BMC Genomics

BioMed Central

Database

Open Access

ChromSorter PC: A database of chromosomal regions associated with human prostate cancer

Ann Etim[1], Guohui Zhou[1], Xinyu Wen[1], Hang Liu[1], Victor Ruotti[2], Simon Twigger[2], Weihong Jin[2], Brian Matysiak[1], Jedidiah Mathis[2], Peter J Tonellato[2] and Milton W Datta*[1]

Address: [1]Departments of Pathology and Bioinformatics Program, Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, WI, 53226, U.S.A and [2]Departments of Physiology and Bioinformatics Program, Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, WI, 53226, U.S.A

Email: Ann Etim - aetim@mcw.edu; Guohui Zhou - guohui@rocketmail.com; Xinyu Wen - xwen@mcw.edu; Hang Liu - hliu@mcw.edu; Victor Ruotti - vruotti@mcw.edu; Simon Twigger - simont@mcw.edu; Weihong Jin - weihong@mcw.edu; Brian Matysiak - bmatysia@mcw.edu; Jedidiah Mathis - jmathis@mcw.edu; Peter J Tonellato - tone@mcw.edu; Milton W Datta* - mdatta@mcw.edu

* Corresponding author

Published: 28 April 2004

BMC Genomics 2004, **5**:27

Received: 30 November 2003

Accepted: 28 April 2004

This article is available from: http://www.biomedcentral.com/1471-2164/5/27

Abstract

Background: Our increasing use of genetic and genomic strategies to understand human prostate cancer means that we need access to simplified and integrated information present in the associated biomedical literature. In particular, microarray gene expression studies and associated genetic mapping studies in prostate cancer would benefit from a generalized understanding of the prior work associated with this disease. This would allow us to focus subsequent laboratory studies to genomic regions already related to prostate cancer by other scientific methods. We have developed a database of prostate cancer related chromosomal information from the existing biomedical literature. The input material was based on a broad literature search with subsequent hand annotation of information relevant to prostate cancer.

Description: The database was then analyzed for identifiable trends in the whole scale literature. We have used this database, named ChromSorter PC, to present graphical summaries of chromosomal regions associated with prostate cancer broken down by age, ethnicity and experimental method. In addition we have placed the database information on the human genome using the Generic Genome Browser tool that allows the visualization of the data with respect to user generated datasets.

Conclusions: We have used this database as an additional dataset for the filtering of genes identified through genetics and genomics studies as warranting follow-up validation studies. We would like to make this dataset publicly available for use by other groups. Using the Genome Browser allows for the graphical analysis of the associated data http://www.prostategenomics.org/datamining/chrom-sorter_pc.html. Additional material from the database can be obtained by contacting the authors (mdatta@mcw.edu).

Background

The biomedical literature is an incredibly rich resource for researchers. Information obtained from previous scientific studies helps researchers focus their own efforts. To obtain the maximal benefit from studies in genetics and genomics there is a need to link this data with the information available in the associated biomedical literature. In particular, microarray gene expression, comparative

genomic hybridization, and genetic mapping studies depend on an integrated pool of information to drive output analysis. In mining the literature to find regions previously associated with Prostate Cancer, one can define focus points for future research efforts. Subsequent analytical methods include actual placement of gene expression patterns on metabolic pathways, and the use of comparative genomic hybridization information along with genetic mapping data to determine localized genomic structure. The latter approach promises the added benefit of associating differential gene expression profiles with chromosomal structure and known genetic mapping data. However, as our knowledge base expands the ability to obtain an integrated working knowledge of these resources diminishes. The biomedical literature has been growing exponentially over the past decades. While the amount of research has increased, the ability to interpret this material becomes increasingly difficult. More and more, the papers being published are highly focused and require special expertise in the given field for a reader to appreciate the work's significance.

Scientific reviews have traditionally been a preferred source of insight into data relating to a specific research area. Articles are annotated by experts in the field, who are usually in a position to determine the significance of the latest information and can establish general trends. However, with the almost daily influx of new data, such reviews can become quickly outdated by these publications. While the Mitleman Database of Chromosomal Aberrations in Cancer http://cgap.nci.nih.gov/Chromosomes/Mitleman has collected and annotated existing biomedical literature for chromosomal aberrations, their work has primarily focused on karyotypic abnormalities to the exclusion of other experimental methods. Furthermore, it does not focus on specific diseases, such as Prostate Cancer, in our case. Software programs are being developed that systematically analyzes chromosomal changes in various tumor types; but these tools ignore the existing biomedical information and are not currently available.

We have set out to build a database of prostate cancer-related chromosomal information from the existing biomedical literature. The input material was based on a broad biomedical literature search with subsequent hand annotation of information relevant to our ongoing prostate cancer research. We have summarized this information and placed the data on the human genome using an open source Generic Genome Browser tool developed as a component of the Generic Model Organism Database. Here we present the database, named ChromSorter PC, and describe some of the associated patterns present in our review of the data.

Table 1: Common data elements and data present in the database. These numbers reflect the data present in the associated references that could be extracted after review of the entire publication

Data Element	Cells Filled	% Total
Year of Publication	785	100.0%
Format Text Available	785	100.0%
First Author	785	100.0%
Institution	784	99.9%
Material Type	593	75.5%
Citations	546	69.6%
Chromosome	780	99.4%
Chrom Arm	574	73.1%
Chrom Region	385	49.0%
Marker	172	21.9%
Gene	104	13.2%
Method	781	99.5%
Material Source	776	98.9%
Type of Gen Alt	525	66.9%
Age Dist	184	23.4%
Age Range	184	23.4%
Geography	778	99.1%
Ethnicity	132	16.8%
Fam History of Cancer	29	3.7%
Male to Male Trans	25	3.2%
Evidence of Association	783	99.7%
Calc of Association	785	100.0%
Calc of Citation	785	100.0%
Subset Totals	200	25.5%
Reference	784	99.9%
Notes	308	39.2%
Average	**513.54**	**65.4%**

Construction and content

Isolation of the associated citations

The biomedical literature present in PubMed http://www.ncbi.nlm.nih.gov/entrez/query.fcgi was searched using the EndNote bibliographic program and the terms "human" "prostate cancer" and "chromosome". The resulting downloaded list of 861 references was then manually parsed by one of the authors (M.D.) to generate a second list of references containing significant information on chromosomal regions. The specific characteristics used to triage these documents were use of some form of human materials, identification of specific chromosomes or chromosomal regions, and identification of experimental methods used in publication. References studying specific genes, but only making casual mention of the associated chromosomal position were not included in this database. These latter references are incorporated into a separate gene-based database (BEAR GeneSifter, M. Datta et. al. unpublished). The culled list was then subjected to further review of both the abstract and the full text article by two of the authors (A.E. and M.D.). The list of references used in the construction of the ChromSorter

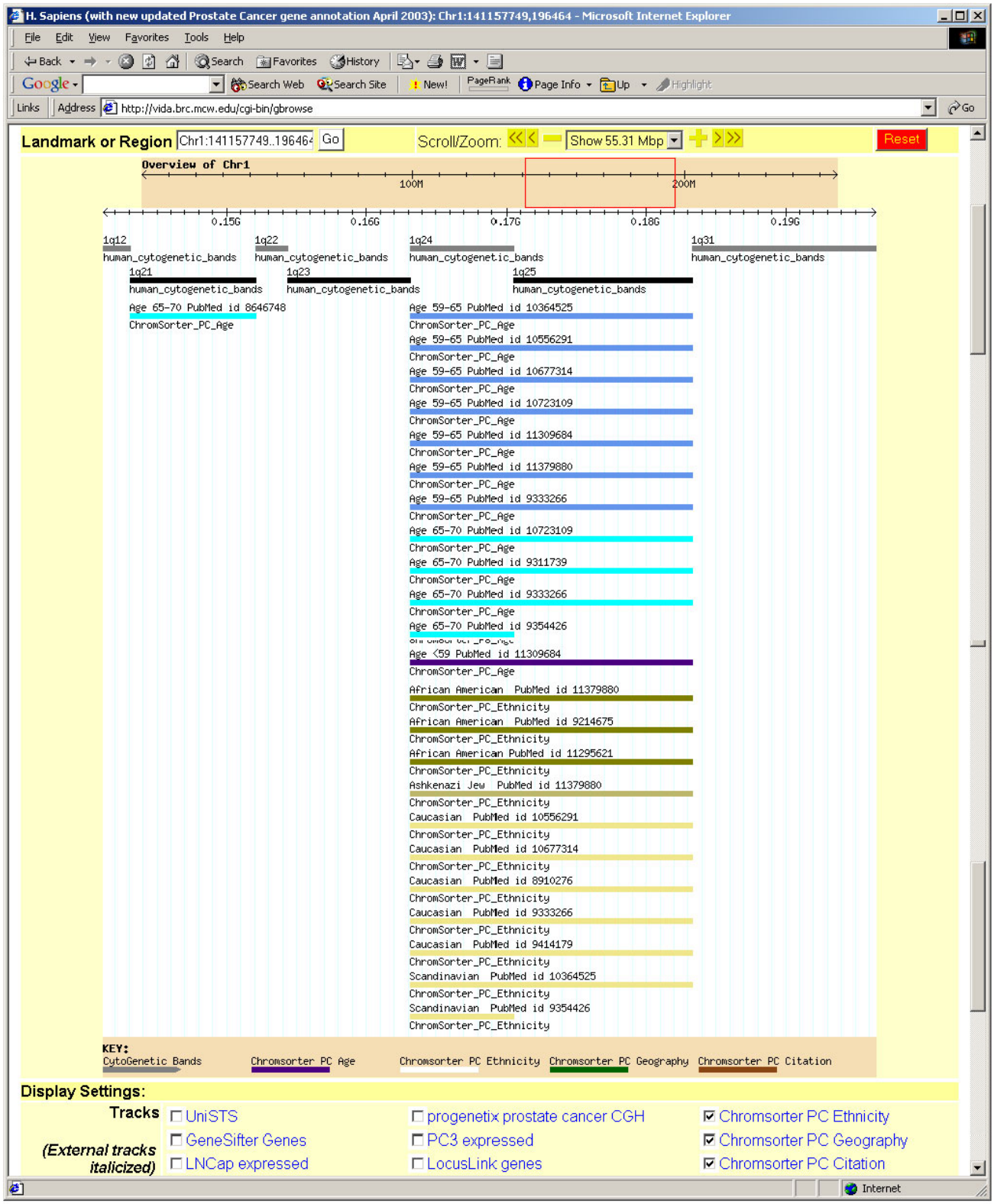

Figure 1
Visualization of the ChromSorter PC data on the human genome. Use of the Genome Browser to visualize chromosome 8 data against the human genome.

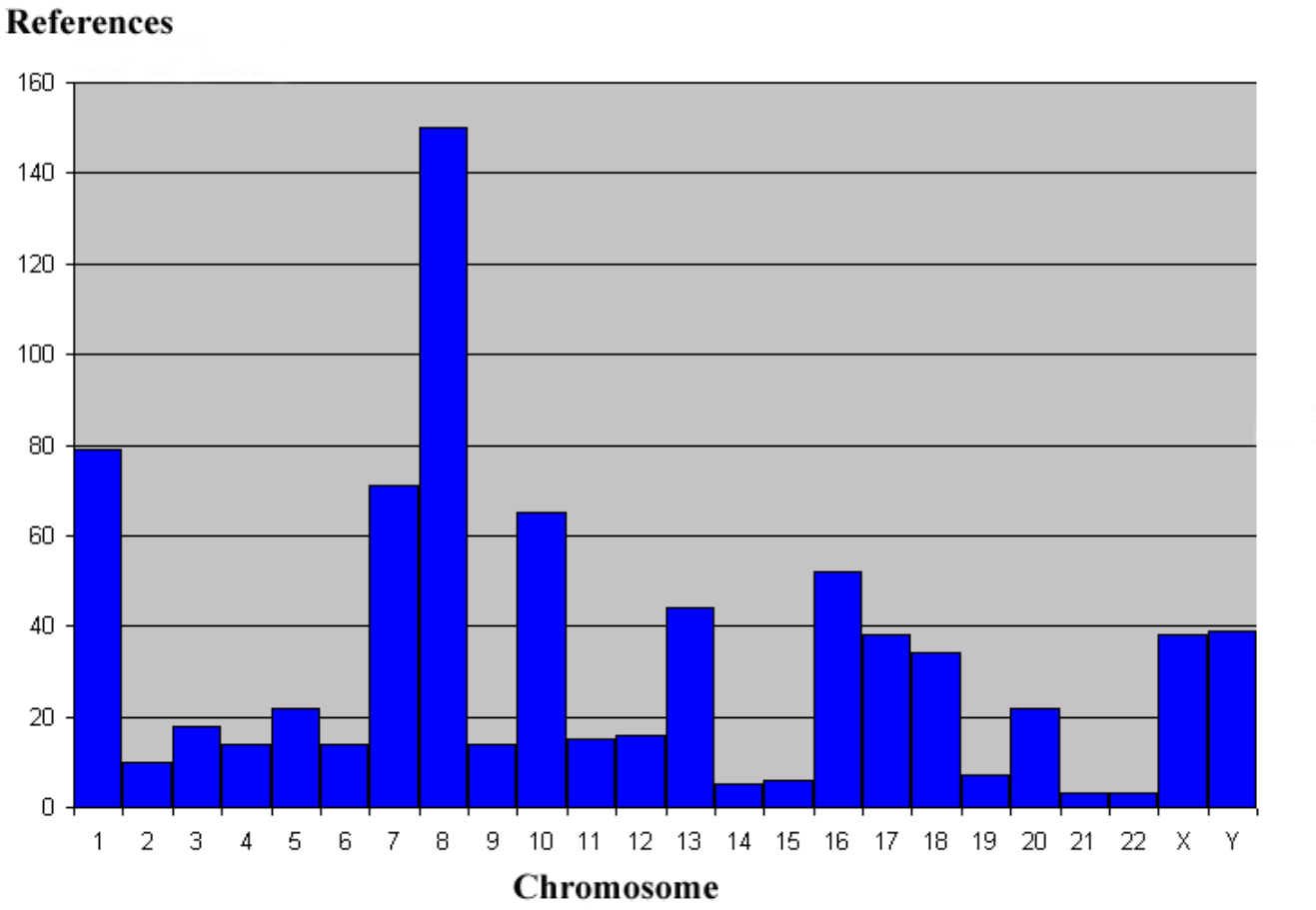

Figure 2
Chromosomal reference data. Individual chromosomes and the associated number of references in which the chromosome is implicated in prostate cancer are shown.

PC database is listed in a references file [see additional file 1].

Annotation of the citations

The abstracts or full-text articles were obtained and the data was annotated into a simple excel spreadsheet. Information provided by the abstract and full text article of each reference were catalogued across a list of 24 common data elements (see table 1). These data elements were chosen to reflect interests in chromosomal regions and focused on source of materials studied, chromosomal location, ethnicity, age, and experimental method (see table 2 for complete description of each data element). Additional data elements were added to reflect common data elements seen in the papers, such as male-to-male transmission. Others were added to facilitate graphical data analysis, such as calculation of evidence of association. Finally, several data elements were incorporated to provide information for future referencing such as First Author, Corresponding Institution, and Citations as provided by ISI's Science Citation Index http://www.isinet.com/isi/products/citation/scie/index.html listing the number of times a reference has been cited by other references. After initial annotation quality assurance of the database was performed by re-review of the entire dataset with the literature, and the data elements were checked for duplications and errors. In this manner, standardization of data entry for gene names and corresponding institutions was established for certain data elements and resulted in a defined dictionary of acceptable entry terms. In the case of the methods data element, we have developed a small glossary to categorize similar laboratory procedures.

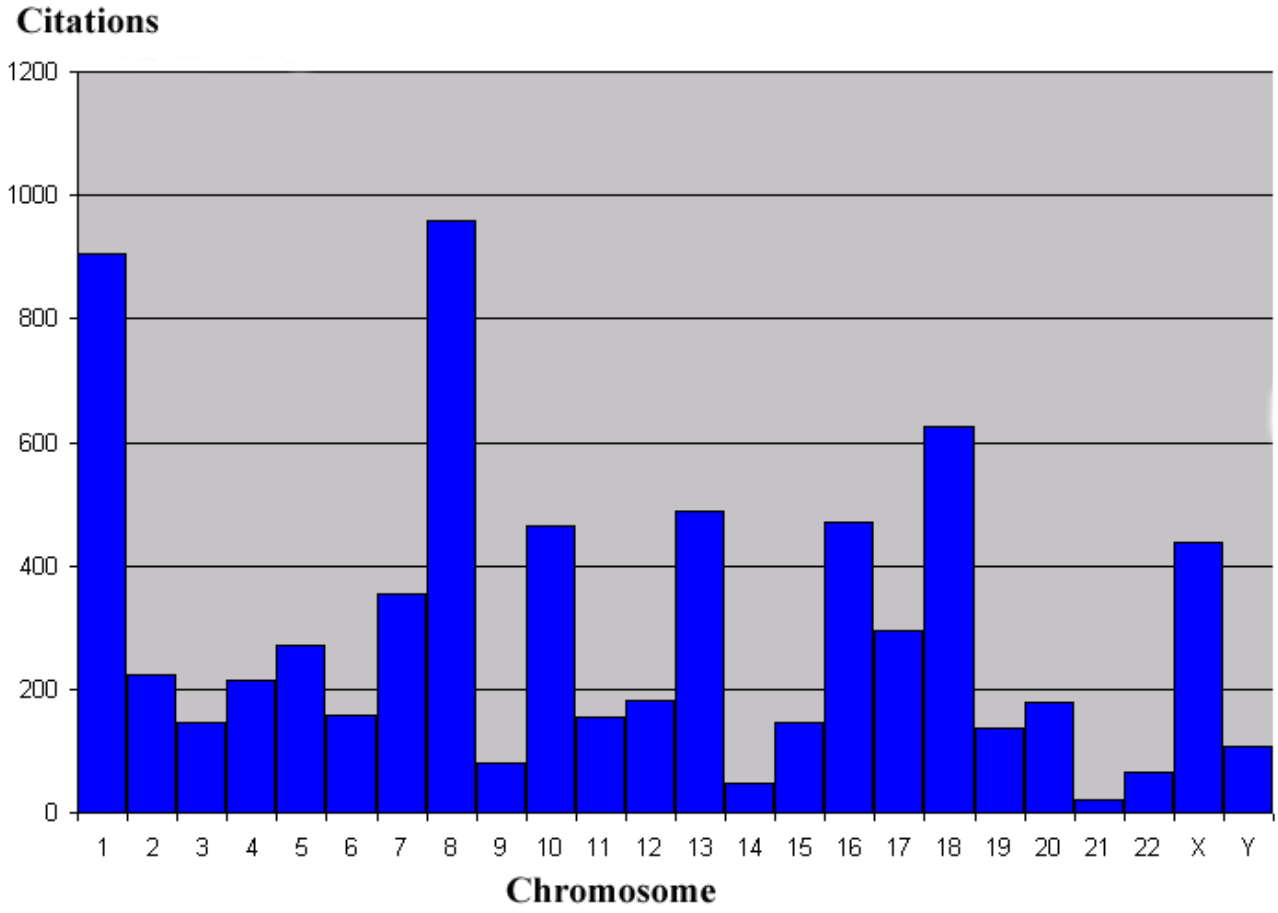

Figure 3
Chromosomal citation data. For the chromosomal data presented in figure 1, the number of citations for the individual references (1998–2001) are listed.

Database design and use

The ChromSorter PC database is currently built on a Microsoft Excel 2000 platform to simultaneously incorporate ease of use with flexibility. For each new chromosomal region, gene, experimental method, ethnic or age group studied a separate entry was made regardless of article. For example, a single article reporting prostate cancer linkage on chromosomes 1, 8 and X would have three separate entries; as would another article that only reported linkage on chromosome 1, but for three separate ethnic groups such as African Americans, European Americans and Ashkenazi Jews.

As previously mentioned, there are 24 common data elements that have been identified in the papers (table 1). A number of data element fields are related to chromosome, or genetic location. These include Chromosome, Chromosomal Arm, Chromosomal Region, Gene, and Marker. A second set of data element fields relates to experimental methodology: Method of Analysis, Type of Genetic Aberration, Evidence of Linkage, Subset Totals, Materials and Material Source. Ethnicity, Age, Family History and Male-to-Male Transmission are all data elements cataloguing demographic information present in the references; whereas Year of Publication, First Author, Corresponding Institution, Reference (PubMed hyperlink) and Citation Score catalogue information about the reference itself. The Calculation of Evidence of Linkage field allows for positive and negative values in pivot charts (described in the following section, Graphical analysis of the dataset).

We also attempted to standardize data entry of the ChromSorter PC database. Certain data element fields have a "defined dictionary", or limited vocabulary of data

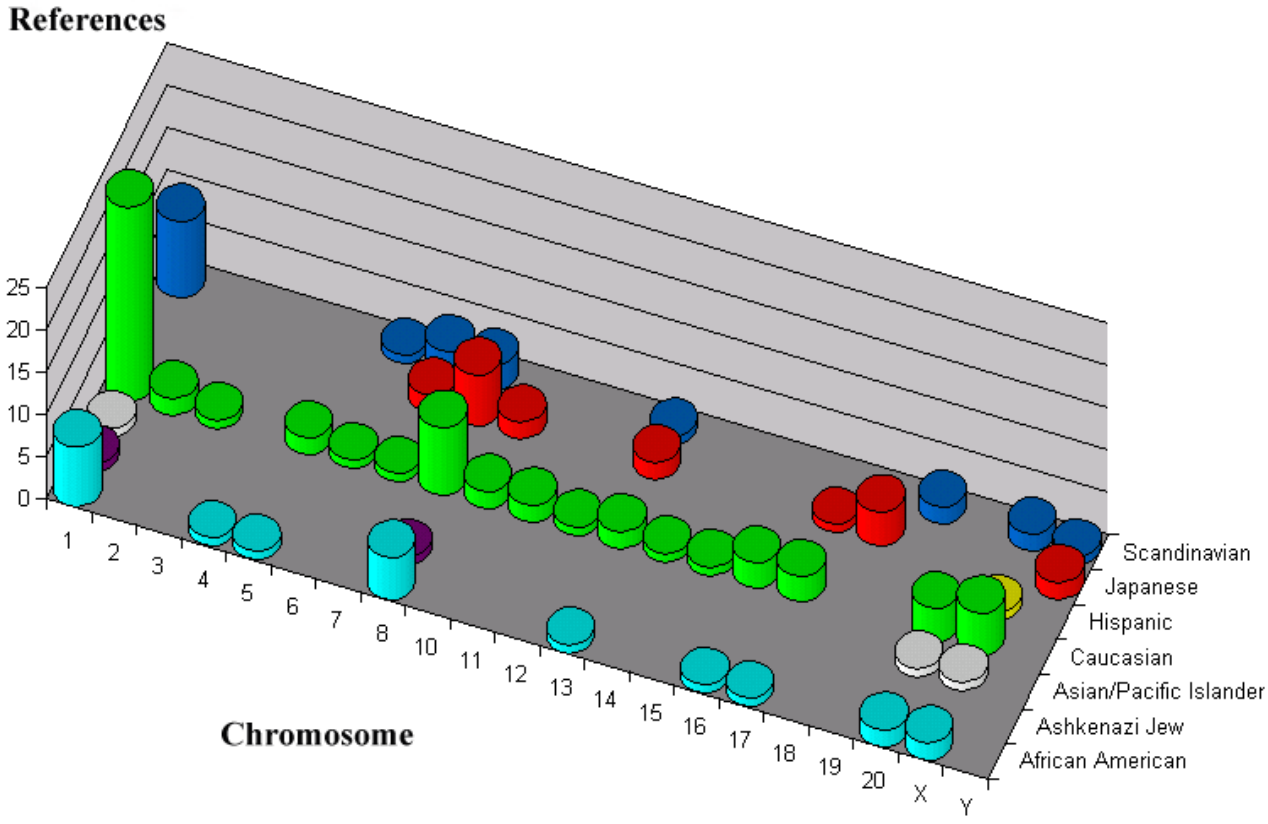

Figure 4
Chromosomal references by ethnicity. For the references that implicated a specific ethnic group chromosome is presented.

entry terms. There are only four acceptable entries in the Material Source field, for example. Only one data element field, the notes field, is open entry. Other data elements are "semi-defined", data entry fields in which only the format, not the content, of the data is specified. Detailed information is present at the help page for ChromSorter PC on http://www.prostategenomics.org/datamining/chrom-sorter_pc.html.

Graphical display of chromosomal prostate cancer data

Using the Genome Browser developed by Dr. Lincoln Stein as a component of the Generic Model Organism Database (GMOD, http://www.gmod.org) we have developed data files from the ChromSorter PC database elements for visualization of the database (figure 1) http://www.prostategenomics.org/datamining/chrom-sorter_pc.html. Data is visualized against the University of California at Santa Cruz human genome sequence data and associated annotated cytogenetic bands as defined in the UCSC Genome Browser http://genome.ucsc.edu/cgi-bin/hgGateway?db=hg12. Each ChromSorter PC data element is linked to the original references in PubMed. Various chromosomal data subsets can be visualized based on age, ethnicity, or geographical region of the publication. These regions can be defined and subsequently analyzed with respect to uploaded user data.

Utility

Identification of the references and reference characteristics

To date, data entry of references from 1998 through 2001 has been completed and is now in its second iteration. A series of charts describing the dataset are available at http://www.prostategenomics.org/datamining/chrom-sorter_pc/summaries.html. Graphical summaries of literature citations across four categories; Ethnicity, Age, Method and Chromosome are summarized in two ways: first by merely counting the number of times a region is identified, and second by adding the citation index score to determine the relative "significance" or importance of

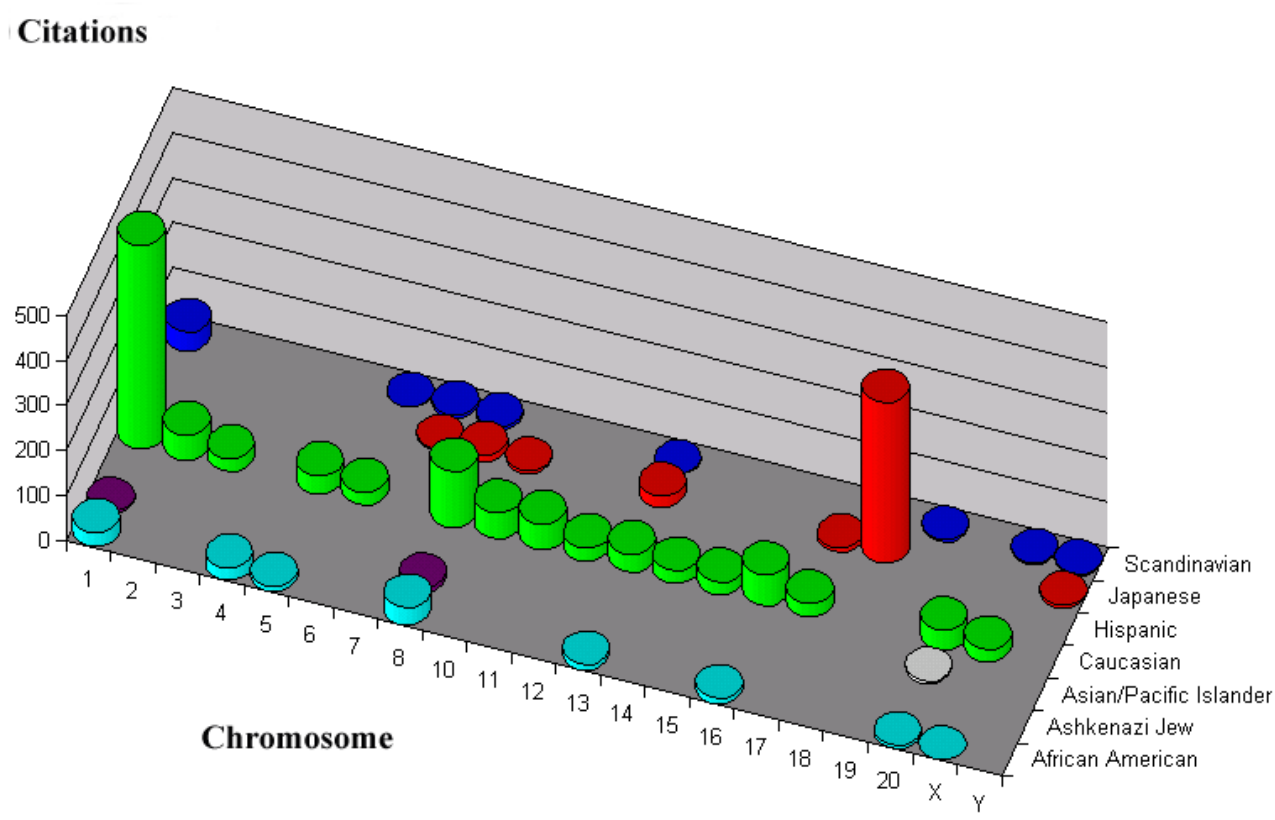

Figure 5
Chromosomal citation data by ethnicity. Data from the previous references (1998–2001) are presented and summarized for citations.

the region. To view the results on the web, the user chooses two categories from the menu and clicks the "Show Result" button. The following is a brief description of all the charts currently available online.

Identification of chromosomes implicated by multiple experimental methods

As evident from the reference count chart, chromosome 8 has the most references and citations, followed by chromosome 1 (figures 2, 3). Chromosome 7 has the 3rd highest reference count, followed by chromosomes 10, 16, 13 and Y respectively. In the citation index chart, it is the 18 chromosome that has the 3rd highest value, followed by chromosomes 13, 16, and 10 respectively.

Prostate cancer chromosomal regions based on ethnicity

Caucasians are by far the most analyzed ethnic group with respect to prostate cancer, followed by Scandinavians and African Americans (figures 4, 5). Japanese patients were the fourth most studied ethnic group, followed distantly by Ashkenazi Jews and Asian/Pacific Islanders. Results for both publications and citations mirrored each other, with Caucasians having a much more significant citation index score than reference count.

The general results are similar when the data is analyzed with respect to ethnicity, where chromosome 1 seems to have both the highest reference count and combined citation index score. In both charts, chromosome 8 is second, but the difference between first and second is much more apparent in the combined citation index score chart. In both charts, chromosomes 1 and 8 are clearly the most studied, and the remaining chromosomes have comparatively low counts or scores. Caucasians are the most common ethnic group studied, and most have an association with chromosome 1. Scandinavians, a subgroup of Caucasians, also have a higher association with chromosome 1. Ashkenazi Jews, another Caucasian subgroup, had an equal number of citations on chromosomes 1 and 8, but chromosome 8 had the highest citation index score. Afri-

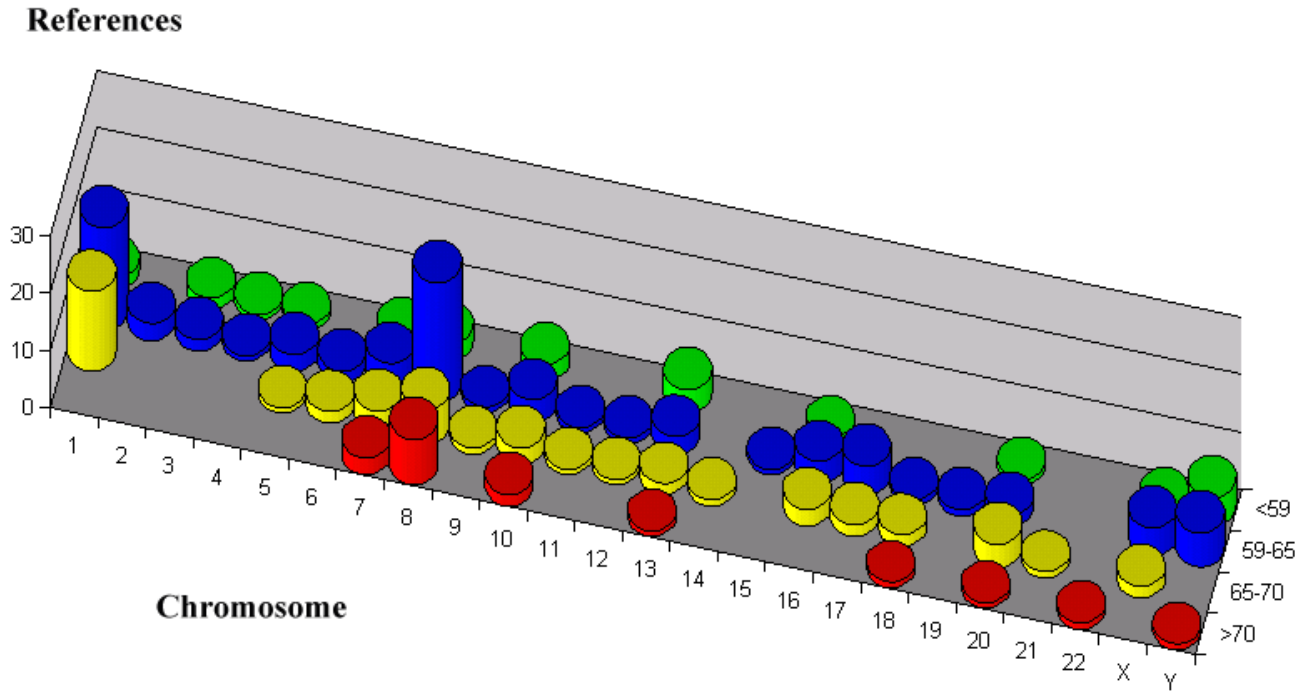

Figure 6
Chromosomal references by age. For the references that implicated a specific defined age group chromosome is presented.

can Americans are studied in references related to chromosomes 1, 4, 5, 8, 13, 16, 20 and X, but it was on chromosome 1 that this group of patients have the highest reference count and chromosome 8 with the highest citation index score. African Americans also had a relatively high combined citation index score on chromosome 5. Japanese patients also had their highest reference counts at chromosome 8, but had their highest citation index scores at chromosome 18. This group had references at chromosomes 7, 8, 9, 13, 17, 18, and Y. The Asian/Pacific Islanders only had one reference at chromosome 20 and Y. Interestingly, these were not chromosomes associated with Japanese patients in this dataset.

Age related chromosomal regions in prostate cancer

In studies of prostate cancer age is often identified as an important associated feature. For this reason we examined the various research articles for indications of age related findings, which were present in 185 referenced data entries. In each case where age demographics were supplied and associated with a specific chromosomal region these results were recorded. Because this resulted in a broad and highly variable grouping of ages we sought to group ages for ease of visualization of the data, and arbitrarily assigned samples to four age categories (> 59 years, 59–65 years, 65–70 years, > 70 years). Using these categories all of the references with age related data could be placed in a specific age category.

Chromosome 8 had the highest number of age-related references, followed closely by chromosome 1 (figure 6). Chromosome Y had the third highest number of age-related references along with the X chromosome. On the combined citation index score chart, we find that chromosome 1 has the highest age-related score, meaning the largest number of references that studied specific age groups (figure 7). This was followed by chromosome 8 and chromosome X.

Both charts demonstrate that patients around 65 years of age had the highest reference count and combined citation index score, most of which were on chromosome 1, followed by chromosome 8. Patients under 65 had the second highest reference count and combined index

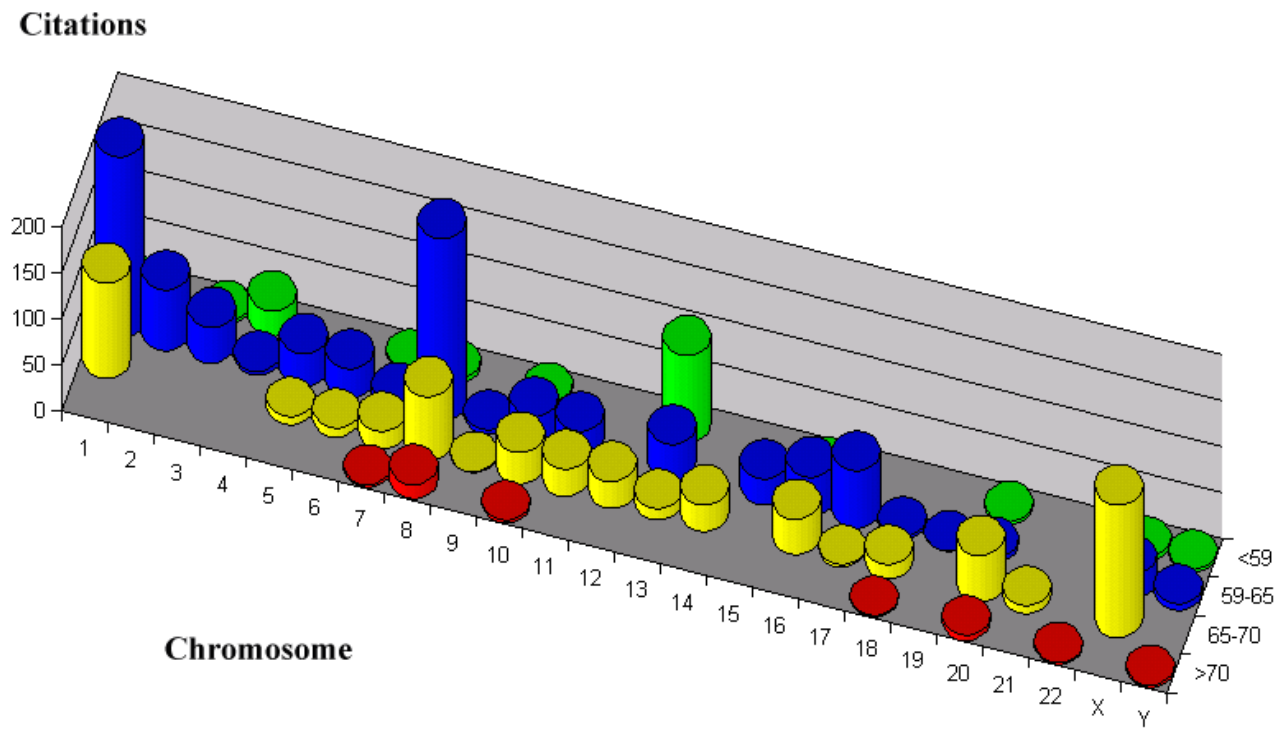

Figure 7
Chromosomal citation data by age groups. Data from the previous references (1998–2001) are presented and summarized for citations.

score. The majority of their associations were again at chromosome 8 and closely followed by chromosome 1 on both charts. Patients under 60 came in at a distant third place, with an equal reference count at chromosomes 13 and Y, closely followed by chromosome 1. Patients between 66 and 70 years of age (subgroup 65–70) had references on chromosomes 8, 20 and X. The differences on the citation index score showed that the X chromosome had the highest score, distantly followed by chromosomes 1 and 20. Patients over 70 years of age had the most references at chromosome 7 and 8, followed closely by chromosome 10. Their highest combined citation index score however was at chromosome 8, followed by chromosomes 20 and 7 respectively. Caucasian patients around 65 years of age are the most studied ethnic group in the dataset. Caucasians under 65 years of age have the second highest reference count, but only the fourth highest citation index score. Caucasians under age 60 have the second highest citation index score, followed by Scandinavians over age 70. The most studied African Americans were under 65. The most studied Japanese patients were under 72. The most studied Ashkenazi Jewish patients were over 65.

Method by reference count and combined citation index score

This chart indicates the total number of references, positive or negative, associated with the nine standardized experimental methods of analysis (figure 8). Comparative Genomic Hybridization (CGH), seems to be the favored method in our dataset. In-Situ Hybridization (ISH/FISH) is the second most popular method. Loss of Heterogeneity (LOH) methods follow closely behind with the third highest number of references. Familial mapping with the fourth highest combined citation index score. Karyotyping is fifth.

Discussion

In general, chromosomes 1 and 8 have the highest reference counts and citation index scores. Specifically for

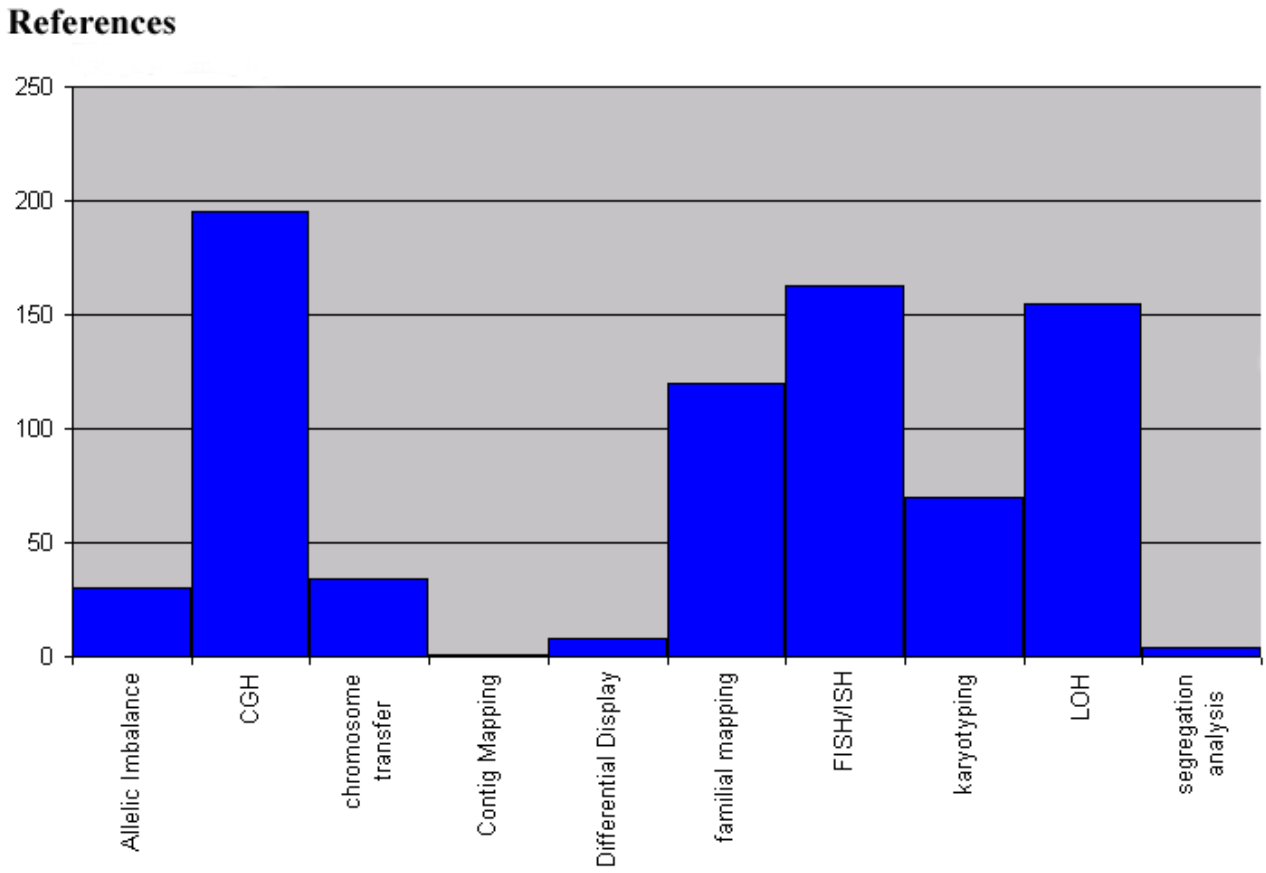

Figure 8
References using a specific experimental method. For the references that are implicated a specific experimental method is presented.

chromosome 1, the HPC1 region, 1q24-25 and the PCAP region, 1q42-43, are the most prevalent regions cited in this dataset. The most prevalent region on chromosome 8 is the PG1 region, 8p22-23. Other frequently cited loci/genes include HPC20, HPCX, RB1, CDKN1B/ETV6, SRY, PTEN/MMAC1, and BRCA1.

Comparative Genomic Hybridization is the most utilized method of analysis, followed by familial mapping. When specific ethnic or age groups are studied however, familial mapping is most frequently used. These methods look at defined chromosomal regions, demonstrating why using only karyotyping data limits the amount of useful information to be gained.

Most studies in this dataset did not provide demographic data on their subjects. Less than 30% of the citations include ethnicity and/or age related information. Nonetheless, Caucasians, including Scandinavians, had their highest reference counts and citation index scores at chromosome 1, although there were also citations around chromosomes 8, 20, X, 17 and 5. Scandinavians also had citations independent of Caucasians on chromosomes 7, 13, 19 and Y. Ashkenazi Jews, another Caucasian subgroup, had an equal amount of reference counts on chromosomes 1 and 8, but chromosome 8 had the highest citation index score. African Americans had their highest scores on chromosome 8, as well as a relatively high citation index score on chromosome 5, a chromosomal region not identified with a similar significance in Caucasians. Japanese subjects also had their highest reference counts at chromosome 8, but their highest combined citation index scores at chromosome 13. The Asian/Pacific Islander subgroup only had one citation at chromosome

20, which was, interestingly enough, not cited in Japanese subjects (Although as discussed below, relatively older patients tend to have significant chromosome 20 disease associations).

In this dataset, the most frequently studied subjects were around 65 years of age, and their disease was most often associated with chromosome 1, followed by chromosome 7. Patients under 65 had the second highest reference count and combined index score and tended to have disease association with chromosomes 1 and 8. Younger patients, (subgroup < 60) also had a disease association on chromosome 1, but also had high reference counts on chromosome 13 and Y. Older patients, those over 65, had, in addition to chromosomes 1 and 8, a significant disease association on chromosome 20. While Caucasians had subjects in all age ranges, African Americans were among the youngest patients. Japanese and Ashkenazi Jews were among the oldest patients.

Conclusions

We have used this dataset for comparison with our prostate cancer genomic data, thus filtering the lists of expressed genes for subsequent validation studies. We would like to make this dataset publicly available for use by other groups. In addition to presenting an interesting summary and evaluation of the prostate cancer chromosomal literature, this dataset also provides a highly structured hand annotated dataset that can be used for subsequent natural language processing studies focused on the extraction of chromosomal data from the biomedical literature. These ongoing studies promise to provide a method for the automated annotation of chromosomal information for various cancers, thus providing a background data for the evaluation of microarray gene expression and genetic mapping data. By comparing across this and similar datasets related to signaling pathways, animal model data, and expression patterns, one can prioritize genes identified through local studies for subsequent investigation, marker validation, and drug design.

Availability and requirements

The ChromSorter PC database is currently available as a web based tool http://www.prostategniomics.org or as a free standing csv files for use in Excel that can be downloaded and modified after contacting the author at mdatta@mcw.edu.

Authors' contributions

A.E., B.M., J.M., and M.W.D. reviewed, annotated and performed data entry, editing, and data quality control for all prostate cancer references, G.Z., X.W., H.L. implemented the Genome Browser with the assistance of W.J., V.R., and S.T. and formatted and uploaded the ChromSorter PC .gff files. This project was conceived through discussions between P.J.T. and M.W.D, who designed and coordinated its implementation. All authors read and approved the final manuscript.

Additional material

Additional File 1

A supplemental data file, references.doc, is provided as a word document, and lists in BMC bibliographic format all the references used in the ChromSorter PC database.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-5-27-S1.doc]

Acknowledgements

The authors would like to acknowledge the support of NCI grant R21CA098032 to MWD and the Milwaukee Breast Cancer Showhouse Foundation Award to MWD in support of this work.