RT2T: A Global Collaborative Project to Study Chromosomal Evolution in the Suborder Ruminantia

Ted Kalbfleisch
Department of Veterinary Science, University of Kentucky, Lexington, KY 40546 https://orcid.org/0000-0002-2370-8189

Stephanie McKay
Division of Animal Sciences, University of Missouri, Columbia, MO, 65211, USA https://orcid.org/0000-0003-1434-3111

Brenda Murdoch
Department of Animal, Veterinary and Food Sciences, University of Idaho, Moscow, ID, 83844, USA https://orcid.org/0000-0001-8675-3473

David L. Adelson
School of Biological Sciences, The University of Adelaide, North Terrace, Adelaide, 5000, SA, Australia https://orcid.org/0000-0003-2404-5636

Diego Almansa
Genomics and Bioinformatics Unit, Departamento de Ciencias Biológicas, CENUR Litoral Norte, Salto, Universidad de la República, Uruguay https://orcid.org/0000-0002-7716-8480

Gabrielle Becker
Department of Animal, Veterinary and Food Sciences, University of Idaho, Moscow, ID, 83844, USA https://orcid.org/0000-0002-1455-6443

Linda M. Beckett
Department of Animal Sciences, Purdue University, West Lafayette, IN 47907, USA https://orcid.org/0000-0003-0176-9500

María José Benítez-Galeano
Genomics and Bioinformatics Unit, Departamento de Ciencias Biológicas, CENUR Litoral Norte, Salto, Universidad de la República, Uruguay https://orcid.org/0000-0002-6332-1632

Fernando Biase
School of Animal Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA, 24061, USA https://orcid.org/0000-0001-7895-0223

Theresa Casey
Department of Animal Sciences, Purdue University, West Lafayette, IN 47907, USA https://orcid.org/0000-0002-8835-3550

Edward Chuong
BioFrontiers Institute, Department of Molecular Cellular and Developmental Biology, University of Colorado Boulder, Boulder, CO 80303 USA https://orcid.org/0000-0002-5392-937X

Emily Clark
The Roslin Institute, University of Edinburgh, EH25 9RG, Edinburgh, UK. https://orcid.org/0000-0002-9550-7407

Shannon Clarke
Invermay Agricultural Centre, AgResearch Ltd, Mosgiel, New Zealand https://orcid.org/0000-0002-4615-8917

Noelle Cockett

Department of Animal, Dairy and Veterinary Sciences, Utah State University, Logan, UT 84322-4815 USA https://orcid.org/0009-0001-2387-3512

Christine Couldrey

Livestock improvement corporation, Hamilton 3286, New Zealand https://orcid.org/0000-0001-7410-9410

Brian W. Davis

Department of Veterinary Pathobiology, School of Veterinary Medicine & Biomedical Sciences, Texas A&M University, College Station, TX 77843, USA https://orcid.org/0000-0002-6121-135X

Christine G. Elsik

Division of Animal Sciences, University of Missouri, Columbia, MO, 65211, USA https://orcid.org/0000-0002-4248-7713

Thomas Faraut

GenPhySE, Université de Toulouse, INRAE, ENVT, 31326 Castanet Tolosan, France https://orcid.org/0000-0001-5156-3434

Yahui Gao

Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD, 20705, USA https://orcid.org/0000-0003-0602-4823

Carine Genet

GenPhySE, Université de Toulouse, INRAE, ENVT, 31326 Castanet Tolosan, France https://orcid.org/0000-0002-3734-9017

Patrick Grady

Institute for Systems Genomics, University of Connecticut, Storrs, CT, 06269 https://orcid.org/0000-0003-0180-7810

Jonathan Green

Division of Animal Sciences, University of Missouri, Columbia, MO, 65211, USA https://orcid.org/0000-0003-4968-8187

Richard Green

Institute for Systems Genomics, University of Connecticut, Storrs, CT, 06269 https://orcid.org/0009-0004-6175-5164

Dailu Guan

Department of Animal Science, University of California Davis, Davis, California, USA https://orcid.org/0000-0001-8800-3158

Darren Hagen

Department of Animal and Food Sciences, Oklahoma State University, Stillwater, OK 74078, USA https://orcid.org/0000-0001-8295-020X

Gabrielle A. Hartley

Institute for Systems Genomics, University of Connecticut, Storrs, CT, 06269 https://orcid.org/0000-0002-5672-2171

Mike Heaton

USDA, ARS, U.S. Meat Animal Research Center, Clay Center, NE, USA https://orcid.org/0000-0003-1386-1208

Savannah J. Hoyt

Institute for Systems Genomics, University of Connecticut, Storrs, CT, 06269 https://orcid.org/0000-0001-7804-3236

Wen Huang

Department of Animal Science, Michigan State University, East Lansing, MI 48824, USA https://orcid.org/0000-0001-6788-8364

Erich Jarvis

Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, USA 10065 https://orcid.org/0000-0001-8931-5049

Jenna Kalleberg
Division of Animal Sciences, University of Missouri, Columbia, MO, 65211, USA https://orcid.org/0000-0003-4505-8516
Hasan Khatib
Department of Animal and Dairy Sciences, The University of Wisconsin-Madison, Wisconsin, USA https://orcid.org/0000-0003-1312-1453
Klaus-Peter Koepfi
Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA 22630, USA https://orcid.org/0000-0001-7281-0676
James Koltes
Department of Animal Science, Iowa State University, Ames, IA 50011, USA https://orcid.org/0000-0003-1897-5685
Sergey Koren
Genome Informatics Section, Center for Genomics and Data Science Research, National Human Genome Research Institute, National Institutes of Health (Bethesda, MD, USA) https://orcid.org/0000-0002-1472-8962
Christa Kuehn
Friedrich-Loeffler-Institute (German Federal Research Institute for Animal Health), 18357 Greifswald-Insel Riems, Germany https://orcid.org/0000-0002-0216-424X
Tosso Leeb
Institute of Genetics, Vetsuisse Faculty, University of Bern, 3001 Bern, Switzerland https://orcid.org/0000-0003-0553-4880
Alexander Leonard
Animal Genomics, ETH Zurich, 8092 Zurich, Switzerland https://orcid.org/0000-0001-8425-5630
George E. Liu
Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD, 20705, USA https://orcid.org/0000-0003-0192-6705
Wai Yee Low
The Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, SA 5371, Australia https://orcid.org/0000-0002-0749-765X
Hunter McConnell
Division of Animal Sciences, University of Missouri, Columbia, MO, 65211, USA https://orcid.org/0009-0008-3490-9518
Kathryn McRae
Invermay Agricultural Centre, AgResearch Ltd, Mosgiel, New Zealand https://orcid.org/0000-0002-4229-9341
Karen Miga
UC Santa Cruz Genomics Institute, University of California, CA 95064 https://orcid.org/0000-0001-9709-4565
Michelle Mousel
USDA, ARS, Animal Disease Research Unit, Pullman, WA 99164, USA and School for Global Animal Health, Washington State University, Pullman, WA 99164, USA https://orcid.org/0000-0003-1367-7005
Holly Neibergs
Department of Animal Science, Washington State University, Pullman, WA https://orcid.org/0000-0001-8121-3072
Rachel O'Neill
Institute for Systems Genomics, University of Connecticut, Storrs, CT, 06269 https://orcid.org/0000-0002-1525-6821
Temitayo Olagunju

Department of Animal, Veterinary and Food Sciences, University of Idaho, Moscow, ID, 83844, USA https://orcid.org/0000-0002-1497-6173

Matt Pennell

Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089 https://orcid.org/0000-0002-2886-3970

Bruna Petry

Department of Animal Science, Iowa State University, Ames, IA 50011, USA https://orcid.org/0000-0001-9559-8559

Mirjam Pewsner

Institute of Fish and Wildlife Health, Vetsuisse Faculty, University of Bern, 3001 Bern, Switzerland https://orcid.org/0009-0000-7139-0032

Adam M. Phillippy

Genome Informatics Section, Center for Genomics and Data Science Research, National Human Genome Research Institute, National Institutes of Health (Bethesda, MD, USA) https://orcid.org/0000-0003-2983-8934

Brandon D. Pickett

Genome Informatics Section, Center for Genomics and Data Science Research, National Human Genome Research Institute, National Institutes of Health (Bethesda, MD, USA) https://orcid.org/0000-0001-8235-4440

Paulene Pineda

The Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, SA 5371, Australia https://orcid.org/0000-0002-7147-3302

Tamara Potapova

Stowers Institute for Medical Research (Kansas City, MO, USA) https://orcid.org/0000-0003-2761-1795

Satyanarayana Rachagani

Veterinary Medicine and Surgery, NextGen Precision Health Institute, University of Missouri, Columbia https://orcid.org/0000-0003-3949-8566

Arang Rhie

Genome Informatics Section, Center for Genomics and Data Science Research, National Human Genome Research Institute, National Institutes of Health (Bethesda, MD, USA) https://orcid.org/0000-0002-9809-8127

Monique Rijnkels

Department of Veterinary Integrative Biosciences , School of Veterinary Medicine & Biomedical Sciences, Texas A&M University, College Station, TX 77843, USA https://orcid.org/0000-0002-8156-3651

Annie Robic

GenPhySE, Université de Toulouse, INRAE, ENVT, 31326 Castanet Tolosan, France https://orcid.org/0000-0003-3071-8614

Nelida Rodriguez Osorio

Genomics and Bioinformatics Unit, Departamento de Ciencias Biológicas, CENUR Litoral Norte, Salto, Universidad de la República, Uruguay https://orcid.org/0000-0002-2235-979X

Yana Safonova

Computer Science and Engineering Department, Huck Institutes of the Life Sciences, Pennsylvania State University, State College, PA 16801, USA https://orcid.org/0000-0002-9634-4216

Gustavo Schettini

School of Animal Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA, 24061, USA https://orcid.org/0000-0003-0415-0201

Robert D. Schnabel

Division of Animal Sciences, University of Missouri, Columbia, MO, 65211, USA https://orcid.org/0000-0001-5018-7641

Nagabhishek Sirpu Natesh

Department of Veterinary Medicine and Surgery, University of Missouri, Columbia, MO 65211, USA. https://orcid.org/0000-0003-3826-1591

Morgan Stegemiller

Department of Animal, Veterinary and Food Sciences, University of Idaho, Moscow, ID, 83844, USA https://orcid.org/0000-0001-9675-3047

Jessica Storer

Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT https://orcid.org/0000-0002-9619-5265

Paul Stothard

Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB, T6G 2P5, Canada. https://orcid.org/0000-0003-4263-969X

Caleb Stull

Division of Animal Sciences, University of Missouri, Columbia, MO, 65211, USA https://orcid.org/0009-0001-5370-9852

Gwenola Tosser-Klopp

GenPhySE, Université de Toulouse, INRAE, ENVT, 31326 Castanet Tolosan, France https://orcid.org/0000-0003-0550-4673

Germán M. Traglia

Genomics and Bioinformatics Unit, Departamento de Ciencias Biológicas, CENUR Litoral Norte, Salto, Universidad de la República, Uruguay https://orcid.org/0000-0002-4780-5311

Chris Tuggle

Department of Animal Science, Iowa State University, Ames, IA 50011, USA https://orcid.org/0000-0002-4229-5316

Curtis P. Van Tassell

Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD, 20705, USA https://orcid.org/0000-0002-8416-2087

Corey Watson

Department of Biochemistry and Molecular Genetics, University of Louisville, Louisville, KY, USA https://orcid.org/0000-0001-7248-8787

Rosemarie Weikard

Institute of Genome Biology, Research Institute for Farm Animal Biology (FBN), 18196 Dummerstorf, Germany. https://orcid.org/0000-0002-0308-9966

Klaus Wimmers

Institute of Genome Biology, Research Institute for Farm Animal Biology (FBN), 18196 Dummerstorf, Germany. https://orcid.org/0000-0002-9523-6790

Shangqian Xie

Department of Animal, Veterinary and Food Sciences, University of Idaho, Moscow, ID, 83844, USA https://orcid.org/0000-0002-8821-2344

Liu Yang

Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD, 20705, USA https://orcid.org/0000-0002-4179-8587

Tim Smith (tim.smith2@usda.gov)

USDA, ARS, U.S. Meat Animal Research Center, Clay Center, NE, USA https://orcid.org/0000-0003-1611-6828

Ben Rosen (ben.rosen@usda.gov)

Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD, 20705, USA https://orcid.org/0000-0001-9395-8346

Short Report

Keywords: T2T, Evolutionary Genomics

Posted Date: February 6th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3918604/v2

Additional Declarations: The authors declare no competing interests.

Abstract

The publication of the first complete, haploid telomere-to-telomere (T2T) human genome revealed new insights into the structure and function of the heretofore “invisible” parts of the genome including centromeres, tandem repeat arrays, and segmental duplications. Refinement of T2T processes now enables comparative analyses of complete genomes across entire clades to gain a broader understanding of the evolution of chromosome structure and function. The human T2T project involved a unique *ad hoc* effort involving many researchers and laboratories, serving as a model for collaborative open science. Subsequent generation and analysis of diploid, near T2T assemblies for multiple species represents a substantial increase in scale and would be daunting for any single laboratory. Efforts focused on the primate lineage continue to employ the successful open collaboration strategy and are revealing details of chromosomal evolution, species-specific gene content, and genomic adaptations, which may be general or lineage-specific features. The suborder Ruminantia has a rich history within the field of chromosome biology and includes a broad range of species at varying evolutionary distances with separation of tens of millions of years to subspecies that are still able to interbreed. We propose an open collaborative effort dubbed the “Ruminant T2T Consortium” (RT2T) to generate complete diploid assemblies for species in the Artiodactyla order, focusing on suborder Ruminantia. Here we present the initial near T2T assemblies of cattle, gaur, domestic goat, bighorn sheep, and domestic sheep, and describe the motivation, goals, and proposed comparative analyses to examine chromosomal evolution in the context of natural selection and domestication of species for use as livestock.

Introduction

The recent, highly successful, project to generate the first complete T2T human genome assembly[1] was the result of a multi-institution and country collaborative effort, led by genomicists at the National Institutes of Health, the University of California at Santa Cruz, and the University of Washington. This effort to produce the highest quality assembly for any mammalian species included the latest 3rd generation long-read sequencing technologies and assembly algorithms at the time; manually closing gaps, curating errors, and analyzing the genome required a wide range of expertise. The human T2T project (HT2T) was distinguished by the open nature of the collaboration and establishment of “working groups” (WG) dedicated to specific tasks such as assembly completion and curation, transcriptional and epigenetic maps and patterns that included heterochromatic regions[2-4], variant analysis[5], and characterization of segmental duplications[6]. The results from these WG were released as a series of companion manuscripts in the April 2022 issue of *Science* [1-6], demonstrating the value and power of the open collaboration model.

The rationale for the HT2T project is that missing data and assembly artifacts in the existing human reference genome interfered with analyses that were based on mapping reads from other individuals to the reference, or comparative analysis with reference genomes of other species. Assembly artifacts in the pre-T2T era include short insertion/deletions, collapses of near identical duplicated sequences, haplotype switching, and misrepresentations of highly repetitive sequence elements such as ribosomal RNA (rRNA) gene arrays and centromeric and telomeric regions. These limitations have long been recognized, but only recently have the long-read sequencing technology and algorithms been developed capable of addressing these shortcomings and making possible the creation of diploid, phased, gapless T2T genomes for most chromosomes[7-11].

The underlying hypothesis for T2T efforts was that new insights into chromosomal biology would be made possible by having a complete representation of the genome. This hypothesis was confirmed when the complete genome added 8% to the assembly size of the human genome and increased the predicted genes in the human genome by 1,956 - of which at least 99 are likely protein coding. A comprehensive view of the segmental duplications (SDs) organization in the complete genome indicated they account for nearly one-third of the additional sequence and increase the estimate of their frequency in the human genome from 5.4% to 7.0%. Epigenetic pattern analysis increased the number of CpG methylation calls by 10% and added 3% to 19% more peak calls compared to the “gold standard” ENCODE annotation[1,2,12]. The analysis detailed chromatin status for ~2,680 genes lacking such annotation. This provided evidence for 57 previously unannotated active promoters in more than one cell type and 82 genes with cell type-specific active promoter signal[2]. Genomic and epigenomic maps of heterochromatin including centromeric/pericentromeric regions and other repeat elements, which constitute 6.2% of the human genome (approximately 190 megabase pairs, Mbp), revealed numerous multi-Mbp structural rearrangements that overlap active centromeric repeat arrays involved in attachment of spindle fibers during cell division[3,4]. The data suggests that centromere position has strong effects on the evolution of surrounding chromosome regions through layered repeat expansions and revealed high degrees of structural, epigenetic, and sequence variation among centromeres of the X chromosome across individuals. The analysis of non-centromeric repeats, including transposable elements, repeat expansions, and repeat-mediated structural rearrangements, revealed the impact of specific repeats on the human genome. The authors concluded that the work “demonstrates the need for T2T-level assemblies of non-human primates and other species to fully understand the complexity and impact of repeat-derived genomic innovations that define primate lineages”.

Human acrocentric chromosomes play a role in nucleolar formation through their highly repetitive ribosomal RNA genes. Analysis of the complete genome revealed that the repeat families including rRNA gene arrays and extended segmental duplications on individual acrocentric chromosomes have unique epigenetic identity, likely contributing to their functional roles. The complete sequences of the p-arms of acrocentric chromosomes (human chromosomes 13,14,15,21,22) also supported work based on a subsequent human pangenome assembly demonstrating recombination between non-homologous chromosomes[13,14] in pseudo-homologous regions (PHR). These PHR include sequences that are known to lie at the breakpoint of Robertsonian translocations[15], which are relatively frequent in cattle[16] due to the acrocentric/telocentric nature of their chromosomes, suggesting they may be important in chromosomal evolution in ruminants. However, little is known about the structure of rRNA genes in ruminant species or their position, making it difficult to generalize the mechanism suggested in the human studies.

A major use of reference genome assemblies has been the identification and genotyping of variants for population analysis and detection of associations between phenotype and particular genetic variation. One WG of the human effort examined the impact of a complete reference genome on the analysis of genetic variation in humans[6]. The analysis quantified the impact of the complete assembly versus the GRCh38 reference assembly on read mapping and variant calling using data from 3,202 short read datasets and 17 long read datasets representing diverse populations. Hundreds of thousands of new variants were detected, and, just as importantly, tens of thousands of variants per haplotype detected relative to the prior GRCh38 human reference were revealed to be spurious by comparison to the T2T-CHM13 assembly. This included elimination of spurious variants in 269 medically relevant genes, reducing the number of variants for consideration in phenotypic effect by up to a factor of 12.

Expanding on this initial foundation of a complete human genome, and a pangenome from different populations, we propose a ruminant T2T (RT2T) project, not just on one species, but of multiple species in this lineage, targeting gapless diploid assemblies. We hypothesize that similar impacts will be obtained in the proposed RT2T project. We follow the open collaboration pioneered by the HT2T effort, which greatly contributed to its success and provides a model for the scientific community to foster inclusivity and effectiveness of the scientific enterprise. The goal of the present report is to publicly announce and describe the RT2T effort and invite participants from the global community interested in evolution, genome biology, and conservation thereupon to join in the project.

Rationale for the RT2T project

Ruminants have long played important roles in agriculture, particularly since the domestication of cattle, sheep, and goats over 10,000 years ago[17]. Their ability to convert cellulose forage into human-digestible protein has provided a key source of nutrition for many thousands of years. Ruminantia is a suborder of the Artiodactyla mammalian order. Domestic livestock from the Artiodactyla underpin a multibillion-dollar global industry growing to meet increasing demand of an increasing population. Later domestication or partial domestication of yaks, buffalo, bison, and some deer species present the opportunity for comparison of the effects of domestication over time. Hybridization between species, such as between cattle and gaur to produce mithun (also known as gayal), presents potential for study of the process of speciation. High-quality genome assemblies also support genetic improvement of livestock for production traits including efficiency, health, nutritive value, and sustainability/environmental impact. Genome assemblies also represent a resource for conservation efforts of endangered species by providing a basis for evaluating diversity and population viability.

There are about 359 recognized species of Artiodactlya, (Mammal Diversity Database, 2023) found in four suborders, Suina (pigs and peccaries; 21 species), Tylopoda (camels and llamas; 7 species), Whippomorpha (hippos, whales, dolphins, and porpoises; 97 species), and Ruminantia (cattle, sheep, goats, antelopes, deer, and giraffes; 234 species; Figure 1). The ruminant clade has a high species richness in terms of geographic distribution and adaptions to a wide variety of environments compared to non-human primate species. The six families in suborder Ruminantia include Tragulidae, Antilocapridae, Giraffidae, Cervidae, Moschidae, and Bovidae (Figure 1).

There is an inexact species count, due in part to controversy surrounding species delimitations, with some potential contraction of species due to collapsing currently separate species into

subspecies (for example banteng and cattle) and some potential expansion due to differing weight on the role of morphology versus genomic relationships (for example riverine and swamp buffaloes). The RT2T results could assist in resolving these controversies in addition to providing a basis for evaluating diversity and population stability.

Ruminant genome evolution is marked by the impact of horizontally transferred transposable elements (TEs), including bovine BovB LINE retrotransposons with direct effect on immune function evolution[19,20]. The original bovine genome assembly was the first mammalian assembly to have TEs annotated based on self-alignment[21,22], which proved critical for robust annotation. Ruminant pregnancy recognition is dependent on placental expression of

interferon-Tau stimulated by expression of endogenous beta-retroviruses (ERVs) that regulate placental growth and differentiation[23]. These features make the ruminant suborder an interesting model for study of the effects of horizontal transfer of TEs and endogenous retroviruses (ERVs) on the evolution of immunity and placental development.

Functional regions of chromosomes demarcated by epigenetic features and enriched for repetitive sequences define evolutionary breakpoints that contribute to chromosome rearrangements, genome instability, and reproductive isolation. Key among these regions are centromeres – a fixed location on a chromosome that carries a modified, centromere-specific histone H3, CENP-A, and nucleates the kinetochore to ensure proper chromosome segregation in mitosis and meiosis. Centromere *locations* are typically fixed within species but not between species, often representing species-specific rearrangements, but centromere *sequences* can be highly variable despite their conserved function across metazoans. In fact, centromeric DNA is among the most rapidly evolving genomic sequences in eukaryotic genomes[9,24]. The recent T2T CHM13 genome assembly highlights the requisite for T2T scale genomics to sequence, assemble, annotate, and study the highly repetitive landscape of functional centromeres[3]. Uncovering the patterns of sequence composition, methylation, and evolution, within and across species, promises to reveal rates of chromosome evolution, mechanisms of rearrangement, and the consequences of centromere divergence.

The immune systems of ruminants have revealed distinct characteristics that have been difficult to study prior to the advent of sequencing technology capable of properly assembling the germline repetitive adaptive immune loci that undergo somatic rearrangements. For example, cattle produce “ultralong” antibodies, which may have evolved as a compensation for a reduced number of antibody-encoding segments in the genome compared to other mammals. Ultralong antibodies likely play a key role in responses to bovine diseases[25] and may have potential therapeutic applications for HIV[26]. It appears that not all ruminants are able to produce these ultralong antibodies and thus the ruminant clade represents an intriguing model for examining evolution of immune systems. Study of the immune systems of ruminants is also crucial to ensure the sustainability of economically significant and endangered species and to control zoonotic diseases that can be transmitted from animals to humans.

Domestic livestock and select wild species maintained in captivity are amenable to genetic analyses that would be difficult for other species including primates. For example, fetal tissues at select developmental stages can be analyzed for chromatin status and conformation and used to annotate the genome. Moreover, correlations between gene expression and chromatin state can be investigated, including heterochromatin and other repetitive parts of the genome. The geographic distribution of ruminant species emergence and extensive adaptation to diverse environments[27] makes this suborder an excellent model for testing theories of chromosomal evolution and speciation. Inclusion of Cetacea could advance understanding of re-adaptation to the marine environment. Comparisons between domesticated species, some of which retain wild relatives, is likely to provide insight into the mechanisms and consequences of domestication. For example, genomic evidence suggests at least two separate domestication events in the history of cattle as they were separated from the ancestral aurochs population, resulting in two subspecies. Comparison of complete genomes of these subspecies and close relatives may shed light on general principles of evolution under artificial selection. The processes of chromosomal fusion/fission and centromere evolution appear to

have played out repeatedly over time in this order and complete genome comparisons can advance studies of these processes.

Genomics of ruminants

There are presently 27 ruminant species with high-quality chromosomal-level assemblies that meet the Vertebrate Genomes Project (VGP) N50 metrics (contig N50 > 1 Mb; scaffold N50 > 10 Mb) [7] and at least 60 species with draft assemblies below these metrics (Supplementary Table 1). The reference assemblies are not always the most contiguous or highest quality achieved for a species, but they represent the assembly agreed upon by the research community as a standard to allow across-study comparisons. In some cases, the listed reference assembly in GenBank has a modest contiguity compared to other assemblies in the database. For example, the yak, where the reference (RefSeq accession GCF_000298355.1)[28] has a contig N50 of 22.8 kilobase pairs (Kbp), has three other assemblies of contiguity with contig N50 as high as 72.3 Mbp (accession *GCA_009493645.1*)[29]. To our knowledge there are no T2T genome assemblies for species in the ruminant clade in the GenBank or ENA databases.

Ruminants generally have higher numbers of acrocentric or telocentric chromosomes than primates. For example, compared to the five acrocentric chromosomes in humans, cattle have 29 acrocentric/telocentric autosomes and sheep have 23 acrocentric/telocentric and 3 submetacentric autosomes. Remarkably, two species in the genus *Muntiacus* have very different chromosome numbers, with the Indian muntjac (*Muntiacus muntjac*) having 2n=6 in females (2n=7 in males) and the Chinese muntjac (*Muntiacus reevesi*) with 2n=46[30]. Dama gazelles (*Nanger dama*) show different numbers of diploid chromosomes or cytotypes within subspecies, ranging from 2n=38 to 40, due to fusions between autosomes and sex chromosomes[31]. These variations within the suborder further support the value of ruminants for the study of mammalian speciation.

Domesticated livestock species retain substantial within and between-breed diversity despite the genetic bottlenecks created during domestication and breed formation[32]. This diversity has been exploited to drive genetic improvement for production traits, accelerated by genomic characterization in the past 15 years. Some wild species have experienced historical population bottlenecks, with some such as moose resulting in very low levels of heterozygosity[33] and others like American bison surviving with surprisingly high diversity[34]. Homozygosity represents an important concern for managing endangered populations and high-quality genome assemblies facilitate efforts to conserve and manage genetic diversity[35,36]

Generating complete genomes of the order Artiodactyla:

The HT2T effort began with the selection of a hydatidiform mole cell line (CHM13) for sequencing. The CHM13 genome resulted from a haploid 23,X sperm that duplicated upon fertilization with the maternal complement lost. The result was a 46,XX genome entirely homozygous, with all of its material inherited from the male. The reduced single-haplotype genome simplified the assembly process, but still involved substantial manual efforts to close gaps and bridge highly repetitive regions. Generating such haploid cell lines is not possible for large-scale projects across phylogenies nor were the initial algorithms optimal for generating assemblies from true diploid samples. Advances in sequencing technology and algorithms have supplanted some of the strategies used for the CHM13 assembly and support complete genome assembly of diploid samples[7-11]. The trio binning approach of resolving haplotypes by

using sequences unique to each parent[37] is considered a “best practice” approach to resolve a haplotype genome assembly, but it is unlikely that samples from both parents would be available for many of the species we wish to target in the RT2T effort. Alternatively, the use of a type of chromatin conformation capture such as Hi-C or Pore-C can substitute for parental data if dam and sire samples are not available[8], allowing the derivation of haplotype phased assemblies.

The pipeline for generation of RT2T assemblies must have flexibility to accommodate available tissues and samples. This may necessarily include some relaxation of the definition of “complete” genome, where insufficient or suboptimal material is available to provide sufficient coverage to close all gaps. The project plans to follow the current guidelines of the T2T consortium, particularly thus far tested on human and non-human primates (in prep), by generating 50x coverage (25x per haplotype) of high accuracy HiFi Pacific Biosciences (PacBio) long reads in the range 18-20 Kbp in length, 50x coverage (25x per haplotype) of “ultra-long” Oxford Nanopore (UL-ONT) reads exceeding 100 Kbp, and 50x Hi-C short reads. Of note, this recipe continues to evolve as lengths increase in UL-ONT data. This recipe requires a relatively large amount of high-quality tissue or blood sample, or cell lines, particularly for ONT-UL. DNA generally works best if extracted shortly before use. Males are preferred to represent the sex chromosomes. When parental samples can be accessed, haplotypes will be resolved by sequencing the parents to identify parent-specific markers. With the trio binning approach, these markers are used to separate sequence reads derived from the progeny by parent of origin for independent haplotype assembly. When collection of both parental samples is not feasible, HiC data generated from an additional sample of the individual can be used to phase haplotypes. The goal is to produce two assemblies per individual containing completely haplotype-assigned, chromosome-length, contigs with telomeric sequences at each end and no gaps (i.e., T2T). However, the experience of the Human Pangenome Consortium suggests that gaps solely related to repeat copy number near the tips of chromosomes and within rDNA arrays require a large manual effort to resolve and may be ignored for the purposes of the project. Close to complete, “near T2T” assemblies would be considered for inclusion in the project if no gaps exist outside of these regions. Fortunately, improvements in both sequencing technology (e.g. high accuracy nanopore sequencing data) and algorithms for assembly (e.g. Verkko[10] and hifiasm[38]) in the recent past have overcome obstacles to true diploid T2T assembly and replaced much of the need for manual intervention, yet still require substantial expertise to achieve gapless, haplotype-resolved assemblies from diploid samples.

The assembly and curation WG face a substantial challenge and represent one potential bottleneck in the pursuit of the goals of the RT2T project. It is common for the assembly graphs to contain “tangles” representing repetitive regions where the assembler is not able to confidently choose the path through the alternative nodes in the graph. Frequently, manual inspection can identify a clear best path or recognize that the supporting data is insufficient to resolve the tangle. In addition, quality control steps are necessary to identify problems such as sequence errors, collapsed duplications, and mis-assemblies. For example, circular RNAs (circRNAs) are formed when a splicing donor and acceptor are linked upstream of a linear RNA in a process named backsplicing[39]. The identification of apparent circRNAs can result from problems in the genome assembly and represent an orthogonal measure of assembly or annotation accuracy. We welcome researchers willing to contribute by using various evaluation methods for moving the assemblies from “near complete” to T2T status.

The RT2T project intends to submit working drafts of each species' assembly to the GenomeArk database (https://genomeark.github.io/) upon sufficient curation that it represents a stable version, barring request for delayed release by the sample provider(s). The assemblies are intended for free use by any researchers interested in genomic analysis of that species. The RT2T WG will enhance annotation performed in collaboration with NCBI to annotate the genomes with gene information, identify centromeres, and characterize repeat classes and distribution, thus we encourage individuals with interest in those activities to participate with these WG rather than pursue separate efforts. At present, we have generated T2T assemblies of seven ruminant species including cattle *Bos taurus*, gaur *Bos gaurus*, sheep *Ovis aires*, goat *Capra hircus*, bighorn sheep *Ovis canadensis*, roan antelope *Hippotragus equinus*, and sable antelope *Hippotragus niger*. These assemblies include the first complete sex chromosomes of these species reported. Statistics of the submitted assemblies are shown in Table 1. When final versions are fully curated, they will be submitted to GenBank and EBI databases for processing.

Each individual sequenced will produce two autosomal haplotype assemblies that can be compared. Crosses between breeds are optimal for expanding haplotype diversity of the data. However, generally only one haplotype per breed of each sex chromosome is produced. Ongoing assemblies of riverine buffalo (*Bubalus bubalis*), American bison (*Bison bison*), moose (*Alces alces*), dama gazelle (*Nanger dama*), scimitar-horned oryx (*Oryx dammah*), Eld's deer (*Rucervus eldii*), domestic pig (*Sus scrofa*), Indian muntjac (*Muntiacus muntjac),*and musk ox (*Ovibos moschatus*) are at various stages of assembly and have underscored the need for flexibility of the pipeline to accommodate sample types and variation in sample characteristics between species. A list of proposed species is illustrated in Figure 1. Initial analyses have provided some unexpected results, for example the rRNA gene clusters do not reside on the short p-arm of any of the chromosomes, instead lie closer to the distal end of the q-arm (data not shown). Therefore, it is unlikely that a mechanism of recombination between non-homologous acrocentric chromosomes is mediated by the rRNA gene clusters as proposed for humans[13].

Assembly	Haploid chr #	Sex chromosome	Length (Gb)	QV	Contig N50 (Mb)	Scaffold N50 (Mb)	NT2T	T2T scaffolds	T2T contigs (ideal)
Kiko goat	30	X	2.88	65.2	101.5	101.5	5	3	19
Saanen goat	30	X	2.89	65.2	100.9	100.9	6	4	17
Bighorn	27	Y	2.83	65.4	103.3	103.3	2	3	15
Polypay sheep	27	X	2.98	65.2	94.8	94.8	20	2	1
Gaur	29	X	3.13	52.2	88.2	108.5	25	1	1
Piedmontese cattle	30	X	3.20	55.6	112.3	112.3	17	1	10
Roan antelope	30	X	2.65	56.1*	106.1	106.1	0	0	30
Sable antelope	30	X	2.65	56.1*	102.5	102.5	0	0	30

Table 1: Current T2T assemblies and statistics released by members of the RT2T project. QV is calculated with merqury[40] using trio short-read sequence. A chromosome is defined as Near T2T (NT2T) if it is acrocentric, gaps are from unresolved graph tangles (sequence content is known but not resolved), and it starts at the centromere and ends at the telomere or an rDNA array. A chromosome is T2T scaffold when there are telomeres on both ends and the only gaps are from unresolved graph tangles. A chromosome is T2T contig (the ultimate goal) when there are telomeres on both ends and no gaps. *QV calculated from ONT duplex data.

Annotation will be supported through generation of full-length transcriptome sequence for available tissues. Data in public databases will be included to identify expressed regions of the genome. Computational gene predictions will also be carried out to identify predicted transcripts not observed in the available RNA sequence data. Focus will be given broadly to genes related to reproduction, development, lactation, ruminant digestion, innate and adaptive immunity, histone deacetylation, and methyl transferases, along with imprinted genes, genes within segmental duplications, retrogenes, and genes within heterochromatin regions. Annotation of transposable elements (TEs) will be emphasized through comprehensive, curated identification of TEs and ERVs and will be evaluated as drivers of structural variation, horizontal gene transfer, genome evolution, gene regulation, and adaptation.

Sample acquisition has been the most challenging aspect of the RT2T project to date and we seek any researchers with access to appropriate samples that are found in this table to become part of the consortium. In some instances, multiple samples from a species could be considered, where subspecies boundaries present potentially important context for evolutionary studies as aforementioned for muntjacs and dama gazelles. It is important to note that generating and curating the assemblies is only step one of the intensive processes of genome analysis and comparative analyses. The goals of the RT2T consortium are to use comparative analyses across the suborder to identify mechanisms of chromosomal evolution (both autosomes and sex chromosomes) and the impacts of selection/environmental pressures on genome sequence and structure. We request that comparative genomic studies that make use of the T2T assemblies we produce be reserved for the consortium, while reiterating that studies within species are encouraged and not in violation of fair use guidelines.

Proposed comparative studies

The RT2T consortium WG of volunteers are dedicated to sample acquisition and processing, sequencing, assembly, curation, cytogenetics, and annotation to prepare the complete genome assemblies for public use. Specific efforts to characterize heterochromatin regions and genomic repeats, non-coding RNA components, small RNAs, centromeres/kinetochores, and TEs are planned. These activities will contribute to the final, annotated, complete genome assemblies. The consortium will make internal decisions on data freezes and the point(s) at which comparative analyses will commence. We recognize that interesting results could be obtained by similar analyses with partial sets of T2T genomes prior to completion of the full set of species but would then dilute the impact of the broader studies envisioned. We request that non-consortium researchers refrain from using these assemblies for similar analyses without written permission from the consortium leadership. We welcome requests from interested researchers to participate within the consortium. A list of WG and lead contacts are found in Table 2 with specific analyses planned described below.

Working Group Name	Working Group Leaders	Contact information
3D genome architecture	Darren Hagen, Brenda Murdoch	darren.hagen@okstate.edu, bmurdoch@uidaho.edu
Annotation	Christine Elsik, James Koltes	elsikc@missouri.edu, jekoltes@iastate.edu
Assemblers and curation	Ben Rosen, Tim Smith	ben.rosen@usda.gov, tim.smith2@usda.gov
Chromosome evolution	Rachel O'Neill, Temitayo Olagunju	rachel.oneill@uconn.edu, tolagunju@uidaho.edu
Comparative methylome	Stephanie McKay, Brenda Murdoch	stephanie.mckay@missouri.edu, bmurdoch@uidaho.edu
Cytogenetics	Tamara Potapova	tpo@stowers.org
Immunogenomics	Yana Safonova, Corey Watson	yana@psu.edu, corey.watson@louisville.edu
Repeat Annotation	Rachel O'Neill, Jessica Storer	rachel.oneill@uconn.edu, jessica.storer@uconn.edu
Variant Discovery and Population Sequencing	Robert D. Schnabel, George Liu	schnabelr@missouri.edu, george.liu@usda.gov

Table 2: Working groups, their respective co-leaders, and contact information for the co-leaders for those with questions, or who wish to contribute.

The chromosome evolution and 3D genome architecture WGs will undertake comparative analysis of chromosomal and centromere structure and evolution. Eukaryotic genomes are not simply a linear compilation of coding and noncoding sequences; rather, genomes are organized into three dimensions that not only define gene regulatory domains[41] and contribute to genome stability[42], but also provide signatures of regulatory evolution[43] over long evolutionary time frames[44-46]. Regulatory elements can be positioned up to millions of base pairs away from the genes they regulate[47] and several levels of organization can be defined including chromosome compartments, topologically-associated domains (TADs), and loops. There are two categories of compartments, open chromatin in A compartments and closed chromatin in B compartments[48]. Up to 36% of the mammalian genome undergoes compartment changes during development[49] and compartments are tissue-specific[48]. TADs display characteristic self-interactivity with boundaries frequently indicated by the presence of CTCF[50]. Loops are generally smaller than TADs and have been shown to be a mechanism for regulation of gene expression, with disruption of loops causing dysregulation associated with disease phenotypes[51].

The 3D genome WG will annotate compartments, TADs, and loops for comparative analyses using chromatin conformation contact assays (Hi-C, Pore-C and/or Micro-C). A previous 3D genome study of carnivore species[52] using Hi-C reported broad conservation at the level of whole chromosomes across three families separated by 54 My since the last common ancestor. High consistency of TADs and compartments in liver samples across livestock species including chickens, pigs, and goats has also been reported[53]. There is little literature describing 3D conformation in and between ruminant species, and tissue and developmental stage-specific aspects of compartments, TADs, and loops complicate comparative studies. The goal of the 3D genome WG is to compare chromatin structure across tissues within and between species where tissue source and developmental stage are similar. This information will be used to annotate the T2T genome assemblies with structural information for all tissues/cell lines used. These analyses will provide additional information for predicting effects of genomic variants, including structural variants and sequence polymorphisms, on phenotype and adaptation. An important contribution of RT2T genome assemblies is to enable a comprehensive analysis of structural variation (SV) among species and within populations. SVs are known to have a significant impact on genome 3D organization and may impact genome function through this reshaping of the 3D structure[54-56]. The study of the relationship between SVs and genome 3D structure will therefore benefit from the different ruminant T2T genome assemblies.

The chromosome evolution WG will use chromosome-wide aspects of 3D genome assays as well as sequence content to examine the evolution of chromosomes since the last common ancestor of the suborder Ruminantia and the order Artiodactyla. Recent work in model species spanning the metazoan phylogeny (human, mouse, *Drosophila*, yeast) has shown that TAD boundaries define evolutionarily conserved gene expression patterns and that lineage-specific rearrangements in response to selection are enriched at TAD boundaries[57]. In a recent study using reconstructed ancestral karyotypes of Artiodactyls, Ruminants, Pecorans, and Bovids, evolutionary breakpoints defining chromosome rearrangements among species were found to be enriched for sequences associated with active or lineage-specific TEs and genes with divergent gene expression patterns[58]. Thus, TADs and linear chromosome organization are implicated in defining gene expression regulatory patterns likely by delineating the regions in which genes interact through insulator activities. Although cell-type specific TAD boundaries within an organism may be variable [59], TADs shared across all tissues are stable across cell types and appear enriched for heritability of complex multigenic traits and evolutionary constraint[45]. Organismal TAD boundaries are linked to chromosome rearrangements, repeat expansions[60] and epigenetic signatures (e.g., DNA methylation[61] and histone modification[62]) and are enriched for adaptive structural variants[57]. While TADs are considered synonymous with regions of conserved synteny and constrained gene regulation, genome organization across mammals beyond model systems is largely unexplored. Moreover, how TAD organization and boundaries imbue constraint on adaptation is unknown. The ability to derive methylation and TAD organizational information from data used in the generation of long-read based genome assemblies affords an unprecedented opportunity in the context of ruminant genome biology and chromosome evolution.

Fixed chromosomal rearrangements among species may be an important driver of species evolution by contributing to species-specific gene regulation patterns, genome organization, selfish element activity, recombination patterns, and faithful Mendelian inheritance. Ruminant species carry the broadest range of chromosome complements among mammals – ranging in chromosome number from the smallest of any mammal, 2n=6/7 in the Indian muntjac[30], to 2n=70 found in many deer species. In this regard, many ruminant species are distinguished by extensive chromosome rearrangements (fissions, fusions, translocations, centric shifts), multiple sex chromosome complements (e.g.,

XX/XYY), and the presence of potentially meiotically driven selfish B chromosomes (several brocket deer species[genus *Mazama*], Siberian roe, and Siberian musk deer)[63]. The chromosome evolution WG aims to perform comparative analysis of chromosome structure, including centromere and kinetochore sequence and position, to reveal the evolutionary pathways leading from the last common ancestor in the order Artiodactyla to the existing species across all ruminant genera and closely related species with widely divergent karyotypes. Among the primary goals are deriving an ancestral karyotype, defining ruminant evolutionary breakpoints, and discriminating general mechanisms of chromosome structure evolution across mammals from clade-specific features. This will entail close interaction between the chromosome evolution WG and the cytogenetics WG, clarifying ambiguous karyotypes, such as in dama gazelle and in species for which cell lines are available or can be established. Fluorescence *in situ* hybridization (FISH) on chromosomal spreads will be used to validate the locations of specific DNA sequences on chromosomes, which can be particularly useful for hard-to-assemble regions such as satellite repeats and rDNA gene arrays. For example, gaps within rDNA gene arrays[4] and tandem repeat arrays of nearly-identical copies can be resolved using high-resolution FISH.

The annotation WG will undertake comparative analysis of gene family contractions and expansions, and identification of targets of positive selection. Correlations of transcript abundance and polymorphisms of cis-regulatory elements can identify expression quantitative trait loci (eQTL) and illuminate principles of functional biodiversity and consequences to evolutionary development, selection, and adaptive responses. Gene retrocopies[64], resulting from reverse transcription and genomic insertion of spliced mRNA by LINE-1 retrotransposition[20,65], will be evaluated for evolution in non-coding pseudogenes[66]. The polled and fleece type traits in sheep represent examples of the impact of retrogenes[16]. Retrocopies identified with RetroScan[67] will provide preliminary classification of retrocopies. Ka/Ks ratios will provide their estimated age distribution.

The annotation WG has also planned specific comparative analysis of lactation-related genes to examine the evolution of genes that regulate milk synthesis and variation in milk constituents. There is a wealth of transcriptome datasets in cattle, buffalo, and sheep related to the mammary gland, and milk represents a rich, non-invasive source of RNA, including non-coding sno-RNA, miRNA, lnc-RNA, and mRNA[68-70]. Both are part of epithelial cells present in milk and within cytoplasmic droplets encapsulated in milk fat globules during apocrine secretion [71-75]. Planned analyses will capitalize on data from public repositories or sequence from milk samples of non-agricultural ruminant species.

One outcome of the human T2T project was the identification of additional genes in the newly assembled and corrected portions of the genome, most of which corresponded to predicted non-coding RNAs. There is limited knowledge of their organization across species, function, and evolution. The annotation WG will use data generated in the project and public transcriptome datasets to annotate non-coding RNA genes, particularly in previously unassembled portions of genomes, to provide new understanding of the evolution and activity of these little-understood genes. Comparative study across the ruminants and correlation of gene expansion/contraction with other aspects of genome biology will provide new insights into the role of non-coding RNAs in genome function and evolution.

The immunogenomics WG has a focus on expressed adaptive immune gene repertoires (antibody repertoires and T-cell receptor repertoires), some of which are in germline loci encoded through somatic genomic recombination[76]. These genomic regions have been difficult to examine prior to the advent of 3rd generation sequencing technology and assembly methods. The WG will annotate germline genes encoding antibodies and T-cell repertoires, identify their eQTL characteristics using expressed repertoire-sequencing data (AIRR-Seq), and perform comparative analyses to reveal species-specific adaptations of ruminant adaptive immune systems related to environment, pathogen exposure, and domestication. The comparative analysis will also make it possible to investigate the evolutionary origin of the ultralong antibodies. Previous studies show that such antibodies are partially encoded by unusually long immunoglobulin diversity (D) and joining (J) genes, selecting one V gene, one D gene, and one J gene and concatenating them together to generate the variable region of a heavy or light chain of the antibody[77]. The resulting VDJ sequences are further diversified by somatic hypermutations and gene conversion. Recent studies showed that cattle cysteine-rich ultralong antibodies likely play a key role in responses to bovine respiratory disease[25]. Orthologs of *IGHD8-2* were found in genomes of cows and its close relatives (e.g., zebu, American bison, gayal), suggesting that ultralong antibodies are common for some bovines[78,79]. A preliminary analysis of existing ruminant genomes performed by the immunogenomics WG revealed *IGHD8-2-like* genes in red deer and giraffe, suggesting that ultralong antibodies either emerged earlier in the ruminant lineage or resulted from convergent evolution (Table 3). The WG will explore the role of ultralong antibodies in immune responses and their therapeutic potential and invite collaborators studying ruminant diseases and developing antibody-based drugs. The WG will also collaborate with immunogenomics societies to deposit AIRR-Seq data in a standardized and open-to-public manner.

The Comparative methylome (epigenetic) WG will make use of the latest sequencing technologies that provide information on 5-methylcytosine (5mC) base modification in the genome, associated with gene regulation. These methylation patterns are generated simultaneously in the HiFi and ONT-UL reads base calls[80-85]. A complete T2T assembly will subsequently yield an accompanying T2T methylome for ruminant species and will assist in realizing the extent to which 5mC influences gene expression, genome regulation, and genome stability. Initially, patterns of 5mC will be characterized throughout the genomes including previously unresolved genomic regions such as rDNA arrays and centromeric regions. Subsequently, comparative epigenomics will be employed to discern molecular insights into domestication and selection as has been studied in dogs and fish[86,87]. Of particular interest is investigation of the change in 5mC over evolutionary distance among species within the suborder Ruminantia. Understanding the epigenetic mechanisms altered as a result of domestication and selection may inform the agricultural genomics community of the potential for marker assisted selection of epigenetically induced phenotypes.

Species	Gene length (nt)	Amino acid translation
Cow (*Bos taurus*)	148	SCPDGYSYGYGCGYGYGCSGYDCYGYGGYGGYGGYGYSSYSYSYTYEY
Red deer (*Cervus elaphus*)	117	YCYSSSSGYYDCSSGYYDCCGSSSYYGYCGSSYYSYYG
Giraffe (*Giraffa camelopardalis*)	104	CHSSSCRSGYSSGYGCRSGYGYGYSYGYGYGCCG

Table 3: Ultralong D genes of three ruminant species: cow, red deer, and giraffe. The corresponding amino acid sequence is shown in the “Amino acid translation” column.

A subset of species used for agriculture, including cattle, sheep, goat, and American bison, present an opportunity to obtain samples from specific developmental stages for T2T assembly and enhanced annotation. Fetal tissues from these species collected after secondary myogenesis were used for genome sequencing, transcriptome profiling with long and short reads, chromatin conformation contact analysis, and methylome characterization. Multi-tissue comparative analysis will be generally confined to these species since similar samples from many of the ruminant species, including some critically endangered, are neither practical nor possible. However, where cell lines can be obtained/created from fibroblast cells, a parallel comparative effort characterizing gene expression, 3D architecture, and DNA modification is planned.

A large amount of population-level SNP-chip and sequence data exists for agricultural species, as well as for some wild species. The Variant Discovery and Population Sequencing (variant) WG has the goal of determining the impact of T2T-level assemblies on the use of short and long read-based variant identification and genotyping. Ruminants used in agriculture have been the subject of projects modeled after the human “1000 Genomes” project utilizing medium or high-density SNP chips, and more recently whole genome shotgun (WGS) sequence data. These resources have been used to establish association with more or less detailed phenotypes. The variant WG has plans to evaluate the impact of T2T-level assemblies on the use of short and long read-based variant identification and genotyping compared to current reference assemblies. Additionally, for species with a sufficient amount of population data available, the variant WG intends to produce standardized resources to enable researchers to use the respective T2T assemblies, thus enabling a transition from previous references to the newly produced T2T assemblies.

Population-level WGS of endangered species of ruminants will enhance understanding of the distribution of genome-wide variation, inbreeding through analyses of the amounts and distribution of runs of homozygosity, and the burden of masked and realized genetic load within the context of the declining populations that often characterize such species. Such information can be incorporated into conservation management programs that seek to ensure the long-term sustainability of endangered species [88].

Conclusion

The RT2T consortium has the ambitious goal to create and annotate complete genomes for as many species in the suborder Ruminantia as possible, along with additional species in the Artiodactyla to enhance the comparative genomic analyses of chromosome structure and evolution. The approach goes beyond that of the current minimum metrics of the VGP[7] or the Earth Biogenome Project (EBP)[89] in the targeting of T2T-level assemblies. The analysis of the complete genomes across the evolutionary history of ruminants will provide the first look at chromosome evolution of this group that encompasses the repeat-dense portions of the genome. The assemblies will contribute to the goals of the broader VGP and EBP projects while having the potential to perform analyses and enabling studies that will not be possible with the current minimum assembly criteria in those projects. We present the initial efforts of the RT2T project, with the assemblies available at the GenomeArk database and including the first complete assemblies of livestock X and Y chromosomes.

The purpose of the current communication is to present the project, its goals, and to announce the intended use of the assemblies produced by the consortium. The RT2T project welcomes participation of the international community and encourages interested researchers to contact us. This is an exciting time in the field of genomics and large, open collaborations appear to be the most productive way to advance the field.

References

1. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44-53 (2022).

2. Gershman, A. *et al.* Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).

3. Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).

4. Hoyt, S.J. *et al.* From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).

5. Vollger, M.R. *et al.* Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).

6. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).

7. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737-746 (2021).

8. Cheng, H. *et al.* Haplotype-resolved assembly of diploid genomes without parental data. *Nat Biotechnol* **40**, 1332-1335 (2022).

9. Jarvis, E.D. *et al.* Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519-531 (2022).

10. Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* **41**, 1474-1482 (2023).

11. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344-354 (2023).

12. Bernstein, B.E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

13. Guarracino, A. *et al.* Recombination between heterologous human acrocentric chromosomes. *Nature* **617**, 335-343 (2023).

14. Liao, W.W. *et al.* A draft human pangenome reference. *Nature* **617**, 312-324 (2023).

15. Jarmuz-Szymczak, M., Janiszewska, J., Szyfter, K. & Shaffer, L.G. Narrowing the localization of the region breakpoint in most frequent Robertsonian translocations. *Chromosome Res* **22**, 517-32 (2014).

16. Wiedemar, N. & Drogemuller, C. A 1.8-kb insertion in the 3'-UTR of RXFP2 is associated with polledness in sheep. *Anim Genet* **46**, 457-61 (2015).

17. Alberto, F.J. *et al.* Convergent genomic signatures of domestication in sheep and goats. *Nat Commun* **9**, 813 (2018).

18. Zurano, J.P. *et al.* Cetartiodactyla: Updating a time-calibrated molecular phylogeny. *Mol Phylogenet Evol* **133**, 256-262 (2019).

19. Walsh, A.M., Kortschak, R.D., Gardner, M.G., Bertozzi, T. & Adelson, D.L. Widespread horizontal transfer of retrotransposons. *Proc Natl Acad Sci U S A* **110**, 1012-6 (2013).

20. Ivancevic, A.M., Kortschak, R.D., Bertozzi, T. & Adelson, D.L. Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. *Genome Biol* **19**, 85 (2018).

21. Bovine Genome, S. *et al.* The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**, 522-8 (2009).

22. Adelson, D.L., Raison, J.M. & Edgar, R.C. Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proc Natl Acad Sci U S A* **106**, 12855-60 (2009).

23. Dunlap, K.A. *et al.* Endogenous retroviruses regulate periimplantation placental growth and differentiation. *Proc Natl Acad Sci U S A* **103**, 14390-5 (2006).

24. Melters, D.P. *et al.* Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol* **14**, R10 (2013).

25. Safonova, Y. *et al.* Variations in antibody repertoires correlate with vaccine responses. *Genome Res* **32**, 791-804 (2022).

26. Sok, D. *et al.* Rapid elicitation of broadly neutralizing antibodies to HIV by immunization in cows. *Nature* **548**, 108-111 (2017).

27. Gilbert, M. *et al.* Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. *Sci Data* **5**, 180227 (2018).

28. Qiu, Q. *et al.* The yak genome and adaptation to life at high altitude. *Nat Genet* **44**, 946-9 (2012).

29. Rice, E.S. *et al.* Chromosome-length haplotigs for yak and cattle from trio binning assembly of an F1 hybrid. *bioRxiv* (2019).

30. Wurster, D.H. & Benirschke, K. Indian muntjac, Muntiacus muntjak: a deer with a low diploid chromosome number. *Science* **168**, 1364-6 (1970).

31. Vassart, M., Seguela, A. & Hayes, H. Chromosomal evolution in gazelles. *J Hered* **86**, 216-27 (1995).

32. Bovine HapMap, C. *et al.* Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* **324**, 528-32 (2009).

33. Kalbfleisch, T.S. *et al.* A SNP resource for studying North American moose. *F1000Res* **7**, 40 (2018).

34. Cherry, S.G., Merkle, J.A., Sigaud, M., Fortin, D. & Wilson, G.A. Managing Genetic Diversity and Extinction Risk for a Rare Plains Bison (Bison bison bison) Population. *Environ Manage* **64**, 553-563 (2019).

35. Theissinger, K. *et al.* How genomics can help biodiversity conservation. *Trends Genet* **39**, 545-559 (2023).

36. Paez, S. *et al.* Reference genomes for conservation. *Science* **377**, 364-366 (2022).

37. Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol* (2018).

38. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170-175 (2021).

39. Lasda, E. & Parker, R. Circular RNAs: diversity of form and function. *RNA* **20**, 1829-42 (2014).

40. Rhie, A., Walenz, B.P., Koren, S. & Phillippy, A.M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245 (2020).

41. Krefting, J., Andrade-Navarro, M.A. & Ibn-Salem, J. Evolutionary stability of topologically associating domains is associated with conserved gene regulation. *BMC Biol* **16**, 87 (2018).

42. Sarni, D. *et al.* 3D genome organization contributes to genome instability at fragile sites. *Nat Commun* **11**, 3613 (2020).

43. Eres, I.E., Luo, K., Hsiao, C.J., Blake, L.E. & Gilad, Y. Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. *PLoS Genet* **15**, e1008278 (2019).

44. Torosin, N.S., Anand, A., Golla, T.R., Cao, W. & Ellison, C.E. 3D genome evolution and reorganization in the Drosophila melanogaster species group. *PLoS Genet* **16**, e1009229 (2020).

45. Fudenberg, G. & Pollard, K.S. Chromatin features constrain structural variation across evolutionary timescales. *Proc Natl Acad Sci U S A* **116**, 2175-2180 (2019).

46. Lazar, N.H. *et al.* Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. *Genome Res* **28**, 983-997 (2018).

47. Hafner, A. & Boettiger, A. The spatial organization of transcriptional control. *Nat Rev Genet* **24**, 53-68 (2023).

48. Fortin, J.P. & Hansen, K.D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol* **16**, 180 (2015).

49. Dixon, J.R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-6 (2015).

50. Oudelaar, A.M. & Higgs, D.R. Publisher Correction: The relationship between genome structure and function. *Nat Rev Genet* **22**, 808 (2021).

51. Behrends, M. & Engmann, O. Loop Interrupted: Dysfunctional Chromatin Relations in Neurological Diseases. *Front Genet* **12**, 732033 (2021).

52. Corbo, M., Damas, J., Bursell, M.G. & Lewin, H.A. Conservation of chromatin conformation in carnivores. *Proc Natl Acad Sci U S A* **119**(2022).

53. Foissac, S. *et al.* Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biol* **17**, 108 (2019).

54. Spielmann, M., Lupianez, D.G. & Mundlos, S. Structural variation in the 3D genome. *Nat Rev Genet* **19**, 453-467 (2018).

55. Shanta, O., Noor, A., Human Genome Structural Variation, C. & Sebat, J. The effects of common structural variants on 3D chromatin structure. *BMC Genomics* **21**, 95 (2020).

56. Anania, C. & Lupianez, D.G. Order and disorder: abnormal 3D chromatin organization in human disease. *Brief Funct Genomics* **19**, 128-138 (2020).

57. Liao, Y., Zhang, X., Chakraborty, M. & Emerson, J.J. Topologically associating domains and their role in the evolution of genome structure and function in Drosophila. *Genome Res* **31**, 397-410 (2021).

58. Farre, M. *et al.* Evolution of gene regulation in ruminants differs between evolutionary breakpoint regions and homologous synteny blocks. *Genome Res* **29**, 576-589 (2019).

59. McArthur, E. & Capra, J.A. Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *Am J Hum Genet* **108**, 269-283 (2021).

60. Sun, J.H. *et al.* Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries. *Cell* **175**, 224-238 e15 (2018).

61. Du, Q. *et al.* DNA methylation is required to maintain both DNA replication timing precision and 3D genome organization integrity. *Cell Rep* **36**, 109722 (2021).

62. Liu, C., Cheng, Y.J., Wang, J.W. & Weigel, D. Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. *Nat Plants* **3**, 742-748 (2017).

63. Vujosevic, M., Rajicic, M. & Blagojevic, J. B Chromosomes in Populations of Mammals Revisited. *Genes (Basel)* **9**(2018).

64. Brosius, J. Retroposons--seeds of evolution. *Science* **251**, 753 (1991).

65. Kelly, C.J., Chitko-McKown, C.G. & Chuong, E.B. Ruminant-specific retrotransposons shape regulatory evolution of bovine immunity. *Genome Res* **32**, 1474-86 (2022).

66. Casola, C. & Betran, E. The Genomic Impact of Gene Retrocopies: What Have We Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses? *Genome Biol Evol* **9**, 1351-1373 (2017).

67. Wei, Z. *et al.* RetroScan: An Easy-to-Use Pipeline for Retrocopy Annotation and Visualization. *Front Genet* **12**, 719204 (2021).

68. Chen, Y. *et al.* Effect of high-fat diet on secreted milk transcriptome in midlactation mice. *Physiol Genomics* **49**, 747-762 (2017).

69. Mumtaz, P.T. *et al.* Mammary epithelial cell transcriptome reveals potential roles of lncRNAs in regulating milk synthesis pathways in Jersey and Kashmiri cattle. *BMC Genomics* **23**, 176 (2022).

70. Munch, E.M. *et al.* Transcriptome profiling of microRNA by Next-Gen deep sequencing reveals known and novel miRNA species in the lipid fraction of human breast milk. *PLoS One* **8**, e50564 (2013).

71. Maningat, P.D. *et al.* Gene expression in the human mammary epithelium during lactation: the milk fat globule transcriptome. *Physiol Genomics* **37**, 12-22 (2009).

72. Lemay, D.G. *et al.* Sequencing the transcriptome of milk production: milk trumps mammary tissue. *BMC Genomics* **14**, 872 (2013).

73. Suarez-Vega, A., Gutierrez-Gil, B., Klopp, C., Tosser-Klopp, G. & Arranz, J.J. Variant discovery in the sheep milk transcriptome using RNA sequencing. *BMC Genomics* **18**, 170 (2017).

74. Beckett, L. *et al.* Mammary transcriptome reveals cell maintenance and protein turnover support milk synthesis in early-lactation cows. *Physiol Genomics* **52**, 435-450 (2020).

75. Arora, R. *et al.* Buffalo milk transcriptome: A comparative analysis of early, mid and late lactation. *Sci Rep* **9**, 5993 (2019).

76. Tonegawa, S. Somatic generation of antibody diversity. *Nature* **302**, 575-81 (1983).

77. Wang, F. *et al.* Reshaping antibody diversity. *Cell* **153**, 1379-93 (2013).

78. Smider, B.A. & Smider, V.V. Formation of ultralong DH regions through genomic rearrangement. *BMC Immunol* **21**, 30 (2020).

79. Ott, J.A. *et al.* Evolution of immunogenetic components encoding ultralong CDR H3. *Immunogenetics* **75**, 323-339 (2023).

80. Clark, T.A. *et al.* Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res* **40**, e29 (2012).

81. Lee, W.C. *et al.* The complete methylome of Helicobacter pylori UM032. *BMC Genomics* **16**, 424 (2015).

82. Payelleville, A. *et al.* The complete methylome of an entomopathogenic bacterium reveals the existence of loci with unmethylated Adenines. *Sci Rep* **8**, 12091 (2018).

83. Rand, A.C. *et al.* Mapping DNA methylation with high-throughput nanopore sequencing. *Nat Methods* **14**, 411-413 (2017).

84. Simpson, J.T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**, 407-410 (2017).

85. Tvedte, E.S. *et al.* Comparison of long-read sequencing technologies in interrogating bacteria and fly genomes. *G3 (Bethesda)* **11**(2021).

86. Janowitz Koch, I. *et al.* The concerted impact of domestication and transposon insertions on methylation patterns between dogs and grey wolves. *Mol Ecol* **25**, 1838-55 (2016).

87. Konstantinidis, I. *et al.* Major gene expression changes and epigenetic remodelling in Nile tilapia muscle after just one generation of domestication. *Epigenetics* **15**, 1052-1067 (2020).

88. Guhlin, J. *et al.* Species-wide genomics of kakapo provides tools to accelerate recovery. *Nat Ecol Evol* **7**, 1693-1705 (2023).

89. Lewin, H.A. *et al.* Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci U S A* **115**, 4325-4333 (2018).

Figures

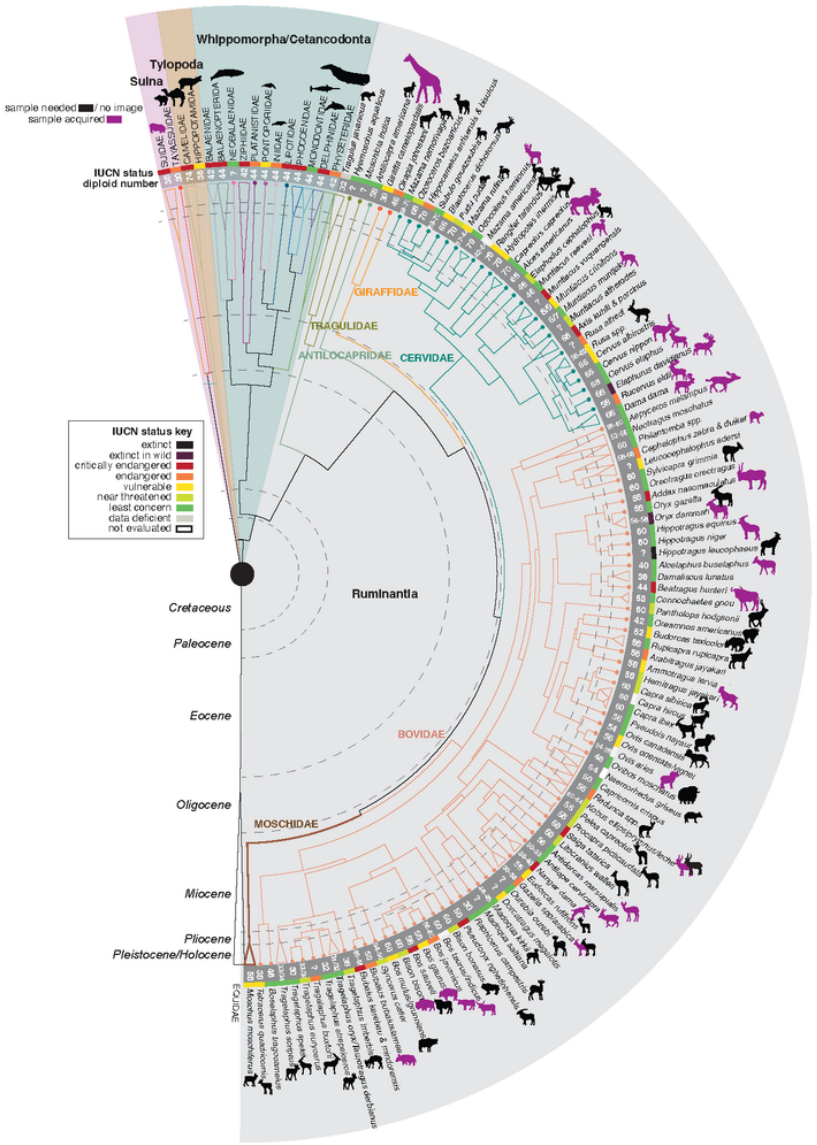

Figure 1

Phylogenetic relationships of target lineages for RT2T. From top left, shown are a single representative family for suborders Suina (pink), Tylopoda (tan), and Whippomorpha/Certancodonta (blue). In the grey block, terminal nodes indicate specific species within suborder Ruminantia; families are indicated (all caps) and color coded to the corresponding branch/lineage. Nodes for branches with several species but for which only one will be targeted for RT2T are collapsed to the most recent common ancestor. The topology and branch length of the tree are derived from Zurano et al., 2019[18]. From outer to inner circle: Silhouettes (taken from Phylopic, public domain license) for some species are shown (not to scale), with black or no silhouette indicating a sample for this lineage or species is needed and purple indicating a sample has been acquired. Under each lineage/species, the IUCN status is indicated, as per the inset key. For branches with a family listed, if any species within the family has a threatened status, the most severe threat is indicated. The number of diploid chromosomes is indicated (unknown =?). If the lineage has a range of

diploid numbers, the minima and maxima are indicated. Dotted lines indicate the start time of each epoch, as listed in the bottom left.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile.xlsx