 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. Green shading represents presence in CTCs and at least one other sample (dark green) or not present in CTCs (light green). Genes highlighted indicate nonsynonymous mutations present in more than two patients from a previous sequencing study in prostate cancer[30]. Of note, one of the cores included regions of both Gleason 3 and Gleason 5 cancer (indicated as '3–5'). (**d**) Dendrogram representing hierarchical clustering by Jaccard index and the timeline of sample acquisition. SSNVs detected in all individual cores of tissue (early trunk) or in all cores that belong to only one of the two branchpoints of the clustering dendrogram are listed. Nonsynonymous mutations are highlighted in bold with asterisks. 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 displayed in colors corresponding to the cluster dendrogram and those in **c**. The regions of the prostate from which the pathology blocks were retrieved are schematized. (**e**) The number of mutations found in the metastasis and at least one core of the primary tumor (metastatic trunk), the early trunk mutations and the overlap of these with CTC mutations excluding sites that were consistently underpowered in greater than half of the samples. The one early trunk mutation not detected in CTCs ($n \geq 3$) was observed in two CTCs.

own (**Supplementary Fig. 9b**). When combining multiple single CTC libraries, however, the false-positive rate of called SSNVs dropped substantially from ~500 per Mb (n = 1 library) to a rate below the expected mutational rate in treated prostate cancer (~2 per Mb) when observed across two or more CTCs (**Fig. 2b**)[14,15,19]. When increasing the number of multiple observations required (**Supplementary Fig. 10**), the false-positive rate further diminished to 0.9 per Mb (n = 3) with an estimated sensitivity of 82.6% (**Fig. 2b**, inset). Analysis of six CTCs from CRPC_10 supported our statistical predictions from CRPC_36 that census-based sequencing also improves specificity (**Supplementary Fig. 11a**).

Applying this analytical method to the 19 CTCs from CRPC_36, we detected 73 SSNVs (n = 3; **Supplementary Table 3**). We found that 51 of these SSNVs (70%) were also present among 9 cores from the matched primary tumor and a lymph node metastasis, confirming that these EpCAM$^+$ cells were genetically related to the primary prostate cancer in this patient (**Fig. 2c**). Similarly applying this analytical method to the 6 CTCs from CRPC_10 sequenced to 89 ± 8× (mean ± s.d.) mean target coverage (**Supplementary Table 2**), 12 of 22 CTC SSNVs called (55%) (**Supplementary Table 4**) were also present among 12 cores from the primary tumor (**Supplementary Fig. 11b**). Overall, the sensitivity of this technique increases with the number of CTC libraries included in the analysis and reached a relative sensitivity of 88% using 10 of 19 libraries from CRPC_36 (**Fig. 2d**).

The results described above suggest that comprehensive sequencing of prostate CTCs and accurate assessments of SSNVs are possible with our approach. We then hypothesized that CTC sequencing could have clinical utility, perhaps providing a reasonable proxy for metastatic sampling in disseminated cancer. Clinical sequencing in metastatic prostate cancer is challenging because metastatic tissue is not routinely sampled[20] and computed topographic–guided biopsy has a poor success rate with a low purity of biopsied lesions[21]. For patient CRPC_36, there was widespread metastatic disease (**Fig. 3a**). Although the vast majority of the metastases (over ten) were not available for sequencing, one neck lymph node had been resected 6 months before CTC collection. Of the 93 SSNVs detected in this metastasis, 47 (51%) were detected in CTCs (**Fig. 3b**). Owing to the timing of the sample acquisition, the CTCs sequenced could not have derived from this particular metastasis, so nonoverlapping mutations could reflect divergent evolution at different sites, as has been demonstrated previously in prostate cancer[22,23].

We next asked whether sequencing CTCs could uncover mutations present early in tumor evolution (early trunk mutations) or in the inferred metastatic precursor (metastatic trunk mutations)[24]. Such founder mutations in other cancer types (for example, mutations in BRAF in malignant melanoma or in KIT in gastrointestinal stromal tumors) are excellent therapeutic targets[25,26]. Detecting such mutations in patients with various types of metastatic tumors through simple blood draws might therefore have considerable clinical utility.

We compared the landscape of mutations in CTCs and the metastatic sample to multiple samples from the patient's primary prostate tumor resected 5.3 years earlier. Because prostate cancer is often histologically multifocal[27], we sequenced nine spatially distinct foci of the primary tumor in regions of uniform Gleason grade (Gleason 4 or 5) with one exception noted (**Fig. 3c**). To assess relationships between these foci, the CTCs and the metastasis sample, we performed hierarchical clustering, excluding sites that were consistently underpowered in more than half of the samples (owing to lack of coverage) (**Fig. 3c**). Indeed, the primary tumor foci exhibited marked heterogeneity, but as expected, foci from similar physical regions of the tumor were more closely related to each other than those from other locations (**Fig. 3c,d**). Notably, we also identified one particular focus that most closely resembled the CTCs and the metastasis, suggesting that this focus may share a more recent common ancestor with the CTCs than other foci. Although this focus had a Gleason score of 5, the score itself did not predict the likely metastatic precursor, as other Gleason 5 regions were not in this evolutionary branch.

We found ten SSNVs, including a mutation in *TP53*, that were ubiquitous among all primary foci and metastasis, suggesting the cancer arose from a single ancestor with divergent evolution thereafter (**Fig. 3d**). The CTCs had nine of ten (90%) of these early trunk mutations (**Fig. 3e**). Notably, despite allelic distortion, these mutations were present in a greater fraction of CTCs (corrected for power) than the non-trunk mutations on average ($P = 0.0012$, Wilcoxon rank sum test). Fifty-six mutations were present in both the metastasis and primary tumor (any foci), and the CTCs had 41 (73%) of these metastatic trunk mutations (**Fig. 3e**). For patient CRPC_10, we found that the CTCs had all three early trunk mutations for which they were powered (out of nine) (**Supplementary Fig. 11c**). Together, these proof-of-concept data support the notion that CTC sequencing can reveal early mutations in tumor evolution and those that could be shared among metastatic sites. As such trunk mutations are likely to be present in the majority of sites in patients with advanced cancer, these results suggest clinical utility for systematic CTC genomics.

Here we have demonstrated the feasibility of sequencing whole exomes of prostate CTCs and confidently calling SSNVs to provide a minimally invasive window into the mutational landscape of metastatic prostate cancer. We implemented a systematic process to obtain, qualify and sequence whole exomes of CTCs and call SSNVs. Applying this process to two individual patients showed that sequencing of multiple independent CTC libraries can achieve full coverage of the exome territory that is accessible in bulk sequencing and a false-positive rate below the expected mutational rate in prostate cancer.

As implemented, the current process works well for patients from whom five or more single CTCs are recovered from 3.75 ml of blood and high-quality libraries generated for sequencing. The numbers of CTCs can vary substantially among different types of cancer[12], so sequencing of CTCs may not apply directly to all patients with cancer. Nonetheless, advancing the individual technologies used to recover and amplify as many CTCs as possible from patients would increase the numbers of cancers for which this approach could benefit. For instance, enriching CTCs on the basis of physical separation or microfluidics rather than expression of EpCAM[6], processing greater volumes of blood or using other means of whole-genome amplification that improve uniformity in genome-wide coverage of the amplified DNA could all increase the numbers of CTC libraries available[15]. The designed modularity of our approach can accommodate new emerging technologies for the enrichment and isolation of CTCs, whole-genome amplification[28] and sequencing platforms. Furthermore, as the costs of library preparation, hybrid selection and sequencing decline, we anticipate that census-based sequencing will become more cost effective for monitoring more patients.

The approach described here does not emphasize private mutations that may be held by individual CTCs. Such analysis of heterogeneity is extremely challenging because variants in individual cells can be absent for either technical (allelic distortion or false positive) or biological (subclonality in the population) reasons. Although we identified many mutations in single-cell libraries, our census-based approach deprioritizes these mutations because we could not distinguish them from false positives.

Exome sequencing of circulating tumor DNA has also demonstrated concordance of variants with tumor biopsies[29]. Although such DNA is fragmented and has a similarly low abundance, these materials may provide a complementary source for reducing false-positive calls or revealing other mutations that are not sampled among sparse CTCs.

Nonetheless, our results suggest that CTC sequencing could augment both large-scale efforts to map the genetics of cancer and clinical sequencing from individual patients with cancer. A focus on evolutionarily early and shared metastatic events such as those identified by the proof-of-concept study here would be critical for precision medicine. The integrated process may also enable longitudinal monitoring of the genetic state of disseminated cancer, revealing important insights in tumor evolution, metastatic dissemination and resistance to therapeutics.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequencing data have been deposited in dbGaP under accession code phs000717.v1.p1.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

J.G.L. was supported by a Conquer Cancer Foundation Young Investigator Award, US National Institutes of Health grant 5P50CA100707-10 (DF/HCC SPORE) and the Wong Family Award. V.A.A. was supported in part by a graduate fellowship from the National Science Foundation. A.D.C. is supported by the Prostate Cancer Foundation Young Investigator Award and the Department of Defense Physician Scientist Training Award. J.C.L. is a Camille Dreyfus Teacher-Scholar. We acknowledge the Arthur and Linda Gelb Center for Translational Research for the acquisition and annotation of clinical samples and A. Abbott and A. Van Den Abbeele from the Dana-Farber Cancer Institute (DFCI) Department of Imaging for positron-emission tomography (PET) images. We also acknowledge P.K. Brastianos (Department of Medical Oncology, DFCI) and I. Dunn (Department of Neurosurgery, Brigham and Women's Hospital) for contributing samples for CTC analysis, D. Peck for help with technology development, O. Voznesensky and S. Balk for purification of DNA from the metastatic tumor for sequencing, C. Whittaker and S.S. Levine for advice on sequencing and analysis and the Broad Genomics Platform for the development of new sequencing approaches used here. This work was also supported in part by the Koch Institute Support (core) grant P30-CA14051 from the National Cancer Institute, and we thank the Koch Institute Swanson Biotechnology Center for technical support, specifically the BioMicroCenter. This work was also supported in part by Janssen Pharmaceuticals, Inc. and the Klarman Family Foundation. We would like to thank Illumina for providing the MagSweeper. The authors dedicate this paper to the memory of Officer Sean Collier, for his caring service to the MIT community and for his sacrifice.

AUTHOR CONTRIBUTIONS

J.G.L. and V.A.A. designed and performed experiments, analyzed data and wrote the manuscript. K.C. and M.R. developed computational methods, analyzed data and wrote the manuscript. A.D.C. provided clinical samples and patient data and analyzed clinical data. P.C.-G., N. Tahirova and S.K. performed experiments for isolating CTCs. J.M.F. developed single-cell sequencing methods and designed experiments. C.-Z.Z. analyzed data and applied the autocorrelation methods. A.K.S., R.S., J.J.T. and D.L. performed single-cell RNA sequencing and data analysis. N. Tallapragada developed code for determining CTCs to recover from nanowells. B.B. performed early technology development. C.S. and D.A. performed sample and data management and gave conceptual advice. A. Lowe and A. Ly performed experiments comparing our process to the Veridex CellSearch System. E.M.V.A. analyzed sequencing data. M.N., G.-S.M.L., T.L. and M.S.C. coordinated and acquired clinical samples. R.T.L. reviewed pathology slides and guided selection of clinical samples. B.W. performed data visualization. T.E.C. provided samples and validated methods for isolating CTCs. M.-E.T., M.L., A.R., M.M., W.C.H. and P.W.K. supervised experiments and sample and data collection and edited the manuscript. T.R.G., G.G., J.S.B. and J.C.L. designed the experimental strategy, supervised the analysis and wrote the manuscript. All authors discussed the results and implications and reviewed the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Garraway, L.A. Genomics-driven oncology: framework for an emerging paradigm. *J. Clin. Oncol.* **31**, 1806–1814 (2013).
2. International Cancer Genome Consortium. *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010); erratum **465**, 966 (2010).
3. Dawson, S.-J. *et al.* Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N. Engl. J. Med.* **368**, 1199–1209 (2013).
4. Cristofanilli, M. *et al.* Circulating tumor cells: a novel prognostic factor for newly diagnosed metastatic breast cancer. *J. Clin. Oncol.* **23**, 1420–1430 (2005).
5. Zhang, L. *et al.* The identification and characterization of breast cancer CTCs competent for brain metastasis. *Sci. Transl. Med.* **5**, 180ra48 (2013).
6. Yu, M. *et al.* Circulating tumor cells: approaches to isolation and characterization. *J. Cell Biol.* **192**, 373–382 (2011).
7. Cohen, S.J. *et al.* Relationship of circulating tumor cells to tumor response, progression-free survival, and overall survival in patients with metastatic colorectal cancer. *J. Clin. Oncol.* **26**, 3213–3221 (2008).
8. Maheswaran, S. *et al.* Detection of mutations in EGFR in circulating lung-cancer cells. *N. Engl. J. Med.* **359**, 366–377 (2008).
9. Heitzer, E. *et al.* Complex tumor genomes inferred from single circulating tumor cells by array-CGH and next-generation sequencing. *Cancer Res.* **73**, 2965–2975 (2013).
10. Ni, X. *et al.* Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proc. Natl. Acad. Sci. USA* **110**, 21083–21088 (2013).
11. Yu, M. *et al.* RNA sequencing of pancreatic circulating tumour cells implicates WNT signalling in metastasis. *Nature* **487**, 510–513 (2012).
12. Allard, W.J. *et al.* Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases. *Clin. Cancer Res.* **10**, 6897–6904 (2004).
13. Swennenhuis, J.F. *et al.* Efficiency of whole genome amplification of single circulating tumor cells enriched by CellSearch and sorted by FACS. *Genome Med.* **5**, 106 (2013).
14. Hou, Y. *et al.* Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**, 873–885 (2012).
15. Zong, C. *et al.* Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626 (2012).
16. Cann, G.M. *et al.* mRNA-Seq of single prostate cancer circulating tumor cells reveals recapitulation of gene expression and pathways found in prostate cancer. *PLoS ONE* **7**, e49144 (2012).
17. El Gammal, A.T. *et al.* Chromosome 8p deletions and 8q gains are associated with tumor progression and poor prognosis in prostate cancer. *Clin. Cancer Res.* **16**, 56–64 (2010).
18. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
19. Grasso, C.S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).
20. Beltran, H. *et al.* New strategies in prostate cancer: translating genomics into the clinic. *Clin. Cancer Res.* **19**, 517–523 (2013).
21. Ross, R.W. *et al.* Predictors of prostate cancer tissue acquisition by an undirected core bone marrow biopsy in metastatic castration-resistant prostate cancer—a Cancer and Leukemia Group B study. *Clin. Cancer Res.* **11**, 8109–8113 (2005).
22. Robbins, C.M. *et al.* Copy number and targeted mutational analysis reveals novel somatic events in metastatic prostate tumors. *Genome Res.* **21**, 47–55 (2011).
23. Nickerson, M.L. *et al.* Somatic alterations contributing to metastasis of a castration-resistant prostate cancer. *Hum. Mutat.* **34**, 1231–1241 (2013).
24. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
25. Chapman, P.B. *et al.* Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.* **364**, 2507–2516 (2011).
26. Heinrich, M.C. *et al.* Kinase mutations and imatinib response in patients with metastatic gastrointestinal stromal tumor. *J. Clin. Oncol.* **21**, 4342–4349 (2003).
27. Lindberg, J. *et al.* Exome sequencing of prostate cancer supports the hypothesis of independent tumour origins. *Eur. Urol.* **63**, 347–353 (2013).
28. Gole, J. *et al.* Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nat. Biotechnol.* **31**, 1126–1132 (2013).
29. Murtaza, M. *et al.* Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* **497**, 108–112 (2013).
30. Barbieri, C.E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.* **44**, 685–689 (2012).

ONLINE METHODS

Patient recruitment. Eligible patients were metastatic patients with CRPC who had (i) progression on a phase 2 study of abiraterone in combination with dutasteride (DFCI Protocol # 10-448, Institutional Review Board (IRB) expiration date 2/7/2014) or (ii) PSA >20 ng/ml and progressive disease based on rising PSA and scan progression to enrich for patients likely to have detectable circulating tumor cells. There is no PSA cutoff for the phase 2 study itself. The Prostate Clinical Research Information System (CRIS) database at DFCI was used to identify metastatic patients with CRPC. The CRIS system comprises data-entry software, a central data repository, collection of patient data, including comprehensive follow-up of all patients, and tightly integrated security measures, as previously described[31]. All patients provided written informed consent to allow the collection of tissue and blood and the analysis of clinical and genetic data for research purposes (DFCI Protocol #01-045, IRB expiration date 3/20/2014). After initial screening of patients with metastatic CRPC, chart review was performed by a physician to identify those patients who had progressive disease as described above. Blood specimens were prospectively collected from eligible patients. **Supplementary Table 1** shows patient information. Blood was drawn in EDTA tubes and transported at room temperature to the Broad Institute within 3 h.

Patient characteristics. The clinical course of one of the patients described here (CRPC_36) was as follows: at the age of 54, the patient was diagnosed with a T3, Gleason 9, PSA 10 prostate cancer. He was treated on a research trial of neoadjuvant docetaxel and bevacizumab. At prostatectomy, he had a pT3b, N1 tumor. He next received adjuvant radiation and androgen deprivation therapy (ADT). Upon rising PSA, continuous ADT was reinitiated, and he was treated with sequential bicalutamide and nilutamide. Seventeen months after initial diagnosis, he developed metastases and was treated with the following therapies until his death 6 years after his presentation: docetaxel, phase 1 trial of PI3-kinase and MEK kinase inhibitor, sipuleucel-T, XL-184, abiraterone, enzalutamide, cabazitaxel, palliative radiation to bone and combined abiraterone and enzalutamide. At the time of CTC isolation, he received abiraterone and enzalutamide, and a lymph node biopsy was performed while he received enzalutamide. A basic summary of the timeline of events is illustrated in **Figure 3d**.

Enrichment of CTCs from blood. For each 3.75 ml of blood, red blood cell (RBC) lysis was first performed using 1× Pharm Lyse solution (BD Biosciences). The RBC-depleted sample was then incubated with four tests of fluorescein isothiocyanate (FITC)-conjugated antibody to CD45 (20 µl of antibody per 100 µl) (eBiosciences; clone HI30) for 30 min at 4 °C, followed by incubation with anti-EpCAM magnetic beads (Illumina)[16] for 30 min at 4 °C. Phycoerythrin (PE)-conjugated antibody to EpCAM (20 µl per 1 ml, BD Biosciences; clone EBA-1) was then added for 30 min at 4 °C before immunomagnetic isolation using the Illumina MagSweeper[16].

Isolation of CTCs. Isolation of CTCs from the enriched samples from the MagSweeper was performed using either the nanowell-based method with automated imaging and robotic retrieval of single cells or a six-well dish with manual imaging and pipetting. For the nanowell-based approach, enriched samples were loaded into the wells of a 1 × 3 in polydimethylsiloxane nanowell device containing a 24 × 72 array of 7 × 7 wells, each of the dimensions 50 × 50 × 50 µm (**Supplementary Fig. 2**)[32]. Automated epifluorescence imaging of the array was performed (Zeiss), and images were processed using a custom software program. After manual review of candidate cells with custom CTC analysis software (EVA), candidate CTCs ($DAPI^-CD45^-EpCAM^+$) were retrieved from individual wells of the device using an automated robotic micromanipulator (CellCelector, ALS) and deposited within 3-µl droplets of Superblock/PBS (Thermo Scientific) into empty wells of a 96-well PCR plate. For the manual approach, candidate CTCs ($DAPI^-CD45^-EpCAM^+$) were recovered from Superblock/PBS by pipetting 3 µl into a 96-well PCR plate. PCR plates were frozen down at −80 °C until ready for further processing.

Lysis and whole-genome amplification of CTCs. Each PCR plate containing frozen CTCs was thawed on ice, and the volume of the individual wells was diluted to 5 µl using UltraPure water (Invitrogen). 5 µl of lysis buffer, containing 0.4 M KOH (Sigma Aldrich) and 80 mM dithiothreitol (DTT) (Qiagen), was added to each well, and the plate was sealed, gently shaken to mix, spun down at 300 r.c.f. for 1 min and incubated for 10 min at 50 °C using a thermal cycler (Eppendorf). After lysis, the plate was spu down at 300 r.c.f. for 1 min, 5 µl of 0.4 M HCl (Fluka) was added to each well and the plate was kept on ice. Master mix for whole-genome amplification by MDA was prepared by adding, for each reaction, 26.25 µl of sterile water, 5 µl of 10× reaction buffer from the RepliPHI kit (Epicentre), 0.5 µl of 10 mg/ml bovine serum albumin (BSA) (NEB), 0.2 µl of 1 M DTT (Qiagen), 0.8 µl of 25 mM dNTPs from the RepliPHI kit (Epicentre), 1.25 µl of 10 mM random hexamers (NNNN*N*N) from IDT and 1 µl of RepliPHI enzyme from the RepliPHI kit (EpiCentre). 35 µl of this master mix was added to each well of the PCR plate containing lysed genomic DNA and incubated for 2 h at 30 °C on a thermal cycler (Eppendorf). After the MDA reaction, clean up was performed using AmpureXP beads (Beckman Coulter). Briefly, 100 µl of AmpureXP beads were added to each sample and incubated for 5 min at room temperature. The samples were then placed on a 96-well plate magnet (Invitrogen) and incubated for 5 min. Supernatant was removed from each sample, and 100 µl of fresh 70% ethanol (Koptec) was added and removed twice to wash the beads. After complete removal of the ethanol and drying for 10 min at room temperature, beads were resuspended in 60 µl of Tris-EDTA buffer, pH 8 (Teknova), incubated for 5 min at room temperature and placed back on the magnet for 5 min, and then cleaned-up products were transferred to a new PCR plate. These MDA products were quantified using the Quant-lT PicoGreen dsDNA assay kit (Invitrogen), and products with concentrations greater than the negative control were selected for low-pass whole-genome sequencing.

Library preparation and low-pass whole-genome sequencing. Whole-genome sequencing libraries were prepared using the Nextera DNA Sample Prep Kit (Illumina), quantified using the Library Quantification Kit for Illumina (Kapa Biosystems) and pooled and loaded at 12 pM onto the Illumina MiSeq sequencer using the MiSeq Reagent Kit v2 (Illumina). Up to 96 libraries can be multiplexed in the same run. The MiSeq Reporter (Illumina) was used to align reads and generate BAM files, and IGV Tools (Broad Institute) was used to bin the genome for coverage at 1-Mb intervals and generate TDF files. The TDF files were viewed in the Integrative Genomics Viewer (IGV) to visually inspect genome-wide uniformity in coverage of each MDA product.

Calculation of autocorrelation coefficient and selection of CTC libraries. Using the data from low-pass whole-genome sequencing, we computed the degree of correlation in single-base coverage over various distances, normalized by the mean target coverage, for each library. The DepthOfCoverage module from GATK was used to compute single-base coverage (http://www.broadinstitute.org/gatk/) using a minimum mapping quality of 5, and the autocorrelation coefficient represents the magnitude (not the length scale or genomics distance of correlation) of the correlation in single base coverage at 1-kb distances normalized by the mean sequencing depth. 1 kb represents a length scale that is well above the average fragment length yet is short enough to capture local biases in coverage due to preferential overamplification in whole-genome amplification. In our study, the autocorrelation analysis was computed over chromosome 1 because it is the largest chromosome, does not have visible copy-number alterations from both whole-genome sequencing and WES read coverage and provides the analysis with the most statistical power. The analysis could have been performed on other chromosomes, too. Chromothripsis, although rare, would only affect the correlation near translocation junctions, which would be a small fraction of the chromosome and would have negligible effects on the analysis.

Libraries were ranked on the basis of autocorrelation coefficient. In this study, we selected libraries for WES that had the logarithm of (1/autocorrelation coefficient) greater than −1.8. In the rare event of insufficient coverage (~0.0001×) for computing of the autocorrelation coefficient, visual inspection of genome-wide read densities may be used to include samples with seemingly uniform genomic coverage, as demonstrated in **Supplementary Figure 6**.

Isolation of genomic DNA from blood and tumor tissue. Genomic DNA was isolated from blood to control for germline variants using the DNeasy Blood and Tissue Kit (Qiagen). 100 µl of anticoagulated blood was added to 20 µl proteinase K and adjusted to a 220 µl volume with PBS. 200 µl of Buffer AL (from the DNeasy

Blood and Tissue Kit) was added, mixed and incubated at 56 °C for 10 min. All subsequent steps were performed per the manufacturer's recommendations.

Genomic DNA and RNA were isolated from primary tumor tissue using the AllPrep DNA/RNA Mini Kit (Qiagen) and from metastatic tumor tissue using the AllPrep DNA/RNA Micro Kit (Qiagen) following the manufacturer's recommendations. Primary tumor tissue consisted of blocks of fresh frozen tissue acquired at the time of radical prostatectomy frozen in optimal cutting temperature (OCT) medium and stored in liquid nitrogen at the Gelb Center for Translational Research. Accompanying slides were reviewed by a pathologist, and tumor boundaries were marked at the time of storage. Slides were re-reviewed at the time of retrieval for the presence of tumor, regions of Gleason 3, 4 and 5 within the tumor were identified, and the areas were marked. Nine representative tumor foci were chosen to maximize the distance between the cores, favoring regions of higher Gleason grade. Each block was removed from the cassette, placed on a Petri dish on dry ice to keep cold and aligned with the accompanying marked slide to identify the selected foci. Blocks were cored using a Miltex 2-mm Disposable Biopsy Punch with Plunger, placed into a DNA LoBind Eppendorf tube and stored at −80 °C until the time of nucleic acid extraction.

Metastatic tumor tissue consisted of blocks of fresh frozen tissue acquired from excision of the left supraclavicular lymph node and frozen in OCT medium. Tumor shavings were obtained at the time of sectioning the OCT block for immunohistochemistry. About 100–300-µm shavings were obtained using a cryostat, placed into a DNA LoBind Eppendorf tube and stored at −80 °C until the time of nucleic acid extraction.

Selection of libraries and WES. WES was performed as previously described[33]. Briefly, 100 ng of DNA from each sample was used for library preparation, which included shearing and ligation of sequencing adaptors. Exome capture was performed using the Agilent v2 Human Exon bait kit. Captured DNA was sequenced using the Illumina HiSeq platform, and paired-end sequencing reads were generated for each sample. Initial alignment and quality control were performed using the Picard and Firehose pipelines at the Broad Institute. Picard generates a single BAM file for each sample that includes reads, calibrated quantities and alignments to the genome. Firehose is a set of tools for analyzing sequencing data from tumor and matched normal DNA. The pipeline uses GenePattern as its execution engine and performs quality control, local realignment, mutation calling and coverage calculations, among other analyses. Complete details of this pipeline can be found in Stransky *et al.*[34] or at http://www.broadinstitute.org/cancer/cga/. Sequencing was performed to an average target coverage of >120×.

Calling of SSNVs from WES data. Reads were aligned to the reference human genome build hg19 through implementation of the Burrows-Wheeler Aligner[35] and processed through Picard[36]. The Firehose pipeline (http://www.broadinstitute.org/cancer/cga) was used to manage input and output files and submit analyses for execution. MuTect was used to identify somatic SSNVs in targeted exons by Bayesian statistical analysis of bases and their qualities in the tumor and normal BAMs at each given genomic locus; the MuTect publication describes the specificity and sensitivity of the method[18]. Reads from all SSNV candidates were then realigned more stringently by disregarding read-pair information to reduce alignment-based artifacts. All SSNVs were subjected to filtering against a large panel of normal samples in order to remove common artifacts that escaped the original calling algorithms[37]. Further, only sites within chromosomes 1–22 and X were considered. Sites in the exomes of primary samples and metastasis were considered to be powered for mutation calling if the number of reads at a site allowed for a 0.9 probability of observing three supporting reads of the alternate allele, considering the purity of the sample and assuming a minimum cancer cell fraction of 0.1. Sites were considered powered in CTCs if five CTCs had coverage of more than three reads (this achieves ~98% power to detect a clonal mutation based on a maximum loss of coverage of 14.4% of the alternate allele only across the CTCs, as determined in **Fig. 2a**).

Calculation of false-positive rate. To calculate the effective false-positive rate of the method, we computed the total number of potential false-positive events and the territory at risk for these events. To calculate an upper bound for the number of events, we assumed that every event was a false positive after removing events seen independently in at least one of the primary tumor cores or metastasis. To calculate the denominator of the false-positive rate, or the number of bases that were at risk for a mistake being made, we considered the effects of biallelic dropout, as this leads to regions of the genome with no coverage and thus no possibility of a false positive occurring. Biallelic drop out was calculated from germline heterozygous single-nucleotide polymorphisms (SNPs) (**Fig. 2b**). We then used a binomial model to calculate the probability of having k or more observations in n CTCs using 1 minus the median value of the estimated fraction of sites without biallelic dropout (0.33) as the probability of each CTC having coverage and multiplied by the approximate size of the original targeted exome territory (~32 Mb) to arrive at the total number of bases at risk.

RNA sequencing of single CTCs. Single CTCs were recovered using the nanowell-based isolation platform into individual wells of a 96-well plate containing 10 µl of buffer TCL (Qiagen) supplemented with 1% 2-mercaptoethanol (Sigma), spun down, snap frozen on dry ice and stored at −80 °C until further processing. Next, RNA from each single CTC was isolated, reverse transcribed and amplified using the SMARTer Ultra-low RNA kit (Clontech) as previously described[38]. Afterwards, cDNA libraries were prepared using Nextera XT DNA Sample preparation reagents (Illumina) as per the manufacturer's recommendations, with minor modifications. Specifically, reactions were run at one-fourth the recommended volume, the tagmentation step was extended to 10 min and the extension time during the PCR step was increased from 30 s to 60 s. After the PCR step, all 96 samples were pooled without library normalization, cleaned twice with 0.9× AMPure XP SPRI beads (Beckman Coulter) and eluted in Tris-EDTA buffer, pH 8 (Teknova). The pooled libraries were quantified using the Quant-IT DNA High-Sensitivity Assay Kit (Invitrogen) and examined using a high-sensitivity DNA chip (Agilent). Finally, samples were sequenced using a MiSeq sequencer (Illumina).

Analysis of RNA sequencing data. Raw sequencing data were processed as described previously[38] except that there was no need to trim SMARTer short and long adaptor sequences because of the Nextera library preparation[39]. Short sequencing reads were aligned to the UCSC hg19 transcriptome. These alignments were used to estimate transcriptomic alignment rates and were also used as input in RSEM v 1.12 (ref. 40) to quantify gene expression levels (transcripts per million) for all UCSC hg19 genes in all samples. Genomic mappings were performed with Tophat v. 1.41 (ref. 41), and the resulting alignments were used to calculate genomic mapping rates, ribosomal RNA contamination and 3′ and 5′ positional bias[36].

Cell spike-in experiments. LNCaP prostate cancer cells (ATCC) were cultured in RPMI-1640 medium (Corning, Cellgro) supplemented with 10% FBS (Sigma). To determine surface expression of EpCAM, LNCaP cells were stained with PE-conjugated antibodies to EpCAM (20 µl per 1 ml, BD Biosciences; clone EBA-1), and the level of expression was determined by flow cytometry on a LSRII (BD Biosciences) compared to unstained control. For technical validation of the MagSweeper enrichment procedure, LNCaP cells were labeled with carboxyfluorescein diacetate succinimidyl ester (CFDA; Invitrogen) and spiked into normal blood (obtained from Research Blood Components, LLC) at the indicated concentrations before MagSweeper enrichment. Successful isolation of LNCaP cells was ascertained by determining the number of CFDA-labeled cells by microscopy.

31. Oh, W.K. *et al.* Development of an integrated prostate cancer research information system. *Clin. Genitourin. Cancer* **5**, 61–66 (2006).
32. Love, J.C. *et al.* A microengraving method for rapid selection of single cells producing antigen-specific antibodies. *Nat. Biotechnol.* **24**, 703–707 (2006).
33. Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* **12**, R1 (2011).
34. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157–1160 (2011).
35. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
36. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
37. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
38. Shalek, A.K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
39. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
40. Li, B. & Dewey, C. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
41. Trapnell, C. *et al.* TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).

doi:10.1038/nbt.2892