Genome Medicine

RESEARCH Open Access

Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing

Ellen Heitzer[1†], Peter Ulz[1†], Jelena Belic[1], Stefan Gutschi[2], Franz Quehenberger[3], Katja Fischereder[2], Theresa Benezeder[1], Martina Auer[1], Carina Pischler[1], Sebastian Mannweiler[4], Martin Pichler[5], Florian Eisner[5], Martin Haeusler[6], Sabine Riethdorf[7], Klaus Pantel[7], Hellmut Samonigg[5], Gerald Hoefler[4], Herbert Augustin[2], Jochen B Geigl[1*] and Michael R Speicher[1*]

Abstract

Background: Patients with prostate cancer may present with metastatic or recurrent disease despite initial curative treatment. The propensity of metastatic prostate cancer to spread to the bone has limited repeated sampling of tumor deposits. Hence, considerably less is understood about this lethal metastatic disease, as it is not commonly studied. Here we explored whole-genome sequencing of plasma DNA to scan the tumor genomes of these patients non-invasively.

Methods: We wanted to make whole-genome analysis from plasma DNA amenable to clinical routine applications and developed an approach based on a benchtop high-throughput platform, that is, Illuminas MiSeq instrument. We performed whole-genome sequencing from plasma at a shallow sequencing depth to establish a genome-wide copy number profile of the tumor at low costs within 2 days. In parallel, we sequenced a panel of 55 high-interest genes and 38 introns with frequent fusion breakpoints such as the *TMPRSS2-ERG* fusion with high coverage. After intensive testing of our approach with samples from 25 individuals without cancer we analyzed 13 plasma samples derived from five patients with castration resistant (CRPC) and four patients with castration sensitive prostate cancer (CSPC).

Results: The genome-wide profiling in the plasma of our patients revealed multiple copy number aberrations including those previously reported in prostate tumors, such as losses in 8p and gains in 8q. High-level copy number gains in the *AR* locus were observed in patients with CRPC but not with CSPC disease. We identified the *TMPRSS2-ERG* rearrangement associated 3-Mbp deletion on chromosome 21 and found corresponding fusion plasma fragments in these cases. In an index case multiregional sequencing of the primary tumor identified different copy number changes in each sector, suggesting multifocal disease. Our plasma analyses of this index case, performed 13 years after resection of the primary tumor, revealed novel chromosomal rearrangements, which were stable in serial plasma analyses over a 9-month period, which is consistent with the presence of one metastatic clone.

Conclusions: The genomic landscape of prostate cancer can be established by non-invasive means from plasma DNA. Our approach provides specific genomic signatures within 2 days which may therefore serve as 'liquid biopsy'.

* Correspondence: jochen.geigl@medunigraz.at; michael.speicher@medunigraz.at
† Contributed equally
[1]Institute of Human Genetics, Medical University of Graz, Harrachgasse 21/8, A-8010 Graz, Austria
Full list of author information is available at the end of the article

Background

Prostate cancer is the most common malignancy in men. In Europe each year an estimated number of 2.6 million new cases is diagnosed [1]. The wide application of PSA testing has resulted in a shift towards diagnosis at an early stage so that many patients do not need treatment or are cured by radical surgery [2]. However, patients still present with metastatic or recurrent disease despite initial curative treatment [3]. In these cases prostate-cancer progression can be inhibited by androgen-deprivation therapy (ADT) for up to several years. However, disease progression is invariably observed with tumor cells resuming proliferation despite continued treatment (termed castration-resistant prostate cancer or CRPC) [4]. CRPC is a strikingly heterogeneous disease and the overall survival can be extremely variable [5]. Scarcity of predictive and prognostic markers underlines the growing need for a better understanding of the molecular makeup of these lethal tumors.

However, acquiring tumor tissue from patients with metastatic prostate cancer often represents a challenge. Due to the propensity of metastatic prostate cancer to spread to bone biopsies can be technically challenging and limit repeated sampling of tumor deposits. As a consequence, considerably less is understood about the later acquired genetic alterations that emerge in the context of the selection pressure of an androgen-deprived milieu [6].

Consistent and frequent findings from recent genomic profiling studies in clinical metastatic prostate tumors include the *TMPRSS2-ERG* fusion in approximately 50%, 8p loss in approximately 30% to 50%, 8q gain in approximately 20% to 40% of cases, and the androgen receptor (*AR*) amplification in approximately 33% of CRPC cases [7-10]. Several whole-exome or whole-genome sequencing studies consistently reported low overall mutation rates even in heavily treated CRPCs [9-14].

The difficulties in acquiring tumor tissue can partly be addressed by elaborate procedures such as rapid autopsy programs to obtain high-quality metastatic tissue for analysis [15]. However, this material can naturally only be used for research purposes, but not for biomarker detection for individualized treatment decisions. This makes blood-based assays crucially important to individualize management of prostate cancer [16]. Profiling of blood offers several practical advantages, including the minimally invasive nature of sample acquisition, relative ease of standardization of sampling protocols, and the ability to obtain repeated samples over time. For example, the presence of circulating tumor cells (CTCs) in peripheral blood is a prognostic biomarker and a measure of therapeutic response in patients with prostate cancer [17-20]. Novel microfluidic devices enhance CTC capture [21-23] and allow to establish a non-invasive measure of intratumoral AR signaling before and after hormonal therapy [24]. Furthermore, prospective studies have demonstrated that mRNA expression signatures from whole blood can be used to stratify patients with CRPC into high- and low-risk groups [25,26].

Another option represents the analysis of plasma DNA [27]. One approach is the identification of known alterations previously found in the resected tumors from the same patients in plasma DNA for monitoring purposes [28,29]. Furthermore, recurrent mutations can be identified in plasma DNA in a subset of patients with cancer [30-32]. Given that chromosomal copy number changes occur frequently in human cancer, we developed an approach allowing the mapping of tumor-specific copy number changes from plasma DNA employing array-CGH [33]. At the same time, massively parallel sequencing of plasma DNA from the maternal circulation is emerging to a clinical tool for the routine detection of fetal aneuploidy [34-36]. Using essentially the same approach, that is, next-generation sequencing from plasma, the detection of chromosomal alterations in the circulation of three patients with hepatocellular carcinoma and one patient with both breast and ovarian cancer [37] and from 10 patients with colorectal and breast cancer [38] was reported.

However, the costs of the aforementioned plasma sequencing studies necessary for detection of rearrangements were prohibitive for routine clinical implementation [37,38]. In addition, these approaches are very time-consuming. Previously it had been shown that whole-genome sequencing with a shallow sequencing depth of about 0.1x is sufficient for a robust and reliable analysis of copy number changes from single cells [39]. Hence, we developed a different whole-genome plasma sequencing approach employing a benchtop high-throughput sequencing instrument, that is, the Illumina MiSeq, which is based on the existing Solexa sequencing-by-synthesis chemistry, but has dramatically reduced run times compared to the Illumina HiSeq [40]. Using this instrument we performed whole-genome sequencing from plasma DNA and measured copy number from sequence read depth. We refer to this approach as plasma-Seq. Furthermore, we enriched 1.3 Mbp consisting of exonic sequences of 55 high-interest cancer genes and 38 introns of genes, where fusion breakpoints have been described and subjected the DNA to next-generation sequencing at high coverage (approximately 50x). Here we present the implementation of our approach with 25 plasma samples from individuals without cancer and results obtained with whole genome sequencing of 13 plasma DNA samples derived from nine patients (five CRPC, four CSPC) with prostate cancer.

Methods

Patient eligibility criteria

This study was conducted among men with prostate cancer (Clinical data in Additional file 1, Table S1) who met the following criteria: histologically-proven, based on a biopsy, metastasized prostate cancer. We distinguished between CRPC and CSPC based on the guidelines on prostate cancer from the European Association of Urology [41], that is: 1, castrate serum levels of testosterone (testosterone <50 ng/dL or <1.7 nmol/L); 2, three consecutive rises of PSA, 1 week apart, resulting in two 50% increases over the nadir, with a PSA >2 ng/mL; 3, anti-androgen withdrawal for at least 4 weeks for flutamide and for at least 6 weeks for bicalutamide; 4, PSA progression, despite consecutive hormonal manipulations. Furthermore, we focused on patients who had ≥5 CTCs per 7.5 mL [19] and/or a biphasic plasma DNA size distribution as described previously by us [33].

The study was approved by the ethics committee of the Medical University of Graz (approval numbers 21-228 ex 09/10, prostate cancer, and 23-250 ex 10/11, prenatal plasma DNA analyses), conducted according to the Declaration of Helsinki, and written informed consent was obtained from all patients and healthy blood donors. Blood from prostate cancer patients and from male controls without malignant disease was obtained from the Department of Urology or the Division of Clinical Oncology, Department of Internal Medicine, at the Medical University of Graz. From prostate cancer patients we obtained a buccal swab in addition. Blood samples from pregnant females and from female controls without malignant disease were collected at the Department of Obstetrics and Gynecology, Medical University of Graz. The blood samples from the pregnant females were taken prior to an invasive prenatal diagnostic procedure.

Plasma DNA preparation

Plasma DNA was prepared using the QIAamp DNA Blood Mini Kit (Qiagen, Hilden, Germany) as previously described [33]. Samples selected for sequence library construction were analyzed by using the Bioanalyzer instrument (Agilent Technologies, Santa Clara, CA, USA) to observe the plasma DNA size distribution. In this study we included samples with a biphasic plasma DNA size distribution as previously described [33].

Enumeration of CTCs

We performed CTC enumeration using the automated and FDA approved CellSearch assay. Blood samples (7.5 mL each) were collected into CellSave tubes (Veridex, Raritan, NJ, USA). The Epithelial Cell Kit (Veridex) was applied for CTC enrichment and enumeration with the CellSearch system as described previously [42,43].

Array-CGH

Array-CGH was carried out using a genome-wide oligonucleotide microarray platform (Human genome CGH 60K microarray kit, Agilent Technologies, Santa Clara, CA, USA), following the manufacturer's instructions (protocol version 6.0) as described [33]. Evaluation was done based on our previously published algorithm [33, 44, 45].

HT29 dilution series

Sensitivity of our plasma-Seq approach was determined using serial dilutions of DNA from HT29 cell line (50%, 20%, 15%, 10%, 5%, 1%, and 0%) in the background of normal DNA (Human Genomic DNA: Female; Promega, Fitchburg, WI, USA). Since quantification using absorption or fluorescence absorption is often not reliable we used quantitative PCR to determine the amount of amplifiable DNA and normalized the samples to a standard concentration using the Type-it CNV SYBR Green PCR Kits (Qiagen, Hilden, Germany). Dilution samples were then fragmented using the Covaris S220 System (Covaris, Woburn, MA, USA) to a maximum of 150-250 bp and 10 ng of each dilution were used for library preparation to simulate plasma DNA condition.

Plasma-Seq

Shotgun libraries were prepared using the TruSeq DNA LT Sample preparation Kit (Illumina, San Diego, CA, USA) following the manufacturer´s instructions with three exceptions. First, due to limited amounts of plasma DNA samples we used 5-10 ng of input DNA. Second, we omitted the fragmentation step since the size distribution of the plasma DNA samples was analyzed on a Bioanalyzer High Sensitivity Chip (Agilent Technologies, Santa Clara, CA, USA) and all samples showed an enrichment of fragments in the range of 160 to 340 bp. Third, for selective amplification of the library fragments that have adapter molecules on both ends we used 20-25 PCR cycles. Four libraries were pooled equimolarily and sequenced on an Illumina MiSeq (Illumina, San Diego, CA, USA).

The MiSeq instrument was prepared following routine procedures. The run was initiated for 1x150 bases plus 1x25 bases of SBS sequencing, including on-board clustering and paired-end preparation, the sequencing of the respective barcode indices and analysis. On the completion of the run, data were base called and demultiplexed on the instrument (provided as Illumina FASTQ 1.8 files, Phred+33 encoding). FASTQ format files in Illumina 1.8 format were considered for downstream analysis.

Calculation of segments with identical $\log_2$ ratio values

We employed a previously published algorithm [46] to create a reference sequence. The pseudo-autosomal region

(PAR) on the Y chromosome was masked and the mappability of each genomic position examined by creating virtual 150 bp reads for each position in the PAR-masked genome. Virtual sequences were mapped to the PAR-masked genome and mappable reads were extracted. Fifty thousand genomic windows were created (mean size, 56,344 bp) each having the same amount of mappable positions.

Low-coverage whole-genome sequencing reads were mapped to the PAR-masked genome and reads in different windows were counted and normalized by the total amount of reads. We further normalized read counts according to the GC content using LOWESS-statistics. In order to avoid position effects we normalized the sequencing data with GC-normalized read counts of plasma DNA of our healthy controls and calculated $\log_2$ ratios.

Resulting normalized ratios were segmented using circular binary segmentation (CBS) [47] and GLAD [48] by applying the CGHweb [49] framework in R [50]. These segments were used for calculation of the segmental z-scores by adding GC-corrected read-count ratios (read-counts in window divided by mean read-count) of all the windows in a segment. Z-scores were calculated by subtracting mean sum of GC-corrected read-count ratios of individuals without cancer (10 for men and 9 for women) of same sex and dividing by their standard-deviation.

$$z_{segments} = \frac{\sum ratio_{GC-corr} - mean\left(\sum ratio_{GC-corr,controls}\right)}{SD \sum \left(ratio_{GC-corr,controls}\right)}$$

Calculation of z-scores for specific regions

In order to check for the copy-number status of genes previously implicated in prostate-cancer initiation or progression we applied z-score statistics for each region focusing on specific targets (mainly genes) of variable length within the genome. At first we counted high-quality alignments against the PAR-masked hg19 genome within genes for each sample and normalized by expected read counts.

$$ratio = \frac{reads_{region}}{reads_{expected}}$$

Here expected reads are calculated as

$$reads_{expected} = \frac{length_{region}}{length_{genome}} * reads_{total}$$

Then we subtracted the mean ratio of a group of controls and divided it by the standard deviation of that group.

$$z_{region} = \frac{ratio_{sample} - mean\left(ratio_{controls}\right)}{SD\left(ratio_{controls}\right)}$$

Calculation of genome-wide z-scores

In order to establish a genome-wide z-score to detect aberrant genomic content in plasma, we divided the genome into equally-sized regions of 1 Mbp length and calculated z-scores therein.

Under the condition that all ratios were drawn from the same normal distribution, z-scores are distributed proportionally to Student's t-distribution with n-1 degrees of freedom. For controls, z-scores were calculated using cross-validation. In brief, z-score calculation of one control is based on means and standard deviation of the remaining controls. This prevents controls from serving as their own controls.

The variance of these cross-validated z-scores of controls is slightly higher than the variance of z-scores of tumor patients. Thus ROC performance is underestimated. This was confirmed in the simulation experiment described below.

In order to summarize the information about high or low z-score that was observed in many tumor patients squared z-scores were summed up.

$$S = \sum\nolimits_{i\ from\ all\ Windows} z_i^2$$

Genome-wide z-scores were calculated from S-scores. Other methods of aggregation of z-score information, such as sums of absolute values or PA scores [38], performed poorer and were therefore not considered. Per window z-scores were clustered hierarchically by the *hclust* function of R using Manhattan distance that summed up the distance of each window.

In order to validate the diagnostic performance of the genome-wide z-score *in silico*, artificial cases and controls were simulated from mean and standard deviations of ratios from 10 healthy controls according to a normal distribution. Simulated tumor cases were obtained through multiplication of the mean by the empirical copy number ratio of 204 prostate cancer cases [9]. Segmented DNA-copy-number data were obtained via the cBio Cancer Genomics Portal [51].

To test the specificity of our approach at varying tumor DNA content, we performed *in-silico* dilutions of simulated tumor data. To this end we decreased the tumor signal using the formula below, where λ is the ratio of tumor DNA to normal DNA:

$$(1 - \lambda) + \lambda \cdot ratio_{segment}$$

We performed ROC analyses of 500 simulated controls and 102 published prostate tumor data and their respective dilutions using the pROC R-package [52]. The prostate tumor data were derived from a previously published dataset [9] and the 102 cases were selected based on their copy number profiles.

Gene-Breakpoint Panel: target enrichment of cancer genes, alignment and SNP-calling, SNP-calling results

We enriched 1.3 Mbp of seven plasma DNAs (four CRPC cases, CRPC1-3 and CRPC5; three CSPC cases, CSPC1-2 and CSPC4) including exonic sequences of 55 cancer genes and 38 introns of 18 genes, where fusion breakpoints have been described using Sure Select Custom DNA Kit (Agilent, Santa Clara, CA, USA) following the manufacturer's recommendations. Since we had very low amounts of input DNA we increased the number of cycles in the enrichment PCR to 20. Six libraries were pooled equimolarily and sequenced on an Illumina MiSeq (Illumina, San Diego, CA, USA).

We generated a mean of 7.78 million reads (range, 3.62-14.96 million), 150 bp paired-end reads on an Illumina MiSeq (Illumina, San Diego, CA, USA). Sequences were aligned using BWA [53] and duplicates were marked using picard [54]. We subsequently performed realigning around known indels and applied the Unified Genotyper SNP-calling software provided by the GATK [55].

We further annotated resulting SNPs by employing annovar [56] and reduced the SNP call set by removing synonymous variants, variants in segmental duplications and variants listed in the 1000 Genome Project [57] and Exome sequencing (Project Exome Variant Server, NHLBI Exome Sequencing Project (ESP), Seattle, WA) [58] with allele frequency >0.01.

We set very stringent criteria to reduce false positives according to previously published values [37]: a mutation had to be absent from the constitutional DNA sequencing and the sequencing depth for the particular nucleotide position had to be >20-fold. Furthermore, all putative mutations or breakpoint spanning regions were verified by Sanger sequencing.

Split-read analysis

Since plasma DNA is fragmented the read pair method is not suitable for identification of structural rearrangements [59] and therefore we performed split-read analysis of 150 bp reads. We used the first and the last 60 bp of each read (leaving a gap of 30 bp) and mapped these independently. We further analyzed discordantly mapped split-reads by focusing on targeted regions and filtering out split-reads mapping within repetitive regions and alignments having a low mapping quality (<25). Reads where discordantly mapped reads were found were aligned to the human genome using BLAT [60] to further specify putative breakpoints.

Data deposition

All sequencing raw data were deposited at the European Genome-phenome Archive (EGA) [61], which is hosted by the EBI, under accession numbers EGAS00001000451 (Plasma-Seq) and EGAS00001000453 (Gene-Breakpoint Panel).

Results

Implementation of our approach

Previously, we demonstrated that tumor-specific, somatic chromosomal alterations can be detected from plasma of patients with cancer using array-CGH [33]. In order to extend our method to a next-generation sequencing-based approach, that is, plasma-Seq, on a benchtop Illumina MiSeq instrument, we first analyzed plasma DNA from 10 men (M1 to M10) and nine women (F1 to F9) without malignant disease. On average we obtained 3.3 million reads per sample (range, 1.9-5.8 million; see Additional file 1, Table S2) and applied a number of filtering steps to remove sources of variation and to remove known GC bias effects [62-64] (for details see Material and Methods).

We performed sequential analyses of 1-Mbp windows (n=2,909 for men; n=2,895 for women) throughout the genome and calculated for each 1-Mbp window the z-score by cross-validating each window against the other control samples from the same sex (details in Material and Methods). We defined a significant change in the regional representation of plasma DNA as >3 SDs from the mean representation of the other healthy controls for the corresponding 1-Mbp window. A mean of 98.5% of the sequenced 1-Mbp windows from the 19 normal plasma samples showed normal representations in plasma (Figure 1a). The variation among the normalized proportions of each 1-Mbp window in the plasma from normal individuals was very low (average, 47 windows had a z-score £-3 or ≥3; range of SD, ±52%) (Figure 1a).

In addition, we calculated 'segmental z-scores' where the z-scores are not calculated for 1-Mbp windows but for chromosomal segments with identical copy number. In order to determine such segments we employed an algorithm for the assignment of segments with identical $\log_2$ ratios [39,46] (Material and Methods) and calculated a z-score for each of these segments (hence, 'segmental z-scores'). As sequencing analyses of chromosome content in the maternal circulation are now frequently being used for detection of fetal aneuploidy [34,36] and as our mean sequencing depth is lower compared to previous studies, we wanted to test whether our approach would be feasible for this application. To this end we obtained two plasma samples each of pregnancies with euploid and trisomy 21 fetuses and one each of pregnancies with trisomies of chromosomes 13 and 18, respectively. In the trisomy cases the respective chromosomes were identified as segments with elevated $\log_2$ ratios and accordingly also increased z-scores (Additional file 2).

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.