Cancer Genetics 207 (2014) 474–488

Cancer Genetics

Recurrent copy number alterations in prostate cancer: an in silico meta-analysis of publicly available genomic data

Julia L. Williams [a], Peter A. Greer [a], Jeremy A. Squire [a,b,*]

[a] *Department of Pathology and Molecular Medicine, Queen's University, Kingston, Ontario, Canada;* [b] *Departments of Genetics and Pathology, School of Medicine of Ribeirao Preto, University of Sao Paulo at Ribeirao Preto, Sao Paulo, Brazil*

We present a meta-analysis of somatic copy number alterations (CNAs) from 11 publications that examined 662 prostate cancer patient samples, which were derived from 546 primary and 116 advanced tumors. Normalization, segmentation, and identification of corresponding CNAs for meta-analysis was achieved using established commercial software. Unsupervised analysis identified five genomic subgroups in which approximately 90% of the samples were characterized by abnormal profiles with gains of 8q. The most common loss was 8p (*NKX3.1*). The CNA distribution in other genomic subgroups was characterized by losses at 2q, 3p, 5q, 6q, 13q, 16q, 17p, 18q, and *PTEN* (10q), and acquisition of 21q deletions associated with the *TMPRSS2-ERG* fusion rearrangement. Parallel analysis of advanced and primary tumors in the cohort indicated that genomic deletions of *PTEN* and the gene fusion were enriched in advanced disease. A supervised analysis of the *PTEN* deletion and the fusion gene showed that *PTEN* deletion was sufficient to impose higher levels of CNA. Moreover, the overall percentage of the genome altered was significantly higher when *PTEN* was deleted, suggesting that this important genomic subgroup was likely characterized by intrinsic chromosomal instability. Predicted alterations in expression levels of candidate genes in each of the recurrent CNA regions characteristic of each subgroup showed that signaling networks associated with cancer progression and genome stability were likely to be perturbed at the highest level in the *PTEN* deleted genomic subgroup.

Keywords Array CGH, genomic instability, pI3 kinase, bioinformatics, tumor suppressor

Prostate cancer is the most commonly diagnosed malignancy in men and a leading cause of cancer deaths in developed countries (1). Emerging prostate cancer genomic data hold great promise not only in stratifying this heterogeneous disease at biopsy, but also in providing the groundwork for future development of targeted therapies (2).

The frequencies of mutated genes in prostate cancer, which are determined by sequence-based methods, are surprisingly low, with *TP53* (17%), *TTN* (15%), *PTEN* (11%), *MUC16* (9%), and *SPOP* (8%) (Catalogue of Somatic Mutations in Cancer (COSMIC)) (3). In contrast, the frequency of large-scale copy number alterations (CNAs) and genomic rearrangements is significantly higher, suggesting that the development and progression of prostate cancer is primarily the result of an accumulation of larger-scale genomic aberrations, such as deletions, gains, and fusion gene formation (4–6), instead of more localized mutational events.

Genomic rearrangements leading to the formation of *TMPRSS2-ETS* gene fusions and deletion of the *PTEN* tumor suppressor (10q23.31) often occur concurrently, and are the most widely reported genomic biomarkers in prostate cancer (7). The *TMPRSS2-ERG* gene fusion is the principle ETS family prostate cancer–specific gene fusion, a characteristic signature in approximately one half of prostatic malignancies. *PTEN* deletions and the *TMPRSS2-ERG* fusion genes are independently associated with aggressive disease; likewise, concurrent exhibition portends a shorter time to biochemical recurrence and decreased prostate cancer–specific survival (8–12). Metastatic disease is portrayed by heightened genomic complexity as well as an increased frequency of CNAs (7). Hormonal therapies for

Received February 16, 2014; received in revised form July 22, 2014; accepted September 7, 2014.

* Corresponding author.

E-mail address: jsquireinsp@gmail.com

http://dx.doi.org/10.1016/j.cancergen.2014.09.003

the treatment of advanced or recurrent disease often result in genomic amplification of the androgen receptor locus (*AR*, Xq11.2–q12), which is one of several mechanisms to overcome androgen ablation regimens that results in the development of castration-resistant prostate cancer (13).

In recent years, several prostate cancer cohort microarray studies have been published, but interpretations have been limited by the relatively small sample sizes, and rarely has there been uniformity in the methods used for data analysis between studies (5,6,14–26). Thus, the objective of this study was to perform a systematic review and selection of public domain prostate cancer genomic datasets, followed by a comprehensive meta-analysis of the pooled data of 662 tumors to derive consensus data on the common CNAs. The combined genomic data was then used to identify distinct subgroups and associated candidate pathways of prostate carcinogenesis that could be inferred from the diverse patterns of genomic imbalance. The subgroups that were the most distinct were tumors positive for *PTEN* deletions and/or *TMPRSS2-ERG* fusion status. *PTEN* deletions were significantly associated with a greater percent of the genome being altered (PGA). Predicted changes in expression levels of candidate genes that mapped to recurrent CNA regions showed that signaling networks and canonical pathways associated with cancer progression were more likely to be perturbed in the *PTEN* deleted subgroups. The large size of this meta-dataset permitted an in-depth survey and examination of concurrent losses and gains that consistently associate within tumors, suggesting that previously unrecognized relationships may exist between specific DNA changes and recurrently targeted signaling pathways.

Methods

Collecting prostate cancer public genomic datasets

High-resolution, human prostate cancer array comparative genomic hybridization (aCGH) or single nucleotide polymorphism (SNP) microarray datasets were collected from supplementary files of published manuscripts and the National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo) using "prostate cancer" with "aCGH", "copy number", or "SNP" as keyword combinations. Querying the ArrayExpress (http://www.ebi.ac.uk) database did not reveal any additional datasets. Six Agilent (Santa Clara, CA) aCGH datasets (6,18,19,22–24) and five Affymetrix (Santa Clara, CA) datasets (14,15,17,20,21) were integrated to create a prostate cancer genomic copy number meta-dataset (Table 1).

A total of 879 raw genomic microarray files were collected, including primary and advanced prostate cancers, HGPIN lesions, cell lines, and xenografts, as well as matched normal tissues (Table S1). The overall workflow for this meta-analysis is shown in Figure 1.

A total of 11 datasets were incorporated to build the meta-dataset, which is referenced by GEO accession numbers and PubMed IDs in Table 1. The platform used in each case is mentioned as well as a breakdown of the sample type. The right side of the table displays the number and type of samples that passed quality control inspection, the number and type of unique samples that were included, as well as the number of samples excluded from the analysis and the reason why. High-grade prostatic intraepithelial neoplasia (HGPIN), cell lines, and xenograft samples were not included in the analysis due to the same sample size and matched benign controls being used as baseline when available for Affymetrix data.

Reported clinical characteristics of patient tumors in the pooled study group

The final pool comprised 568 primary prostate cancer tumor samples from a total of 545 men who were diagnosed with clinically localized prostate cancer after radical prostatectomy. The primary unique tumors analyzed in this study were derived from eight published manuscripts (6,14,15,17–20,24). Further details specific to the different samples, including the patient with more than one unique tumor, can be found in the Supplemental Methods. In most cases, the tumors were staged using the 2002 TNM classification of malignant tumors (27) and graded according to the revised Gleason Grading system (28). The distribution of the Gleason Grade (available for 350 (61.6%) primary tumors) was as follows: 149 of 350 (42.6%) had Gleason Grade 5 or 6, 156 of 350 (44.6%) had Gleason Grade 7, and 37 of 350 (10.6%) had Gleason Grade 8 or higher. The 161 advanced prostate cancers (158 (98.1%) were distant metastatic lesions) included in this study were derived from 115 men obtained from five published manuscripts (6,19,21–23). Of these advanced cases, 89 (77.4%) were castration-resistant metastatic prostate cancer. Further details, as available, of the clinical characteristics of the 729 tumors included in this study are provided in Table S1.

Sample quality control, inclusion, and exclusion criteria

The raw copy number data files from 568 primary, 161 advanced, 13 HGPIN, and 120 benign control samples, as well as 17 cell line or xenograft samples, were downloaded, which amounted to 879 raw data copy number profiles (Figure 1, Tables 1 and S1). Sample exclusion criteria included: (1) corrupt raw data files (2 samples); (2) raw data of poor quality (5 samples); and, for benign samples, (3) profiles that exhibited large or prostate cancer–like CNAs believed to be potentially contaminated with adjacent tumor tissue (10 samples). A total of 563 primary, 161 advanced, 12 HGPIN, and 110 matched benign samples remained after quality control verification. Cell lines and xenografts were excluded from the analysis. To ensure the integrity of the meta-dataset, only unique samples were included in the analysis. Unique samples were defined as a single sample per patient tumor type. In cases where multiple primary, HGPIN, or metastatic foci were collected for a particular patient, incorporation of the sample harboring the greatest number of aberrations was selected (14,15,21,22). Therefore, final exclusion criteria also included one tumor focus/patient tumor type (62 samples). In only two cases (one primary and one advanced) were two samples kept for an individual patient tumor type, because each sample exhibited an aberrant yet different copy number profile (Figure S1)