National Cancer Institute (NCI)
Integrated Canine Data Commons (ICDC) Steering Committee (SC)

Teleconference
Wednesday, June 17, 2020

Participants (*Present)
External Committee Members
Matthew Breen*
Renee Chambers*
Dawn Duval*
Heather Gardner*
Allison Heath*
Will Hendricks*
Warren Kibbe*
Debbie Knapp, ICDC-SC Chair*
Cheryl London*
Jeff Trent*
Roel Verhaak*
Shaying Zhao*

Internal Committee Members (NCI, National Institutes of Health, and Frederick National Laboratory for Cancer Research [FNLCR])
Matthew Beyers*
Allen Dearry
Toby Hecht*
Paula Jacobs*
Tony Kerlavage*
Erika Kim*
Amy LeBlanc*
Christina Mazcko*
Philip Musk*
Elaine Ostrander*
John Otridge*
Ralph Parchment, ICDC-SC Managing Secretary*
Connie Sommers*
Greg Tawa*

Others
Anju Singh*
Lori Lydard*
Tara Whipp
Mary Cerny (writer)*

Opening and Welcome

Drs. Knapp and Parchment opened the meeting at 12:05 p.m. ET and welcomed attendees.

Minutes of the May 2020 Meeting

The minutes of the May 20, 2020, ICDC-SC meeting were accepted as written.

Reports from the Working Groups (WGs)

Data Governance Advisory Board (DGAB) Chair's Report

As discussed during the May ICDC-SC meeting, the DGAB realized during the submission review process that the ICDC submission and intake form needed to be revised to ask all submitters whether they are getting any specimens or data from a repository/biobank and/or from more than one institution and, if so, to provide that information. The proposed changes to the intake form were approved by the scientific advisory group, and the form and submission process have been updated to include this additional request for information. The DGAB is also going back to query prior submissions about this issue. Having these details will help facilitate cross-referencing animals, data, and specimens across repositories and research organizations within the ICDC.

As part of the submission review and prioritization process, the DGAB recommended prioritizing the multi-omics osteosarcoma submission over the transcriptomic mammary cancer submission due to the much smaller sample size of the latter dataset. The NCI Executive Team has approved the DGAB's prioritization recommendation.

Best Practices Subcommittee (BPS) Chair's Update

Dr. Trent introduced and welcomed Dr. Gardner to the ICDC-SC. Dr. Gardner officially joined the BPS as an *ex officio* member as of the June BPS meeting. BPS working group (WG) updates included the following.

Genomics WG

Drs. Duvall, Hendricks, and Zhao will finalize the best practices document for genomics analysis as a one-page document for investigators. Drs. Gardner, Heath, and Hendricks will take the lead on drafting a two-page guidance on creating harmonized and validated genomic datasets with input from multiple stakeholders and potential cloud assessment pipelines to synergize with other similar efforts that, in turn, could be leveraged by the ICDC. The Genomics WG hopes to have draft versions of the two documents available for review at the next ICDC-SC meeting.

Imaging WG

The Imaging WG continues to reach out to researchers who have MRI files for the canine glioma dataset. Some of the feedback has slowed as a result of logistical and other operational changes at academic institutions because of the COVID-19 situation.

Other BPS WG Updates

The other WGs continue to focus on the issues discussed at previous meetings. No additional updates on the other WGs were provided during the teleconference.

Presentation: Breed-Specific Reference Genomic Data

Dr. Ostrander highlighted the work of her lab and labs and consortia around the world that are focused on cataloging and characterizing genetic variation across specific populations of dogs.

The Dog 10,000 Genome Consortium (Dog 10K) is an international collaboration of researchers established to address major research questions regarding the genetic underpinnings of domestication, breed formation, aging, behavior, and morphologic variation and to advance understanding of human and canine health. The consortium includes participants at 18 institutions from nine countries. Founding sponsors include Dr. Ya-Ping Zhang at the Kunming Institute of Zoology in China, Dr. Robert Wayne at UCLA, and Dr. Ostrander.

The primary goals of this collaborative endeavor are to generate whole genome sequences (WGS) from 10,000 canine/canids, including both modern and ancient dog populations, in 5 years; refine the existing reference genome from a boxer; create new reference genomes from additional canids; and apply the data to a broad range of scientific questions. The resulting catalog will contain comprehensive high-density genomic data, including single nucleotide variants (SNVs); single–nucleotide polymorphisms (SNPs); indels; structural variants (SVs), including copy number variations (CNVs); and mobile element insertions (MEIs). Data will be used to estimate mutation rates for SNVs, CNVs, and MEIs to test whether mutational events cluster along the genome and improve understanding of mutational processes. Ultimately, the Dog 10K catalog will provide a comprehensive catalog of rates of gene gain/loss in distinct canid populations and individual dog breeds relative to the canine phylogeny.

The Dog 10K initiative aims to sequence each genome to at least 20× coverage using *de novo* assembly, which results in fewer gaps and errors. "*De novo* genome assembly will be undertaken for both breed dogs and wild canids using a variety of technologies, including Pacific Biosciences long-reads (PacBio) (100×), bacterial artificial chromosome (BAC)-end sequencing, optical mapping (Bionano Saphyr), phased haplotypes (60x by the 10x genomics platform), and chromosome conformation (Hi-C)" (Wang et al., *Natl Sci Rev* 2019).

Data from the approximately 450 breeds recognized worldwide as well as from non-breed and geographically isolated populations that meet the statistical definitions of breeds (e.g., the Patagonian sheep dog) are being captured by this initiative. Good reference populations are being collected through the consortium from partners in Europe and Asia. To date, the project has cataloged approximately 3,000 WGS.

The Dog 10K program has placed a high priority on disease mapping resources. For breed dogs, Dog 10K will ideally be collecting samples from aged healthy dogs to benefit genetic disease studies. Dr. Ostrander pointed out, however, that many public databases lack information on disease status, necessitating the duplication of already sequenced breeds as a reference for disease mapping studies. Variants in breeds not at increased risk for a particular disease but at high frequency across breeds are unlikely to be disease–associated. Variants in at-risk breeds that are otherwise rare are candidates for human and dog studies. The challenge in these cases involves setting thresholds, given that both neutral and disease alleles of varying types are nearing fixation in a variety of breeds. This is especially difficult in breeds with a high incidence of disease.

Dr. Ostrander's team used WGS, SNP chips, and pedigree structure to determine the number of dogs to sequence per breed to capture the greatest amount of information to optimize genetic-trait mapping. The team looked at homozygosity measures for a large number of dogs and breeds. Analysis of decay rates, measured as a function of total length of homozygosity (in megabases (Mb)), showed that for the majority of breeds, four individual dogs per breed (two male, two female) are needed to capture the most information from that breed (adapted from Dreger et al., 2016). For non-breed and geographically isolated populations, the goal is to sequence three dogs per population. Further analysis using genome-wide association studies (GWAS) and selective sweep analyses identified variants of strong impact associated with 16 phenotypes (e.g., ear morphology, body size, coat color, longevity) in domestic breeds.

Dr. Ostrander and her colleagues have put together a catalog of 722 WGS from a sample of 538 domestic dogs (144 breeds), 54 wild canids, and 104 village dogs. This dataset, which is separate from Dog 10K, is available and ready to use for mapping and includes approximately 91 million variants (Plassais et al., *Nat Comm* 2019). About 1% of the variants are in exons. The rest are intergenic (51%), introns (31%), downstream (7%), and upstream (6%), while splice sites, UTR 3 prime, and UTR 5 prime constitute the remaining variants. A smaller catalog of domestic dogs was created, and removing the wild canids and village dogs and keeping (a) only higher-quality genomes greater than 10x, (b) no more than two males and two females per breed, and (c) only biallelic variants (SNPs and indels) yielded approximately 78 million variants in 287 dogs. Use of GWAS and genome-wide efficient mixed-model association algorithm (GEMMA) on this catalog has been successful in identifying regulatory variants, enhancers, microRNAs, and other coding sequences because testing is done on such a large pool of variants, compared with, for example, use of SNP chips, which only get within Mb of the loci. These catalogs are public and available to everyone (links below).

The original reference genome for the dog, from a female boxer named Tasha, was published in 2005. Despite methodological advances and improvements, assembly is often still based on many short reads that use repeats and, in turn, introduce a lot of errors. To address this shortcoming, many individual groups in addition to Dog 10K are doing new *de novo* assemblies that independently align the genome instead of sequencing and aligning the genome based on the reference genome. Key to this approach has been the use of long read sequence technologies, which eliminate many of the errors and gaps when short read assembly methods are used. *De novo* assembly of the original boxer genome using long read sequencing is complete and generated an improved reference genome by gap filling and sequence error correction. The new Tasha genome assembly has been uploaded to National Center for Biotechnology Information (NCBI); public release of the genome is pending finalization of details such as nomenclature by NCBI.

Other newly available *de novo* assemblies in dogs include two German shepherds, a Labrador retriever, a dingo, a Great Dane, and a Basenji. One complete wolf assembly is also available, as are partial assemblies in a Tasmanian wolf and a gray wolf from China. Dr. Ostrander noted that there are at least 10 additional ongoing *de novo* assembly efforts within the research community. One of the next steps is to generate a single *de novo* assembled reference pan-genome for each of the following, many of which are being developed outside the United States: gray wolf, golden wolf, coyote, African village dog, South Chinese village dog, Siberian or Alaskan dog, and golden retriever. Dr. Ostrander and her team are also working on creating a singular high-quality pan-genome reference sequence for each breed in their lab.

Additional information about Dog 10K can be found here. Other resources include:

- The iDog database, an integrated resource for domestic dogs and wild canids that aims at providing a public warehouse to release, annotate, and update whole genome sequences on dogs.
- The Dog Genome SNP Database (DoGSD) is a data container for variation information of dog/wolf genomes that is also designed and constructed as an SNPs detector and visualization site.
- The NHGRI Dog Genome Project is run by Dr. Ostrander's laboratory and focuses on the genetics of health and body structure in the domestic dog.
- The Canid sequence dataset includes 722 genomes sequenced via whole genome sequencing from various wild canids, dingo, and domesticated dogs.
- The Sequence Read Archive (SRA) database, the largest publicly available repository of high throughput sequencing data, is available through multiple cloud providers and NCBI servers.
- The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.
- The Dog Biomedical Variant Database Consortium (DBVDC) is a comprehensive list of functionally annotated genome variants identified with whole genome sequencing of 582 dogs from 126 breeds and eight wolves (described here).

Dr. Ostrander's presentation will be uploaded to the ICDC-SC Box folder. Questions for Dr. Ostrander should be forwarded to Mr. Beyers, who will compile the queries and comments and send them to Dr. Ostrander.

Discussion/Q&A

Committee members agreed that the Dog10K consortium and other datasets and catalogs are tremendous resources.

In response to a question about annotation of these catalogs for disease demographics, Dr. Ostrander said that all available health information (e.g., disease/non-disease status, age, sex, body measurements, tumor location, time/age of blood draws) is provided. The Dog 10K project is also collecting health updates on disease status, co-morbidities, and cause of death. Some earlier samples may not have complete demographics, however. Although resources for these catalogs are available, level of funding can impact the amount of clinical and demographic data obtained.

Some reports indicate 100% homology for expression of immune response genes (e.g., *PD-1*, *PD-L1*) across breeds. A Committee member asked if any of the datasets or catalogs support these reports. Dr. Ostrander noted that this is not her area of expertise and did not know if these analyses have been done. However, she pointed out that 1,575 WGS are available for free and that the sequences can be searched by breed and subsets of breeds to answer this question. In addition, many researchers in the dog genome community are interested are interested in expression and are building expression datasets (germline, tissue, tumor), often with a focus on cancer.

Another question was how investigators who focus on somatics/cancer genomics could use these resources to build workflows and pipelines for breed-specific and/or pan-genomic alignments. Dr. Ostrander said that the gold standard for human cancer genetics is to sequence normal and tumor tissue from different locations (e.g., in the primary tumor, recurrence, metastases) in the same individual.

Dr. Ostrander's connection was lost at this point during the discussion/Q&A.

Other questions were how the sequences in these datasets and catalogs can be accessed, how resources cited during Dr. Ostrander's presentation can be linked and integrated into the ICDC, and what power and cohort sizes are needed for somatic cancer landscapes based on current knowledge.

Regarding linking these datasets and catalogs to the ICDC, one option might be through the Seven Bridges Platform, which is used with the NCI's Genomic Data Commons (GDC).

Additional questions and comments should be forwarded to Mr. Beyers, who will compile and send them to Dr. Ostrander.

Other Issues

Revisiting Topics Discussed in Prior ICDC-SC Meetings

The ICDC-SC has met 15 times since its inaugural meeting in February 2019. The ICDC-SC's activities and discussions over this time have focused on building the canine commons prototype, developing the framework for data submissions, establishing best practices, and reviewing initial submissions to the ICDC. Now that many of these foundational tasks are complete, or nearly so, it was suggested that the Committee go back and review past minutes to identify ideas and topics that were set aside during the early stages of the ICDC. The following issues were raised during the current meeting:

- Mapping canine diseases to human diseases, which is one of the goals of the ICDC.
  - A lot is known about differences between dogs and humans for certain cancers, such as B cell lymphoma, but investigations comparing disease status, progression, and outcomes across species are limited. Clinical trials that test the same treatment/agent in humans and dogs should reveal differences in responses between species. However, results depend on factors such as specific questions being asked about the disease, the treatment regimen, and the type of response being investigated. This field of research is evolving but is not quite at the point of informing definitive conclusions.
  - For any dog model of cancer, comparative studies and analyses should look at specific common features or targets instead of expecting all aspects of disease to be the same for humans and dogs. For example, the driver of many bladder cancers in dogs, but not humans, is *b-ras* mutation. A test that detects *b-ras* in dog urine might serve as a platform for a diagnostic test for a different gene. Thus, the technology can be developed and adapted to the research questions and the specific needs of the disease.
  - Another task involves identifying subsets of characteristics for canine cancers that can then be compared to human disease. Although genotypic divergence between species is seen for some cancers (e.g., bladder cancer, osteosarcoma), other cancers have similar genotypic characteristics (e.g., gliomas). Data also support phenotypic convergence at the transcriptional, proteomic, pathology, and disease behavior levels.

Being able to translate genotypic divergence and phenotypic convergence data into harmonized datasets for the community will help answer questions more rigorously than looking at siloed studies and data.
    - Using imaging to identify and characterize disease subsets.
    - Integrating canine characteristics into the larger NCI GDC. Creating algorithms to query the data commons by disease characteristics, structural or copy variance, transcription pathways, mutation landscape (instead of histopathology), or specific key search terms.

Drs. Hecht and Parchment and Mr. Beyers will go through past minutes to identify other topics for further discussion at future ICDC-SC meetings.

Administrative Items

July ICDC-SC Meeting
The next meeting of the ICDC-SC will be held via teleconference on Wednesday, July 29, 2020, from 11:30 a.m. to 1:30 p.m. ET. Dr. Parchment will forward the meeting information and materials ahead of the July teleconference.

DGAB and BPS Meetings
Dates and times for other upcoming meetings are still to be determined. Details will be distributed as they become available.

Honoraria
External members should continue to forward the paperwork for their honoraria to Ms. Lydard. Anyone having problems with the form or reimbursement should contact Dr. Parchment.

Action Items

- Mr. Beyers will upload Dr. Ostrander's slide presentation to the ICDC-SC Box folder.
- Committee members should send questions about Dr. Ostrander's presentation and research to Mr. Beyers will compile and forward them to Dr. Ostrander.
- Drs. Hecht and Parchment and Mr. Beyers will review past minutes to identify topics for discussion at the next ICDC-SC meeting.
- Topics for future meetings should be forwarded to Drs. Knapp and Parchment.
- Dr. Parchment will forward the logistics information and materials for the next teleconference as they become available.

Adjournment

The meeting was adjourned at 1:03 p.m. ET.