Data Submission Request to the Integrated Canine Data Commons

1. Name/Identifier of Study:

PI: Deborah Knapp, Purdue University

Study name: Genomic characteristics and molecular subtypes in treatment-naive canine invasive urinary bladder cancer

2. Grant ID and funding source (if applicable):

P30CA023168 Supp Grant to the Purdue University Center for Cancer Research

3. IACUC/IRB approval numbers (if applicable):

Purdue Animal Care and Use Committee Approval Number: #1111000169

4. Scientific Point of Contact (Name, Phone, Email):

Deborah W. Knapp
765 494-0578
knappd@purdue.edu

5. Data Manager Point of Contact (Name, Phone, Email):

Deepika Dhawan, Research Scientist, Purdue Comparative Oncology Program
765 496-2372
ddhawan@purdue.edu

Sagar Utturkar, Senior Bioinformatician, Purdue University Center for Cancer Research
765 494-9129
sutturka@purdue.edu

6. Data access policy (choose one): Open-access – no-embargo, Controlled-access – no embargo, Open-access – embargo, Controlled-access – embargo

The data will need to be embargoed until publication is in print. We anticipate submitting the publication by March 2020.

7. Cancer type(s) included in study:

Canine urinary bladder cancer, specifically invasive urothelial carcinoma, also referred to as transitional cell carcinoma

8. Number of subjects included in study:

56 dogs with treatment-naive bladder cancer (invasive urothelial carcinoma), 4 control dogs with no evidence of bladder disease

9. Data types included in study (check all that apply): Imaging, genomics, proteomics,

immunology, clinical, other (specify)

The data include:

- pathology (reports including tumor grade)
- imaging reports
- genomics (RNA-seq, molecular subtype, immune type, BRAF mutation status)
- immunology (IHC for CD3 positive infiltrating lymphocytes, RNA-seq patterns defining tumors as immune “hot” vs “cold”)
- clinical data (breed, sex, neuter status, date of birth, method of tissue sample collection, date of diagnosis, age at diagnosis, TNM stage at diagnosis and at death, treatment summary including tumor response and PFI, date of death, age at death, survival, necropsy findings, cause of death)

Also, please note, limited WES or WGS data from these cases are expected to become available in the future. This is not part of this current submission to ICDC, but could be added to the Commons in the future if a mechanism for adding data exists.

10. Amount of data (in TB):

This will depend on which type of file that is preferred: Fastq.gz, fastq, bam?

For the Fastq.gz files, approximately 4.5 TB of memory will be needed.

11. The overall scientific benefit of including this study in the ICDC prototype:

There are 3 especially compelling pieces of scientific information that have been derived from this data set that will greatly increase the value of the naturally-occurring canine model of muscle invasive bladder cancer in humans. First, the analyses of this data set have confirmed the clear existence of molecular subtypes (luminal, basal) defined by expression patterns in RNA-seq data in the canine bladder tumors that closely mimic those in muscle invasive bladder cancer in humans. The importance of this is that there is strong emerging evidence that the molecular subtype affects the aggressiveness of the cancer and the response to specific treatments including immunotherapy in humans. Therefore, it is essential that animal models used to test new treatment strategies in humans possess these subtypes. The analyses of the data in this data set clearly demonstrate that the luminal and basal subtypes (and most likely subgroupings within those) exist in canine bladder cancer. Second, the analyses of this data set have demonstrated strong breed association for molecular subtype. This novel finding sets the stage for further work to determine heritable factors and potentially gene-environment interactions that lead to bladder cancer across species, and how specific factors can drive the cancer towards one subtype or another. Third, the analyses of this data set have revealed important links between molecular subtype and level of immune infiltration of the cancer, i.e. the assignment of immune "hot" (immune infiltrated) vs "cold" (nonimmune infiltrated) tumors. There is mounting evidence that the "hot" tumors respond better to immunotherapy and other therapies. This demonstrates that the canine tumors can serve as models for the varied local immune states in studies to test strategies to exploit the "hot" immune state, and strategies to convert "cold" tumors to "hot" tumors.

In addition, human bladder cancer data from TCGA have already been mined for the dog-human comparison in this data set, and key questions have been answered. This sets the stage for further expanding the comparative cancer value of this data set and the overall ICDC.

And, finally, this data set will be excellent for "trouble shooting" and refining the process of preparing and transferring data, and for incorporating the data into the Commons.

12. Any publications associated with this study, if any:

Manuscript to be submitted to: Molecular Cancer Research

Please note, an initial analysis of 29 tumors in this data set has been published (Dhawan, et al. PLoS Genet. 2018 Aug 8;14(8):e1007571. doi: 10.1371/journal.pgen.1007571). Some of the data from this initial analysis was submitted to the GEO data base. Our intent is to provide the full data on all 56 cases for inclusion in the ICDC.

13. Time constraints on processing/loading/releasing the data:

We anticipate being able to start transferring data by early March, 2020.

14. Data standards used, if any (e.g., SEND):

NA

15. Anticipated budget needed to prepare data set for submission:

If this data set is accepted for submission, could we discuss what the process will involve? The data set has been generated, and the clinical findings captured in an Excel file. Should extensive time be required at Purdue to transfer the data or put it in a different format, we may need to request funds to cover the staff's time for this.

The following documentation should be attached to the submission:

1. Data Dictionary:

Tissue samples of invasive urothelial carcinoma were collected in pet dogs and from normal bladder mucosa (controls) from dogs with no history of bladder disease who were euthanized for a different reason. All tumor samples included were high grade tumors and had >95% epithelial cellularity. For the normal samples, the mucosal layer was separated from the bladder wall using fine dissection. Samples were preserved at -80^{o}C in trizol and RNA extracted, purified (RNeasy kit, Qiagen, Valencia), and submitted for RNA-sequencing (Nationwide Children's Hospital, Columbus, OH). Briefly, RNA was isolated, treated with DNAse and purified. Ribosomal RNA was removed using Ribo-ZeroTM rRNA removal kit (Illumina) and libraries constructed (ScriptSeqTM v2 RNA-Seq library preparation kit, Epicentre Biotech, Madison, WI). Following purification, di-tagged cDNA were amplified by limit-cycle PCR and purified using AMPure XP System (Beckman Coulter). Paired-end 150 bp sequence reads were generated using Illumina HiSeq 4000 platform, to obtain an average of 50 million reads/sample. Raw reads were cleaned for PCR artifacts and adapter trimmed, then aligned to the canine genome (Strand NGS, Bangalore, India) using COBWeb to obtain expression levels for annotated genes and isoforms against CanFam3.1 reference genome.

Raw data were normalized independently using TMM and DESeq, filtered by read metrics, and subjected to quantification. Pairwise comparison was performed (Normal vs tumor) using edge R (for TMM normalized data) and DESeq2 (for DESeq normalized data) with Benjamini-Hochberg FDR multiple testing correction ($p < 0.05$; 2-fold or higher change) comparing normal

mucosa/urothelial layer samples. Differentially expressed genes identified using both methodologies were pooled for subsequent analyses. Unsupervised clustering was performed on initial tumor samples using a list of genes that segregate tumors into basal and luminal subtypes (n=2015). Based on the segregation of the canine tumors in this unsupervised manner, a class prediction model was developed which was applied to the larger dataset (n=56) to assign basal and luminal subtypes.

In addition, the data were also interrogated using a list of genes that are enriched in T-cell inflamed bladder tumors in humans (n=1984). This allowed us to assess the immune state (immune “hot” or “cold” or mixed) for each sample and how that related to subtype.

Histopathology (H&E) slides were read by one pathologist to assign tumor grade.

CD3 immunohistochemistry was used to visualize tumor infiltrating lymphocytes in the tumors, and this data studied in the context of the RNA-seq findings.

Clinical data were assembled in an Excel spreadsheet. The data include: breed, sex, neuter status, date of birth, method of tissue sample collection, date of diagnosis, age at diagnosis, TNM stage at diagnosis and at death, treatment summary including tumor response and PFI, date of death, age at death, survival, necropsy findings, and cause of death.

2. Data Model/Schema diagram indicating how collected data relates to subjects,

visits, samples, etc:

The samples were collected from dogs presenting for evaluation of bladder cancer in the Canine Bladder Cancer Clinic in the Purdue University Veterinary Teaching Hospital. Cases were included for which ample tissues were available and owners provided consent for the inclusion of the dog in the study. The dates of participation (diagnosis to death) are included in the Excel spreadsheet.

Each dog was given a unique identifier number linked (in a secure location at Purdue University) to the dog/owner’s name and contact information.