Transcript

KATE: Let's get started. Welcome to today's webinar Revolutionizing Biological Research with the NIH Comparative Genomics Resource. Our presenter for today Dr. Valerie Schneider has been at the National Center for Biotechnology Information (NCBI) at the National Library of Medicine since 2007 and is currently the Acting Chief of the Information Engineering Branch. In this newly assumed role, she overseas NCBI's collection, creation, analysis, organization, curation, and dissemination of data and analysis tools in the areas of molecular biology and genetics, as well as the collection and management of bibliographic information. Previously, Dr. Schneider served as the head of the Sequence Plus Program and the Deputy Director of Sequence, where she coordinated efforts in curation, enhancement, and organization of sequence data and managed tools and resources that enable the public to access, visualize, and analyze biomedical data. In 2020, Dr. Schneider inaugurated development of the NIH Comparative Genomics Resource (CGR), which you'll hear about today. CGR offers new possibilities for scientific advancement by facilitating reliable comparative genomics analysis for all eukaryotic organisms through community collaboration and an NCBI toolkit of genomic resources. Dr. Schneider earned a PHD in Biological and Biomedical Sciences from Harvard University in 2001, followed by a postdoctoral fellowship at the University of Pennsylvania. In her former life as a wet lab biologist, she studied Tetrahymena thermophila, Drosophila melanogaster, Xenopus laevis, chicken, zebrafish, and other research organisms to answer questions relevant to human development. Thank you, Dr. Schneider, for joining us today. I will pass the virtual microphone to you.

DR. SCHNEIDER: Great. Thanks so much, Kate. I'm going to go ahead and start sharing my screen and get my slides up here. And hopefully everybody is now seeing some slides. All right. With that, I'd like to go ahead and get started and thank Kate for the invitation and opportunity to speak with all of you today. As she said, I'm Valerie Schneider and I'm the acting Chief of the Information Engineering Branch at NCBI, which really designs and builds NCBI's production software. And this afternoon, I'll be introducing you to the NIH Comparative Genomics Resource, which is new work that's happening at NCBI in the National Library of Medicine to support comparative genomics for eukaryotic organisms.

And as you listen to this presentation today, I hope that you'll consider the questions that I've got on this slide here. And that's because community collaboration is really critical to CGR's success. It's my hope that the questions here will guide your mind towards that and that you, your colleagues and your clients will subsequently reach out to share your feedback and your thoughts in these areas because they will inform improvements that are made to existing NCBI data, tools and interfaces, as well as new developments in these areas.

I'm going to be covering the following items associated with CGR today. I'll start by giving a brief introduction to comparative genomics, talk about the value of research organisms and some of the challenges facing comparative genomics researchers to really prevent the full realization of this value, the solution that CGR offers, and at a very high level how CGR can impact comparative genomics. And I'll relate this impact to a couple of use cases that introduce you to some of the data resources and tools that are being newly developed or improved through CGR to create an NCBI toolkit for comparative genomics. And then I'll be wrapping up with some information on what's coming next for CGR and also some information on how you and your clients can get more involved with it.

So let's start with the question of what is comparative genomics? And comparative genomics is really the comparison of genetic information both within and across research organisms to understand the evolution structure and function of genes, proteins and non-coding genomic regions. And notably, this approach can inform nearly any area of biological study. That's because all biological processes rely on genes, proteins, and the biochemical pathways that they participate in that are really shared across the tree of life due to evolution. And so, for the purposes of this presentation, I'm also going to be defining research organisms as the broad collection of all eukaryotic taxa on which research is done. And as a reminder, eukaryotes could be thought of as all organisms that are not bacteria or viruses. So comparative analysis can be leveraged to systematically explore and evaluate the biological relationships and evolution between species, aid in understanding the structure and function of genes, and to gain a better understanding of disease and potential drug targets. And comparing genomes is really essential to understanding species specific adaptations and how those adaptations contributed to evolutionary success. But comparative genomics can also help us identify emerging model organisms among a broader span of the tree of life, which then positively impacts human health. And so, as more data becomes available and technology advances to permit more thorough analysis, comparative genomics findings in distantly related species in addition to closely related ones, can really be extrapolated again to impact human health.

So you know who does comparative genomics? Well it turns out many people use a comparative genomics approach to the research, even if they don't call it by name. A translational researcher might be using this approach to study hereditary disorders, do zoonotic disease research, therapeutics development, or even microbiome research, while an organismal biologist might be using it to gain greater insight into a biological process that they're interested in. And many evolutionary and developmental biologists are members of various large consortia that are working to sequence all organisms on the planet. On the other hand, a bioinformatics core scientist might be supporting translational researchers in their work or be affiliated with the data management team for one of these large sequencing consortia, while a hard-core computational biologist might be using comparative genomics data as a test set for an algorithm that they're developing. And comparative techniques are commonly taught in biology classes. So basically what you can think of is that most types of biological research can make use of comparative genomics approaches in one or more ways.

But you know, why is it important to use research organisms and do comparative genomics? And I'll just say that the value of research organisms is really twofold. First, they help us understand basic biological processes, but they additionally help us understand human disease. And as I've got illustrated on this slide here, comparative analysis have been used to study areas as diverse as host pathogen interactions, including those involving the SARS-CoV-2 virus, cancer, xenotransplantation, and it's also informed our understanding of biochemical pathways, vision, metabolism and aging, and even supported areas like pharmacogenomics and drug development. But today, we're really poised to make these bold new strides through comparative genomics as advancements in sequencing technology and algorithmic methods are democratizing the generation and the analysis of genomics data, which is resulting in this explosion in the number of organisms that are used for research and this rapidly expanding collection of sequence genomes.

But while these research organisms in comparative genomics offer great promise, the current landscape of individualized data resources for the greater universe of these organisms presents investigators for several limitations and challenges when it comes to using the data, and one big one is the data itself. We have exponential growth in the number of genomes and the corresponding taxa submitted to GenBank, and that data varies in quality, impacting its suitability for use in comparative genomics analysis. Other data related issues that we see include such things as there being multiple different user interfaces, having only a limited number of organisms that have corresponding robust data resources, the data in the applications themselves may be siloed, and users typically have to download data to apply tools. And so there's really just a limited scalability in this approach given the data and the number of organisms.

So considering all of those challenges, in 2020 NIH initiated the comparative genomics resource to maximize the impact of eukaryotic research organisms and their genomic data resources to biomedical research. As you've heard, NCBI is charged with leading CGR development and maintenance and CGR facilitates reliable comparative genomics analysis through collaboration with the genomics community and also with an NCBI genomics toolkit, which includes high quality data, tools and interfaces that are used to connect community provided resources with NCBI. And as we're proceeding with CGR, we're keeping in mind the very broad and diverse base of users who encounter comparative genomics in their work and we really value feedback from all of them. So this image over here on the right represents how CGR works. In the middle is NCBI with that toolkit of data, tools, and interfaces that can be used for comparative genomics. And in the outer ring are community provided resources that with community collaboration can be brought into this NCBI ecosystem through standardized interfaces that are illustrated by these arrows. And I'll just say that having an organism agnostic central connection point with core content that can also offer connectivity to community provided content has the potential to bring together data from this expanding universe of research organisms in a way that we haven't seen before. So in short, CGR facilitates reliable comparative genomics analysis for all eukaryotic organisms through again, community collaboration and NCBI toolkit of interoperable resources.

But I also want to mention to you as librarians that CGR can be integrated as part of our librarian's training program to help users stay current with the enormous amount of data generated by sequencing projects. And CGR also provides access to a biological knowledge repository and associated scientific publications that enable both experienced and novice researchers to examine, visualize and analyze the underlying data for themselves. And having a strong understanding of CGR's many facets and applications will allow you as to demonstrate to new users how CGR can act as a discovery hub and amplify research outcomes.

The CGR achieved its mission through three key components that I'm going to go over on this slide here. The first is the NCBI Toolkit which you've heard me mention. This NCBI Toolkit offers high quality genomics related resources that are impactful to users. Research in this image here over on the right we can take a look at how the three components of CGR work together. And listed in the inner circle are some of NCBI's data and tools that can be used for comparative genomics including these interconnected databases that have access points that enable seamless navigation of NCBI content, as well as these interoperable tools and data that can integrate into users' workflows. And so the Toolkit here allows users to compare genomic sequences, to explore NCBI sequence data, to visualize data, and also to improve data quality before submission. And this list here just shows you some of the NCBI resources that are under development or improvement as part of CGR that really comprised this toolkit.

Now, the second CGR component is community collaboration. And as I've mentioned, community collaboration is really critical to CGR success. It identifies opportunities to connect more genome related data and metadata with the Toolkit and additionally feedback from the genomics community will inform new developments and improvements to that toolkit, ensuring that CGR is meeting users' needs. And in the image here on the outside you can see what we mean by community collaboration and resources which includes such things as organism specific knowledge bases, genotype/phenotype data, variation data, and even image collections.

So the third CGR component I want to mention are standardized interfaces for connectivity. So CGR helps communities that have their own organism specific resources connect with NCBI Toolkit and these connections make it easier for researchers at large to use data from sources they may not have known about. So we encourage you to encourage researchers to contact NCBI to explore opportunities to connect curated genome related content through an application programming interface, or API as you may know them. We'll provide links to resources that can be featured in the NCBI Toolkit. And in this image here, these smaller circles with the arrows illustrate that NCBI through CGR is providing the standardized interfaces and tools for the data connectivity that enable these opportunities for tie-ins of the community resources and content to NCBI. So together, these three components work harmoniously to support new discoveries and scientific advancement in the field of comparative genomics by maximizing the impact of the comparative genomics research and the work that happens as part of that.

We also believe that CGR is going to be impactful to users in a variety of ways, which I really just outlined in this pyramid here. And starting down here at the bottom, the foundation of CGR is high quality genomic data. And that's because uncontaminated and consistently annotated genomes are foundational to the success comparative genomics analysis. But we also know that many genomes that are submitted to GenBank today don't meet those criteria. So the NCBI Foreign Contamination Screening tool, which is now available and I'll tell you a little bit more about shortly, allows assembly producers to remove this contamination prior to submission. And likewise, we'll be making the Eukaryotic Genome Annotation Pipeline publicly available to help the community create and submit consistent high-quality annotations. NCBI is also making protein records more usable for comparative analysis through increased annotation of features like protein domains and their associated metadata.

But moving up the pyramid here, next we see that CGR offers new and improved comparative genomics tools. So in addition to providing these new tools for improving genomic data quality, CGR is also supporting the improvement of existing NCBI tools for genome analysis and the development of new ones. And we apply community collaboration to make the changes in these tools that most benefit comparative genomic stakeholders. And these tools provide researchers with really a seamless experience in exploring, analyzing, and retrieving eukaryotic genome related content. These efforts then improve the integration and the visualization of genome related data and provide better access to data from a wider range of organisms. And I'll be telling you about a number of these tools as a part of the case studies we look at shortly. CGR also offers the benefit of supporting scalable analysis. The number of genomes in GenBank and the species that they represent are growing annually and at an exponential rate. And the NCBI Toolkit's cloud ready tools and organism agnostic databases help CGR to scale with this data growth. You'll be hearing me talk about NCBI Datasets which provides resources that improve data discovery and retrieval. And central to this are APIs and structured data packages for all genomic related sequence and metadata that are interoperable with common genomics workflows. So this helps in creating artificial intelligence (AI) ready data sets that can be used to ask and answer new questions.

And CGR really also emphasizes the application of FAIR principles, and FAIR data principles make it easier to search, browse, download, and use NCBI's genome associated data with a wide range of standard bioinformatics platforms and tools. I'll also mention here that CGR amplifies opportunities for new discoveries that can be made through comparative genomics analysis. The Toolkit provides organism agnostic tools that offer equal access to datasets from across the tree of life and these expose biological information and can reveal patterns in the data that can spur really just new hypothesis for future research. And additionally with these standardized interfaces CGR can connect NCBI's Toolkit with communities they currently have or want to develop their own organism specific resources and the exposure of those data to more researchers can amplify its use and again contribute to new discoveries. So just putting all of this together, by maximizing the impact of eukaryotic research organisms and their genomic data resources to biomedical research, we think we're offering really new possibilities for scientific advancement.

So CGR can be impactful to biological research at all the levels I just mentioned. But I'm now going to go ahead and share two case studies with you. And these examples should provide you with a more tangible appreciation of all those abstract concepts that I just went over on the prior slides. And the first is an example focusing on how CGR is contributing to higher data quality and improved data access. And the second is an example that demonstrates how CGR can help users in analyzing and interpreting data.

So I'm going to begin with a concrete example of how CGR facilitates a very common comparative genomics research task, which is finding what kind of quality genomic resources exist for a focal species of interest. And in this case here, as an example, let's say you're contacted by a researcher who's working for a local public health agency who wants to learn more about a new and invasive species of tick that has started appearing on the East Coast of the US, which is where NIH is based. And so the tick Haemaphysalis longicornis or the Asian longhorned tick was deemed a pest of concern for livestock and pets. But we really need more study to understand whether it can also contribute to tick borne illness spread in people. And so this researcher needs to identify the available genomic data at NCBI for tick species and also to improve the quality of genomic data that they have generated for the tick species themselves. And so over the next couple of slides, I'll be introducing you to the specific CGR related resources that support both broad levels of tasks.

So since we're starting with a single species in this analysis, a natural place to get started for this user is on NCBI's new Taxonomy search page. And so once you've typed in the name of the species, you can see that you end up with a table that shows you there are a few genome assemblies. And as noted over here on the left, you can search in this table by common name, species name, or higher taxonomic groupings in addition to the scientific name of the species. Looking at our results here, we see that there are four available genome sequences for our focal species, H. longicornis and 44 genomes for ticks in general. Now if we want to learn more broadly about the sequence data available for any taxon level, we can click on the taxon name shown in the table and that will take you to the corresponding taxonomy pages. So here we're going to go ahead and do this for H. longicornis.

These taxonomy pages serve as your one stop shop for all sorts of sequencing related data available at NCBI for this organism. This runs from SRA experiments to protein sequences to even gene expression data. These pages also provide a direct link to the designated reference genome for an organism, and they're also very useful because they include links to other CGR resources such as new BLAST databases and visualization tools. But here to explore the four longicornis genomes in an organized way, what we can do is follow this link right here in the Taxonomy page. And on clicking it, it takes you directly to a genome table that summarizes the information we have about those genome assemblies. This includes metadata like genome size, the quality, as well as the type of annotation on the assembly. This helps researchers identify the assembly that's most suitable for their particular research needs, and to modify the information that's in the table, you can click on the Select columns button, which then brings up the options menu. And looking at our example right away, we can see that none of them are RefSeq genomes and only one of them has annotation which was provided by the genome submitter. And while perhaps those annotations are sufficient, we might also want to compare with other species of ticks that have genomes available at NCBI's GenBank.

So going back to that Taxonomy landing page, from our initial search result, you probably noticed there are entries and links for every taxonomy level that NCBI has listed above our focal species. And we can start by going up to the next highest taxonomic level, the hardback ticks, that have additional genomes. And you can see that this family of ticks to which our focal species belongs, has 30 Genomes available at GenBank. And clicking on that genome count will take us to a table like the one we saw before. But it now has all 30 genomes. Now you could be interested in this much larger set, since genome data from related species can be really useful in making inferences about our focal species. But that's also a lot of genomes to do an analysis on. And fortunately, that same genome table that I introduced you to before also allows you to do some filtering by useful assembly characteristics, such as whether something is the reference genome for the organism or whether or not NCBI's RefSeq group has provided annotations for it, or whether or not the assembly is highly contiguous. These kinds of filters can help folks identify the genomes that are most likely to be informative for the analysis that need to be performed. And so applying the filters that I've selected here takes us down to just five different assemblies that meet these standards.

And then after this filtering for genome quality and annotation availability, I also used information from additional table columns shown here to identify two genomes to study further. The Download button right here allows you to download bulk data in packages, making the data retrieval and subsequent analysis of these genomes much more efficient. And these packages can be configured to contain the corresponding sequence and annotation data for each assembly, as well as a metadata report that consolidates information that's spread over multiple file locations on NCBI's FTP sites as well as from different NCBI databases, which previously would have required you to make multiple requests or have a really deep understanding of our databases and FTP structure in order to get. So the files on this package also use industry standard formats for the corresponding data types and they're compatible with many tools in a number of bioinformatics pipelines. But in addition to the genomes, we can also access the genes that are annotated on a specific assembly by navigating to the corresponding genome record page and then selecting the corresponding link for gene data.

So we're going to head over to that gene data and here we see now that gene table. And in our example, one of the genomes that I decided to use for my analysis comes from a southern tick called the southern cattle tick, which is a much better studied relative of the focal species that has a lot more corresponding data including our RefSeq annotation. And so this genes table, it really serves as the resource for downloading the gene transcript and protein sequence data along with the corresponding metadata that are annotated on a specific assembly. I'll just emphasize that these tables are available for more than 7000 eukaryotic genomes in RefSeq and GenBank that have annotation provided either by NCBI or by the submitter. And you can download all the annotation data in bulk. Or again you can use filters to find specific genes by name, type or location. And as shown here, you can see that we got one result when searching the R. microplus annotation set for evasin, which is a protein in tick saliva that helps evade the host immune system. But again from here I can download the data or even follow a link to the gene page to learn more about the gene itself.

Now, NCBI Datasets is the resource in CGR’s NCBI Toolkit that provides the web-based information that we just looked at for genes and genomes, and it also has a command line interface whose organization is shown here. And in addition to the download command which I just introduced you to through the web page, there's also a command line feature called summary which allows you to browse metadata without having to download it. I'll emphasize that the Datasets' command line tool can also be installed as an Anaconda package, and NCBI Datasets is also available in Galaxy, which if you've heard of that is a popular bioinformatics ecosystem. So really any workflow that requires the input of NCBI genome data can now be run entirely in that environment. And I'll also just mention that there is an NCBI Datasets REST API that can be used to make programmatic requests for all this content.

So to summarize what we've just covered, we were able to use CGR to find genomic data for a focal species and for a related species, and we also sorted and filtered those assemblies and their annotated features using metadata. But what if you are working with someone who wants to improve or annotate their own tick genome assemblies? And this is where the data quality tools of CGR's NCBI Toolkit came into play, namely the Foreign Contamination Screen and the Eukaryotic Genome Annotation Pipeline.

So we can define contamination as sequences present in a dataset that don't originate biologically from the indicated organism. And contamination can occur at many steps in the genome assembly process, starting from the initial DNA extraction all the way to data submission. And that's because there are multiple potential sources of contamination in a genome sequencing project. So for example, in an insect genome like our tick, there could be contamination coming from a toast organism, it could be from bacterial contamination from endosymbionts that coexist with the insect, or it could be introduced during the lab sample preparation. And unfortunately we know today that, at least in 2022, 1 out of every 3 eukaryotic genomes submitted to GenBank had detectable contamination. And so to help users deal with these contamination issues which can really negatively influence the analysis that happened downstream that use the data, we decided to improve the legacy GenBank screening process and we made it available as a public tool. And so this cloud ready software is now available on GitHub and it includes FCS adapter which is a tool that allows the identification and trimming of contamination from sequencing adapters, which are a very common source of contamination, as well as FCS-GX which is a brand new cross species aligner that runs two to three orders of magnitude faster than the legacy process and it also has improved sensitivity and specificity. This is all very straightforward to run. A user needs only to provide a genome assembly and the corresponding NCBI Taxonomy Identifier, and then the user gets in return a contamination summary report actions for cleaning the genome, as well as a clean genome and a file of the contaminants themselves should they have reason to want to do analysis on that additional set of sequences.

I also want to tell you about the Eukaryotic Genome Annotation Pipeline (EGAP), which has been used by NCBI to annotate more than 1000 eukaryotic RefSeq assemblies in the past few years. And what's really exciting about EGAP are two developments. The first is that we're soon planning to make EGAP publicly available as a containerized piece of software that folks can run on their own assemblies with all kinds of input data before it even gets to GenBank at NCBI for submission and QC. And as part of that, we're also working to make the pipeline support the annotation of a wider range of species such as fungi, nematodes, and protists, all of which are often pathogens to humans. And in the next couple of months there's going to be opportunities for researchers to serve as alpha testers of the standalone package. We're expecting that to begin happening in January. So if you're in contact with folks that might be interested in testing that out, we'd love for you to put them in contact with us.

So summarizing our first case study in which we use CGR to find and contribute genomic data for an invasive tick species to find these data, we used the following resources. With the Taxonomy Browser, we found four available genome sequences for H. longicornis. We used the Improved Taxonomy Pages to view gene expression, raw sequencing and overarching projects and more. Through the Datasets Genome Table, we evaluated and filtered all 30 genomes in the Ixodidae tick family, and with the Datasets Command Line Interface and API, we were able to do either a bulk download and parsing of genome and annotation data. The annotation table let us search for and download data for a specific gene and the genome of closely related tick species and we also use two items from the NCBI Toolkit to improve our own data prior to submission to NCBI. The Foreign Contamination Screen can be used to remove sequencing adapters and contamination from other genomes to improve genome data and quality and utility. And I told you about how the EGAP Eukaryotic Genome Annotation Pipeline will be coming soon, which would allow us to produce annotation of coding and non-coding genes, transcripts and proteins for genome assembly.

So we're now going to turn to our second case study to learn how CGR can help users analyze and interpret data. So as compared to genomics advances, we know that an increasing number of model organisms, both established and emerging, are being relied upon for studying human health. And dog genomes have been extensively studied and characterized, and many genetic variations have been identified that are associated with canine hereditary diseases. And among these are several cancers which are common in both species, making dogs really an ideal model for studying human disease progression and treatment. Several genetic variants in the TP53 or tumor protein P53 gene have been shown to cause leaf from any syndrome in humans, which increases the risk of developing cancer. And in dogs, the development of osteosarcoma and histiocytic sarcoma have been strongly correlated with similar T53 variants. So understanding how these variants cause disease can guide the creation of next generation diagnostics and therapeutics to improve the health of both dogs and humans. So in this case study, we'll learn about several CGR and other NCBI resources that I've got listed here that can help with this type of research. With NCBI Gene, we'll learn what's known about the dog TP53 gene. The Comparative Genome Viewer helps us explore synteny for the dog and human TP53 genomic regions, while the Genome Data Viewer allows us to examine TP53 gene and sequence details for the dog and human genome assemblies alongside other annotation tracks. With the Multiple Sequence Alignment Viewer or MSA viewer, we can compare similarities and differences in dog, human, and other organisms TP53 proteins with a quick multiple sequence alignment. And NCBI Orthologs lets us understand the evolutionary history of the TP53 protein. And finally, iCn3D will let us visualize the human TP53 protein's 3D structure and directly map aligned human and dog sequences to known clinical variants that are described in ClinVar. So we'll now just take a closer look at each one of those.

So NCBI's Gene records provide summaries of many aspects of an organism's gene specific information, and that includes such things as sequences, expression data, published literature, functional domains, homologs, structures. These web pages also offer access to a variety of other NCBI resources containing relevant data. And so we can then explore NCBI Gene records for the TP53 gene in both humans and dogs. And in doing so, what we would find is that the human TP53 gene record here contains much more information than is displayed for the dog version. So if I'm someone interested in TP53's role in dog cancer, this data here may help fill in knowledge gaps for me as I explore that role.

There's also a new web tool called the Comparative Genome Viewer that was released in 2022. It provides a visual comparison of two genomes based on the alignment of the sequences to one another. And what this tool does is it helps users explore genomic structural changes in the context of annotations as well as explore syntactic relationships. And I'll just give a quick shout out here. If you're demonstrating this tool to someone or using it yourself and you don't see the organisms or assemblies that you're looking for, we really encourage you to use the feedback button that's found on the bottom of the page to contact us and let us know because we're still actively adding data into this resource. The alignments here have been generated with the NCBI Assembly Alignment Pipeline, as well as from the UCSD Genome Browser and the Human Pangenome Research Consortium, as well as by a few other methods. The interactive display that we've got allows you to zoom in and out to search for specific gene annotations, customize your display and download an image or the underlying data. And you can right click within the CGV to get access to other key visualization resources for such things like that MSA viewer that I mentioned to you, the Genome Data Viewer, and other tools. And so when we consider our specific case study, what we can see here is that while at the at greater level the TP 53 genomic regions appear largely similar for human and dog, if we dive down into the gene itself, we can see a few gene structural differences in there.

Now, taking a look at one of those other tools, the Genome Data Viewer (GDV), which is shown here, allows you to explore and analyze genomic regions by annotations as well as alignments. And this interactive display let's you zoom in and out, search for annotations and features, you can customize the display, and you can even download an image or the underlying data that's in the browser. And I want to emphasize that you can explore available data tracks that are provided by NCBI, or you can upload them from the track hub registry, which is a resource served from UCSC that contains third party data. Or you can even upload your own data into GDV. And like CGV, this is also a continually developing resource and new data tracks are added as additional data becomes available. So again, taking a look at our specific example, what we would have found is that the human and dog genomic alignments in the TP53 gene region enable really a direct comparison of differences in known transcript variants in this tool. And in addition, we can use the tool to make comparisons with other track annotations, including, as I've shown here, RNA seq expression data for specific tissues.

Another visualization tool that you could access from CGV, the Comparative Genome Viewer, is the Multiple Sequence Alignment Viewer (MSA). And again, it supports the visual analysis of aligned nucleotide or protein sequences. And it's interactive display lets you zoom in and out. You can even color residues based on several different characteristics. You can really customize this display and again, you can download an image or you can download the data that sits beneath it and the resource can be used to visualize your own sequence alignments as well as alignments that you get out of a blast search. And for those folks who are involved with people who are building their own web pages, I want to shout out that the MSA viewer is also available with an API call so you can embed it in your own web page. And again, returning to our case study, we can see that in comparison with the human TP53 protein sequence, the dog and other mammals have a region that shows significant sequence diversity. And using MSA Viewer, we can really dive in and examine that region in more detail to see how they've changed, which in turn might allow us to identify whether there are parts of the protein that may contribute to cancer in one species versus another.

NCBI Orthologs is also a resource I want to give a shout out to. It helps you find related genes across organisms and provides you the opportunity to do quick sequence comparisons. So it's really easily accessed from any NCBI gene record page if the gene belongs to an ortholog set. And it's a great way to explore the wealth of data at NCBI from across the taxonomic tree, so you can use it to draw connections that help you really understand your understanding of genes in some of these lesser studied organisms. So within NCBI Ortholog you can select specific transcript or protein sequences for download or alignment. You can refine the results that you get with this taxonomic tree. So if you didn't want to see all 400 members of the TP53 ortholog set, you could whittle that down to attacks that you're interested in. And you can access relevant PubMed citations for orthologs. And so let's just say that when it comes to our interest in the role that TP53 plays in cancer, you know, as I just mentioned, our annotation pipeline at NCBI has identified TP53 ortholog sequences for over 400 organisms. And so the sequence alignments followed by variant analysis, perhaps done in MSA, that tool I just mentioned, can help reveal those individual bases that play a critical role in the function of this gene in one species or the other, or in both of them.

And in addition to sequence level comparisons such as those that are made possible by the definition of orthologous sets of genes, I also want to give a shout out to protein structures because protein structures are another important and valuable means to studying biological questions from across different species. iCn3D is a web-based tool that supports the visualization and mapping of a protein's key sequence residues to its 3D structure along with NCBI's annotations such as the positions of known clinical variants. And as I mentioned here, in iCn3D you've got an interactive display, you've customized the image, and again download data. One of the really cool things is you can align multiple structures through it and look at it. And the source code for this is available on GitHub. So it's a tool that for a tool developer is a really great playground. And so when exploring TP53 in iCn3D we can start here by looking at these multiple sequence alignments for the core DNA binding domain. And what we see is that in this region the two proteins have sequence very similar to one another in human and in dog. And so then we can use a tool like iCn3D to map some known ClinVar pathogenic variants to the structure in order to predict whether a specific genetic variant would have a similar impact in dog.

So summarizing our second case study, we use CGR and other NCBI resources to help us analyze and explore data that would help us make discoveries about the role of TP53 in cancer is common to both humans and dogs. From NCBI Gene we learned what's known about the dog TP53 gene and what information we might infer from it. And by studying its human version. From CGV, we're able to align and explore the relationship of the dog and human TP53 genomic regions. On GDV, we examined the TP53 gene annotations from dog and human alongside that of other annotation tracks, including RNA seq expression data. While, if we had gone and used MSA viewer, we would have been able to explore and directly compare the human, dog, and other mammalian TP53 protein sequences down at the sequence level from NCBI orthologs. We found the orthologous sequences for over 400 organisms and then we're able to have a resource that let that let us easily download and align data sets for the organisms that we selected. Like in this case we might focus on human and dog but add in a few other as well. And with iCn3D, with this web tool, we were able to interactively visualize the human TP53 protein 3D structure and we could also map aligned human and dog sequences and annotations such as ClinVar variants onto those structures to explore how those variants might be impactful in different species.

So with that, I want to talk a little bit about what's coming next for CGR. It is an active project and there are ongoing resource improvements based on community feedback, which is why we really value having you check out the tools that we have in CGR and alerting the folks that you work with or your clients to CGR, having them try it out, not only so it helps their research, but gets us some feedback on how we can continue to make it better. As I've mentioned, we're working on making that Eukaryotic Genome Annotation Pipeline publicly available and expanding its taxonomic skill and we are certainly interested in getting alpha testers for testing that should be starting around January. And again, we're also continuing to make more data available in the Comparative Genome Viewer. We've also got a set of tools that I didn't have time to tell you about today, which you can learn about from our website, which I'll introduce you to in just a second, where we also provide information on our current activities.

So with that, you know where do you send researchers, you know where you're sending yourselves to learn more about CGR and the research that you know may be benefiting from it? So as I've just mentioned, we're continuing to work on CGR and we're also researching and developing ways to meet the objectives of the project. And so there really are what we think of as considerable opportunities for the genomics community to share feedback and inform future enhancements and to connect with us at NCBI. And so I've just listed on this slide here some ways that you can get involved with CGR. And I always sort of jokingly say that we have pretty thick skin at NCBI and we're really eager to learn how we can improve whether it's our data or tools or anything to support the needs of our users and with CGR in particular comparative genomics research. So you can e-mail us directly at the address shown here or you can use the link that's on our website. The website URL is shown here, to contact us for feedback. You can also use that website to access all the development I told you about today, as well as ones I didn't have time to talk about. And the website's also a great place to sign up for the CGR e-mail newsletter to stay up to date on the latest news and information. The website and the e-mail list will alert you to when we're holding feedback sessions, product testing, and other opportunities to inform the development process. And I'll just say our mailing list, we only send out a couple of times a month so you're not going to get spammed, but it really is a great way to stay on top of what's happening.

But we have a variety of different communication channels that are available to engage you or your clients and they really fall into three main categories. We have direct channels which we like to think of as push communications that deliver information about CGR to you that's timely and relevant and resonant. And so this includes things like that e-mail list I told you about before, but it's also our social media surveys that we do and a couple of other products I'll introduce in just a second. We also have indirect channels which are more sort of pull communications, which help provide access to information about CGR that stakeholders can proactively retrieve. So you can go in and get this yourself. You can visit our website, you can try out our products, or you can use any of our training materials like our fact sheets or our facts. And then finally, I'll just mention that again, we do have a number of interactive channels that we're using to provide engagements with the diverse community out there that use the CGR. And so that includes things like again, feedback sessions, user testing of our products. We are setting up codeathons and also doing workshops kind of like the one we have today. And folks are always welcome to reach out to the help desk.

And so in addition to all these things I've mentioned today, we're going to be providing you with four pieces of information that we hope you can take back with you to help you start exploring

CGR and introducing it to clients and other folks that you know. One of these is our Impact Spotlights, which are short pages that describe how CGR could help research that's already been published. So it describes a recent publication that attempts to answer a very common type of comparative genomics question and then shows the specific CGR resources that could be used to benefit that particular type of research or answer that type of question. And we have these for a couple of different types of comparative analysis. We've also got a trifold flyer which you know you can print out and share with others. We'll also have a PDF of today's publication. And finally, now available on the NLM Product Guides for Training and Outreach page, there is an NLM guide for CGR. So we hope that you'll check that out.

So today we covered a number of topics relevant to CGR. We learned about comparative genomics and the value for research organisms, challenges that users face, and the solution that CGR offers. And then we went through a number of case studies to look at some specific CGR resources that really demonstrate how CGR is providing that solution. And I told you about what's coming next. So with that, I'm just going to conclude my presentation by thanking many of the people that were involved in the work I told you about today. This really takes a village to build and this slide doesn't even do justice to the number of hands that have been involved in this project. It's been a fantastic group of people working on it and we really welcome your feedback so we can continue to make it better. And thank you very much.

KATE: Thank you so much, Dr. Schneider. Thank you for that excellent presentation and explanation of the vision and rationale behind and the very practical features of the CGR. I particularly love the features that let you select and filter for the data that you want to view in the browser. The beautiful and flexible visualization tools including that API functionality, the summary option in the command line interface, the availability of the contaminant screening tool, and the upcoming alpha test of the publicly available Eukaryotic Genome Annotation Pipeline. That's very exciting and this all should be very exciting information for anyone interested in finding and viewing biological data.

So let's now turn to the questions that our participants have about CGR. We do have one question in so far, from Gareth. Gareth would like to know **horizontal genome transfer can occur from endosymbionts to the host genome. Is there a way to let the FCS pass certain quote contaminants through? Or is there another workflow that you might recommend?**

DR. SCHNEIDER: That's a really good question. And yeah, when we see sequences from more than one organism in a submitted genome, you're right, it could be more than just contamination. It really could be that situation like horizontal gene transfer. So while we don't at the moment have a way to say, you know, ignore sequences that come from species Y, if I'm looking at species X, what you do get from the Foreign Contamination Screen tool is a report that basically says, you know, here's what we found and it makes recommendations and it doesn't automatically take those sequences out of the genome for you. That is a separate additional step. So if you look at that report and you're like, hey, I think this thing kind of overstepped, this is horizontal gene transfer. We know about this relationship between the two organisms. Then you can leave them in place.

KATE: Thank you. Excellent. Molly asks. Well, she says, **all of these new NCBI resources are great to know about, but I'm struggling to find them from ncbi.nlm.nih.gov, the home page. Are there plans to add these to the popular resources menu or in the drop down database menu on the NCBI search box? What's the easiest way to find and access all of these new resources?**

DR. SCHNEIDER: That's a good question too. Yeah, we know that we have a lot of resources in NCBI and the things that make up CGR's NCBI Toolkit come from a lot of different places. And so if you're looking for a centralized point to just find what's in CGR, I would encourage you to go to the CGR website at NCBI. So it's just the NCBI homepage slash CGR and on that you'll find links to all of the different tools and they're sort of organized by whether it's a data quality tool or you know a data resource in and of itself.

At this point in time, we're still looking at different ways that we can expose CGR throughout NCBI's website. If you look at the new NCBI Datasets tool that I mentioned when I was talking about all those genome tables, you'll see down at the footer there that we've got some markup on there to let you know that you're at CGR at that point in time. And we're considering that for some of the other resources. So we'd like feedback on and how difficult it is especially using that CGR homepage to navigate.

KATE: Great, thank you. Yes, we love to get feedback. A question from Chris. Chris would like to know **are there specific resources for mitochondrial genome analysis?**

DR. SCHNEIDER: That is a good question as well. So we are working on having an organelle annotation available through that NCBI Datasets tool that I mentioned early on. So it's not there right now, but it will be coming I believe within the next couple of months. So can be on the lookout for that.

KATE: Great. Thanks. So we've come to the end of the questions in chat.